Elisabeth André
Laila Dybkjær
Wolfgang Minker
Heiko Neumann
Michael Weber (Eds.)

LNAI 4021

Perception and Interactive Technologies

International Tutorial and Research Workshop, PIT 2006
Kloster Irsee, Germany, June 2006
Proceedings

Springer

Lecture Notes in Artificial Intelligence 4021

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Elisabeth André Laila Dybkjær
Wolfgang Minker Heiko Neumann
Michael Weber (Eds.)

Perception and Interactive Technologies

International Tutorial and Research Workshop, PIT 2006
Kloster Irsee, Germany, June 19-21, 2006
Proceedings

Springer

Volume Editors

Elisabeth André
University of Augsburg
Faculty of Applied Informatics
Eichleitnerstr. 30, 86159 Augsburg, Germany
E-mail: andre@informatik.uni-augsburg.de

Laila Dybkjær
University of Southern Denmark
Natural Interactive Systems Laboratory (NISLab)
Campusvej 55, 5230 Odense, Denmark
E-mail: laila@nis.sdu.dk

Wolfgang Minker
University of Ulm
Department of Information Technology
Albert-Einstein-Allee 43, 89081 Ulm, Germany
E-mail: wolfgang.minker@e-technik.uni-ulm.de

Heiko Neumann
University of Ulm
Department of Neural Information Processing
Oberer Eselsberg, 89069 Ulm, Germany
E-mail: heiko.neumann@uni-ulm.de

Michael Weber
University of Ulm, Department of Media Informatics
James-Franck-Ring, 89069 Ulm, Germany
E-mail: michael.weber@uni-ulm.de

Library of Congress Control Number: 2006926721

CR Subject Classification (1998): I.2, H.5.2-3, I.4

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-34743-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-34743-9 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11768029 06/3142 5 4 3 2 1 0

Preface

The Tutorial and Research Workshop on Perception and Interactive Technologies (PIT 2006) is the continuation of a successful series of workshops that started with an ISCA Tutorial and Research Workshop on Multimodal Dialogue Systems in 1999. This workshop was followed by a second one focusing on mobile dialogue systems (IDS 2002) and a third one exploring the role of affect in dialogue (ADS 2004). All workshops took place at Kloster Irsee in Bavaria and attracted participants from both sides of the Atlantic as well as from Asia.

Given the state of ongoing research in perceptive interfaces, we found it timely to organize a Tutorial and Research Workshop on Perception and Interactive Technologies (PIT 2006) at Kloster Irsee in Germany.

Due to its interdisciplinary nature, the workshop attracted submissions from researchers with very different backgrounds, such as computer science, neurology, medicine, bioinformatics and engineering. Overall, PIT 2006 embodied 16 long papers, 4 short papers and 6 demonstration papers. The programme was enhanced by an industrial session.

Enhancing interfaces with perceptive capabilities is a necessary prerequisite to render human–computer communication more natural and engaging. Papers represented at the PIT 2006 workshop focus on a broad range of perception technologies from head pose and eye gaze tracking in order to guide the users' level of attention to the analysis of speech and physiological data in order to acquire information on their emotional state. Furthermore, attempts are being made to integrate information from various input channels as well as from the situative context to compensate for errors and to resolve ambiguities. The workshop papers focus not only on the input side, but also explore how to guide the output of interactive systems by perceptive principles. In addition to enabling technologies for perceptive interfaces, the workshop features first prototypes of adaptive dialogue systems that consider, for example, a driver's mental load when providing help. Another class of applications represented at the workshop are interfaces with embodied conversational agents that integrate perceptive technologies with animation capabilities.

We would like to thank all authors for the effort they made on their submissions, and the Program Committee - nearly 50 distinguished researchers from industry and academia - who worked very hard to tight deadlines and selected the best contributions for the final program.

In addition, we would like to express our thanks to several people who assisted us in organizing the workshop. Torben Kruchov Madsen took care of the Web page for uploading papers. Angela Rittinger, Brigitte Waimer-Eichenauer and Claudia Wainczyk provided worthwhile administrative support. A number of organizations supported PIT 2006 including ACL Sigmedia, ACL/ISCA Sigdial and Gesellschaft für Informatik (GI). In particular, we gratefully acknowledge GI

for their valuable assistance in handling the financial matters. Last, but not least, we are grateful to Springer for publishing the proceedings in their LNCS/LNAI series.

April 2006

Elisabeth André
Laila Dybkjær
Wolfgang Minker
Heiko Neumann
Michael Weber

Organization

Organizing Committee

Elisabeth André (University of Augsburg, Germany)
Gregory Baratoff (Siemens VDO Automotive AG, Germany)
Laila Dybkjær (University of Southern Denmark, Denmark)
Markus Hennecke (Harman/Becker Automotive Systems, Germany)
Wolfgang Minker (University of Ulm, Germany)
Heiko Neumann (University of Ulm, Germany)
Michael Weber (University of Ulm, Germany)

Scientific Committee

Jan Alexandersson
Erhardt Barth
Andy Beall
Niels Ole Bernsen
Dirk Bühler
Andreas Butz
Rolf Carlson
Edward Y. Chang
Gabriel Christobal
Cristina Conati
Klaus Dorfmüller-Ulhaas
Ellen Douglas-Cowie
Sadaoki Furui
Hans-Werner Gellersen
Martin Giese
Jim Glass
Silke Goronzy
Jon Gratch
Joakim Gustafson
Eli Hagen
Gerhard Hanrieder
Ludwig Hitzenberger
David House
Thomas Kleinbauer
Marina Kolesnik
Stefanos Kollias
Ralf Kompe
Jeff Krichmar
Marc Latoschick
Rainer Lienhart
Jean-Claude Martin
Joseph Mariani
Dominic Massaro
Elmar Nöth
Sharon Oviatt
Tim Paek
Ana Paiva
Catherine Pelachaud
Alex Pentland
Christopher Peters
Robert Pieraccini
Helmut Prendinger
Alex Rudnicky
Marc Schroeder
Rainer Stiefelhagen
Oliviero Stock
David Traum
Ipke Wachsmuth
Wayne Ward

Sponsoring Institutions

University of Augsburg
University of Southern Denmark
University of Ulm - Competence Centre Perception and Interactive Technologies
ACL/ISCA Sigdial
ACL Sigmedia
Gesellschaft für Informatik (GI)

Table of Contents

Head Pose and Eye Gaze Tracking

Guiding Eye Movements for Better Communication and Augmented Vision
Erhardt Barth, Michael Dorr, Martin Böhme, Karl Gegenfurtner, Thomas Martinetz 1

Detection of Head Pose and Gaze Direction for Human-Computer Interaction
Ulrich Weidenbacher, Georg Layher, Pierre Bayerl, Heiko Neumann 9

Modelling and Simulation of Perception

Modelling and Simulation of Spontaneous Perception Switching with Ambiguous Visual Stimuli in Augmented Vision Systems
Norbert Fürstenau 20

Neural Network Architecture for Modeling the Joint Visual Perception of Orientation, Motion, and Depth
Daniel Oberhoff, Andy Stynen, Marina Kolesnik 32

Integrating Information from Multiple Channels

AutoSelect: What You Want Is What You Get: Real-Time Processing of Visual Attention and Affect
Nikolaus Bee, Helmut Prendinger, Arturo Nakasone, Elisabeth André, Mitsuru Ishizuka 40

Emotion Recognition Using Physiological and Speech Signal in Short-Term Observation
Jonghwa Kim, Elisabeth André 53

Visual and Auditory Displays Driven by Perceptive Principles

Visual Attention in Auditory Display
Thorsten Mahler, Pierre Bayerl, Heiko Neumann, Michael Weber . . . 65

A Perceptually Optimized Scheme for Visualizing Gene Expression Ratios with Confidence Values
Hans A. Kestler, Andre Müller, Malte Buchholz, Thomas M. Gress, Günther Palm 73

Spoken Dialogue Systems

Combining Speech User Interfaces of Different Applications
Dongyi Song 85

Learning and Forgetting of Speech Commands in Automotive Environments
Alexander Hof, Eli Hagen 97

Help Strategies for Speech Dialogue Systems in Automotive Environments
Alexander Hof, Eli Hagen 107

Multimodal and Situated Dialogue Systems

Information Fusion for Visual Reference Resolution in Dynamic Situated Dialogue
Geert-Jan M. Kruijff, John D. Kelleher, Nick Hawes 117

Speech and 2D Deictic Gesture Reference to Virtual Scenes
Niels Ole Bernsen 129

Combining Modality Theory and Context Models
Andreas Ratzka 141

Integration of Perceptive Technologies and Animation

Visual Interaction in Natural Human-Machine Dialogue
Joseph Machrouh, Franck Panaget 152

Multimodal Sensing, Interpretation and Copying of Movements by a Virtual Agent
Elisabetta Bevacqua, Amaryllis Raouzaiou, Christopher Peters, George Caridakis, Kostas Karpouzis, Catherine Pelachaud, Maurizio Mancini 164

Poster Session

Perception of Dynamic Facial Expressions of Emotion
Holger Hoffmann, Harald C. Traue, Franziska Bachmayr, Henrik Kessler 175

Multi-level Face Tracking for Estimating Human Head Orientation in Video Sequences
Tobias Bausch, Pierre Bayerl, Heiko Neumann 179

The Effect of Prosodic Features on the Interpretation of Synthesised Backchannels
Åsa Wallers, Jens Edlund, Gabriel Skantze 183

Unsupervised Learning of Spatio-temporal Primitives of Emotional Gait
Lars Omlor, Martin A. Giese 188

System Demonstrations

Talking with Higgins: Research Challenges in a Spoken Dialogue System
Gabriel Skantze, Jens Edlund, Rolf Carlson 193

Location-Based Interaction with Children for Edutainment
Matthias Rehm, Elisabeth André, Bettina Conradi, Stephan Hammer, Malte Iversen, Eva Lösch, Torsten Pajonk, Katharina Stamm 197

An Immersive Game - Augsburg Cityrun
Klaus Dorfmueller-Ulhaas, Dennis Erdmann, Oliver Gerl, Nicolas Schulz, Volker Wiendl, Elisabeth André 201

Gaze-Contingent Spatio-temporal Filtering in a Head-Mounted Display
Michael Dorr, Martin Böhme, Thomas Martinetz, Erhardt Barth ... 205

A Single-Camera Remote Eye Tracker
André Meyer, Martin Böhme, Thomas Martinetz, Erhardt Barth 208

Miniature 3D TOF Camera for Real-Time Imaging
Thierry Oggier, Felix Lustenberger, Nicolas Blanc 212

Author Index 217

Guiding Eye Movements for Better Communication and Augmented Vision

Erhardt Barth[1], Michael Dorr[1], Martin Böhme[1],
Karl Gegenfurtner[2], and Thomas Martinetz[1]

[1] Institute for Neuro- and Bioinformatics, University of Lübeck,
Ratzeburger Allee 160, D-23538 Lübeck, Germany
{barth, dorr, boehme, martinetz}@inb.uni-luebeck.de
http://www.inb.uni-luebeck.de
[2] Allgemeine Psychologie, Justus-Liebig-University,
Otto-Behaghel-Str. 10, D-35394 Gießen, Germany
Karl.R.Gegenfurtner@psychol.uni-giessen.de

Abstract. This paper briefly summarises our results on gaze guidance such as to complement the demonstrations that we plan to present at the workshop. Our goal is to integrate gaze into visual communication systems by measuring and guiding eye movements. Our strategy is to predict a set of about ten salient locations and then change the probability for one of these candidates to be attended: for one candidate the probability is increased, for the others it is decreased. To increase saliency, in our current implementation, we show a natural-scene movie and overlay red dots very briefly such that they are hardly perceived consciously. To decrease the probability, for example, we locally reduce the temporal frequency content of the movie. We here present preliminary results, which show that the three steps of our above strategy are feasible. The long-term goal is to find the optimal real-time video transformation that minimises the difference between the actual and the desired eye movements without being obtrusive. Applications are in the area of vision-based communication, augmented vision, and learning.

1 Introduction

An important property of human vision is that we must constantly shift our gaze between objects of interest because only the central part of the retina provides high visual acuity. Moreover, we can only attend to a very limited number of features and events in the visual environment. These facts have severe consequences for visual communication, because what is communicated depends to a large degree on those mechanisms in the brain that deploy our attentional resources and determine our eye movements.

The message that is conveyed by an image is thus determined not only by the image itself, but by the image in conjunction with the observer's gaze pattern, which may vary considerably from person to person and with context. Therefore, gaze is as important an attribute as brightness or colour for defining the message that reaches the observer.

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 1–8, 2006.

We propose that future information and communication systems should be designed to optimise gaze patterns and the use of the user's limited attentional resources. We believe that in future communication systems, images and movies will be defined not only by brightness and colour, but will be augmented with a recommendation of where to look, of how to view the images.

Gaze guidance can also be used to create new kinds of vision aids that fuse the strengths of human and computer vision to improve human visual capabilities. Such augmented-vision systems are of particular interest for automotive applications. For example, the driver's attention can be directed towards a pedestrian, who has been detected by sensors looking out of the car, in cases when the driver would otherwise fail to see the pedestrian.

Gaze guidance can be used for a further kind of application in which novices can be taught to view images with the eyes of experts. It is known that experts, for example experienced pilots, scan their environment in a way that substantially differs from how inexperienced vi ewers would. We believe that by applying the gaze pattern of experts to novices, we can evoke a sub-conscious learning effect.

To reach such goals, however, a considerable amount of basic research and technological development is still required. During the workshop we will demonstrate the eye-tracking and display technology that we currently use and continue to develop [1, 2]. In the remainder of this paper we report on a few results that address selected problems of gaze guidance.

2 Guidance of Eye Movements

The final goal of our gaze guidance system is to direct the user's attention to a specific part of a scene, ideally without the user noticing this guidance. Our strategy is to (i) predict a few candidate locations for saccade targets, (ii) increase the probability of being attended for one candidate, and (iii) decrease the probability for the other candidates. We have not yet implemented a system that integrates all components of this strategy, but will show some results related to the individual points (i-iii) above.

2.1 Eye Movement Predictions

We have investigated the variability of eye movements with dynamic natural scenes and found that 5-15 clusters (frequently looked at regions that, when taken together, cover 2-5 % of the viewing area) account for 60 % of all fixations [3]. This justifies our approach of predicting a small set of candidate locations based on low-level features of the visual input.

Like other authors, e.g. [4], we base our approach on a saliency map that assigns a certain degree of saliency to every location in every frame of a video sequence. Various techniques exist for computing saliency maps, but they are all based, in one way or another, on local low-level image properties such as contrast, motion or edge density, and are intended to model the processes in the human visual system that generate saccade targets. We assume that the human visual system uses low-level features such as those used in saliency maps to

generate a list of candidate locations for the next saccade target, and that top-down attentional mechanisms then select one of the candidate locations as the actual saccade target (although in reality, bottom-up and top-down processing will most likely be hard to segregate). This selection mechanism is probably very difficult to model algorithmically. In our view, a more realistic goal is therefore to predict a certain number of candidate locations, say ten, that will with high probability include the actual saccade target, and our above mentioned results on gaze analysis show that a small number of target locations usually covers most of the variations in the eye movements made by different observers.

As described previously [5, 6], our approach to saliency is based on the concept of intrinsic dimension that was introduced for still images in [7]. The intrinsic dimension of a signal at a particular location is the number of directions in which the signal is locally non-constant. It fulfils our requirement for an alphabet of image changes that classifies a constant and static region with low saliency, stationary edges and uniform regions that change in time with intermediate saliency, and transient patterns that have spatial structure with high saliency. We also note that those regions of images and image sequences where the intrinsic dimension is at least 2 have been shown to be unique, i.e. they fully specify the image [8]. The evaluation of the intrinsic dimension is possible within a geometric approach that is plausible for biological vision [9] and is implemented here by using the structure tensor $\mathbf{J}$, which is well known in the computer-vision literature [10]. The structure tensor is defined in terms of the spatio-temporal gradient ∇f of the image intensity function

$$\mathbf{J} = \omega * \nabla\mathbf{f} \otimes \nabla\mathbf{f} = \omega * \left(\nabla\mathbf{f}\nabla\mathbf{f}^{\mathbf{T}}\right) \tag{1}$$

where ω denotes a convolution kernel that performs a local averaging of the product terms. Our saliency measure that was used for the results presented in Fig. 1 is K, the determinant of $\mathbf{J}$, which indicates an intrinsic dimension of 3, i.e., there is no spatio-temporal direction along which intensity is constant. Typical features with high K-values would be blobs or corners that appear or dissappear. To extract salient features on different spatial and temporal scales, we construct a 4-level spatio-temporal Gaussian pyramid from the image sequence and compute the saliency measures on each level. Details of how we select the candidate locations from the saliency map are given in [11]. We have used other invariants of the structure tensor (trace and sum of minors that indicate intrinsic dimensions of at least one or two respectively) and found prediction results that were slightly worse. We have not performed more comprehensive comparisons for two reasons: (i) the results are good enough for our purpose of guiding eye movements, and (ii) we are not aware of similar attempts to predict a set of locations based on a spatio-temporal measure.

As a baseline for assessing the saliency maps computed analytically using the K measure, we use empirical saliency maps, i.e. saliency maps computed from the actual eye movements of the test subjects. These give us an idea of what the saliency map should look like for a given video sequence, and they can serve as a basis for judging what the best possible results are that we can expect for

predictions made solely on the basis of a saliency map generated from the image data, without taking individual top-down strategies into account (the empirical saliency actually does even better than the best possible saliency because it is derived from the data that are then predicted). To generate the empirical saliency map for a video frame, we determine the current gaze position of each observer and place a Gaussian with a standard deviation of 16 pixels at each of these positions. The superposition of these Gaussians then yields the empirical saliency map. For a detailed description see [6].

Note that the attempt of predicting ten candidate locations based on a simple saliency measure is quite successful in the sense that it approaches the limit given by the empirical predictor. However, when using only one instead of ten candidates, the errors obtained for both the empirical and the analytical saliency measures are still better than chance but unacceptably high.

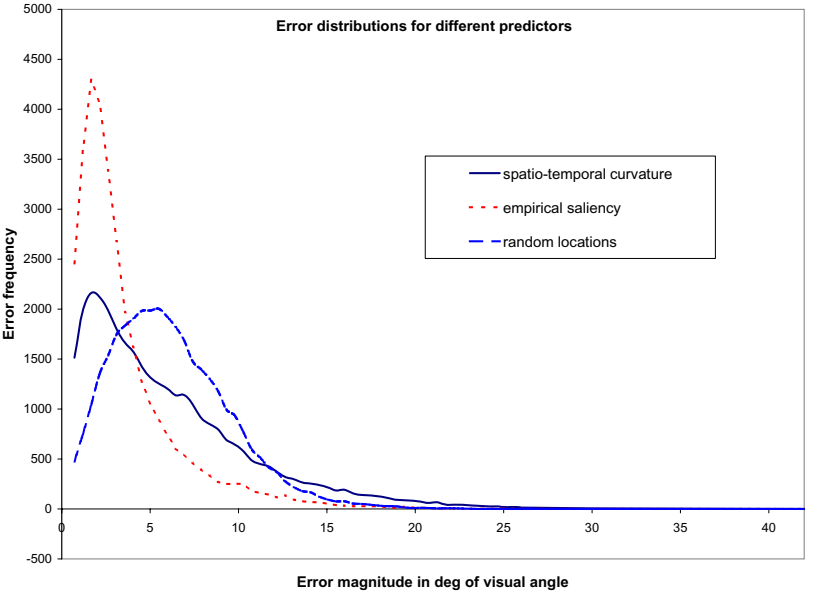


Fig. 1. Saliency prediction results. Histogram of error (distance of saccade target to closest salient candidate location) for ten candidate locations. The horizontal axis plots the error magnitude in degrees, the vertical axis plots the number of saccades per histogram bin. Plots are shown for the K saliency measure, the empirical saliency measure, and locations chosen at random.

2.2 Effect of Gaze-Contingent Red Dots

We now address the problem of increasing the probability of candidate locations to be attended. In a first set of experiments, we were motivated by the well-known fact that sudden object onsets in the visual periphery can attract attention. We therefore chose to briefly superimpose small bright red dots on the displayed

movie. Depending on the eccentricity of the dots at the time of their flashing, in up to about 40 % of trials, saccades were initiated towards the location of the flashed red dot when subjects were asked to just watch the movie. The results of these experiments are shown in Fig. 2. Note that the red-dot stimuli have a considerable effect when presented at 10 degrees eccentricity. The effect is smaller at 15 and 20 degrees partly because of the limited field of view and partly due to the fact that the stimulus size was not scaled (enlarged by the cortical magnification factor) with eccentricity and was therefore less effective.

Because the typical saccadic latency of about 200 ms exceeds the presentation time of the dot, which was set to 120 ms, the red dot was already switched off by the time the saccade was finished. In another set of experiments, where (other) subjects watched the same movies and were instructed to look for red dots and press a button when they detected one, the stimulation remained invisible in about 50 % of cases. Stimuli at 15 and 20 degrees were less visible than those at 10 degrees.

Similar effects were obtained in an experiment where the red dot was replaced by a looming stimulus, the looming stimulus being harder to detect than the red dot. Nevertheless, the exact parameters for an optimal guidance effect, such as size, contrast, duration, or the timing with regard to previous saccades, still need to be determined.

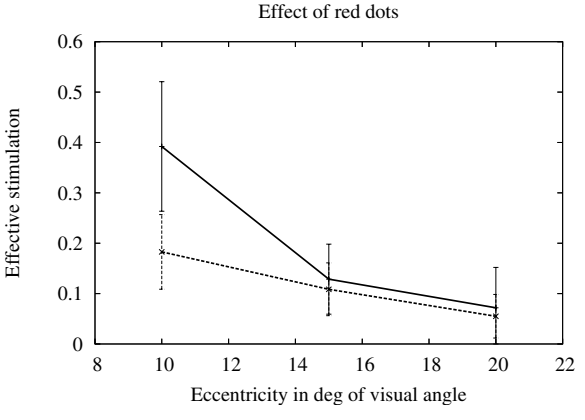


Fig. 2. Effect of red-dot stimulation. We show the number of cases (in % of all trials) in which subjects made a saccade to where the red dot had been shown (the saccade started in a time window of 100 to 300 ms after stimulus onset and landed within a circle of 5 deg radius around stimulus location). As a reference we show (with a dashed line) the same results for cases in which the red dots had not been displayed.

2.3 Effect of Gaze-Contingent Spatio-temporal Filtering

We now address the problem of decreasing the probability of candidate locations to be attended. To do this, we have developed a gaze-contingent display that can in real time change the spatio-temporal content of an image sequence as a function of where the observer is looking [12].

Gaze-contingent displays manipulate some property of the (static or moving) image as a function of gaze direction (see [13] for a review). The image property that is most commonly manipulated in a gaze-contingent display is spatial resolution. A popular type of manipulation is to foveate an image or video, i.e. to simulate the effect of the variable resolution of the human retina, which is highest at the fovea and falls off towards the periphery. If the foveation is adjusted to match the resolution distribution of the retina, the effect is not noticeable for the observer, but the resulting images can be compressed more efficiently because they contain less high-frequency content [14, 15]. The current state-of-the-art algorithm for gaze-contingent spatial filtering of video is due to Perry and Geisler [16]. Unlike previous algorithms, which introduced artifacts in the filtered images, their algorithm produces smooth, artifact-free results.

Based on this work, we have developed a gaze-contingent display that manipulates not the spatial, but the temporal resolution of a video. The basic effect of temporal filtering is to blur the moving parts of an image while leaving the static parts unchanged (demonstrations will be shown at the workshop). Our motivation for performing this type of manipulation is that we want to examine the effect that it has on eye movements; movement or change in the periphery of the visual field is a strong cue for eye movements. The results presented in Fig. 3 show that gaze-contingent temporal filtering reduces the number of saccades into the periphery where the temporal frequencies are reduced by our gaze-contingent display. To

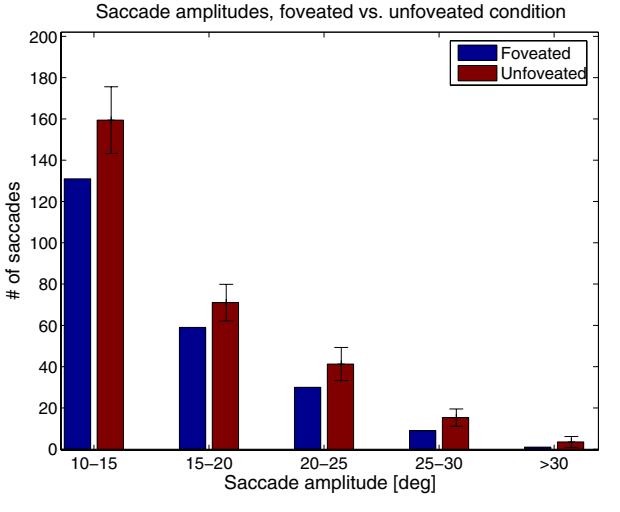


Fig. 3. Effect of spatio-temporal filtering shown as histograms of saccade amplitudes with and without gaze-contingent temporal filtering. Only the histogram bars for saccades of 10 degrees and greater are shown since no effect could be expected in the central field of view due to the shape of the foveation function. Since the number of subjects who watched the unfiltered movies was much larger, error bars show s.d. over population samples with the same number of subjects as in the filtered condition.

further improve the effect of the gaze-contingent display, we plan to specifically change the spatio-temporal content only at certain locations in an image.

3 Discussion

We have shown that a rather small set of locations where people may look while watching a natural video can be predicted with acceptable errors based on simple low-level dynamic saliency measures. This fits well with our previous analysis of eye movements on high resolution natural videos, which shows that eye movements tend to cluster in rather few (10-20) locations [3]. We then reported on simple experiments that were meant to increase the saliency by a brief gaze-contingent presentation of red dots in the periphery. Concerning methods for decreasing saliency, we have shown that peripheral temporal blur changes the gaze pattern by inhibiting saccades.

Since we are looking for unobtrusive ways of guiding gaze, we also analysed the visibility of red dots and of the temporal blur. The visibility is hard to determine since it depends on the task. However, our results [17] indicate that there exists a set of manipulations (of which we do not yet know how large it is) that are effective in guiding eye movements but are not perceived consciously.

We therefore conclude that gaze guidance seems possible in principle. However, more work is required to better understand the most efficient ways of performing it. Eventually, we will have to show that gaze guidance can improve human vision capabilities in behavioural tasks and thus justify the applications that we have in mind.

Acknowledgements

Research was supported by the German Ministry of Education and Research (BMBF) under grant number 01IBC01B with acronym *ModKog*. We thank the reviewers for valuable remarks.

References

1. Dorr, M., Böhme, M., Martinetz, T., Barth, E.: Gaze-contingent spatio-temporal filtering in a head-mounted display. In: Perception and Interactive Technologies. (same volume)
2. Meyer, A., Böhme, M., Martinetz, T., Barth, E.: A single-camera remote eye tracker. (Same volume)
3. Dorr, M., Böhme, M., Martinetz, T., Gegenfurtner, K.R., Barth, E.: Analysing and reducing the variability of gaze patterns on natural videos. In Groner, M., Groner, R., Müri, R., Koga, K., Raess, S., Sury, P., eds.: Proceedings of 13th European Conference on Eye Movements. (2005) 35
4. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 1254–1259

5. Barth, E., Drewes, J., Martinetz, T.: Individual predictions of eye-movements with dynamic scenes. In Rogowitz, B., Pappas, T., eds.: Electronic Imaging 2003. Volume 5007., SPIE (2003) 252–259
6. Böhme, M., Dorr, M., Krause, C., Martinetz, T., Barth, E.: Eye movement predictions on natural videos. Neurocomputing (2006) (in press).
7. Zetzsche, C., Barth, E.: Fundamental limits of linear filters in the visual processing of two-dimensional signals. Vision Research **30** (1990) 1111–1117
8. Mota, C., Barth, E.: On the uniqueness of curvature features. In Baratoff, G., Neumann, H., eds.: Dynamische Perzeption. Volume 9 of Proceedings in Artificial Intelligence., Köln, Infix Verlag (2000) 175–178
9. Barth, E., Watson, A.B.: A geometric framework for nonlinear visual coding. Optics Express **7** (2000) 155–165
10. Jaehne, B., Haußecker, H., Geißler, P., eds.: Handbook of Computer Vision and Applications. Academic Press (1999)
11. Böhme, M., Dorr, M., Krause, C., Martinetz, T., Barth, E.: Eye movement predictions on natural videos. Neurocomputing (2005) (in press).
12. Böhme, M., Dorr, M., Martinetz, T., Barth, E.: Gaze-contingent temporal filtering of video. In: Eye Tracking Research and Applications (ETRA). (2006) (in press).
13. Duchowski, A.T., Cournia, N., Murphy, H.: Gaze-contingent displays: A review. CyberPsychology & Behavior **7** (2004) 621–634
14. Kortum, P., Geisler, W.: Implementation of a foveated image coding system for image bandwidth reduction. In: Human Vision and Electronic Imaging, SPIE Proceedings. Volume 2657. (1996) 350–360
15. Geisler, W.S., Perry, J.S.: A real-time foveated multiresolution system for low-bandwidth video communication. In Rogowitz, B., Pappas, T., eds.: Human Vision and Electronic Imaging: SPIE Proceedings. (1998) 294–305
16. Perry, J.S., Geisler, W.S.: Gaze-contingent real-time simulation of arbitrary visual fields. In Rogowitz, B.E., Pappas, T.N., eds.: Human Vision and Electronic Imaging: Proceedings of SPIE, San Jose, CA. Volume 4662. (2002) 57–69
17. Dorr, M., Böhme, M., Martinetz, T., Barth, E.: Visibility of temporal blur on a gaze-contingent display. In: APGV 2005 ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization. (2005) 33–36

Detection of Head Pose and Gaze Direction for Human-Computer Interaction

Ulrich Weidenbacher, Georg Layher, Pierre Bayerl, and Heiko Neumann

University of Ulm, Germany
{Ulrich.Weidenbacher, Georg.Layher, Pierre.Bayerl,
Heiko.Neumann}@uni-ulm.de

Abstract. In this contribution we extend existing methods for head pose estimation and investigate the use of local image phase for gaze detection. Moreover we describe how a small database of face images with given ground truth for head pose and gaze direction was acquired. With this database we compare two different computational approaches for extracting the head pose. We demonstrate that a simple implementation of the proposed methods without extensive training sessions or calibration is sufficient to accurately detect the head pose for human-computer interaction. Furthermore, we propose how eye gaze can be extracted based on the outcome of local filter responses and the detected head pose. In all, we present a framework where different approaches are combined to a single system for extracting information about the attentional state of a person.

1 Introduction

1.1 Motivation

Interaction between humans and computers is commonly restricted to typing on a keyboard or pointing and clicking the mouse button. This type of interaction is very unnatural for humans since human-human interaction is commonly based on multi-modal interaction. For example, in a conversation auditory and visual information is important to be interpreted properly in order to react to a dialog partner. In such a conversation important visual cues can be facial expression, head pose and particularly the eye gaze for getting feedback about the attentional and mental state of a dialog partner [Emery, 2000].

1.2 Previous Work on Eye Gaze Estimation

One of the first applications that utilizes eye gaze as a computer interface was developed by [Hutchinson et al., 1989] where computer users could interact by directing their gaze to specific areas on the monitor. Similar to recent eyetracker applications [Eyelink, 2006] their system requires infrared light to illuminate the eye region. In general, state-of-the-art eyetracker applications are highly accurate and reliable, however most of them require complex hardware (helmet with a mounted camera) or the user has to be in a fixed position (e.g. with a chin chest). Other approaches which are not constrained by specific

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 9–19, 2006.

hardware, referred to as 'non-intrusive' systems, have to compensate for motion affects of the user (e.g. by using tracking methods [Baluja and Pomerleau, 1994, Ji and Zhu, 2003, Zhu and Ji, 2005, Yoo and Chung, 2005]), though they still employ active sensing tools (e.g. by illuminating the eyes with infrared light). In this contribution, we concentrate on purely vision based methods ([Steifelhagen et al., 1997, Heinzmann and Zelinsky, 1998]) where eye gaze is estimated passively (i.e. without special illumination).

1.3 Combining Head Pose and Eye Gaze

There is a large amount of work present for the detection of head pose [Gee and Cipolla, 1994, Krüger et al., 1997, Rae and Ritter, 1998, Wang et al., 2003] and for eye gaze estimation, but there is only few work present where both information, head pose and eye gaze are combined. For example, [Matsumoto and Zelinsky, 2000] present a three stage system that combines head pose and eye gaze information to accurately estimate a person's point of attention. The particular point of their system is that they use a 3D head model and a 3D eye model to accurately detect head pose and eye gaze based on stereo vision. However, our proposed methods are based on monocular images.

1.4 Overview

In this contribution, we focus on methods for the estimation of head pose and eye gaze for human-computer interaction purposes. We compare two methods for the estimation of head pose. (1) a view-based approach where a small set of prototypical views of a head is used to determine the pose of a presented test head [Krüger et al., 1997] and (2) a model-based approach where geometrical information about the face is utilized to determine the head pose of a person [Gee and Cipolla, 1994]. In addition to head pose, eye gaze is an important cue for the detection of attention [Emery, 2000]. Thus, we present a novel method that uses phase information of simple biological motivated filters to determine the direction of gaze from a person.

2 Methods

2.1 Image Acquisition and Ground Truth Generation

We created a dataset of images showing different head pose / eye gaze conditions acquired from 5 subjects. The procedure of acquiring the images was as follows: we fixated a laser pointer on a tripod equipped with an angular meter. The device was used to accurately attach marks to the walls in horizontal steps of 10° ranging from $-90°$ to 90° as shown in Fig. 1a. Then, a subject had to sit on a swivel chair facing the camera. According to Fig. 1c, two clips were used to mount the laser pointer on top of the subjects heads. To calibrate the position of the laser pointer on the subjects heads the height of the chair was

first adjusted so that the laser pointer was located at a level according to the marks (the height of the marks was 140 cm; see Fig. 1b). Then, the subject was told to look straight into the camera. Retaining this state, the position of the laser pointer was corrected until the laser beam spot coincided with the 0° mark. After successful calibration, images with ground truth data of different head pose/eye gaze configurations were recorded by asking the subject to align the laser pointer spot to a specific mark on the wall (head pose) while focusing another mark with the eyes (eye gaze). In this manner, 285 images from five subjects were taken with a digital camera (Casio QV-5700, maximal optical zoom to avoid perspective distortion effects).

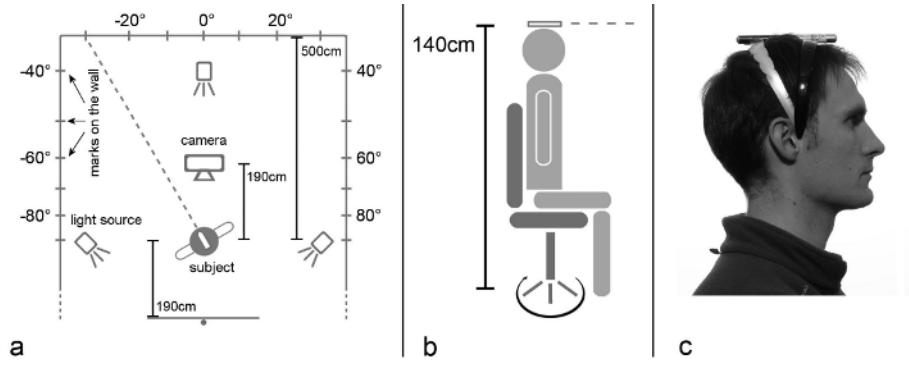


Fig. 1. (a) shows the setup used for image acquisition. The distance between subject and camera as well as the distance between the canvas and the subject was 190 cm. The range between subject and the opposite wall was 500 cm. Three spotlights were used to generate uniform background illumination and to prevent cast shadows. A laser pointer was mounted on a tripod equipped with an angular dimension to mark positions on the walls. These marks are later used to orient the head to a specific direction. (b) side view of a subject during image acquisition. The height of the chair was adjusted until the subject's laser pointer was located at a height of 140 cm. (c) visualizes the fixation of the laser pointer on the head of a subject.

2.2 Head Pose Detection

We compare two different computational approaches to determine the rotation of the head around its vertical axis (yaw or heading), namely a view-based approach and a model-based approach.

Our first method is a view-based approach which employs a simplified version of Elastic Graph Matching (EGM) proposed by [Krüger et al., 1997]. Seven images of different head poses from one individual (from $-90°$ to $90°$ in steps of $30°$ around the vertical axis) are utilized as prototype poses. On each image we manually select 10 landmarks on the face covering the nose, eyes and mouth in frontal view and also the ears in profile view (see Fig. 2a). The prototype pose

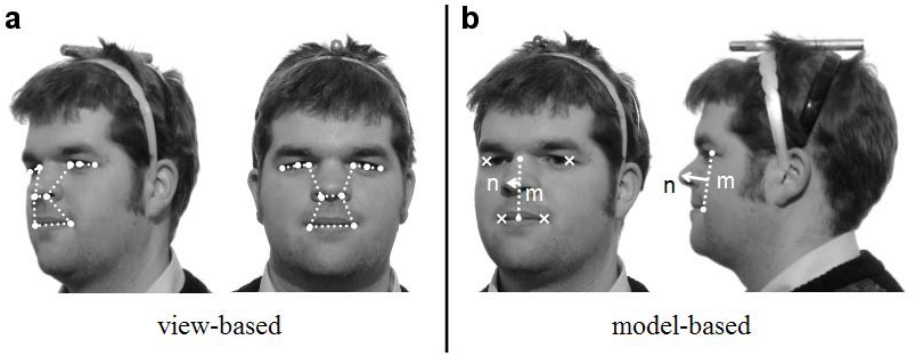


Fig. 2. Labeled head for the view-based approach (a) and the model-based approach (b). In the view-based approach we labeled 7 head poses from -90° to 90° in steps of 30°. Moreover, for each head pose we selected 10 facial features leading to different graphs for each pose. On each node in the graph Gabor filter responses of different orientation and scale are extracted. In the model-based approach the positions of the eye corners and the mouth corners are manually labeled to determine the symmetry axis of the head. Furthermore, the position of the nose tip is labeled manually. The projected nose length n relative to the height of the face m gives information about the pose of the head.

images were convolved using a family of 40 DC-free Gabor wavelets (5 frequencies x 8 orientations). On each landmark the set of 40 complex Gabor coefficients (Gabor jet) is extracted and stored together with the relative positions of the landmarks. This is done for each pose leading to 7 prototypical pose representations (bunch graphs). To detect the position and pose of a novel face, the bunch graphs are shifted over the new image while on each position in the image a similarity value is computed by a normalized cross-correlation between the Gabor responses stored in the graph and the present Gabor responses in the image. The position of the face is determined by the location in the image with the maximal correlation result. For pose estimation we consider the responses of all prototype graphs at this location. Here, we fit a quadratic function onto the prototype responses and determine the estimated head pose at the maximum of this function.

The second method employs a model-based approach for the estimation of head pose proposed by [Gee and Cipolla, 1994]. Here, we assume that localized features, namely the corners of the eyes, the mouth and the nose tip have been already detected in the image. As in the first approach we restrict the possible movements of the head to rotations around the vertical axis. Under the assumption of weak perspective projection described in [Trucco and Verri, 1998] the distance n from the nose tip to the symmetry axis of the face relative to the length m (height of the face) is proportional to the sine of the head angle (see Fig. 2b). Thus, the head angle is computed by sin^{-1} of the projected nose length n relative to the projected height m of the face.

2.3 Detection of Gaze Direction

For the estimation of gaze direction we propose to employ Gabor filter responses similar to the EGM method used for pose estimation. Here, we evaluate the phase of Gabor responses [Gabor, 1946] in facial sub-regions around the eyes [Langton et al., 2000]. The idea is that the iris region is always darker than the remaining regions on the sclera (see Fig. 3; [Sinha, 2000, Langton et al., 2000]). Thus, gradual eye movements imply a gradual change of the Gabor phase. Therefore, we conclude that there exists a direct relation between Gabor phase and eye gaze dependant on the head pose.

The phase representing the optimal Gabor pattern matching the underlying image pattern at one specific location is described by Eq. 1:

$$\text{phase} = \text{atan2}(I * G_{\sin}, I * G_{\cos}) \tag{1}$$

where I is the input image, $*$ is the convolution operator, and $G_{\cos}$ and $G_{\sin}$ are the real and complex parts of the Gabor filter as follows (Eq. 2):

$$G_{\cos} + iG_{\sin} = \exp(i\frac{2\pi}{\lambda}x)\exp(\frac{-(x^2+y^2)}{2(0.45\lambda)^2}) \tag{2}$$

where λ is the Gabor wave length and x, y are image coordinates. Note that our Gabor filters are self-similar which means that the number of wave cycles under the Gaussian envelope function remains constant for values of λ.

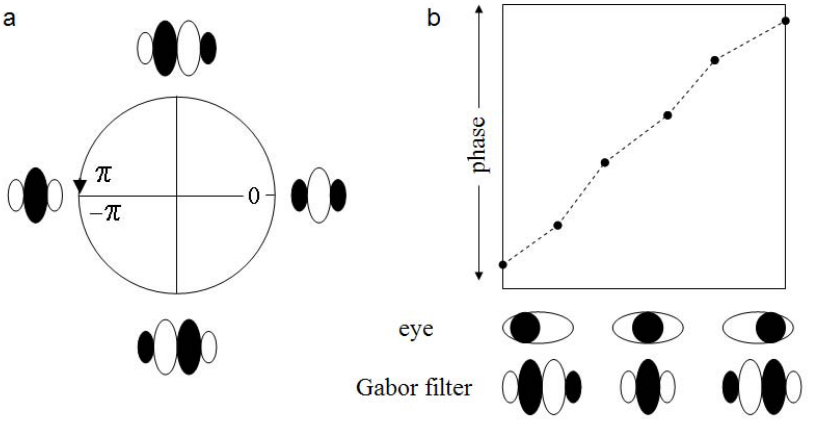


Fig. 3. Gabor filters consist of a wave function multiplied with a Gaussian envelope. (a) changing the phase of the wave function leads to different Gabor filter shapes. White indicates positive filter components while black stands for negative filter components (displayed are the real parts of the Gabor filter). (b) the filter that best matches the underlying eye pattern determines the phase for a specific eye gaze direction. Note that the phase is a circular measure which may lead to discontinuities when visualized in Cartesian coordinates. To avoid discontinuities the phase is shifted for visualization purposes where necessary.

For the estimation of gaze direction we generate a lookup table describing the relation between extracted image phase, gaze and head pose. This lookup table (LUT) is then utilized to map the extracted phase to the appropriate gaze direction for a given head pose[1].

3 Simulations

3.1 Head Pose Estimation

For the view-based approach the prototype bunch graphs were obtained from images of one person excluded from the test dataset as described in the previous section (see Fig. 5). For the model-based approach the length of the nose relative to the face length was measured manually for each person. Fig. 5 illustrates the estimated pose for 19 presented head poses of one person ranging from $-90°$ to $90°$. Despite that both approaches for pose estimation are rather simple in their

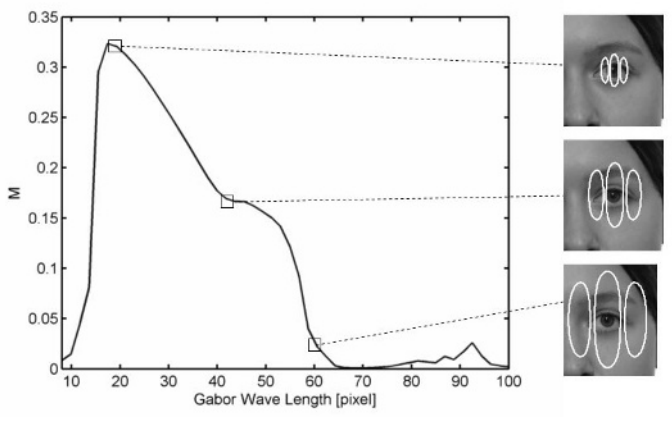


Fig. 4. Analysis of different Gabor filter proportions. The measure M (averaged over all heads in frontal pose) is plotted across different Gabor filter sizes. High values of M indicate phase linearity in combination with a high gradient of the phase. On the right we show three Gabor filters of different size belonging to different positions in the graph. The graph illustrates that there is an optimal Gabor filter size of about 20 pixels per cycle where slope and linearity of the phase are both high. As the filter size increases more and more adjacent parts of the eye are covered by the filter which results in a gradual drop of M towards zero. Note that the average width of the eye region within the image was about 45 pixels.

implementation and that not much effort was put in training or calibration we obtain results which allow to determine the horizontal head orientation up to an accuracy of approximately $10°$. Our experimental investigations show that 75%

[1] This operation requires a one-to-one mapping between gaze direction and phase for each given head pose (compare section 3.2).

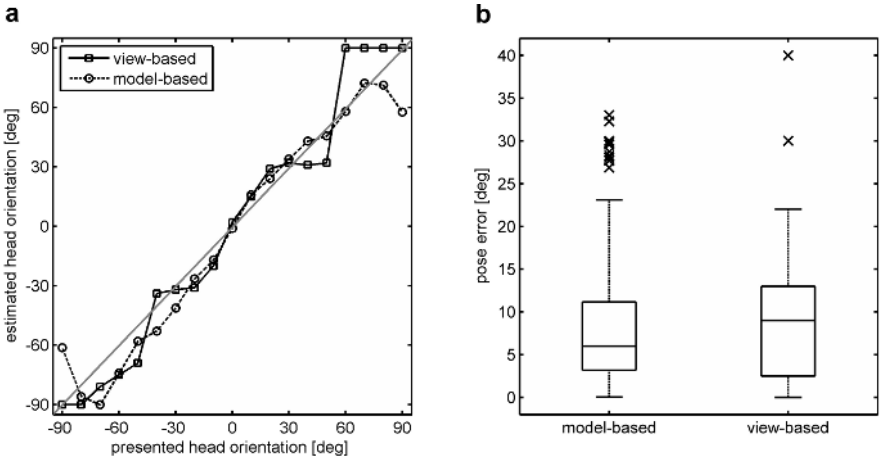


Fig. 5. (a) shows the estimated head pose for one head using the view-based and the model-based approach respectively. The grey line indicates optimal head pose estimation. The result demonstrates that both approaches yield good estimation results for presented head poses between $-30°$ and $30°$. Results near $-90°$ and $90°$ show significantly higher errors for both approaches. (b) summarizes the distribution of the error over all sample cases for both approaches (excluding the labeled head for the view-based approach). In both cases 75% of the errors are smaller than $14°$.

of all estimated head poses over all tested input images are smaller than $14°$ for the view-based method and smaller than $12°$ for the model-based approach (in accordance with the investigations in [Gee and Cipolla, 1994]).

3.2 Eye Gaze Estimation

Gabor parameters. Given the head pose and the location of the eyes it is possible to investigate the properties of the phase of Gabor responses. To determine the optimal wave length λ of the Gabor filter we introduce a measure M that describes the quality of the Gabor filter for gaze estimation.

Fig. 6 exemplifies that a gradual change of the eye gaze direction leads to a gradual change of the Gabor phase. Therefore, we conclude that a good choice of the filter could yield a near linear dependency of the phase from the gaze direction. Moreover, a large slope of the linear dependency helps to prevent ambiguities in the mapping between phase and gaze. In other words, if the angular distance between phases is very small (i.e. low slope) then the discriminative power of the phase LUT gets lost. To investigate the phase linearity across eye gaze direction we fit a linear function to the measured phases and consider the sum of residuals as a quantitative measure for the linearity of the phase. Thus, we define M as follows (Eq. 3):

$$M = \frac{m^2}{1 + \sum r_i^2} \tag{3}$$

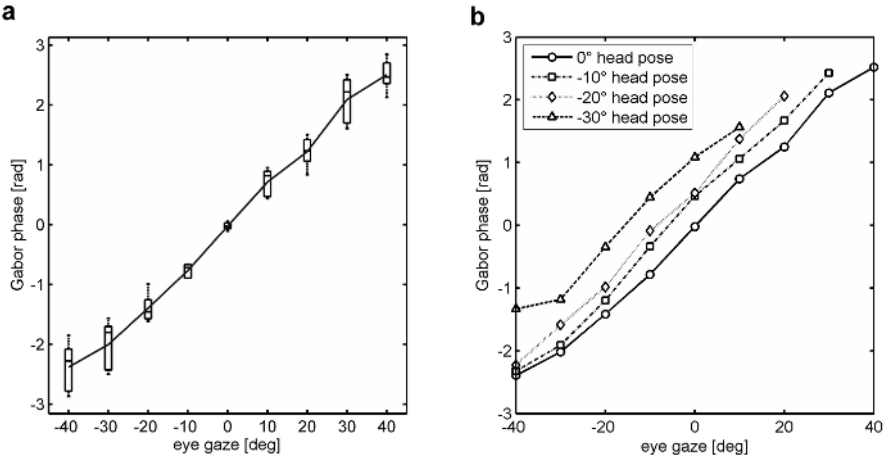


Fig. 6. (a) shows the average Gabor phase extracted from all heads (solid line) for frontal head poses across different eye gaze directions (gaze angles are given in world coordinates). Phase variances are smaller for central gaze directions than for peripheral gaze directions. (b) shows Gabor phases (averaged over all heads) acquired from four different head pose conditions reaching from -30 to 0 degrees. The extracted phases suggest a linear relation between gaze direction and Gabor phase. Note that we place the Gabor filters on the eye that is within the facial part of better visibility (the left eye for negative head angles and the right eye for positive head angles).

where m is the gradient of the linear fit function and r_i is the residual error. M should be maximal if both conditions (linearity and large slope) are present in the phase responses. In Fig. 4 we show M across different Gabor filter sizes. Small Gabor filters (smaller than the eye region) produce phase discontinuities resulting in a drop of M. Filter sizes in the proportion of the eye region lead to maximal values of M followed by a gradual decrease of M for larger filter sizes. We therefore set the size of our Gabor filters to 20 pixel per cycle for our images (corresponding to the size of the eye; see Fig. 4b).

Gabor phase results and gaze estimation. Fig. 6a illustrates the phase averaged over all heads in frontal head pose. The variance is very small when the eye looks straight and increases gradually when the eye looks to the left or to the right. Fig. 6b shows the extracted phase information for different gaze directions and four different head poses. As expected, the outcome suggests that for all presented head poses the phase can directly be mapped to the gaze direction.

In line with perceptual investigations [Sinha, 2000, Langton et al., 2004] our approach suggests that based on the head pose, the gaze can be determined by the local distribution of luminance values within the eyes. Consistent with experimental observation of [Gibson and Pick, 1963] the gaze direction is determined by the eye pattern (the phase) relative to the face configuration (head pose).

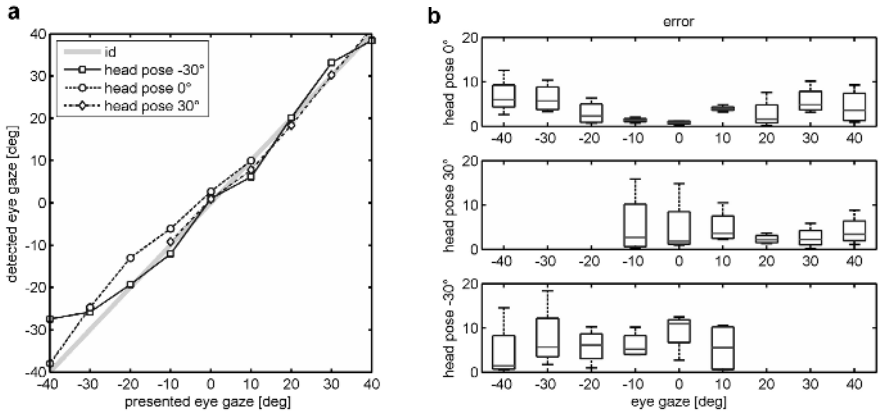


Fig. 7. (a) shows estimated eye gaze directions for one single head for three different head poses. The gray line indicates optimal gaze estimation. (b) shows errors across all heads for three different head poses ($-30°/0°/30°$). Overall the results demonstrate that in nearly all conditions 75% of the estimation errors are below 10 degrees. In frontal head pose estimation errors are specifically low for $0°$ gaze direction (corresponding to the mutual gaze condition).

Thus, Gabor phase responses are learned from one head by creating a simple lookup table between gaze direction, head pose, and the extracted phase. For a given head pose the gaze of a test face can now be determined by matching the detected phase to the linearly interpolated phases in the lookup table. The results are illustrated in Fig. 7a where we show the estimated eye gaze based on the learned phases for one test head. Fig. 7b shows the gaze detection errors for all test heads (excluding the head used for generating the LUT). Errors are mostly under 10 degrees and are significantly small (almost zero) for frontal head pose and straight eye gaze.

4 Discussion and Conclusion

In this contribution we describe how a database of images is generated showing faces from different head poses with different gaze directions. In contrast to other image databases [Phillips et al., 2000, Sim et al., 2003] our experimental setup allows to provide a ground truth for both gaze direction and head pose. We compare two approaches for head pose estimations and present how the gaze direction can be extracted from the local phase of a given gray-level image of a person's face.

For gaze detection we present a quality measurement to determine the parameters of the employed Gabor filter. Our measurement is based on the linearity and the slope of the relation between gaze direction and extracted Gabor phase. Note that ambiguities between gaze configurations can occur if the complete range of possible phases is covered by this relation caused by the circularity of

the phase (e.g., $-\pi = \pi$, see Fig. 3). To take this into account we choose the parameters of the Gabor filters so that the angular distance between minimal and maximal detected Gabor phase is no less than $\frac{\pi}{8}$.

We claim that all approaches for head pose estimation as well as for gaze detection that we investigated here either utilize information that is represented in the visual system or induce perceptual effects observed in experiments with human observers. (1) The correlation of Gabor responses for head pose estimation represents a pattern matching of filter responses similar to neural responses of cells in early visual cortex [Hubel and Wiesel, 1968]. (2) The length or asymmetry of the projected nose has a direct effect on the perception of the head and gaze direction [Langton et al., 2004]. (3) The perception of the gaze direction is highly dependent on the luminance distribution of the presented face, in particular the perceived gaze is inverted if the polarity of the image is inverted [Sinha, 2000]. Furthermore, the employed filter responses proposed for gaze estimation are also expected in visual cortex [Langton et al., 2000].

Thus, we propose an extended framework in which all visual information described in this contribution are merged to determine the facial configuration concerning head pose and gaze direction. Particularly, we show how simple view or model based approaches can be utilized for determining the head pose and illustrate how gaze can be extracted in a framework for human-machine interaction.

Acknowledgements

This work has been supported by a grant from the Ministry of Science, Research and the Arts of Baden-Württemberg (Az: 23-7532.24-13-19/1) to Heiko Neumann and Ulrich Weidenbacher.

References

[Baluja and Pomerleau, 1994] Baluja, S. and Pomerleau, D. (1994). Non-intrusive gaze tracking using artificial neural networks. *Technical Report CMU-CS-94-102, Carnegie Mellon University.*

[Emery, 2000] Emery, N. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, 24:581–604.

[Eyelink, 2006] Eyelink (2006). http://www.eyelinkinfo.com.

[Gabor, 1946] Gabor, D. (1946). Theory of communication. *Journal of IEE*, 93: 492–457.

[Gee and Cipolla, 1994] Gee, A. H. and Cipolla, R. (1994). Determining the gaze of faces in images. *Image and Vision Computing*, 12(10):639–647.

[Gibson and Pick, 1963] Gibson, J. J. and Pick, A. D. (1963). Perception of another persons looking behaviour. *American Journal of Psychology*, 76:386–394.

[Heinzmann and Zelinsky, 1998] Heinzmann, J. and Zelinsky, A. (1998). 3-d facial pose and gaze point estimation using a robust real-time tracking paradigma. *Intern. Conf. on Automatic Face and Gesture Recognition.*

[Hubel and Wiesel, 1968] Hubel, D. and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Psychology*, 195:215–243.

[Hutchinson et al., 1989] Hutchinson, T., Jr., K. W., Reichert, K., and Frey, L. (1989). Human-computer interaction using eyegaze input. *IEEE Transactions on Systems, Man, and Cybernetics*, 19:1527–1533.

[Ji and Zhu, 2003] Ji, Q. and Zhu, W. (2003). Non-intrusive eye gaze tracking for natural human computer interaction. *MMI-Interactive*, 6.

[Krüger et al., 1997] Krüger, N., Pötzsch, M., and von der Malsburg, C. (1997). Determination of face position and pose with a learned representation based on labelled graphs. *Image Vision Comput.*, 15(8):665–673.

[Langton et al., 2004] Langton, S. R., Honeyman, H., and Tessler, E. (2004). The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception & Psychophysics*, 66(5):752–771.

[Langton et al., 2000] Langton, S. R., Watt, R., and Bruce, V. (2000). Do the eyes have it? cues to the direction of social attention. *Trends in Cognitive Science*, 4(2):50–59.

[Matsumoto and Zelinsky, 2000] Matsumoto, Y. and Zelinsky, A. (2000). An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. *4th Intern. Conf. on Face and Gesture Recognition*, pages 499–505.

[Phillips et al., 2000] Phillips, P., Moon, H., Rauss, P., and Rizvi, S. (2000). The feret evaluation methodology for face recognition algorythems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104.

[Rae and Ritter, 1998] Rae, R. and Ritter, H. (1998). Recognition of human head orientation based on artificial neural networks. *IEEE Transaction on Neural Networks*, 9(2):257–265.

[Sim et al., 2003] Sim, S., Baker, S., and Bsat, M. (2003). The cmu pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618.

[Sinha, 2000] Sinha, P. (2000). Last but not least. heres looking at you, kid. *Perception*, 29:1005–1008.

[Steifelhagen et al., 1997] Steifelhagen, R., Yang, J., and Waibel, A. (1997). Tracking eyes and monitoring eye gaze. *Proc. of the Workshop on Perceptual User Interfaces*, pages 98–100.

[Trucco and Verri, 1998] Trucco, E. and Verri, A. (1998). *Introductory Techniques for 3-D Computer Vision*. Prentice Hall.

[Wang et al., 2003] Wang, K., Wang, Y., Yin, B., and Kong, D. (2003). Face pose estimation with a knowledge based model. *IEEE Int. Conf. Neural Networks and Signal Processing*, pages 1131–1134.

[Yoo and Chung, 2005] Yoo, D. H. and Chung, M. J. (2005). A novel non-intrusive eye gaze estimation using cross-ration under large head motion. *Computer Vision and Image Understanding*, 98:25–51.

[Zhu and Ji, 2005] Zhu, Z. and Ji, Q. (2005). Robust real-time eye detection and tracking under variable lighting conditions and various face orientations. *Computer Vision and Image Understanding*, 98:124–154.

Modelling and Simulation of Spontaneous Perception Switching with Ambiguous Visual Stimuli in Augmented Vision Systems

Norbert Fürstenau

German Aerospace Center, Institute for Flight Guidance, Lilienthalplatz 7
D-38108 Braunschweig, Germany
norbert.fuerstenau@dlr.de

Abstract. A behavioral nonlinear dynamics model of multistable perception due to ambiguous visual stimuli is presented. The perception state is formalized as the dynamic phase variable v(t) of a recursive process with cosinuidal transfer characteristic which is created by superposition (interference) of neuronal mean fields. The two parameters μ = difference of meaning of alternative percepts and G = attention parameter, control the transition between unambiguous and ambiguous stimuli, e.g. from stimulus off to stimulus on, and attention fatigue respectively. Mean field interference with delayed phase feedback enables transitions between chaotic and limit cycle attractors v(t) representing the perception states. Perceptual reversals are induced by attention fatigue G(t) (~ adaptive gain g(v)) with time constant γ, and attention bias which determines the relative duration of the percepts. The coupled attention – perception dynamics with an additive stochastic noise term reproduces the experimentally observed Γ-distribution of the reversal time statistics. Mean reversal times of typically 3 – 5 s as reported in the literature, are correctly predicted if delay T is associated with the delay of 40 ms between stimulus onset and primary visual cortex (V1) response. Numerically determined perceptual transition times of 3 – 5 T are in reasonable agreement with stimulus – conscious perception delay of 150 – 200 ms [11]. Eigenfrequencies of the limit cycle oscillations are in the range of 10 – 100 Hz, in agreement with typical EEG frequencies.

1 Introduction

From the use of (transparent) monocular head mounted displays (HMD) in military helicopters it is well known that different visual input into both eyes may involve adverse perceptual and attentional effects like binocular rivalry which may lead to significant reduction of reaction times [1]. Laramee et.al. [2] determined more than 100% response time increase in a HMD based visual search task due to binocular rivalry and visual interference effects. Binocular rivalry is the spontaneous involuntary switching of conscious awareness between the different percepts corresponding to the different stimuli of both eyes [3]. It belongs to a larger class of cognitive multistability effects as observed with ambiguous stimuli such as perspective reversal (e.g. the Necker cube [4][5]) or figure-ground reversal. Also dynamic stimuli may give rise to cognitive multistability, e.g. ambiguous motion displays such as plaids as induced by

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 20 – 31, 2006.

moving groups of crossed lines [6]. By determining the correlation dimension, e.g. of corresponding time series, Richards et.al [7] found experimental evidence indicating common nonlinear dynamical cognitive processes underlying a surprisingly diverse range of visual phenomena such as cognitive multistability with ambiguous pictures, saccade intervals in visual search, and movie scene durations. Binocular rivalry, however appears to be a predominantly stochastic phenomenon, in agreement with other authors [10][28][29].

The present macroscopic model provides an approach for explaining the experimental finding that deterministic (even chaotic) as well as stochastic dynamics determines the measured reversal time statistics for different multistability phenomena [7]. It contributes to the ongoing controversial discussion on the deterministic [8] [9] versus purely stochastic character [10] of cognitive multistability. In agreement with the widely accepted view of recursive interactions between distant neural groups leading to conscious perception (e.g. [11]), the model assumes a reentrant process which appears to be related to the dynamical core hypothesis of Tononi et.al. [12]. The model relies on the mean field phase oscillator theory of coupled neuronal columns in the visual cortex [13]. The latter was used for describing the synchronization of neuronal oscillations as the physiological basis of dynamic temporal binding which in turn is thought to be cruical for the selection of perceptually or behaviorally relevant information [14][15][16][18]. Self oscillation of neuronal groups within columns and coupling between columns is excited when the external stimulus exceeds a certain threshold [13]. Single columns exhibit multistable characteristics of the neuronal mean field as function of the stimulus, similar to the present model. Within the phase synchronization theory phase locking between different groups of neurons is described by means of the circle (sin) map. Phase oscillator dynamics is the basis of the phase attractive circle map [17] which was used for describing human coordination dynamics as well as multistable perception.

A multistability model of Ditzinger & Haken [8] is based on the continuous polynomial dynamics of two separate coupled perception state equations. In accordance with the experimentally supported satiation (neuronal fatigue) hypothesis [4] spontaneous transitions between different percepts are induced by the time variation of two attention (control) parameters due to perception – attention coupling. Recently published experimental results of Nakatani et.al.[19] support the perception – attention coupling approach. The present model follows the perception – attention coupling and attention fatigue approach in [8]. It takes into account, however, the reentrant character of the neuronal processes [12] by including a finite delay time T, which results in limit cycle and chaotic attractor states defined as "percepts". In contrast to [8] a single differential – delay perception state equation together with a attention fatigue equation is formalized via the phase dynamics [13][17] of a recursive cosinuidal map as originating from superposition (interference) of neuronal mean fields. The stimulus ambiguity is quantified by a difference-of-meaning parameter which controls the stimulus-on/-off switching. The present model provides an explicit quantitative confirmation of the proposed catastrophe topology of the cognitive multistability dynamics [20].

In the following section 2 I describe the theoretical approach, with details of the recursive interference model in subsection 2.1, and an analysis of the stationary behavior in section 2.2. Results of computer experiments are presented in section 3, with simulated time series and attention – perception phase space plots in subsection 3.1

and a statistical analysis of the reversal time intervals in section 3.2 with different reentrant delay times. The results are discussed with respect to published experimental data and alternative theoretical models in section 4. A conclusion and outlook is presented in section 5.

2 Theory

2.1 The Recursive Mean Field Interference Model

As a kind of minimum architecture allowing for the emergence of discontinuous state transitions, I have proposed in previous papers coupling of the attention and perception dynamics via delayed phase feedback interference, and attention satiation[21][22]. Formally this is achieved analogous to multistable optical systems [23][24]. Interference with contrast μ is the superposition of (electromagnetic) fields $a_{0i} \exp\{ j(\omega t + \Phi_i)\}$, i = 1, 2, 3, ... with ω = circular frequency, Φ = phase, and a_0 = amplitude. The superposition yields extinction or amplification of each other, depending on the relative phase shift $\Delta\Phi = \Phi_1 - \Phi_2$. It may be compared with the phase shift between the coupled self - oscillating neuronal columns of the mean field theory [13]. A simplified block diagram is depicted in Figure 1.

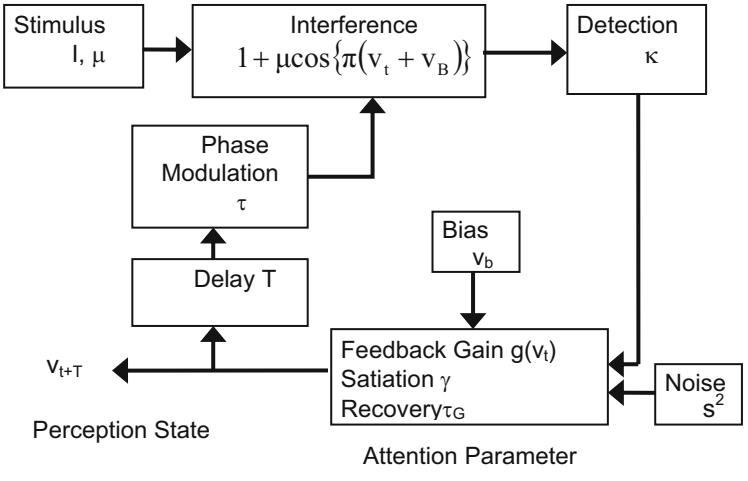


Fig. 1. Simplified block diagram of perception – attention recursive interference model. Symbols explained in the text.

The feedback loop describing the perception state order parameter dynamics v(t) may be compared with the reentrant thalamo - cortical and basal ganglia dynamics presented by Edelman [25] within his dynamical core hypothesis of consciousness: the interference and feedback gain blocks represent the thalamo – cortical complex and the sub – cortical basal ganglia with attention related tasks respectively. In the schematic an ambiguous stimulus with strength I and difference of meaning

μ (= interference contrast $0 \le \mu \le 1$) of the two possible percepts P1, P2, excites two corresponding hypothetical mean fields with phase difference $\Delta\Phi = \pi\, v_t$. Interference contrast μ depends on the coherence and polarization of the two fields. It creates the typical cosinuidal dependence of the output (= squared modulus of the sum of the field amplitudes, with detection conversion factor κ) on the phase difference $\Delta\Phi$ as mapping function. A recurrent process is established by feedback of the output after amplification (feedback gain g, attention bias v_b, satiation (fatigue) and recovery time constants γ and τ_G) with low pass filtering (time constant τ) and delay T into $\Delta\Phi$ via a phase modulation mechanism. As a quantitative estimate for T I chose the stimulus – primary visual cortex response delay ($\approx$ 40 ms, [11]). One possibility for phase feedback is frequency modulation of the input field [24], comparable to the stimulus induced modulation of the neuronal mean field limit cycle oscillations [13]. The normalized output $v_t = U_t / U_\pi$ of the feedback interference circuit with $d\Phi/dU = \pi/U_\pi$ and U ~ percept intensity = |superimposed percept field strength|^2 defines the perception state v as synergetic order parameter. The phase attractive circle map as proposed by Kelso et.al. [9][17] represents a similar recursive (discrete) phase oscillator mapping function which, however attempts to model multistability without feedback gain as control parameter and insofar cannot be mapped to the reentrant loops as proposed by Edelman [25].

According to Hillyard et.al. [26] stimulus-evoked neuronal activity can be modified by an attentional induced additive bias or by a true gain modulation (present model parameters v_b, G(t) ~ g). Increase of gain is correlated with increased blood flow through the respective cortical areas. Recent experimental evidence on perception – attention coupling with ambiguous stimuli was presented by Nakatani & van Leeuven [19] using EEG recording of frontal theta and occipital alpha bands and eye blink rate measurement. Accordingly in the present model, like in [8], the feedback gain serves as adaptive control parameter (~ attention parameter G) which induces the quasi - discontinuous transitions between the alternative stationary perception states P1 and P2, through attention satiation or fatigue [4]. A strongly damped (overdamped) feedback system is assumed with time constant τ >> coefficient of d^2v/dt^2 which is neglected. Formally the model is described by coupling a nonlinear 1st order differential delay equation for v(t) with a linear equation for the control parameter dynamics G(t). In a first approach to model the unavoidable random disturbances due to dissipative processes, a stochastic force L(t) with Gaussian white noise (variance s^2) is added to the attention equation G(t), similar to[8].

$$\tau \dot{v}_{t+T} + v_{t+T} = G\left[1 + \mu\cos\left(\pi\left(v_t + v_B\right)\right)\right] . \tag{1}$$

$$\dot{G}_t = \left(v_b - v_t\right)/\gamma + \left(G_{off} - G_t\right)/\tau_G + L_t \ . \tag{2}$$

The rhs. of equ. (1a) describes the conventional interference between two coherent fields. In what follows I assume the phase bias $v_B = 0$ mod 2. The attention parameter $G(t) = \kappa\, I\, g(t)/\, U_\pi$ with phase – voltage modulation factor $d\Phi/dU = \pi/U_\pi$ is the product of feedback gain g(t) and input (stimulus strength) I (=1 in what follows). The attention dynamics is determined by attention bias v_b (determining the relative preference of P1 and P2), satiation speed $1/\gamma$, recovery time constant τ_G and G_{off} = attention (gain) parameter for stimulus off, defined by $\mu = \mu_{off} < 0.18$ (see below).

The detailed block diagram of the model in Fig. 2 represents the highest hierarchy of an implementation in the graphical programming dynamical systems tool Matlab-Simulink:

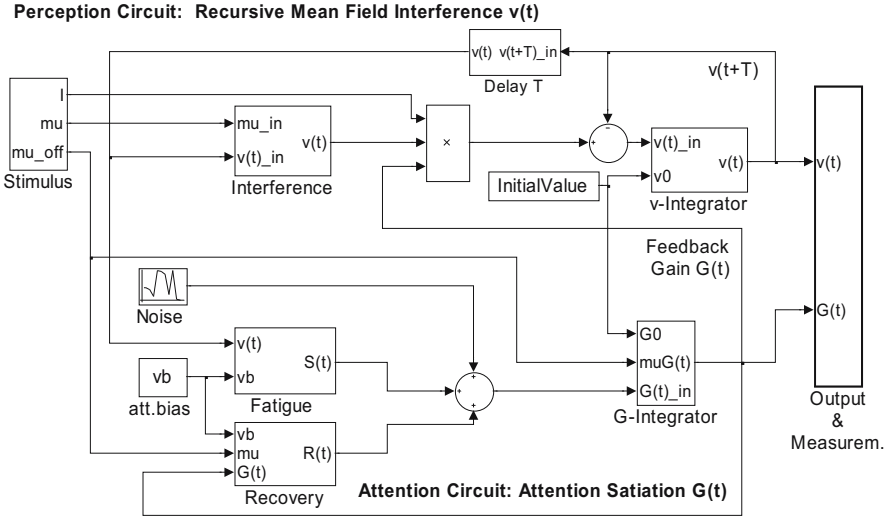


Fig. 2. Matlab – Simulink implementation (highest hierarchy level) of recursive interference model displaying subroutines (blocks) of the reentrant loops of *perception circuit v(t)* in the upper half and *attention circuit G(t)* in the lower half of the diagram. *Stimulus of strength I and difference of meaning mu (=μ)* are fed into perception circuit as control parameters with *interference term* and *integrator loop with time constant* τ. *Attention circuit G(t) with satiation term (fatigue) (vb – v(t))/*γ *and recovery term (Goff – G(t))/*τ_G controls as gain factor the perception dynamics. v(t) and G(t) output into the *data evaluation block* at the right.

2.2 Stationary Solutions of the Recursive Interference Equation

Two types of instabilities are observed with recursive systems described by equation (1): period doubling and node bifurcation. Figure 3 depicts the stationary solutions (dv/dt = 0) including period doubling up to period 8, $v_{t+iT} = v_t = v^*$, i = 1, 2, 4, 8.

Period doubling pitchfork bifurcations are observed on both positive slope regions. The graph yields the control parameter values at the first three bifurcation points providing a first approximation to the Feigenbaum constant $\delta_\infty = 4.6692$ via $\delta^{\mathrm{I}}{}_\infty \approx (G_2 - G_1)/(G_3 - G_2)$. The period doubling behavior proves that within certain parameter ranges (μ, τ) any system noise has chaotic contributions. This is confirmed by numerical evaluation of the Lyapunov coefficient [22].

In [21] I have shown the stationary solution of v*(μ, G) to exhibit a topology similar to a cusp catastrophe. This finding agrees with a proposal of Poston & Stewart [20] who developed a qualitative deterministic model of cognitive bistability based on catastrophe theory.

At the critical value, $\mu_n = 0.18$, node bifurcation is observed and the slope of the stationary system state v* as function of G becomes infinite. For $\mu < \mu_n$ both percepts are fused into a single meaning. For $\mu > \mu_n$ the stationary solution v*(G) becomes multivalued. For maximum contrast $\mu = 1$ the horizontal slope $(dG / dv)^{-1} = 0$ yields $v_i^{\infty} = 2i - 1$, i = 1,2,3,… as stable perception levels in the limit $G \to \infty$.

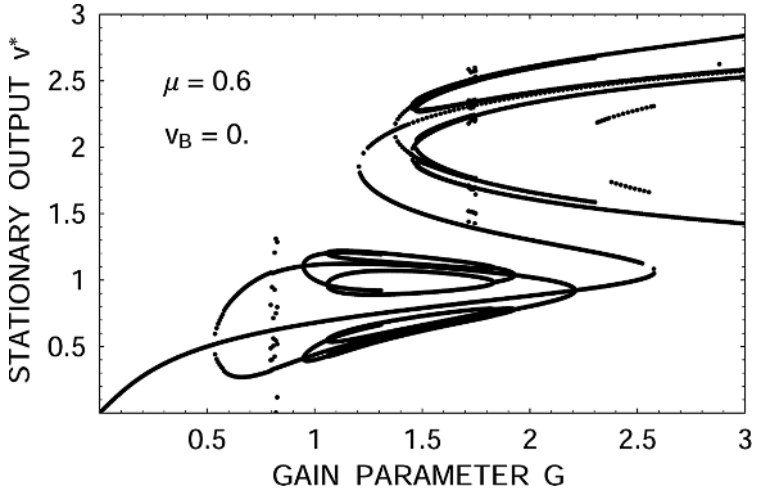


Fig. 3. Stationary solutions of equation (1) exhibiting v* - hysteresis and period doubling bifurcations. Percept 1 (P1) = lower positive slope v*-level; Percept 2 (P2) = higher v*-level.

Node bifurcation is required for explaining the existence of ambiguous perception within the present model. Under increasing stimulus strength I or feedback gain g the stationary (1st order) perception state v* jumps discontinuously from P1 to P2 at the turning points of the S-shaped hysteresis curve (= extrema of the inverse curve G(v*)). The transition of P2 back to P1 occurs at a lower stimulus or gain parameter (~attention) value due to the hysteresis. The width of the instable negative slope section and the multivalued G – range is controlled by μ. A similar hysteresis is observed for the coupling constants of columns of the visual cortex within the neuronal mean field theory [13].

A linear stability analysis of equation (1) yields the regions of instability, i.e. limit cycle and chaotic oscillations as dependent on the ratio τ/T of damping time constant and feedback delay time [21][22]. Eigenfrequencies $\beta = 2\pi f$ are obtained via $\beta\tau = -\tan(\beta T)$. The analytical approximation for $\tau << T$ yields $f \approx f_0\, i / (1 - \tau/T)$, i = 0, 1, 2, … with $f_0 = 1 / 2T$, i.e. half of the inverse feedback delay time. With T = 40 ms (delay between stimulus onset and V1 - response, see above) we obtain f_0 = 12.5 Hz. With large damping time τ of the order of T, period doubling oscillations are suppressed.

3 Computer Experiments

3.1 Simulated Perception – Attention Dynamics

In this section I present numerical evaluations of the coupled differential – delay equations (1, 2) as obtained with the dynamical systems tool Matlab – Simulink (solver ode23tb for stiff problems). Figure 4 shows time series v(t) and G(t) for I = 1, $\mu = 0.6$, T = 2 T_S = 40 ms, time scale in units of the simulation interval T_S, $\tau/T = 0.1$, $\gamma = 60$, $\tau_G = 500$, attention bias $v_b = 1.5$, noise variance $s^2 = 0.001$, with stimulus – off sections ($\mu_{off} = 0.1$, $G_{off} = 1.5$) at the beginning and end of the time series.

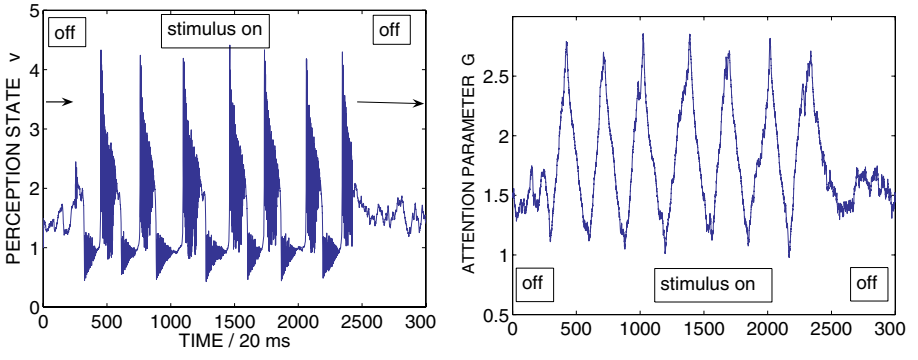


Fig. 4. Numerical evaluation of equations (1)(2) with stochastic noise variance $s^2 = 0.001$. Left: *perception state time series v(t)*; right: *attention parameter G(t)*; Stimulus off ($\mu = 0.1$) during initial and final simulation phases. See text for simulation parameters.

The time series of the perception state v(t) exhibits the spontaneous transitions between stationary perception states P1 (near $v^* \approx 1$) and P2 (near $v^* \approx 2.5$) with the expected superimposed limit cycle and chaotic oscillations. The transition time between P1 and P2 is of the order of 5 - 10 $T_S \approx 100$ - 200 ms, in reasonable agreement with the time interval between stimulus onset and conscious perception [11].

The phase space plot v vs. G in Fig. 5 exhibits separated regions of the stimulus – off and stimulus – on (P1 and P2) states with trajectories of fast oscillations superimposed on the slow satiation (fatigue) dynamics.

The reversal time period is determined by the slow G(t) variation with satiation and recovery time constants γ, τ_G , with an absolute scale given by $T = 2T_S$. Limit cycle oscillations and deterministic chaos within P1, P2 is a characteristic of the individual perception states and has its origin in the finite delay time T.

The effects of decreasing T and variation of attention bias v_b are depicted in Figure 6. The phase space trajectories in the left plot clearly show that with zero delay (T = 0) the limit cycle and chaotic oscillations vanish which are superimposed on the stationary perception states of the time series of Fig. 4 and the corresponding hysteresis loop in Fig. 5.

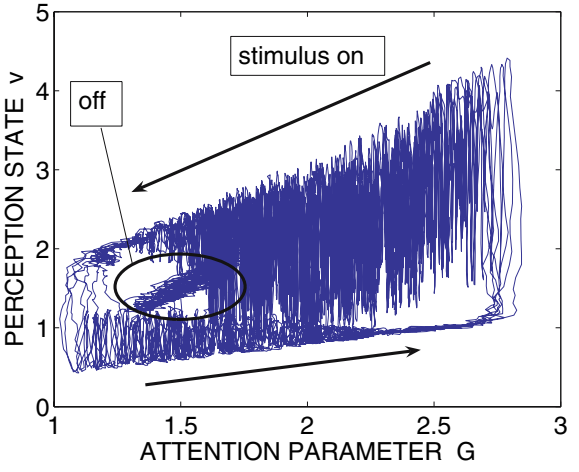


Fig. 5. Phase space trajectories v vs. G of the time series in Fig. 4 with the two control parameter values $\mu = 0.1$ (=stimulus off) and $\mu = 0.6$ (= stimulus on)

The v – G – phase space plots exhibit a clear separation of stimulus – off ($\mu = 0.1$) and – on ($\mu = 0.6$) states due to the node bifurcation at $\mu = 0.18$. The scattering of the reversal time period, however as indicated by the scattering of the P1 – P2 transitions, appears not to be significantly effected.

The right graph in Fig. 6 shows how the attention bias v_b determines the relative dominance of one of the two percepts. In this example (noise variance $s^2 = 0.001$), after stimulus on (percept 0 with $\mu = 0.1$ is switched to $\mu = 0.6$) the offset $v_b = 0.9$ forces the perception to iterate to the lower perception state P1 with suppression of P2. Other parameters are the same as in Figure 5. v_b may be used as a control parameter to model experimental results with perception biased towards one of the two percepts as reviewed in [3].

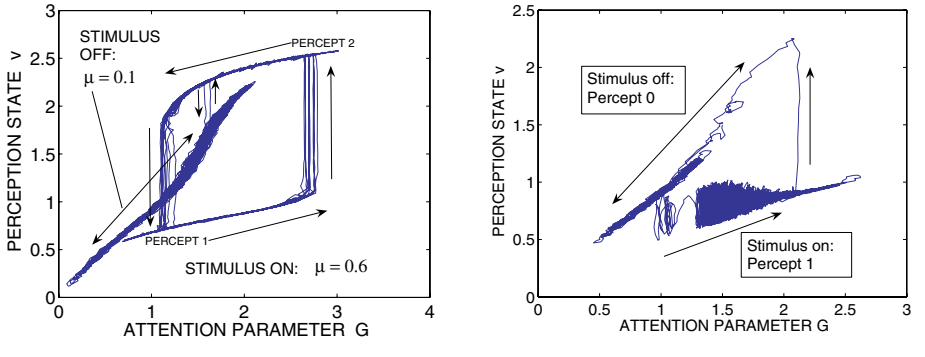


Fig. 6. Left: Perception – attention dynamics with zero feedback delay (T = 0) exhibiting vanishing limit cycle and chaotic oscillations. Phase space plot of perception – attention dynamics with attention bias $v_b = 0.9$ and T = 2.

3.2 Reversal Time Statistics

Figure 7 depicts the relative frequencies of the perceptual duration times of simulations as obtained by averaging 10 time series of N = 50000 iterations each, with T = 2, $\tau = 0.2$, $\gamma = 60$, $\tau_G = 500$, $v_b = 1.5$ and $s^2 = 0.03$. Time series differ by noise (random number generator) seed value and perception state initial value v(t=0).

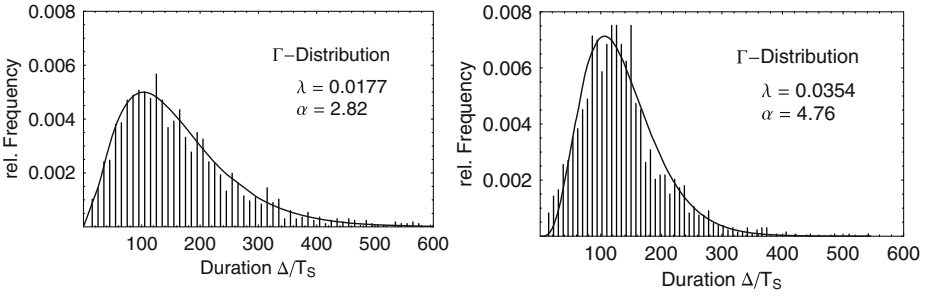


Fig. 7. Relative frequencies of perceptual duration time Δ, in units $T_S = T/2$. Simulation parameters $\mu = 0.6$, $v_b = 1.5$, $\tau/T = 0.1$, $\gamma = 60$, $\tau_G = 500$, $s^2 = 0.03$. Fit with Γ-distribution (solid line). Left: Percept P1 ($v^* \approx 1$), Right: Percept P2 ($v^* \approx 2.5$).

Plotted are the two distributions of the perceptual durations Δ(P1) of percept 1 and Δ(P2) of percept 2. As suggested by a number of experimental results (e.g. [5][19] [27][29]) the relative frequencies are fitted by a Γ – distribution as probability density with shape parameter α and scale parameter λ. Mean and variance are given by $\Delta_m = \alpha/\lambda$ and $\sigma^2 = \alpha / \lambda^2$ respectively. For percept P1 and P2 mean and standard deviation are respectively $\Delta_m = 159\ T_S$, $\sigma = 95\ T_S$ and $\Delta_m = 134\ T_S$, $s = 62\ T_S$. The ratios $s/\Delta_m \approx 0.4 – 0.6$ are in good agreement with the experimental findings reported in the literature.

4 Discussion

In contrast to [21] with simulation of a purely deterministic recursive process, and in agreement with [22] the addition of the stochastic attention noise L(t) leads to a significant increase of the variance, whereas the mean reversal times remain roughly the same as without noise, indicating the dominating influence of the deterministic dynamics on Δ_m. The limit cycle and chaotic contributions to the total variance in the present delay – differential model is significantly smaller as compared to the recursive approximation [22] without differential term. A dominating influence of stochastic processes was reported for binocular rivalry [7][28][29] where no significant chaotic contribution in the reversal time statistics was detected. It appears that with the given choice of model parameters the deterministic "noise" of the perception states P1, P2 is hardly detectible by analysis of reversal time measurements because the variance of the latter is dominated by the stochastic (attention) contribution.

Table 1 lists results of Γ-distribution fits to the relative frequencies of perceptual duration times as obtained from computer simulations with delay times $T/T_S = 0, 1, 2$ (noise variance $s^2 = 0.03$).

Table 1. Shape parameter, mean (in seconds), relative standard deviation and correlation coefficient for Γ – distribution fit to relative frequencies of perceptual duration times (separate for P1, P2) as obtained with different delay times

	T/T_S	α	Δ_m / s	σ / Δ_m	R^2
P2	0	6.6	1.89	0.39	0.97
P2	1	7.5	2.62	0.37	0.98
P2	2	5.0	2.64	0.45	0.96
P1	0	3.5	2.69	0.53	0.95
P1	1	3.2	3.27	0.56	0.94
P1	2	2.5	3.22	0.63	0.93

Fitting of the data by means of a Γ – distributions for all T values yields correlation coefficients which show that only 3 – 7 % of the total scattering is not explained by the Γ – density function. The relative standard errors of the shape and scale parameters α and λ respectively are around 3% for all simulations. An important result is the fact that even with zero delay (T = 0) the mean reversal time $\Delta_m(T=0)$ and the variance σ^2 is of the same order of magnitude as with finite delay. This indicates that the unavoidable contribution of the deterministic limit cycle oscillations and chaos (due to T > 0 for any realistic nonlinear physical system) to the reversal time variance is small as compared to the stochastic noise, in agreement with [28] and [10]. The shape parameter α is of interest with regard to the question if the dynamics is dominated by a Poisson process as suggested by Levelt [29]: in this case the α's should cluster around natural numbers, a prediction which was verified in experiments by Murata et.al. [30].

5 Conclusion

A behavioral recursive nonlinear and stochastic phase oscillator model of spontaneous perceptual switching is presented which is related to previously published dynamical models [8][9]. The model is expected to help clarifying attention and perception related problems of augmented vision based human machine interfaces (e.g. [1][2]). The perception state is assumed to originate from interference between stimulus induced phase synchronized perception fields as proposed by the neuronal mean field theory [13], and it is coupled to the dynamics of an attention control parameter. Experimental results of Nakatani & van Leeuwen [19] support the assumption that attentional effort which is expressed by eye blinking and saccade frequencies controls switching rates. By associating feedback delay time T with the stimulus onset - primary visual cortex (V1) response delay of ~ 40 ms [11] absolute values of mean perceptual duration times of $\Delta_m \approx 3$ s are obtained, in reasonable agreement with published experimental results (1 – 10 s, e.g. [5][27]).

The large inter – subject variations of Δ_m can be modeled by suitable choice of contrast (difference of meaning) parameter μ, satiation (fatigue) and recovery time constants γ and τ_G respectively, and noise variance s^2. The relative duration of dominance vs. suppression times is determined by the attention bias parameter v_b. The magnitude of limit cycle and chaotic oscillations with eigenfrequencies < 100 Hz is controlled by the ratio τ/T of perceptual damping time constant and delay time. Because the reversal time statistics is only weakly dependent on T it is concluded that the limit cycle and chaotic oscillations which are superimposed on the stationary perception states, also contribute only weakly to the reversal time statistics, in agreement with results of other authors [28][10]. The present model thus supports the proposal of Poston & Stewart [20] of a deterministic catastrophe topology as the basis of the perception reversal dynamics, with the higher moments of the statistics determined by a stochastic process which in certain cases (binocular rivalry) hides the deterministic contribution. Ongoing work aims at quantifying the amount of long range correlations of the time series by evaluating the Hurst parameter as proposed by Gao et.al.[31] and the reproduction of the proposed Poisson process [29][30] superimposed on the deterministic dynamics by generating a sufficient statistical basis of shape parameters α of the Γ – distributions.

Acknowledgement

I am indebted to Monika Mittendorf for help in writing the Mathematica and Matlab code of the numerical simulation and analysis software and in performing the statistical analysis. I would like to thank H. Nakatani of Riken Brain Science Institute for information on some recent experimental results.

References

1. Peli, E.: Visual issues in the use of a head-mounted monocular display. Optical Engineering 29 (1990) 883-892
2. Laramee, R.S., Ware, C.: Rivalry and Interference with a Head Mounted Display. ACM Transactions on Computer-Human Interactions 9, (2002) 238-251
3. Engel, A.K., Fries, P., König, P., Brecht, M., Singer, W.: Temporal binding, binocular rivalry, and consciousness. Consciousness and Cognition 8 (1999) 128–151
4. Orbach, J., Ehrlich, D., Heath, H.A: Reversibility of the Necker Cube: An examination of the concept of satiation of orientation, Perceptual and Motor Skills 17 (1963) 439-458
5. Borsellino, A., de Marco, A., Allazetta, A., Rinesi, S., Bartolini, B.: Reversal time distribution in the perception of visual ambiguous stimuli. Kybernetik 10 (1972) 139–144
6. Hupe, J.-M. , Rubin N.: The dynamics of bistable alternation in ambiguous motion displays: a fresh look at plaids, Vision Research 43, (2003) 531–548
7. Richards, W., Wilson, H.R., Sommer, M.A.: Chaos in percepts. Biol. Cybern. 70 (1994) 345-349
8. Ditzinger T., Haken H.: A Synergetic Model of Multistability in Perception. In: Kruse, P., Stadler, M. (eds.): Ambiguity in Mind and Nature. Springer-Verlag, Berlin: (1995) 255-273.

9. Kelso J.A.S., Case P., Holroyd T., Horvath E., Raczaszek J., Tuller B., Ding M.. Multistability and metastability in perceptual and brain dynamics. In: Kruse, P., Stadler, M. (eds.): Ambiguity in Mind and Nature. Springer-Verlag, Berlin: (1995) pp. 255-273
10. Merk, I. L. K., Schnakenberg, J.: A stochastic model of multistable perception. Biol.Cybern. 86, (2002) 111-116
11. Lamme, V.A.F.: Why visual attention and awareness are different. Trends in Cognitive Sciences , (2003) 12–18
12. Tononi, G. , Edelman, G.M.: Consciousness and Complexity. Science 282 (1998) 1846-1851
13. Schuster, H.G., Wagner, P.A.: A Model for Neural Oscillations in the Visual Cortex: 1. Mean field theory and the derivation of the phase equations. Biol. Cybernetics 64 (1990) 77-82
14. Blake, R., Logothetis, N.K.: Visual competition. Nature Reviews / Neuroscience 3 (2002) 1– 11
15. Engel, A.K., Fries, P., Singer, W.: Dynamic Predictions: Oscillations and Synchrony in Top-Down Processing. Nature Reviews Neuroscience, 2 (2001) 704–718
16. Engel, A.K., Fries, P., König, P., Brecht, M., Singer, W.: Temporal binding, binocular rivalry, and consciousness. Consciousness and Cognition 8 (1999) 128–151
17. deGuzman, G. C., Kelso, J. A. S.: Multifrequency behavioral patterns and the phase attractive circle map. Biological Cybernetics 64 (1991) 485–495
18. Srinavasan, R., Russel, D.S., Edelman, G M, Tononi, G.: Increased synchronization of magnetic responses during conscious perception, J. Neuroscience 19 (1999) 5435 – 5448
19. Nakatani H., van Leeuwen C.: Individual differences in perceptual switching rates; the role of occipital alpha and frontal theta band activity. Biol. Cybern. 93 (2005) 343-354
20. Poston T., Stewart, I.: Nonlinear Modeling of Multistable Perception. Behavioral Science 23, (1978) 318-334
21. Fürstenau N.: Nonlinear dynamics model of cognitive multistability and binocular rivalry, Proceedings IEEE 2003 Int. Conf. on Systems, Man and Cybernetics, IEEE cat. no. 03CH37483C (2003) 1081-1088
22. Fürstenau N.: A chaotic attractor model of cognitive multistability. Proceedings IEEE Int. Conf. on Systems, Man and Cybernetics, IEEE cat. no. 04CH37583C (2004) 853- 859
23. Watts, C., Fürstenau, N.: Multistable fiber-optic Michelson Interferometer exhibiting 95 stable states. IEEE J. Quantum Electron 25 (1989) 1-5
24. Fürstenau, N.: Bistable fiber-optic Michelson interferometer that uses wavelength control. Optics Letters 16 (1991) 1896–1898
25. Edelman, G.: Wider than the Sky. Penguin Books (2004) pp. 87-96
26. Hillyard, S.A., Vogel, E.K. Luck, S.J.: Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. In: Humphreys, G.W. Duncan, J.& Treisman, A. (eds.): Attention, Space, and Action. Oxford University Press (1999) 31-53
27. Zhou, Y.H., Gao, J.B., White, K.D., Merk, I., Yao, K.: Perceptual dominance time distributions in multistable visual perception. Biol. Cybern. 90 (2004) 256-263
28. Lehky S. R.: Binocular rivalry is not chaotic, Proc. R. Soc. Lond. B 259 (1995) 71–76
29. Levelt, W.J.M.: Note on the distribution of dominance times in binocular rivalry. Br. J. Psychol. 58 (1967) 143-145
30. Murata, T., Matsui, N., Miyauchi. S., Kakita, Y., Yanagida, T.: Discrete stochastic process underlying perceptual rivalry. NeuroReport 14 (2003) 1347-1352
31. Gao, J.B., Merk, I., Tung W W, Billok V, White, K.D., Harris J G, Roychowdhury V P.: Inertia and memory in ambiguous visual perception. preprint Phys. Rev. Lett., (2005) submitted

Neural Network Architecture for Modeling the Joint Visual Perception of Orientation, Motion, and Depth

Daniel Oberhoff, Andy Stynen, and Marina Kolesnik*

Fraunhofer Institute for Applied Information Technology Schloss Birlinghoven
53754 Sankt Augustin
Germany
marina.kolesnik@fit.fraunhofer.de
http://viswiz.imk.fraunhofer.de/~marina

Abstract. We present a methodology and a neural network architecture for the modeling of low- and mid-level visual processing. The network architecture uses local filter operators as basic processing units which can be combined into a network via flexible connections. Using this methodology we design a neuronal network that models the joint processing of oriented contrast changes, their motion and depth. The network reflects the structure and the functionality of visual pathways. We present network responses to a stereo video sequence, highlight the correspondence to biological counterparts, outline the limitations of the methodology, and discuss specific aspects of the processing and the extent of visual tasks that can be successfully carried out by the suggested neuronal architecture.

1 Introduction

Substantial neurophysiological evidence suggests that low- and mid-level visual processing is based on local operators. This means that the outcome of the processing at a certain image location depends only on a neighborhood of that point and no global processing is employed. The local processing is also supported by retinotopic mapping [Schwartz, 1977], which preserves the topographical relative location of neurons so that neighboring regions on the retina project to nearby regions in the visual cortex. It becomes logical that the visual processing is performed by local connections between neurons arranged in systematic spatial organization preserving neighborhoods between image points. This local processing is modeled by convolutions of the image with filter masks of finite spatial and temporal extent modeling spatiotemporal receptive fields. Choosing suitable parameter for these filters and combining their outputs enables us to generate a population code for directed motion and local contrast orientation. Systematic arrangement of filter outputs allows to preserve neighborhoods between image points along dimensions of the filter parameters (i.e. orientation,

* Supported by the European Commission (Contract no.:12963 NEST, project MCCOOP).

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 32–39, 2006.

depth etc.). Further incorporation of binocular disparity completes the joint encoding of orientation, motion and depth supported by neurophysiological studies [Ohzawa et al., 1996, Ohzawa et al., 1996, Anzai et al., 2001]. Furthermore cortical layout suggests that cells tuned to different features at spatially close image points cooperate strongly [Koulakov and Chklovskii, 2001, Swindale, 1998]. This suggests that interactions between cell populations tuned to different features already take place during low level processing.) It can be assumed that local interactions are important for effective perception of the relevant visual information in our environment.

2 Processing Networks

Linear filters and mathematical point operators can be combined to construct processing networks for various visual tasks. Since different tasks often yield different models, a software framework has been developed, which allows the efficient design, implementation and testing of such models. This framework, written in c++, is portable due to the use of open-source libraries such as Qt, fftw, and blitz++. It implements linear filters, mathematical point operations, multidimensional matrices encoding feature dimensions on top of the two spatial image dimensions. Arbitrary connections can be set up between these components to guide signal flow. A GUI allows a user to construct the computational network such as the one shown in Figure 1 and to change its parameters without having to recompile the software system. Different algorithms are being designed and implemented into this framework with respect to the different processing channels and the fusion of these channels on a low processing level.

The symbols/nodes shown in Figure 1 represent different parts of neuronal processing layers such as connectivity patterns (modeled by convolution filters), synapses (modeled by linear, geometric, and multiplicative combinations), activation, and small neural sub-circuits (such as the and-like combinations used for binocular edge fusion, see [Neumann and Mingolla, 2002]).

2.1 Preprocessing

It is well known that the first stage of image processing in the mammalian visual system takes place even before any signals enter the brain. Cells positioned on the retina fetch signals from rods and cones and perform center- surround-processing of image intensity in time and space as well as color (hence: extracting contrast information). In a computational approach this process corresponds to a band pass filtering of an image in space and time (color is not yet processed by the current network) with filter kernels of the well known Mexican hat shape. This is implemented in the framework as a common input stage (Figure 1,a). After this stage, different processing channels are designed and modeled.

2.2 Stereo Motion

Based on [Adelson and Bergen, 1985], [Shmuel and Grinvald, 1996], [Sabatini et al. 2003] and [Sabatini and Solari, 2004] a motion perception channel (Figure 1,d) has

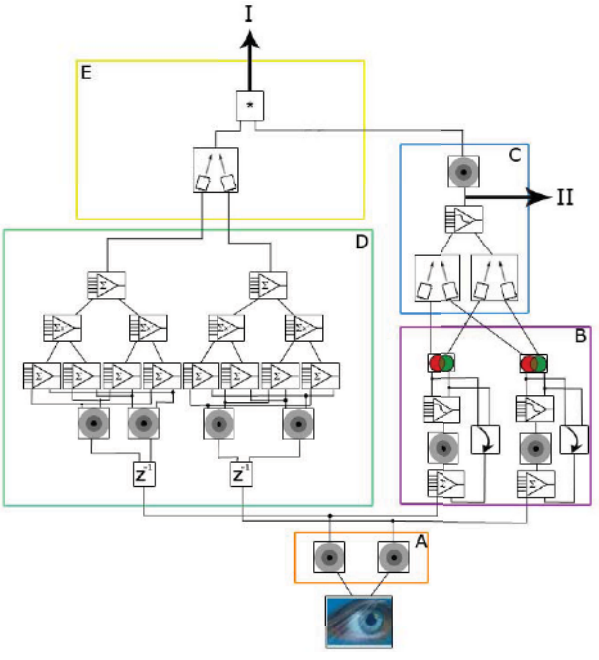


Fig. 1. Processing network combining stereo motion and contour processing. The eye symbol denotes binocular image input. Arrows denote system output:(I) delivers binocular moving contours and (II) delivers static binocular contours. Linear, geometric, and product combinations are denoted by the respective mathematical symbols. The concentric circles denote a convolution filter operation. The partially overlapping circles and the binocular fusion units both use a nonlinear and-like function, operating on on- and off-contrast and edge information from left and right eye respectively. The curved arrow is a placeholder for one of the two feedback mechanisms discussed in Section 2.3. (a) Common input stage performing spatial bandpass filtering. (b) Contour extraction at zero crossings of bandpass filtered signal employing elongated Gaussian filtering and feedback connections based on [Kolesnik and Barlit, 2003]. (c) Stereo fusion of edges as in [Cao and Grossberg, 2004] and subsequent blur prior to the masking operation. (d) Implementation of motion sensitive receptive fields as in [Sabatini et al., 2003]. (e) Stereo fusion of motion signals.

been implemented which takes the preprocessed stereo images from the common input stage. This channel models motion sensitive receptive fields, one for each eye, by sequentially processing the images by quadrature pairs of temporal and spatial filters and suitably combining the resulting channels.

2.3 Contour Detection

Contour detection is based on non-linear combination of oriented on- and off-subfields (see [Kolesnik and Barlit, 2003], and [Neumann et al., 1999]). Furthermore we employcooperative mechanisms to improve the contour image and

achieve a certain level of contrast invariance (Figure 1,b). Here we have the choice between the feedback mechanism employed in [Kolesnik and Barlit, 2003] and anisotropic diffusion (see e.g. [Perona and Malik., 1990]). Where the latter is more stable, the former yields better contrast invariance. Both mechanisms enhance continuous contours.

2.4 Binocular Fusion

Binocular fusion is executed on sparse edges maps obtained from the contour module by simple and-like combination of edge maps from the left and right eye at different disparities (see [Cao and Grossberg, 2004]). The sparsity of the edge map guarantees good performance on vertical edges while no unique depth can be assigned to horizontal edges.

3 Multichannel Fusion

We believe that cooperation of channels tuned to different features on all levels of visual processing contributes toward the final percept of our environment. To implement an example for task oriented multichannel interaction we multiplicatively combined the outputs of the binocular motion and contour channels. The output of this operation yields per pixel information of depth and motion state on contours in a population code which allows easy separation of the contours of consistently moving objects from the background by selecting its motion and depth range. We stress that the basic building blocks of the presented network are not new but that task-oriented performance can be enhanced greatly through the cooperation of feature channels.

4 Results

To test the network performance for figure-ground-separation we generated a synthetic stereo sequence of a moving rendered cube in front of a static background. In Figure 2 we demonstrate separation by disparity and by fused motion information. Comparing the results of segmentation with (second row) and without cue fusion (third row) clearly demonstrates the benefit of cue fusion. It also demonstrates how the use of multi-channel interaction can overcome the shortcomings of the individual channels (here the failures to assign depth information to horizontal edges). The fused motion signal shows some spurious responses for parts of the background contours. This is a consequence of the apparent motion signal, generated in the receptive fields of the motion perception channel, yielded by the background appearing and disappearing due to the movement. For the motion sequence the feedback in the contour detector was switched off because it would smear out moving edges.

To demonstrate depth selectivity on a natural image we applied the network on stereoscopic images of a hand (Figure 3). It shows how the depth selectivity is effective only for vertical lines. Here we applied the feedback in the contour detector to enhance contour information and make it more contrast invariant.

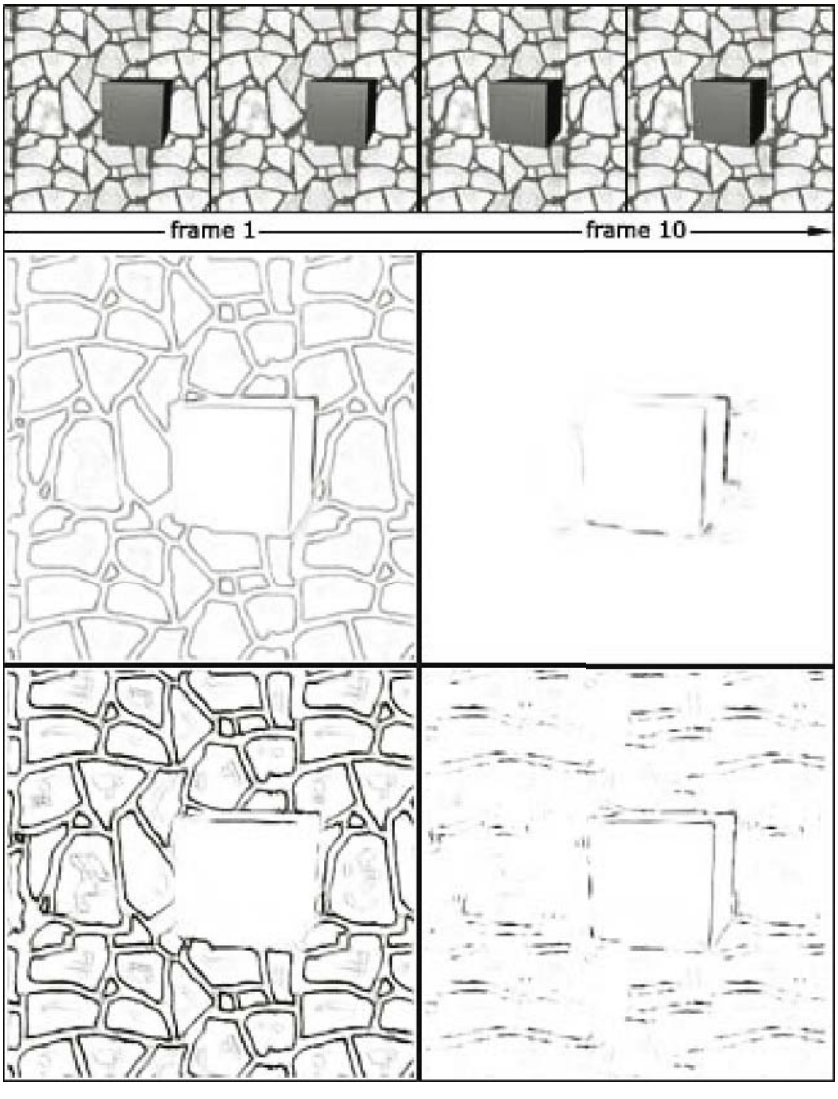


Fig. 2. Top: Two stereo frames of a rendered 3d cube moving in front of a stone wall. Middle Left: static contours of the 6th frame of this animation (left input). Middle Right: moving binocular contours of the sixth frame. The cube is nearly completely isolated from the background. Note that some of the background contours close to the cube enter the fused image. This is caused by occlusion related spurious responses in the motion sensitive receptive fields. Bottom Left: Stereo contours of the sixth frame with disparities between -3 and -1 pixels. Bottom Right: Stereo contours of the sixth frame with disparities between 0 and 3 pixels. Not that discrimination of contours by depth information is only effective on vertical contours.

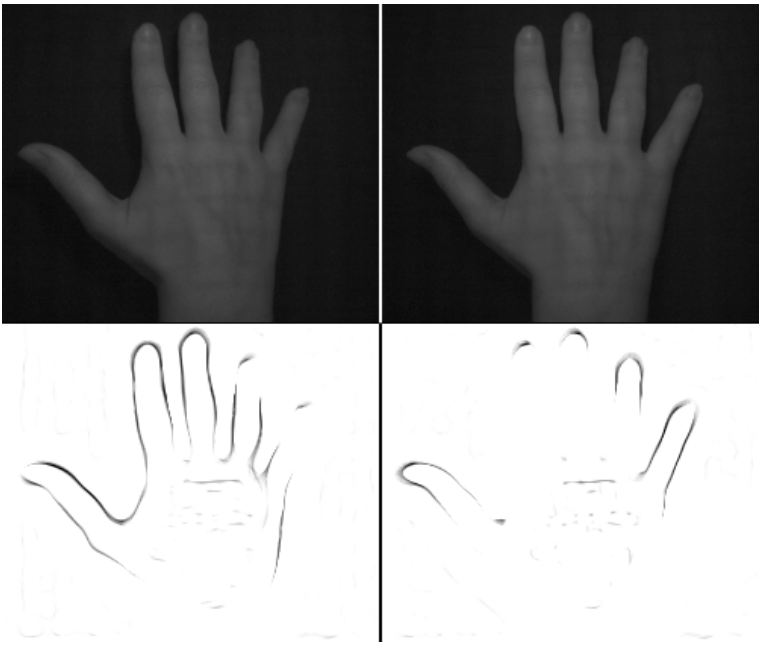


Fig. 3. Top:left and right input images. Bottom left: combined static output for disparities between 0 and 1 pixels. Bottom right: combined static output for disparities between 2 and 4 pixels. Unique detection of disparity is only possible for vertical edges, thus horizontal or near horizontal edges show up for several disparities. The feedback mechanism from [Kolesnik and Barlit, 2003] was employed to enhance contours and achieve contrast invariance.

5 Conclusions

We presented the neuronal network architecture for simultaneous processing of oriented edges, motion and binocular depth. The network uses only linear local operators and in this sense represents a plausible model of low-level visual processing on the retina and in the visual cortex. The neuronal network in its current form generates the complex response, which is an initial stage in encoding the perception of form, depth and motion. This stage can be formally associated with response properties of simple cells, which as argued in the neurophysiological study of [Ohzawa et al., 1996] carry information potentially useful for the perception of all these functions. It further enhances this information by combining information from different feature channels. In the presented network this yields superior performance in the task of segmenting an object from a textured background above that of the individual components.

The presented network can generate a stereoscopic contour image with depth and motion information on all edges. It cannot, however, generate contour

information from pattern- or depth-discontinuities that are not accompanied by a strong change in luminance. Cooperation of processing channels only takes place at a relatively high level. It should be beneficial to share the initial stage by establishing a "simple cells' level" providing a large pool of data from which in a next stage a "complex cell level" can extract relevant depth, motion and contour information. Specifically, inputs from the initial simple cells can be combined in an output that retains selectivity to a smaller set of parameters of interest, but is insensitive to all other parameters. For instance, by using an explicitly binocular input stage, thereby performing stereo fusion on a lower level, it should also be possible to reliably detect motion in depths for image contours.

The current neural network forms a unified source of information from which higher order processing units can draw to serve a multitude of visual functions including selective form and object perception. Our future work will be directed at the incorporation of a form of associative memory to facilitate recognition and attention to a smaller set of parameters of interest. This would make the presented approach useful for real life applications.

References

[Adelson and Bergen, 1985] Adelson, E. H. and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc.*, 2:284.

[Anzai et al., 2001] Anzai, A., Ohzawa, I., and Freeman, R. D. (2001). Joint-encoding of motion and depth by visual cortical neurons: neural basis of the pulfrich effect. *Nature Neurosci.*, 4(5):513–518.

[Cao and Grossberg, 2004] Cao, Y. and Grossberg, S. (2004). A laminar cortical model of stereopsis and 3d surface perception: Closure and da vinci stereopsis. *CAS/CNS Technical Report.*

[Kolesnik and Barlit, 2003] Kolesnik, M. and Barlit, A. (2003). Iterative orientation tuning in v1: a simple cell circuit with cross-orientation suppression. *Lecture Notes in Computer Science.*

[Koulakov and Chklovskii, 2001] Koulakov, A. A. and Chklovskii, D. B. (2001). Orientation preference patterns in mammalian visual cortex: A wire length minimization approach. *Neuron*, 29:519–527.

[Neumann and Mingolla, 2002] Neumann, H. and Mingolla, E. (2002). *Handbook of brain theory and neural networks*, chapter Contour and Surface Perception. MIT Press.

[Neumann et al., 1999] Neumann, H., Pessoa, L., and Hanse, T. (1999). Interaction of on and off pathways for visual contrast measurement. *Biol. Cybern.*, 81:515–523.

[Ohzawa et al., 1996] Ohzawa, I., DeAngelis, G. C., and Freeman, R. D. (1996). Encoding of binocular disparity by simple cells in the cat's visual cortex. *J. Neurophys.*, 75(5):1779–1805.

[Perona and Malik., 1990] Perona, P. and Malik., J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639.

[Sabatini and Solari, 2004] Sabatini, S. P. and Solari, F. (2004). Emergence of motion-in-depth selectivity in the visual cortex through linear combination of binocular energy complex cells with different ocular dominance. *Neurocomp.*, 58-60:865.

[Sabatini et al., 2003] Sabatini, S. P., Solari, F., Cavalleri, P., and Bisio, G. (2003). Phase-based binocular perception of motion in depth: Cortical-like operators and analog vlsi architectures. *J. Appl. Sig. Proc.*

[Schwartz, 1977] Schwartz, E. L. (1977). Spatial mapping in the primate sensory projection: analytic structure and relevance to perception. *Biological Cybernetics*, 25(4):181–194.

[Shmuel and Grinvald, 1996] Shmuel, A. and Grinvald, A. (1996). Functional organization for direction of motion and its relationship to orientation maps in cat area 18. *J. Neurosci.*, 16(21):6945.

[Swindale, 1998] Swindale, N. V. (1998). Cortical organization: Modules, polymaps and mosaics. *Curr. Biol.*, 8.

AutoSelect: What You Want Is What You Get: Real-Time Processing of Visual Attention and Affect

Nikolaus Bee[1], Helmut Prendinger[2], Arturo Nakasone[3], Elisabeth André[1], and Mitsuru Ishizuka[3]

[1] Institute of Computer Science, University of Augsburg
Eichleitnerstr. 30, D-86135 Augsburg, Germany
nikolaus.bee@gmail.com, andre@informatik.uni-augsburg.de
[2] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
helmut@nii.ac.jp
[3] Graduate School of Information Science and Technology, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
arturo@mi.ci.i.u-tokyo.ac.jp, ishizuka@i.u-tokyo.ac.jp

Abstract. While objects of our focus of attention ("where we are looking at") and accompanying affective responses to those objects is part of our daily experience, little research exists on investigating the relation between attention and positive affective evaluation. The purpose of our research is to process users' emotion and attention in real-time, with the goal of designing systems that may recognize a user's affective response to a particular visually presented stimulus in the presence of other stimuli, and respond accordingly. In this paper, we introduce the *AutoSelect* system that automatically detects a user's preference based on eye movement data and physiological signals in a two-alternative forced choice task. In an exploratory study involving the selection of neckties, the system could correctly classify subjects' choice of in 81%. In this instance of AutoSelect, the gaze 'cascade effect' played a dominant role, whereas pupil size could not be shown as a reliable predictor of preference.

1 Introduction and Motivation

A significant amount of recent research in human–computer interaction and interactive systems is driven by the aim of making the interaction with computers more intuitive, natural, and sensitive to the user's affective state. A prominent example is the multi-modal mixed-initiative dialogue system developed in the SmartKom project, where speech, gesture, and facial expression were synchronized, integrated, and interpreted in real-time, yielding robust understanding of the user's information state and intention [Wahlster, 2003]. The core purpose of the SmartKom dialogue system was to support users in task-oriented and collaborative applications, e.g. for receiving directions at an information kiosk.

A key assumption of interactive systems like SmartKom is that users have a particular (communication) goal in mind, which they want to achieve by successful interaction with the system, primarily through verbal utterances. Our

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 40–52, 2006.

work relaxes this assumption in that we do not assume that users express their cognitive state verbally. Instead, we will rely solely on a user's unconscious (or non-conscious) and non-cognitive behavior, specifically eye movements and autonomic nervous system activity, which provides rich evidence on the user's focus (and shift) of visual attention [Duchowski, 2003] and affective state [Levenson, 1988], respectively.

We believe that visual attention and physiology based interactive technology is of high relevance to infotainment and e-learning applications, handicapped people, as well as to future interfaces as propagated in various ambient intelligence and ubiquitous computing projects [Streitz and Nixon, 2005]. In fact, many decisions of our daily life cannot be easily explained in terms of overt reasoning on premises. In a restaurant, for instance, we choose between different types of dishes. Unless price or dietary considerations are of primary importance, our decision for a particular dish might be based on our taste, our expectation of a specific (eating) experience, or even our current mood. Similarly, we might not be able to explain why we choose a particular piece of clothes to wear. The same situation seems to hold true even for products that provide clearly distinguishable functionality, such as digital cameras or cellular phones. In this context, the Japanese term "Kansei" is often used, which refers to a person's feeling or image of a product [terremoto.net, 2006]. The related discipline of Kansei engineering, sometimes translated as "sensory engineering" or "emotional usability", involves the design of products that use sensory attributes in order to elicit desired consumer responses. We take the experience of a product's Kansei as further evidence for the importance of interactive technologies that are aware of a person's affective evaluation of a product, and might respond accordingly, e.g. by providing further information on the product or by reinforcing a customer's non-conscious (positive) evaluation.

In this paper, we will describe a model (and system) that aims at detecting a user's positive affective evaluation of a visually presented item (preference) based on eye movements and physiological signals. Our notion of "positive affective evaluation" is loosely related to the notions of "liking" and "wanting", but also "interest". However, given that those notions have several meanings in the psychological literature [Feldman-Barrett, 2006], we preferred to use a more neutral term that is intended to refer to non-conscious biasing steps involved in decision making [Bechara et al., 1997]. The key technical challenge of our work is to process and interpret physiological signals and eye movements in real-time, given their individual characteristics, such as different latency and duration. Although affective computing [Picard, 1997] is an active research area, quite surprisingly, approaches to fusing physiological signals in real-time are almost non-existent (see [Prendinger and Ishizuka, 2005] for an early attempt).

In the rest of the paper, we first describe relevant background research, and then propose our own model for affective evaluation that is based on biometric signals and eye movements as input modalities. We will demonstrate the operation of our model by implementing an automatic necktie selector, and present our initial results that were obtained from an exploratory study.

2 Background and Related Work

In this section, we will report on related work that aims at interpreting physiological information as affective states, and eye movements as an indicator for preference. Recently, many studies were conducted to recognize affective states, typically the six basic emotions proposed by [Ekman, 1992] (or a subset of them), based on speech, facial expression, and physiological signals, or a combination of signals [Busso et al., 2004]. Others used physiological signals, sometimes in combination with speech, to detect emotions based on the two-dimensional model of [Lang, 1995] (see [Kim et al., 2005]). It is important to note that most of those approaches are *offline* methods that use feature extraction and machine learning techniques for affect recognition. We wish to emphasize that our goal is different in that we want to recognize affective states, specifically (positive) affective evaluation, in *real-time*, and therefore will rely on a knowledge-based approach that is informed by the psychophysiological literature.

2.1 Biometric Signals and Affective Evaluation

There are two famous emotion models available for real-time affect recognition from physiological signals, the *two-dimensional emotion* model [Lang, 1995] and the *autonomic specificity of emotions* approach [Levenson, 1988].

The two-dimensional emotion model advocated by [Lang, 1995] claims that all emotions can be characterized by two bipolar, but independent dimensions: (i) (judged) valence (pleasant or unpleasant, or: positive or negative); (ii) arousal (calm or aroused). Here, named emotions can be conceived as coordinate points in the arousal–valence space. While this model works well for basic emotions of Ekman, such as happy or sad, it is not obvious how to determine the location of positive affective evaluation (or other affect-related states such as confusion).

[Levenson, 1988], on the other hand, argues that (some) emotions can be distinguished by their associated pattern of autonomic nervous system activity, i.e. (some) emotions have "autonomic signatures". For instance, in an early study, [Ekman et al., 1983] found that there was a larger increase of heart rate with anger and fear than with happiness, and that skin temperature decrease was stronger with anger than with happiness, among other findings.

Unfortunately, literature on directly relating biometric signals to affective evaluation is sparse. The following paragraphs summarize findings from the literature. (Unless otherwise indicated, results are based on [Andreassi, 2000]).

Galvanic Skin Response (GSR). The GSR signal is an indicator of skin conductance (SC), which increases linearly with a person's level of overall arousal or stress. While an increased level of SC is associated with an orienting response to new or interesting stimuli, a very high level of SC is a good indicator of (negative) arousal (stress). [Bechara et al., 1997] also report on an interesting relation between SC responses and conscious decision making. In their study using a gambling task, (non-patient) subjects showed *anticipatory* SC responses before they knew explicitly that their choice was risky (as opposed to the patient subjects with prefrontal damage).

Blood Volume Pulse (BVP). The BVP signal is an indicator of blood flow. Since each heart beat (or pulse) presses blood through the vessels, BVP can also be used to calculate heart rate and inter-beat intervals. Heart rate increases with negatively valenced emotions, such as anxiety or fear.

Pupillary Response. Pupil size is affected by human emotional and cognitive processes [Hess, 1972]. Increase in pupil size is a good indicator of novelty, interest, and positive evaluation, but also of cognitive load, whereas decrease in pupil size indicates increased fatigue, and possibly negative stimuli ("perceptual avoidance"). However, [Hess, 1972, p. 509] also reports on subjects whose pupils dilated when shown pictures containing 'shock' content. He further speculates that increased levels of skin conductance might have an impact on pupil dilation. Presumably the most important finding for our research is the significant correlation between persons' (actual) attitude towards items such as food and consumer goods and pupil dilation. [Krugman, 1964] conducted studies demonstrating that goods (e.g. silverware patterns and greeting cards) inducing pupil responses indeed outsold goods without such responses.

Eye Blinks (EB). EBs occur throughout the day, with an average of 15–20 times per minute for a relaxed person. From a physiological point of view, only 2–4 blinks are necessary for an adult; while reading, the blink rate can drop to three blinks per minute. From a psychological perspective, blink frequency reflects negative affective states, such as nervousness, stress, and fatigue. Eye blink magnitudes were shown to be larger and latencies faster during negative as opposed to positive imagery. Moreover, higher arousal resulted in larger magnitude and shorter latency of eye blinks.

For detecting affective evaluation of visually presented stimuli, eye movements are of particular importance, and will be discussed next.

2.2 Eye Movements and Preference Formation

When presenting pairs of human faces to subjects and giving the instruction to decide on their attractiveness, [Shimojo et al., 2003] observed a phenomenon they called gaze 'cascade effect'. This phenomenon involves the gradual gaze shift toward the face that was eventually chosen (as more attractive), while gaze bias was initially distributed evenly between the two presented faces. The results of the two-alternative forced choice (2AFC) task used in their study demonstrated a progressive bias in subjects' gaze toward the chosen stimulus (preference formation), which was measured by the gaze time spent on the selected stimulus. However, the strong correlation between choice and gaze duration occurred only in the last one and half seconds before the decision was made.

A finding that [Shimojo et al., 2003] declared as surprising relates to the result that a larger cascade effect was found in the 'difficult' task, where the comparison between the attractiveness of faces was difficult, while intuitively, subjects were expected to more evenly distribute their gaze between stimuli in this case, in order to compare stimuli in as much detail as possible. The result was be explained by a theory claiming that gaze would significantly contribute

to decision-making when cognitive bias is weak. The importance of this result for our research derives from the fact that a large number daily choices, e.g. regarding consumer products, are also deficient of a strong cognitive bias, and hence contributes to the importance of investigating non-conscious human decisions.

We recently proposed an approach to estimating user interest that is based on both physiological signals and eye movements [Kon et al., 2006]. In order to achieve higher accuracy of estimating interest, gaze duration time was combined with skin potential level (SPL) based arousal detection. Here, arousal data allowed for more precise segregation into interest and non-interest regions in a two-dimensional space by using a machine learning method. A shortcoming of this approach was that SPL data were not aligned with gaze direction, such that we could not determine the visual 'source' of SPL changes. We will address this problem by explaining our model in the following section.

3 A Model for Affective Evaluation

The proposed model is, to our knowledge, the first attempt to estimate (positive) affective evaluation (preference) from physiological information (skin conductance, blood volume pulse, pupil size, blink rate) and focus of attention. The current model is confined to situations where the user is shown two visual stimuli (pictures), one to the left and one to the right of the screen, from which one will be chosen automatically depending on the user's gaze and physiological behavior. This can be conceived as an automated version of the two-alternative forced choice task described in [Shimojo et al., 2003]. It should be noted that most previous research on emotion recognition used stimuli with a single fixed location (showing e.g. pictures or videos) to elicit affective responses.

A key problem is to attribute affective meaning to visual stimuli, i.e. modeling a user's affective evaluation, given an ongoing stream of physiological activity. Typically, a person who chooses between two pictures will alternately look at the stimuli while physiological changes might occur. In order to explain the notion of *temporal matching* between physiological activity and visual stimuli, we will introduce the notions of 'time window' and '(un)directed signal'.

A *time window* refers to a temporal segment of sufficient duration to estimate relevant changes in bio-signal activity. Time windows differ depending on the latency (duration of onset), apex, and offset of each signal. Pupil size can be considered as a *directed* signal, since it reflects physiological changes directly influenced by the stimulus attended to. On the other hand, SC, BVP, and EBs are seen as *undirected* signals in our setting (where persons will select from two stimuli); i.e. the time window for a significant rise of the level of SC may not be included in the time period a person is attending to one single stimulus, but may become evident after multiple shifts between the two stimuli. Our model for affective preference is encoded as a Bayesian network using Netica^{TM} software from [Norsys, 2003] (see Fig. 1).

The network is based on the findings reported in Sects. 2.1 and 2.2. Specifically, the "Gaze Bias" node encodes the gaze cascade effect that was determined

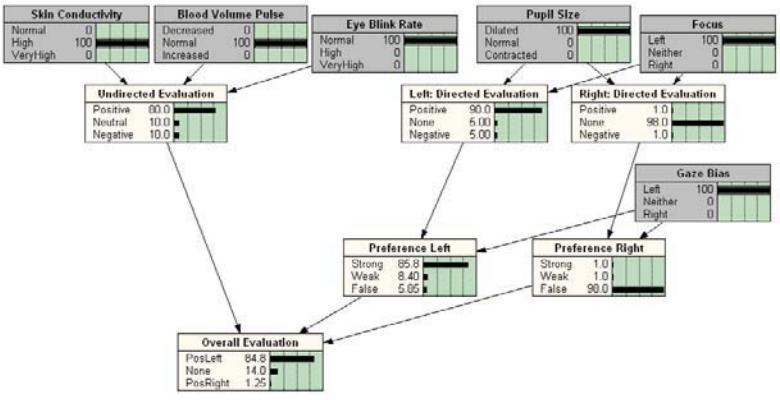


Fig. 1. Bayesian network for recognizing affect based preference for the two stimuli case, assumed as shown to the left and to the right of a screen

empirically in [Shimojo et al., 2003], i.e. the value is "Left" ("Right") if the user continuously fixates the left (right) stimulus for more than one second. A threshold rules out cases where users initially fixate a stimulus longer without having inspected the other stimulus.

Undirected evaluation may contribute to the overall evaluation (preference). If positive, this is seen as evidence that the selected stimulus received (positive) approval. A negative value is assumed to contribute to the (negative) evaluation of the non-selected stimulus. Those interpretations are model assumptions and highly speculative. It is important to notice that the instance of the AutoSelect system (as a tie selector) depends primarily on gaze. However, affect-related signals were included in the model for the more general case of automatically selecting between stimuli that are more emotionally arousing than ties.

4 Algorithms for Real-Time Processing of Affect and Visual Attention

Online feature analysis requires methods that are different from those commonly used in off-line processing since we cannot rely on the availability of the entire signal history built up during some interaction session, including (global) baselines, minima, or maxima.

A core technique underlying most of our algorithms is the derivative as we are only interested in changes over time. We consider different time windows for detecting tendencies (BVP rate, 5 s; eye blink rate, 10 s; pupil response, 1 s) and counting absolute values (skin conductance startle, 1 s; focus of attention, 300 ms). Our analysis is thus independent from a global baseline taken prior to the interaction session. For the task of selecting ties, it was not deemed necessary

to employ a dynamic Bayesian network, and hence each choice is treated as independent from the previous one.

Our algorithms partly depend on the hardware available to us. We used the ProComp Infinity^{TM} encoder from [Thought Technology, 2005] to process bio-signals, and faceLAB^{TM} v4 from [Seeing Machines, 2005] for eye related processing. The ProComp Infinity^{TM} encoder was set to 20 samples/s that was sufficiently high for the investigated signals, e.g. pupil change occurs with 200 ms latency and develops a local maximum within 1500 ms. The faceLAB^{TM} eye tracker has a sampling rate of 60 Hz. Calculation of the head and eye positions with the faceLAB^{TM} software amounted to an average latency of 30 ms.

Skin Conductance. For arousal detection, we modified the algorithm originally developed by [Healey, 2000] by removing slight jitters through smoothing the signal with a basic mean filter. For robust threshold-based startle detection, the square of the raw smoothed signal was calculated. A further improvement was achieved by adapting the threshold based on the past signal during operation.

Blood Volume Pulse. As BVP depends on time and is subject to rise and fall, the derivative of the signal was used as an index of increased heart rate. As soon as the derivative exceeds a certain threshold, a heart beat is detected. Robustness was increased by using a heuristic that is based on the assumption that there is no pulse rate below 45 or above 120 beats/minute.

Eye Blink Rate. Eye blink frequency is obtained from eye closure which is calculated by the faceLAB^{TM} software. If the eye is closed for less than one second a blink is detected. Eye blinks last for 400 ms on average.

Pupil Size. The faceLAB^{TM} software also provides information about the pupil size. Using the method proposed in [Schultheis and Jameson, 2004], artifacts due to EBs are eliminated by suppressing the signal 200 ms before and 1 s after a blink. Since our interest is in the dilation and contraction of the pupil size over time, the derivative is used as a feature.

Gaze Bias and Focus. The gaze bias is calculated from the gaze point distribution regarding two visually presented objects. [Shimojo et al., 2003] analyzed the gaze cascade effect by considering the last 1.5 seconds before a decision was made. We adopted this value for our calculation of gaze bias. Focus was calculated by counting gaze points in time windows of 300 ms. (The implementation of a fixation algorithm was not mandatory for our application.)

5 Exploratory Study

A system that may automatically detect users' choices seems to break new ground. We therefore conducted an exploratory study using the AutoSelect system. Our first application is an automatic necktie selector, where subjects are shown a pair of ties and the AutoSelect system tries to detect the preferred tie. Subjects were given no instruction other than having to choose a tie for themselves or their friend for a graduation party.

5.1 Procedure and Design

Eight subjects (4 female, 4 male), all students or researchers from NII, participated and received an award of 1,000 Yen. Subjects entered the experimental room individually and were provided written instructions about their task.

Subjects were seated in front of an 20.1 inch display with attached infrared lights (see Fig. 2) and their head and eyes were calibrated. This procedure has to be performed for each individual once, and takes approximately 5 minutes. (If a person wears glasses, the procedure might be prolonged to due reflection from the infrared light.) Next, the biometrical sensors of the ProComp Infinity encoder were attached to the subject. The SC sensor was attached to the index finger and the ring finger of the non-dominant hand, and the BVP sensor was fixed to the thumb of the same hand.

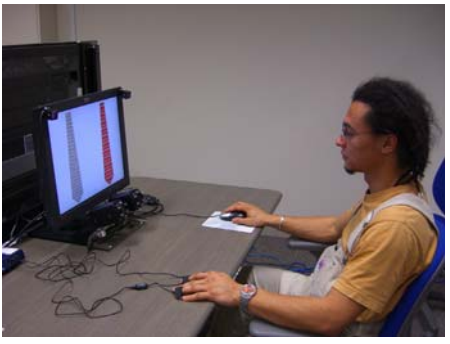


Fig. 2. Experimental setup

A session was initialized by subjects pressing a 'start' button in a web page based interface. Then the following procedure was iterated for 62 pairs of ties. First, a center located 'dot' was shown on the screen for 2.5 s in order to eliminate any initial gaze bias. Next, a pair of ties was presented, located left and right on the screen. In order to guarantee that subjects actually compare the ties, automatic selection was suppressed within the first 2.5 s. (This value was based on the empirically determined decision time of 4 s in [Shimojo et al., 2003].) After the system decision, the selected tie was presented and subjects were asked to indicate whether the system choice is correct by clicking on a 'yes' or 'no' button. Then the next iteration started with the initial view of a center dot.

One initial set of 32 tie pairs was prepared, and the chosen ties were put back into the tie pool, which was used to create the subsequent set of 16 pairs, and so on. Eventually, subjects were shown a single pair of ties they presumably liked best. Hence, subjects were exposed to 63 pairs and performed 62 decisions in total. In the initial set of tie pairs, two partitions were created with 13 pairs each. One partition contained pairs of 'different' type ties, i.e. formal (decent) vs. 'entertainment' (adventurous) style ties, whereas the other partition contained 'similar' type ties that differed only in color or had a slightly different pattern but the same color. The motivation of this grouping was to investigate differences in subjects' decision behavior for presumably 'easy' vs. 'hard' decisions. All sessions were logged and lasted for about 10 minutes.

5.2 Results

The primary result concerns the classification accuracy of the AutoSelect system. In our study, the system could detect subjects' choices correctly in 81% of the cases. The worst recognition rate was 68%. Given a chance level of 50%, the

system performed very well. (One subject was excluded from analysis because of distorted values due to starting a conversation during the experiment.) It is noted that we did not aim at validating the network in which case we would simply have compared user decisions with system decisions. Instead, we wanted to investigate the users' interactive experience with a running system, which can reveal e.g. issues related to the latency between user decision and system decision. Informal comments on the system indeed indicated that subjects were surprised about the system's reliability to timely identify which tie they liked more. Some of the misclassifications were related to a design problem, i.e. when subjects moved their face out of the camera range. The next version of AutoSelect should alert subjects in those situations.

In the following, we will discuss results related to gaze and bio-signals. As to gaze, we were particularly interested in results comparable to the 'difficult' vs. 'easy' choice finding reported in [Shimojo et al., 2003]. We hence compared recognition rates and decision times for 'different' vs. 'similar' tie pairs (see Fig. 3). In line with Shimojo, Simion and colleagues, the decision time for different ties was significantly longer than for similar ties ($t(180) = -1.66$; $p = 0.049$). We used a one-tailed t-test assuming unequal variances in our analysis. This result supports the hypothesis that a choice between unlike items relies on (time consuming) cognitive processing, whereas similar items might be chosen based on non-conscious ('intuitive') preference. We also note that the system calculated the choice between similar ties more accurately.

Although of secondary importance (and certainly speculative given the small number of subjects), we also investigated gender differences in terms of recognition rate and decision time (see Fig. 4). As opposed to a sometimes heard (joking) prejudice, female subjects seemingly decided significantly faster than male subjects ($t(432) = -3.79$; $p = 0.00009$). We should emphasize that this result relates to the decision time calculated by the system, and not necessarily to the decision time of the subjects. It can also be observed that the recognition rate is higher for female subjects.

Our analysis of biometric activity in this paper is confined to the directed signal of pupil size, which is closely related to perception but also to emotion (valence). Based on the findings in [Krugman, 1964, Hess, 1972], we predicted that pupil dilation would occur whenever subjects view the tie they declared as their choice. (The data of one subject were accidentally overwritten and hence missing.) However, only three subjects (S4, S6, S7) supported this hypothesis,

	rec. rate	/ dec. time (s)
different	75%	6.80
similar	81%	7.65
overall	**78%**	**7.22**

Fig. 3. Recognition results and mean decision time for similar/different type ties

	rec. rate	/ dec. time (s)
female	89%	6.15
male	75%	7.43
overall	**81%**	**6.88**

Fig. 4. Recognition results and mean decision time dependent on gender

Subj	C	non-C	t(df)	df	p
S3	3.08	3.13	−3.50	4612	0.0002
S4	3.23	3.19	2.43	5098	0.0076
S5	2.35	2.32	0.91	5680	0.179
S6	3.54	3.51	2.41	4114	0.008
S7	3.27	3.22	5.50	7358	<0.0001
S8	3.13	3.14	−0.69	5393	0.75

Fig. 5. One-tailed t-test for pupil size of preferred (choice) and non-choice ties (in millimeter)

Subj	C	non-C	t(df)	df	p
S3	3.02	3.10	−3.60	2443	<0.0001
S4	3.17	3.24	−3.49	1785	0.0003
S5	2.19	2.30	−2.38	3378	0.0087
S6	3.55	3.56	−1.01	2829	0.155
S7	3.34	3.24	7.83	3414	<0.0001
S8	3.13	3.18	−1.03	1737	0.15

Fig. 6. One-tailed t-test for pupil size of choice vs. non-choice ties when seen for the first time (in millimeter)

whereas S3 supports the opposite, i.e. a significant contraction of pupil size for the preferred tie, while the results for S5 and S8 are not significant (see Fig. 5). Total dwell time on the preferred tie was always greater than on the non-choice tie, with a significant difference in four cases. This result, however, is not surprising given that the gaze cascade effect implies that more time will eventually be spent on the choice.

In view of those results, our initial guess was that *habituation*, i.e. the progressive diminution of behavioral response probability with repetition of a stimulus, might have been responsible for the non-significance of some data. We hence analyzed only those situations where the subjects saw the ties for the first time (ties in the initial 32 pairs). Much to our surprise, we found the opposite effect prevailing, i.e. pupil size for the preferred tie was mostly smaller than for the choice tie, and in three cases (S3, S4, S5), significantly smaller (see Fig. 6).

Finally, we tested a conjecture put forth in [Simion, 2005], namely that pupil size might be a physiological precursor of preference or choice *before* any gaze particular pattern occurs. Specifically, pupil size and dwell time was compared for choice and non-choice ties for the single instance where subjects looked at a tie for the (very) first time. (Note that the analysis in the previous paragraph included multiple situations effected by the comparative task.) However, only two subjects showed significantly smaller pupil size for the preferred tie, and one subject had significantly longer dwell time for the choice.

The present study could thus not establish pupil size (or duration) as a reliable predictor for preference, and leaves open many questions. Looking at the pupil size literature [Hess, 1972], it can be observed that studies often used 'strong' visual stimuli, including the display of 'shock content' but also nudity. The stimuli chosen for our study (neckties) are obviously of a different quality. (URL: `http://research.nii.ac.jp/~prendinger/autoselect`)

6 Conclusions

The contribution of this paper is two-fold. First, we introduced a model for the joint interpretation of attention and affect. The model is encoded as a Bayesian

network and may infer a user's affective evaluation or preference based on eye movements and bio-signal activity. The model is used for AutoSelect, a system that primarily exploits the gaze cascade effect in order to predict the choice of users in a two-alternative forced choice task. Biometric signals may change the probability of the positive left (right) choice in the network, but not the computed preference for the 'left' or 'right' stimulus.

Second, we conducted an exploratory study to test whether AutoSelect can correctly predict the choice of a user. The accuracy of the system with a limited number of subjects (7) is reasonably high (81%). Pupil size could not be established as a reliable predictor of affective evaluation in our study. Certainly, if our sole interest would be tie selection, biometric signals could be removed from the network altogether. However, we plan a follow-up study that employs life-like characters as interaction partners [Prendinger and Ishizuka, 2004] and thus provides an emotionally more arousing interaction than the tie selection task. In this way, we hope to obtain a clearer picture of the utility of pupil size and other biometric input signals, which were of little importance in the tie selection task.

Acknowledgments

The first author was supported by an International Internship Grant from NII under a Memorandum of Understanding with the Faculty of Applied Informatics at the Univ. of Augsburg. We would like to thank Dr. Takamasa Koshizen from Honda Research Institute, Japan, and Dr. Claudiu Simion from California Institute of Technology for valuable discussion and inspiration. The research was supported by the Research Grant (FY1999–FY2003) for the Future Program of the Japan Society for the Promotion of Science (JSPS), by a JSPS Encouragement of Young Scientists Grant (FY2005–FY2007), and an NII Joint Research Grant with the Univ. of Tokyo (FY2005).

References

[Andreassi, 2000] Andreassi, J. L. (2000). *Psychophysiology. Human Behavior & Physiological Response.* Lawrence Erlbaum Associates, Mahwah, NJ, 4 edition.

[Bechara et al., 1997] Bechara, A., Damasio, H., Tranel, D., and Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, 275:1293–1295.

[Busso et al., 2004] Busso, C., Deng, Z., Yildirim, S., Buut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., and Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of 6th International Conference on Multimodal Interfaces (ICMI-04)*, pages 205–211. ACM Press, New York.

[Duchowski, 2003] Duchowski, A. T. (2003). *Eye Tracking Methodology: Theory and Practice.* Springer, London, UK.

[Ekman, 1992] Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3–4):169–200.

[Ekman et al., 1983] Ekman, P., Levenson, R. W., and Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*, 221:1208–1210.

[Feldman-Barrett, 2006] Feldman-Barrett, L. (2006). Emotions as natural kinds? *Perspectives on Psychological Science*, 1:28–58.

[Healey, 2000] Healey, J. A. (2000). *Wearable and Automotive Systems for Affect Recognition from Physiology*. PhD thesis, Massachusetts Institute of Technology.

[Hess, 1972] Hess, E. H. (1972). Pupillometrics: A method of studying mental, emotional and sensory processes. In Greenfield, N. and Sternbach, R., editors, *Handbook of Psychophysiology*, pages 491–531. Holt, Rinehart & Winston, New York.

[Kim et al., 2005] Kim, J., André, E., Rehm, M., Vogt, T., and Wagner, J. (2005). Integrating information from speech and physiological signals to achieve emotional sensitivity. In *Proceedings 9th European Conference on Speech Communication and Technology*.

[Kon et al., 2006] Kon, M., Koshizen, T., and Prendinger, H. (2006). A new user–machine interface using cross-modal computation for deep interest estimation. Towards quantifying user satisfaction. In *Proceedings of IUI-06 Workshop on Effective Multimodal Dialogue Interfaces*, pages 25–34.

[Krugman, 1964] Krugman, H. (1964). Some applications of pupil measurement. *Journal of Marketing Research*, 1:15–19.

[Lang, 1995] Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, 50(5):372–385.

[Levenson, 1988] Levenson, R. W. (1988). Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity. In Wagner, H. L., editor, *Social Psychophysiology and Emotion: Theory and Clinical Applications*, pages 17–42. John Wiley & Sons, Hoboken, NJ.

[Norsys, 2003] Norsys (2003). Norsys Software Corp. Netica.
URL: `http://www.norsys.com`.

[Picard, 1997] Picard, R. W. (1997). *Affective Computing*. The MIT Press, Cambridge, MA.

[Prendinger and Ishizuka, 2004] Prendinger, H. and Ishizuka, M., editors (2004). *Life-Like Characters. Tools, Affective Functions, and Applications*. Cognitive Technologies. Springer Verlag, Berlin Heidelberg.

[Prendinger and Ishizuka, 2005] Prendinger, H. and Ishizuka, M. (2005). The Empathic Companion: A character-based interface that addresses users' affective states. *International Journal of Applied Artificial Intelligence*, 19(3):267–285.

[Schultheis and Jameson, 2004] Schultheis, H. and Jameson, A. (2004). Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioral methods. In *Proceedings Adaptive Hypermedia and Adaptive Web-based Systems (AH-04)*, pages 225–234, Berlin. Springer.

[Seeing Machines, 2005] Seeing Machines (2005). Seeing Machines.
URL: `http://www.seeingmachines.com/`.

[Shimojo et al., 2003] Shimojo, S., Simion, C., Shimojo, E., and Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, 6(12):1317–1322.

[Simion, 2005] Simion, C. (2005). *Orienting and Preference: An Enquiry into the Mechanisms Underlying Emotional Decision Making*. PhD thesis, California Institute of Technology.

[Streitz and Nixon, 2005] Streitz, N. and Nixon, P. (2005). The Disappearing Computer. Guest editors' introduction to Special Issue. *Communications of the ACM*, 48:33–35.

[terremoto.net, 2006] terremoto.net (2006). Kansei engineering: Incorporating affection and emotion into the design process. URL: `http://terremoto.net/kansei/`.

[Thought Technology, 2005] Thought Technology (2005). Thought Technology Ltd. URL: `http://www.thoughttechnology.com`.

[Wahlster, 2003] Wahlster, W. (2003). Towards symmetric multimodality: Fusion and fission of speech, gesture and facial expression. In *Proceedings 26th German Conference on Artificial Intelligence*, pages 1–18, Berlin Heidelberg. Springer LNAI 2821.

Emotion Recognition Using Physiological and Speech Signal in Short-Term Observation

Jonghwa Kim and Elisabeth André

Institute of Computer Science
University of Augsburg, Germany
{Kim, Andre}@informatik.uni-augsburg.de

Abstract. Recently, there has been a significant amount of work on the recognition of emotions from visual, verbal or physiological information. Most approaches to emotion recognition so far concentrate, however, on a single modality while work on the integration of multimodal information, in particular on fusing physiological signals with verbal or visual data, is scarce. In this paper, we analyze various methods for fusing physiological and vocal information and compare the recognition results of the bimodal recognition approach with the results of the unimodal approach.

1 Introduction

Recent work by Picard and others [1] has aroused considerable awareness for the role of emotions in human-computer interaction. Indeed, there is evidence that human-computer interaction is more likely to be accepted by the user if it is sensitive towards the user's affective states. An important prerequisite to realize affective interfaces is a reliable emotion recognition system which guarantees acceptable recognition accuracy, is robust against artefacts, and easily adapts to pragmatic constraints. Most research so far has focused on the analysis of a single modality or an integrated analysis of audio-visual information (see [2] for a comprehensive overview). On the one hand, the integration of multiple modalities raises the expectation of higher recognition rates compared to those obtained from a single modality. On the other hand, more complex classification problems arise.

In this paper, we concentrate on the integration of physiological measures (biosignals) and speech signals for emotion recognition based on short-term observations. Several advantages can be expected when combining biosensor feedback with affective speech. First of all, biosensors allow us to continuously gather information on the users' affective state while the analysis of emotions from speech should only be triggered when the microphone receives speech signals from the user. Secondly, it is much harder for the user to deliberately manipulate biofeedfack than external channels of expression which allows us to largely circumvent the artifact of social masking. Finally, an integrated analysis of biosignals and speech may help to resolve ambiguities and compensate for errors.

When combining multiple modalities, the following questions arise: (1) How to handle conflicting cases between the single modalities? For instance, a user

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 53–64, 2006.

may consciously or unconsciously conceal his/her real emotions by external channels of expression, but still reveal them by internal channels of expression. (2) At which level of abstraction should the single modalities be fused in order to increase the accuracy of the recognition results? (3) How should the window sizes of different modalities be synchronized when same emotional cues in the modalities occur with a time discrepancy?

In the next section, we discuss selected previous work. Section 3 reports on the data set we used and describes the features we extracted from 5-channel biosignal and speech signal. Several classification methods are presented including feature-level fusion, decision-level fusion, and a hybrid fusion scheme. In Section 4, we analyze the classification results with respect to the effect of bimodal integration. We conclude this work with a short outlook on future work.

2 Related Work

There is a vast body of literature on the automatic recognition of emotions. With labelled data collected from different modalities, most studies rely on supervised pattern classification approaches for automatic emotion recognition.

Following the long tradition of speech analysis in signal processing, many efforts were taken to recognize affective states from vocal information. As emotion-specific contents in speech, suprasegmental prosodic features including intensity, pitch, and duration of utterance have been widely used in recognition systems. To exploit the dynamic variation along an utterance, Mel-frequency cepstral coefficients (MFCC) are extensively employed. For example, Nwe and colleagues [3] achieved an average accuracy of 66% for six emotions acted by two speakers using 12 MFCC features as input to a discrete hidden Markov model (HMM). A rule-based method for emotion recognition was proposed by Chen [4]. The data used in this work contained two foreign languages (Spanish and Sinhala) for the judges who did not comprehend either language and were therefore able to make their judgment based on vocal expression without being influenced by linguistic/semantic content. Batliner et al. [5] achieved about 40% for a 4-class problem with elicited emotions in spontaneous speech.

Relatively little attention has been paid so far to physiological signals for emotion recognition compared to other channels of expression. A significant series of work has been conducted by Picard and colleagues at MIT Lab. For example, they showed that certain affective states may be recognized by using physiological measures including heart rate, skin conductivity, temperature, muscle activity and respiration velocity [1]. Eight emotions deliberately elicited from a subject in multiple weeks were classified with an overall accuracy of 81%. Nasoz et al. [6] used movie clips to elicit target emotions from 29 subjects and achieved the best recognition accuracy (83%) by applying the Marquardt Backpropagation algorithm. More recently, Wagner et al. [7] presented an approach to the recognition of emotions elicited by music using 4-channel biosignals which were recorded while the subject was listening to music songs, and reached an overall recognition accuracy of 92% for a 4-class problem.

In order to improve the recognition accuracy obtained from unimodal recognition systems, many studies attempted to exploit the advantage of using multimodal information, especially by fusing audio-visual information. For example, De Silva and Ng [8] proposed a rule-based singular classification of audio-visual data recorded from two subjects into six emotion categories. Moreover, they observed that some emotions are easier to identify with audio, such as sadness and fear, and others with video, such as anger and happiness. Using decision-level fusion in bimodal recognition system, a recognition rate of 72% has been reported. A set of singular classification methods was proposed by Chen and Huang [9], in which audio-visual data collected from five subjects was classified into the Ekman's six basic emotions (happiness, sadness, disgust, fear, anger, and surprise). They could improve the performance of decision-level fusion by considering the dominant modality, determined by empirical studies, in case significant discrepancy between the outputs of each unimodal classifier has been observed. Recently, a large-scale audio-visual database was collected by Zeng et al. [10], which contains five HCI-related affective responses (confusion, interest, boredom, and frustration) in addition to seven affects (the six basic emotions + neutral). To classify the 11 emotions subject-dependently, they used the SNoW (Sparse Network of Winnow) classifier with Naive Bayes as the update rule and achieved a recognition accuracy of almost 90% through bimodal fusion while the unimodal classifiers yielded only 45-56%.

Most previous studies have shown that the performance of emotion recognition systems can be improved by the use of audio-visual information. However, it should be noted that the achieved recognition rates depend rather on the type of the underlying database, whether the emotions were from acted, elicited or real-life situation, than the used algorithms and classification methods. Moreover, apart from our previous work [11], work on the integration of biosignals and speech is rare. In this paper, we will investigate in how far the robustness of an emotion recognition system can be increased by integrating both vocal and physiological cues. We will evaluate two fusion methods that combine bimodal information at different levels of abstraction as well as a hybrid integration scheme. Particularly we focus on shorter observations compared to or earlier work.

3 Methodology

3.1 Dataset

We use the same Quiz data set as in our prior work [11]. The dataset contains speech (sampled with 48Kz/16Bit), physiological (using 6-channel biosensors[1]), and visual information from three male German-speaking subjects in their twenties.

To acquire a corpus of spontaneous vocal and physiological emotions, we used a slightly modified version of the quiz "Who wants to be a millionaire?". Questions along with options for answers were presented on a graphical display whose

[1] ECG (electrocardiogram), BVP (blood volume pulse), EMG (electromyogram), RSP (respiration), SC (skin conductivity), Temp (finger temperature).

design was inspired by the corresponding quiz shows on German TV. In order to make sure that we got a sufficient amount of speech data, the subjects were not offered any letters as abbreviations for the single options (as very common in quiz shows on TV), but were forced to produce longer utterances. Furthermore, the users current score was indicated as well as the amount of money s/he may win or loose depending on whether his/er answer is correct or not. Each of the session took about 45 minutes to complete. The subjects were equipped with a directed microphone to interact with a virtual quiz master via spoken natural language utterances. The virtual quiz master was represented by a disembodied voice using the AT&T Natural Voices speech synthesizer. While the users interacted with the system, their bio and speech signals as well as the interaction with the quiz master were recorded.

The quiz experiment was designed in a Wizard-Of-Oz fashion where the quiz agent who presents the quiz is controlled by a human quiz master who guides the actual course of the quiz, following a working script to evoke situations that lead to a certain emotional response. The wizard was allowed to freely type utterances, but also had access to a set of macros that contain pre-defined questions or comments which made it easier for the human wizard to follow the script and to get reproducible situations (see Fig. 1). The wizards working script can be roughly divided into four situations which serve to induce certain emotional states in the user. We make use of a dimensional emotion model which characterizes emotions in terms of the two continuous dimensions of arousal and valence (see [12]). Arousal refers to the intensity of an emotional response. Valence determines whether an emotion is positive or negative and to what degree. Apart from the ease of describing emotional states that cannot be distributed into clear-cut fixed categories, the two dimensions valence and arousal are well suited for emotion recognition. The four phases of the experiment correspond to extreme positions on the axes of the emotion model: (1) low arousal, positive valence, (2) high arousal, positive valence, (3) low arousal, negative valence and (4) high arousal, negative valence.

First, the users are offered a set of very easy questions every user is supposed to know to achieve equal conditions for all of them. This phase is characterized by a slight increase of the score and gentle appraisal of the agent and serves to

Fig. 1. Interface for the wizard (left) and for the user (right)

induce an emotional state of positive valence and low arousal in the user. In phase 2, the user is confronted with extremely difficult questions nobody is supposed to know. Whatever option the user chooses, the agent pretends the users answer is correct so that the user gets the feeling that s/he hits the right option just by chance. In order to evoke high arousal and positive valence, this phase leads to a high gain of money. During the third phase, we try to stress the user by a mix of solvable and difficult questions that lead, however, not to a drastic loss of money. Furthermore, the agent provides boring information related to the topics addressed in the questions. Thus, the phase should lead to negative valence and low arousal. Finally, the user gets frustrated by unsolvable questions. Whatever option the user chooses, the agent always pretends the answer is wrong resulting in a high loss of money. Furthermore, we include simple questions for which we offer similar-sounding options. The user is supposed to choose the right option, but we make him/er believe that the speech recognizer is not working properly and deliberately select the wrong option. This phase is intended to evoke high arousal and negative valence.

3.2 Synchronized Segmentation of the Bimodal Signals

In our previous work [11], we segmented and labelled the data based on the four experimental phases taking into account that the agreement between coders annotating material of everyday emotions is usually not very high [13]. All speech and physiologial signals that may be interpreted as a response to the same question have been segmented into one chunk and labelled with the emotion corresponding to the experimental phase in which they occurred.

For the analysis described in this paper, the segmentation and labelling was refined by two expert labellers considering the situative context as well as the audio-visual expression of the subjects. In this way, we tried to handle cases where we did not succeed in eliciting the intended emotion. To segment speech and physiological data, we started from verbal phrases. The borders of the segments for both modalities were chosen to lie in the middle of two verbal phrases so that they cover the same time span. For the analysis of speech, we only consider the part of the segment when the verbal phrase occurs while for the analysis of physiological data the complete segment is taken. As a consequence, the observations for speech are usually shorter than the observations for the physiological data, but the length of the corresponding segments is the same which facilitates the later fusion proces. In total, we got 343 samples for classification (343 x 6 channels = 2058 segments in total) from the data set. Based on the four phases of the experiments, our labellers relied on dimensional rating (i.e. labelling within the 4 quadrants of the 2D emotion model). Disagreements between the ratings of the two labellers were discussed and resolved after the annotation process.

Fig. 2 shows a sample segmentation for data from the used channels. The length of the observations varies from 2 to 6 seconds for the speech and from 3

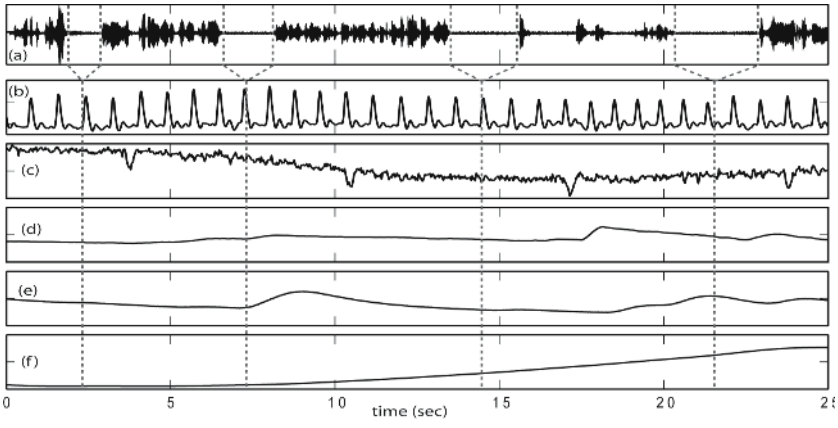


Fig. 2. Segmentation of bimodal signals based on verbal phrases: (a) speech, (b) BVP, (c) EMG, (d) RSP, (e) SC, (f) Temp

to 15 seconds for the biosignals. That is the observations are rather short-term compared to previous studies that start from a segment length between 50 and 300 seconds[2] [15].

3.3 Feature Extraction

An essential step in pattern classification is to extract class-relevant features (preferrably in a compressed form) from the raw signal. Moreover the classification of short-term observations requires more reasonable treatments in signal processing stages, e.g. extracting spectral features in biosignals (containing very low frequencies) within limited bandwidth due to the very short window size.

From physiological data: To remove noisy signals, all segments of the 5-channel biosignals (BVP, EMG, SC, RSP, Temp) are lowpass-filtered using pertinent cut-off frequencies that are empirically determined for each biosensor channel. Differing from [11], we employ the BVP signal instead of the ECG signal and use the Temp signal as an additional channel from the data set. Generally the ECG is measured by using electrodes which do need a firm skin contact, whereas the BVP is measured by using a photoplethysmograph. Hence, using the BVP signal has some advantages such as robustness against motion artefacts during recording process and stable baseline in the signal flow. From the raw signal, we first calculated the 8 subband spectral powers using the conventional 512 points short-time Fourier transform (STFT). To capture the irregularity and the local

[2] Haag et al. [14] used 2 seconds observation of 6-channel biosignals and classified arousal and valence by using a range of specified distance. However, their observation length might be difficult to be compared to our synchronized segmentation of bimodal signals. Moreover using such short length of segment restricts range of usable features, e.g. spectrum features and HRV. They used a limited feature set including 7 fundamental features from each channel for the classification.

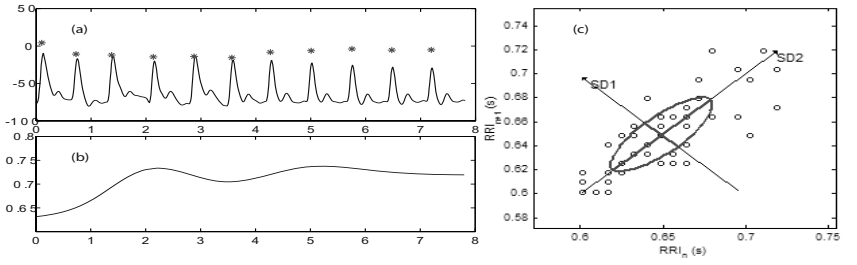


Fig. 3. Example of BVP Analysis: (a) detected pulse interbeats, (b) interpolated PRV like series, (c) Poincaré plot of the PRV

spectral distribution, the spectral entropy is calculated from each subband by converting the spectrum into a PMF-like (Probability Mass Function) form.

Heart rate variability (HRV) is the most frequently used characteristic of the heart activity in biomedical engineering to assess cardiac health. Using the QRS detection algorithm of Pan and Tompkins [16], the HRV like time series (we refer to as PRV)[3] is obtained and typical statistics (mean value, standard deviation, slope, etc.) are calculated from the time series. By calculating the standard deviations in different distances of pulse-pulse interbeats, we also added the Poincaré geometry in the feature set to capture the nature of pulse interval fluctuations. Figure 3 shows an example plot of the geometry. Lastly from the spectrum of the PRV time series, power spectrum densities (PSD) from three subbands are calculated from the ranges of VLF(0-0.04Hz), LF(0.05-0.15 Hz), and HF (0.16-0.4 Hz), respectively and the ratio of LF/HF. Since the RSP signal is quasi periodic we calculated similar types of features like the BVP features including the typical statistics, except for the geometric features and the PSDs. After appropriate detrending the signals using mean value and lowpass filter, we calculated the BRV (time series of the breathing rates) by detecting the peaks using the maxima ranks within zero-crossing. From the SC and EMG signal respectively we calculated 10 features including the mean value, standard deviation, and mean values of first and second derivations. Particularly because of the nature of the signal, the EMG signal required additional pre-processing, such as deep smoothing. The number of transient changes (occurrences) within 4 seconds in SC and EMG signals are calculated from two low-passed signals, very low-passed (SC: 0.08 Hz, EMG: 0.3 Hz) and low-passed signals (SC:0.2 Hz, EMG: 0.8 Hz) respectively. From the Temp signals, three statistical features are calculated: mean value, standard deviation, and ratio of max/min. Finally, we obtained a total of 77 features from the 5-channel biosignals.

From the speech signal: For all segments, the conventional statistics in time domain are calculated, such as mean, absolute extremum, root mean square, standard deviation, energy/power, intensity in dB etc. In frequency domain, three spectrum contents are obtained using the STFT; pitches using a window

[3] Strictly speaking, it is the pulse rate variability (PRV) we use when relying on the BVP instead of the ECG signal.

length of 40 ms,energy spectrum, and formant object using a window length of 25 ms. In addition, 10 MFCCs from each segment are calculated using a window length of 15 ms. From pitch and energy spectrum, also the series of the minima and maxima, and of the distances, magnitudes and steepness between adjacent extrema were obtained. For the MFCCs, we first exponentiated the cepstral coefficients to obtain non-negative values and calculated the spectral entropy as in the case of the biosignal in order to capture the distribution of cepstral energy. From each feature content above, we tried to extract single features (i.e., mean, standard deviation, mean of first and second derivative) representing characteristics (i.e., variance and slop) of each time series vector of spectrum, instead of taking all feature vectors. As a result, we obtained a total of 61 features from the speech segments.

3.4 Feature Selection and Classification

In the next step, we tried to determine which features are most relevant to differentiate each affective state. Reducing the dimension of the feature space has two advantages. First of all, the computational costs are lowered and secondly the removal of noisy information may lead to a better separation of the classes. In all cases, we achieved indeed considerably higher accuracy rates (an increase of about 30 %) when applying sequential backward selection (SBS) to reduce the set of features. Of course, the success of the selection process heavily depends on the employed classifier. Several features were selected by SBS for all three subjects, e.g., the subband spectral entropy from BVP, the number of occurrences in SC and EMG, and the mean values of the MFCCs in the speech features. However, due to the small number of subjects, these findings should not be generalized.

After testing several classification schemes, such as kNN (k-nearest neighbour), MLP (multilayer perception), and LDA (Linear discriminant analysis), we have chosen the LDA classifier which gave the highest accuracy in our case and which we already used for emotion recognition from physiological data in [7].To combine multiple modalities, we need to decide at which level the single modalities should be fused. A straightforward approach is to simply merge the features calculated from each modality (feature-level). An alternative would be to fuse the recognition results at the decision-level based on the outputs of separate unimodal classifiers (decision-level). Finally, we may combine both methods by applying a

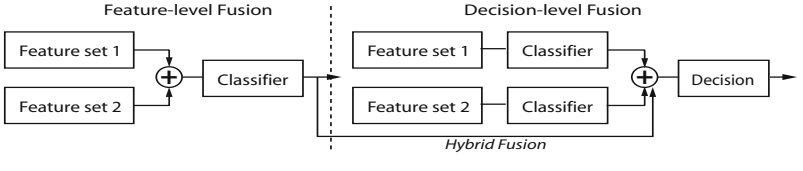


Fig. 4. Considered fusion schemes for integrating bimodal information

Table 1. Recognition results in rates (1.0=100% accuracy) achieved by using SBS, LDA, and leave-one-out cross validation

System	**high/pos**	**high/neg**	**low/neg**	**low/pos**	***Average***
Subject A					
Biosignal	0.95	0.92	0.86	0.85	0.90
Speech signal	0.64	0.75	0.67	0.78	0.71
Feature Fusion	0.91	0.92	1.00	0.85	**0.92**
Decision Fusion	0.64	0.54	0.76	0.67	0.65
Hybrid Fusion	0.86	0.54	0.57	0.59	0.64
Subject B					
Biosignal	0.50	0.79	0.71	0.45	0.61
Speech Single	0.76	0.56	0.74	0.72	0.70
Feature Fusion	0.71	0.56	0.94	0.79	**0.75**
Decision Fusion	0.59	0.68	0.82	0.69	0.70
Hybrid Fusion	0.65	0.64	0.82	0.83	0.73
Subject C					
Bio Single	0.52	0.79	0.70	0.52	0.63
Speech Single	0.55	0.77	0.66	0.71	0.67
Feature Fusion	0.50	0.67	0.84	0.74	**0.69**
Decision Fusion	0.32	0.77	0.74	0.64	0.62
Hybrid Fusion	0.40	0.73	0.86	0.71	0.68
All: Subject-independent					
Bio Single	0.43	0.53	0.54	0.52	0.51
Speech Single	0.40	0.53	0.70	0.53	0.54
Feature Fusion	0.46	0.57	0.63	0.56	**0.55**
Decision Fusion	0.34	0.50	0.70	0.54	0.52
Hybrid Fusion	0.41	0.51	0.70	0.55	0.54

hybrid integration scheme (see Figure 4). We performed both feature-level fusion and decision-level fusion using LDA in combination with SBS. Feature-level fusion is performed by merging the calculated features from each modality into one cumulative structure, selecting the relevant features using SBS, and feeding them to the LDA classifier. Decision-level fusion caters for integrating asynchronous, but temporally correlated modalities. Each modality is first classified independently by the LDA classifier, and the final decision is obtained by fusing the output from the modality-specific classification processes. Three criteria, maximum, average, and product (see [17]) were applied to evaluate the posterior probabilities of the unimodal classifiers at the decision stage. As a further variation of decision-level fusion, we employed a new hybrid scheme of the two fusion methods in which the output of feature-level fusion is also fedas an auxiliary input to the decision-level fusion stage. In Table 1 the best results are summarized that we achieved by the classification schemes we described above. We classified the bimodal data subject-dependently (Subject A, B, and C) and subject-independently (All) since this gave us a deeper insight on what terms the multimodal systems could improve the results of unimodal emotion recognition.

4 Analysis of Results

Table 1 shows that the performance of the unimodal systems varies not only from subject to subject, but also for the single modalities. During our experiment, we could observe individual differences in the physiological and vocal expressions of the three test subjects (see Table 1). As shown in Table 1, the emotions of subject A were more accurately recognized by using biosignals (90 %) than by his voice (71 %) whereas it was inverse for subject B and C (70 % and 67 % for voice and 61 % and 63 % for biosignals). In particular for subject A, the difference between the accuracies of the two modalities is sizable. However, no suggestively dominant modality could be observed in the results of subject-dependent classification in general, which may be used as a decision criterion in the decision-level fusion process to improve the recognition accuracy. Different accuracy rates were also obtained by using the single fusion methods. Overall, we obtained the best results from feature-level fusion. Generally, feature-level fusion is more appropriate for combining modalities with analogous characteristics. For instance, we got an acceptable recognition accuracy of 92 % for subject A when using feature-level fusion which considerably went down, however, when using decision-level or hybrid fusion. As our data show, a high accuracy obtained from one modality may be declined by a relatively low accuracy from another modality when fusing data at the decision level. This observation may indicate the limitations of the decision-level fusion scheme we used, which is based on to a pure arithmetic evaluation of the posterior probabilities at the decision stage rather than a parametric assessment process. Actually, the design of optimal strategies for decision-level fusion, such as the integration of a parametric refinement stage, is still an open research issue. As expected, the accuracy rates for subject-independent classification were not comparable to those obtained for subject-dependent classification. Figure 5 illustrates

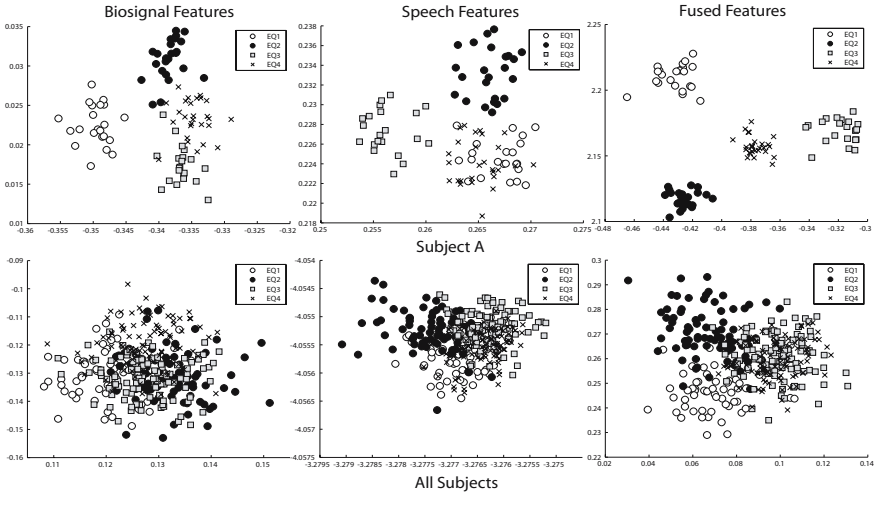


Fig. 5. Fisher projection examples for Subject A and all subjects (person-independent)

examples of Fisher projection which is often used to preview the distribution of the features. Obviously, merging the features of all subjects does not refine the information related to target emotions, but rather leads to scattered class boundaries.

5 Conclusion

In this paper, we treated all stages of emotion analysis, from data collection to classification using short-term observations, and evaluated several fusion methods as well as a hybrid decision scheme. We also compared the results from multimodal classification with the unimodal results. As in our earlier work [11] where we relied on longer observation phases and a different set of features, the best results were obtained by feature-level fusion in combination with feature selection. In this case, not only user-dependent, but also user-independent emotion classification could be improved compared to the unimodal methods.

We did not achieve the same high gains that were achieved for audio-visual data which seems to indicate that speech and physiological data contain less complementary information. Furthermore, in a natural setting like ours, we cannot exclude that the subjects are inconsistent in their emotional expression. Inconsistencies are less likely to occur in scenarios where actors are asked to deliberately express emotions via speech and mimics which explains why fusion algorithms lead to a greater increase of the recognition rate in this case. Ambiguities in emotional expressions are also reflected by work on corpus annotation. For instance, Cowie and colleagues [13] noticed that the agreement between human coders labelling multimodal corpora of everyday emotions was lower when considering both audio and video than when relying on a single modality.

Furthermore some important problems are pointed out, such as the use of posterior probabilities when fusing information with high disparity in accuracy. Most of the existing classifiers used in the literature are generalized methods based on statistics or estimating linear regression of given data. Such classifiers may not be able to capture emotion-specific features and to apply self-adapting decision rules that consider contextual information, for instance. Therefore, the design of an emotion-specific classification scheme is one of the most important issues for the future, and this issue becomes even more critical when classifying combined multimodal observations. To overcome these problems, we need to develop a multilayer fusion scheme with parametric refinement stages in each decision layer.

Acknowledgements

We would like to thank Olena Kuzik for her help with the annotation of the bimodal corpus. All stages from feature extraction to classification are implemented using Matlab/Statistics Toolbox (www.mathworks.com), except for speech feature calculation using Praat (www.praat.org).

References

1. Picard, R., Vyzas, E., Healy, J.: Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Trans. Pattern Anal. and Machine Intell. **23** (2001) 1175–1191
2. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal Processing Mag. **18** (2001) 32–80
3. Nwe, T.L., Wei, F.S., Silva, L.C.D.: Speech based emotion classification. In: IEEE Region 10 International Conference on Electrical Electronic Technology. Volume 1. (2001) 297–301
4. Chen, L.S.: Joint processing of audio-visual information for the recognition of emotional expression in human-computer interaction. PhD thesis, University of Illinois at Urbana-Champaign, Dept. of Electrical Engineering (2000)
5. Batliner, A., Zeissler, V., Frank, C., Adelhardt, J., Shi, R.P., Nöth, E.: We are not amused-but how do you know? user states in a multi-modal dialogue system. In: EUROSPEECH'03, Geneva (2003) 733–736
6. Nasoz, F., Alvarez, K., Lisetti, C., Finkelstein, N.: Emotion recognition from physiological signals for presence technologies. International Journal of Cognition, Technology, and Work - Special Issue on Presence **6(1)** (2003)
7. Wagner, J., Kim, J., André, E.: From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In: ICME'05, Amsterdam (2005)
8. De Silva, L.C., Ng, P.C.: Bimodal emotion recognition. In: IEEE International Conf. on Automatic Face and Gesture Recognition. (2000) 332–335
9. Chen, L.S., Huang, T.S.: Emotional expressions in audiovisual human computer interaction. In: ICME-2000. (2000) 423–426
10. Zeng, Z., Tu, J., Liu, M., Zhang, T., Rizzolo, N., Zhang, Z., Huang, T.S., Roth, D., Levinson, S.: Bimodal HCI-related affect recognition. In: ICMI 2004. (2004)
11. Kim, J., André, E., Rehm, M., Vogt, T., Wagner, J.: Integrating information from speech and physiological signals to achieve emotional sensitivity. In: INTERSPEECH-2005, Lisbon, Portugal (2005) 809–812
12. Lang, P.: The emotion probe: Studies of motivation and attention. American Psychologist **50(5)** (1995) 372–385
13. Douglas-Cowie, E., Devillers, L., Martin, J.C., Cowie, R., Savvidou, S., Abrilian, S., Cox, C.: Multimodal Databases of Everyday Emotion: Facing up to Complexity. In: InterSpeech, Lisbon (2005)
14. Haag, A., Goronzy, S., Schaich, P., Williams, J.: Emotion recognition using bio-sensors: First step towards an automatic system. In: Affective Dialogue Systems, Tutorial and Research Workshop, Kloster Irsee, Germany (2004)
15. Kim, K.H., Bang, S.W., Kim, S.R.: Emotion recognition system using short-term monitoring of physiological signals. Med Biol Eng Comput **42** (2004) 419–27
16. Pan, J., Tompkins, W.: A real-time qrs detection algorithm. IEEE Trans. Biomed. Eng. **32** (1985)
17. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.H., Kazemzaden, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expression, speech and multimodal information. In: ICMI'04, State College, Pennsylvania, USA (2004) 205–211

Visual Attention in Auditory Display

Thorsten Mahler[1], Pierre Bayerl[2], Heiko Neumann[2], and Michael Weber[1]

[1] Department of Media Informatics
[2] Department of Neuro Informatics
University of Ulm, Ulm, Germany
{thorsten.mahler, pierre.bayerl, heiko.neumann,
michael.weber}@uni-ulm.de

Abstract. The interdisciplinary field of image sonification aims at the transformation of images to auditory signals. It brings together researchers from different fields of computer science like sound synthesizing, data mining and human computer interaction. Its goal is the use of sound and all its attributes to display the data sets itself and thus making the highly developed human aural system usable for data analysis. Unlike previous approaches we aim to sonify images of any kind. We propose that models of visual attention and visual grouping can be utilized to dynamically select relevant visual information to be sonified. For the auditory synthesis we employ an approach, which takes advantage of the sparseness of the selected input data. The presented approach proposes a combination of data sonification approaches, such as auditory scene generation, and models of human visual perception. It extends previous pixel-based transformation algorithms by incorporating mid-level vision coding and high-level control. The mapping utilizes elaborated sound parameters that allow non-trivial orientation and positioning in 3D space.

1 Introduction

Human actions, natural occurrences, movements of any kind produce unique acoustic events. Every acoustic event results from an action which is tightly bound to this exact effect. In the natural world this effect is taken (or should be taken) into account whenever a new product is developed: New motors are designed which reduce noise, new street delimiters cause sounds when run over etc. The interesting aspect here is that the product itself emits characteristic sound as feedback and signals for certain events. This occurrence is addressed in the relatively new field of sonification. In the computer science domain sonification becomes used more often, as well. Three classes of sonification approaches have been proposed previously [Hermann et al., 2000b]:

First, parameter mapping, where image data (e.g. position, luminance) is directly mapped to the parameters of the sound signal (e.g. amplitude, frequency, duration); second, model-based sonification, where virtual sound objects (e.g. instruments) are controlled by the visual input; third, auditory scene generation, where the input data is utilized to control the behavior of defined auditory objects which distinguishes auditory scene generation from simple parameter mapping.

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 65–72, 2006.

2 Related Work

The starting point for our work is the sonification system vOICe introduced by Meijer [Meijer, 1992]. In his work he aims on the substitution of a missing visual sense by sound and introduces a sensor substitution device for the blind. This system is a variation of the parameter mapping approach, where image luminance steers the sound amplitude, vertical image location the sound frequency and horizontal location time and stereo. A drawback of this approach is that the entire data contained in the image is sonified, regardless of the relevance of the information.

Other researchers in this domain sonify not only two dimensional images but high-dimensional data sets. Hermann et al. present in [Hermann et al., 2000a] a model based approach in which significant structures in large datasets are detected. These significant points are integrated in a so called principal curve. This one dimensional description of the dataset is then used for sonification and thus presents a general acoustic overview over the data set.

Rath and Rocchesso present another aspect of sonification in [Rath and Rocchesso, 2005] by introducing a tangible interface. A bar is used as an natural interface for the task of balancing a virtual ball. They use an underlying physical model to immediately and accurately predict and produce rolling ball sounds whenever the bar is agitated. A moving ball is shown on the screen and the bar is used to control its motion and every motion triggers an immediate natural sound event. Thus they combine visual feedback on the screen with aural feedback and thereby could reach a significant increase in usability (average task time).

In [Martins et al., 2001] Martins et al. focus on texture analysis and point out the similarity between speech and texture generation models. Their results especially of the conducted user studies show promising results in that computer vision algorithms are of great use for sonification approaches.

The presented work has to be distinguished from other sonification approaches, in that we focus on sonification of data generated by attention driven perceptual models from images. This allows us to generate intuitively comprehensible auditory displays sonifying highly compressed visual data. Our overall investigations lead us to enhanced visual displays which can be used to support user interaction in a future user interface.

3 Image Preparation

We aim to sonify images of any kind and propose two strategies how this can be achieved. The first strategy is to model visual attention particularly concerning orientation and on this basis a second strategy uses visual grouping to dynamically select relevant visual information to be sonified. We restrict the objects of interest to be represented by elongated regions of similar contrast orientation, such as the borders of a road or the bar of a road sign. Alternatively, other approaches such as [Marr, 1982] [Krüger and Wörgötter, 2005]

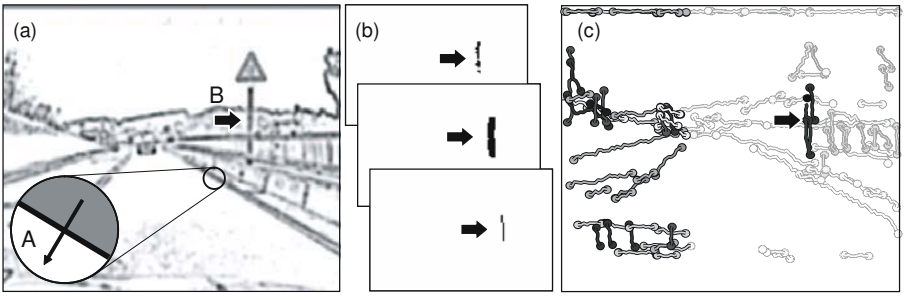


Fig. 1. Example demonstrating initial feature extraction and grouping. (a) Initial features are extracted by a center surround mechanism highlighting isolated contrasts. A feature is located at each position in the image and consists of a salience value (encoded by darker pixels) and an orientation (depicted in inlay A). The arrow (B) points to a vertical contrast which is used to illustrate the grouping scheme in (b-c). (b) Our grouping scheme aims to bind nearby features of similar orientation to individual objects of elongated shape. Selected features are merged and finally thinned by morphological operations in order to generate a thin line depicting the object. (c) shows a set of automatically detected objects represented by connected lines of different luminance. The luminance is dependent on the underlying orientation. The vertical bar extracted in (b) is highlighted and indicated by an arrow.

[Fischer et al., 2004] [Weidenbacher et al., 2005] could also be used to obtain data for sonification.

Our approach is divided in two stages: local feature extraction (one feature per pixel) and the generation of more abstract object descriptions by binding different local features together. The first stage utilizes a local principal component analysis of the luminance gradient in the input image to extract local contrast orientation and amplitude (structure tensor approach [Trucco and Verri, 1998]). Then, attentional mechanisms are borrowed from models of visual perception, such as contrast detectors [Itti et al., 1998] and local normalization [Simoncelli and Heeger, 1998] to enhance perceptual relevant image features. For sonification, we apply a threshold to generate a sparse map of salient oriented contrasts. Individual features are described by their spatial location and orientation (Fig. 1a).

In the second stage, we apply a simple grouping scheme which aims to bind nearby features with similar properties to individual objects. Selected local features are grouped accordingly to local contrast orientation and spatial proximity. The implementation of the grouping employs morphological operations [Gonzalez and Woods, 2001] to achieve the combination of spatial connected locations (Fig. 1b). As a result we extract an image sketch describing connected lines or repeated patterns of similar orientation (Fig. 1c).

Thus, our approach generates (1) sparse features describing the underlying image scene by localizing contrasts described by the position and orientation of these contrasts. Then (2) more abstract objects are generated as combinations

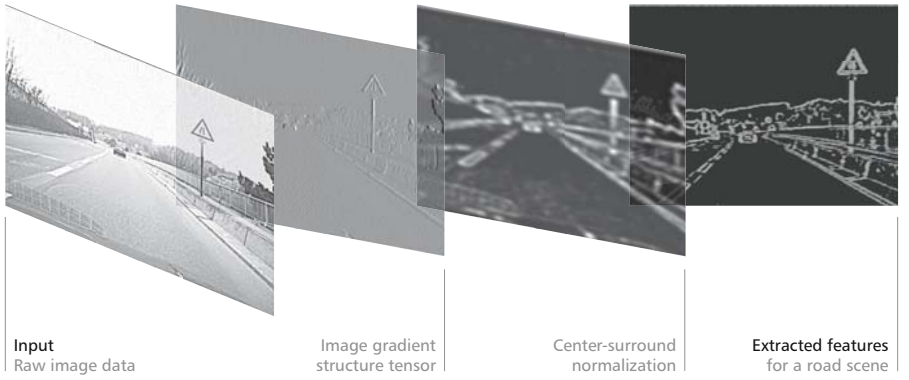


Fig. 2. Visual attention: Salient feature extraction

of local features belonging together described by the size and the shape of the objects in addition to orientation and location of underlying features. Based on the image preparation we apply two models of sonification.

4 Sound Space Modeled Sonification

Of the many different approaches to image sonification we want to present two promising ones here. The first one is an approach where the auditory space is the starting point. We use a grid to divide the auditory space into cells each representing a position in 2D sound space (a plane orthogonal to the vertical axis of the auditory observer). Now an image can be sonified by moving a cursor through the image left to right, line after line, from bottom to top. This left to right direction is used because of the most obvious reason, our natural reading order. The direction of reading an image bottom to top since it seems more natural to display foreground before background. Therefore the objects and streaks in the lower image half are naturally nearer to the spectator and thus more important and seen and therefore sonified first.

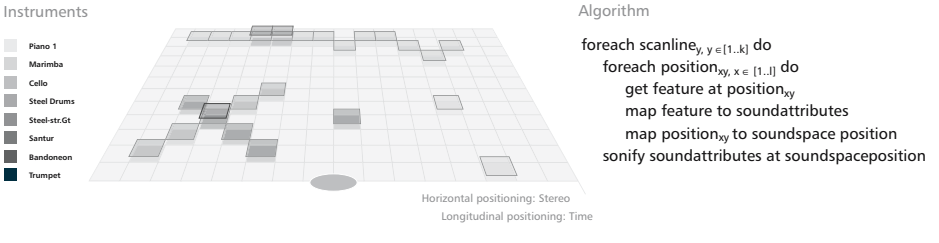


Fig. 3. Sound space modeled sonification: n channels per cell

As a first step in this sonification we transform the original image. We use the thresholded and thus sparse output of the first stage of the presented feature extraction algorithm. The features' orientations (Fig. 2) are mapped to instruments. We use 8 different MIDI instruments corresponding to 8 different orientations. Our pilot investigations showed that raw parameter variations such as changes of tone and pitch are much more difficult to distinguish than complex sound characteristics of known instruments.

After feature extraction our sonification algorithm constantly runs through the image line by line bottom to top. For each line every cell is sonified simultaneously. Corresponding to each orientation found in the image the predefined instrument for this orientation is played at the certain position. The sounds of one scan line are played in parallel using a stereoscopic sound device for auditory localization (Fig. 3).

Using this method the user gets the impression that he is located at the edge of the soundscape. A longitudinal complex stereo sound travels into depth with constantly decreasing volume.

5 Object Modeled Sonification

Sound space modeled sonification presented in the previous section does not take into account individual objects. In contrast to this object modeled sonification uses visual features bound to individual objects in the image to generate auditory features bound to individual sound objects. Visual objects are extracted from the original picture as described above (individual steps of the algorithm are depicted in Fig. 4). Object information is stored as a separate feature set containing a consecutive number and a set of points for each object.

These extracted features are the basis for our object modeled sonification. The idea is to consecutively draw the extracted object strokes into soundspace as continuous sound streams. This is done by first mapping the object stroke to

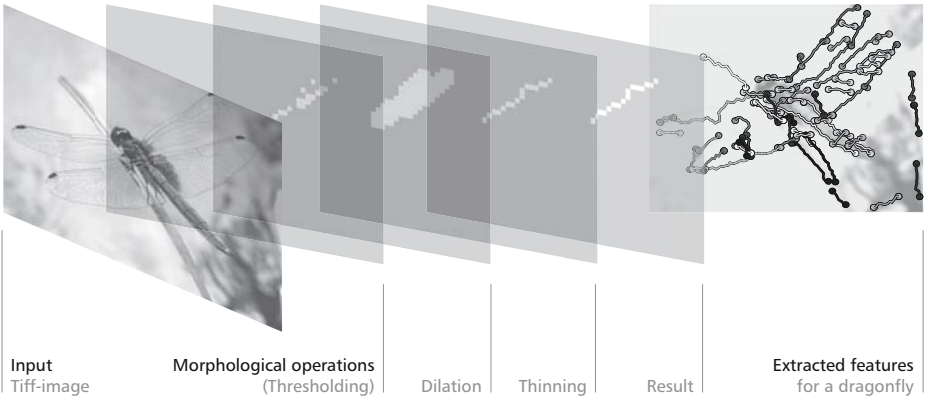


Fig. 4. Visual grouping: Perceptual binding of relatable features

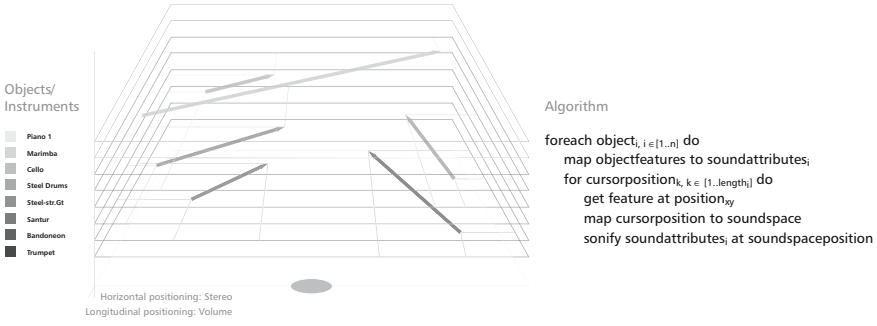


Fig. 5. Object modeled sonification: Free positioning in sound space

an instrument. For example we simply use the given number of the extracted object as an index to the instrument table. Second, the discrete spatial positions within the stroke are mapped to the sound space taking into account that the listener is centered on the horizontal axis and on the edge of the longitudinal axis. This again introduces the possibility to use stereo sound which is very intuitive for horizontal positioning. The longitudinal positioning is indicated by increasing or decreasing volume.

Thus, the listener gets the impression that a virtual instrument is moved through sound space. We again used Midi instruments for sonification after experimenting with drawing multiple strokes simultaneously into sound space (Fig. 5). This is only possible because of the sparseness of the salient feature sets we extract from the images. The bottleneck in sonification is the amount of auditory signals distinguishable by an observer. Compared to the sound space modeled approach we are able to present more compact object representations in the object modeled approach. As a consequence more objects can be sonified simultaneously with this approach.

6 Future Work and Conclusion

In contrast to classical sonification approaches which seek to replace visual displays we plan to enrich visual displays with auditory cues in line with the visual contents. We do not believe that the huge potential of sonification lies not in the substitution of the visual sense but in the extension and enhancement. Sonification techniques can be of great use in for example in user guidance or multidimensional data analysis. Spots of interest can be indicated otherwise lost in the vast amount of data. In surveillance or tracking systems small but important changes can be recognized by the application of the aural system.

We presented how salient image features can automatically be detected by computational models of visual attention to be used for sonification. We depicted different possibilities how such features can be transformed into auditory signals and discussed other sonification approaches previously presented. Our major

contribution is the generation of a comprehensible auditory signal generated from an image. We believe that the strong compression of the visual data necessary to allow an user to interpret a sonified image is possible only by employing perceptual models to extract salient and thus relevant visual information. The two presented approaches are applicable for images of any kind. However we believe that the comprehensibility of the sonification and thus which approach to favor depends on the nature of the original image.

Therefore in the near future we plan to evaluate our approaches in application scenarios in surveillance or in peripheral sensing.

In addition we plan the extension of visual display systems in the future. To overcome major limitations of two and three dimensional displays with limited space and vast degree of detail to display sonification attempts can be used for attentional guidance. They can draw attention to and sonify special information of information units. Furthermore we believe that sonification is a proper way to communicate global background information which can for instance be used with ambient displays.

The nature of sound, its linearity, its strictly increasing character, providing a natural order makes it even more interesting for interactive scenarios. Thus in the long run we plan to incorporate sonification approaches into user interfaces to provide a new way of human computer interaction.

References

[Fischer et al., 2004] Fischer, Bayerl, Neumann, Christobal, and Redondo (2004). Are iterations and curvature useful for tensor voting? In *European Conference on Computer Vision 2004*, volume 3023, pages 158–169. LNCS.

[Gonzalez and Woods, 2001] Gonzalez, R. C. and Woods, R. E. (2001). *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Hermann et al., 2000a] Hermann, T., Meinicke, P., and Ritter, H. (2000a). Principal curve sonification. In Cook, P. R., editor, *Proc. of the Int. Conf. on Auditory Display*, pages 81–86. Int. Community for Auditory Display.

[Hermann et al., 2000b] Hermann, T., Nattkemper, T., Schubert, W., and Ritter, H. (2000b). Sonification of multi-channel image data. In Falavar, V., editor, *Proc. of the Mathematical and Engineering Techniques in Medical and Biological Sciences (METMBS 2000)*, pages 745–750. CSREA Press.

[Itti et al., 1998] Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, 20(11):1254–1259.

[Krüger and Wörgötter, 2005] Krüger, N. and Wörgötter, F. (2005). Symbolic pointillism: Computer art motivated by human brain structures. *Leonardo, MIT Press*, 38(4):337–340.

[Marr, 1982] Marr, D. (1982). *Vision*. W. H. Freeman and Company, New York.

[Martins et al., 2001] Martins, A. C. G., Rangayyan, R. M., and Ruschioni, R. A. (2001). Audification and sonification of texture in images. *Journal of Electronic Imaging*, 10(3):690–705.

[Meijer, 1992] Meijer, P. B. (1992). An experimental system for auditory image representation. *IEEE Transactions on Biomedical Engineering*, 39(2):112–121.

[Rath and Rocchesso, 2005] Rath, M. and Rocchesso, D. (2005). Continuous sonic feedback from a rolling ball. *IEEE Multimedia*, 12(2):60–69.

[Simoncelli and Heeger, 1998] Simoncelli, E. P. and Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research*, 38(5):743–761.

[Trucco and Verri, 1998] Trucco, E. and Verri, A. (1998). *Introductory Techniques for 3-D Computer Vision*. Prentice Hall PTR, Upper Saddle River, NJ, USA.

[Weidenbacher et al., 2005] Weidenbacher, U., Bayerl, P., Fleming, R., and Neumann, H. (2005). Extracting and depicting the 3d shape of specular surfaces. In *Siggraph Symposium on Applied Perception and Graphics in Visualization*, pages 83–86. ACM.

A Perceptually Optimized Scheme for Visualizing Gene Expression Ratios with Confidence Values

Hans A. Kestler[1,2,⋆], Andre Müller[2], Malte Buchholz[2], Thomas M Gress[2,3], and Günther Palm[1]

[1] Department of Neural Information Processing, University of Ulm, 89069 Ulm, Germany
[2] Department of Internal Medicine I, University Hospital Ulm, Robert-Koch-Str. 8, 89081 Ulm, Germany
[3] Division of Gastroenterology and Endocrinology, Department of Internal Medicine, Philipps University, Marburg, Germany

Abstract. Gene expression data studies are often concerned with comparing experimental versus control conditions. Ratios of gene expression values, fold changes, are therefore commonly used as biologically meaningful markers. Visual representations are inevitable for the explorative analysis of data. Fold changes alone are no reliable markers, since low signal intensities may lead to unreliable ratios and should therefore be visually marked less important than the more trustworthy ratios of larger expression values.

Methods: We propose a new visualization scheme showing ratios and their confidence together in one single diagram, enabling a more precise explorative assessment of gene expression data. Basis of the visualization scheme are near-uniform perceptible color scales improving the readability of the commonly used red-green color scale. A sub-sampling algorithm for optimizing color scales is presented. Instead of difficult to read bivariate color maps encoding two variables into a single color we propose the use of colored patches (rectangles) of different sizes representing the absolute values, while representing ratios by a univariate color map. Different pre-processing steps for visual bandwidth limitation and reliability value estimation are proposed.

Results and Conlusions: The proposed bivariate visualization scheme shows a clear perceptible order in ratio and reliability values leading to better and clearer interpretable diagrams. The proposed color scales were specifically adapted to human visual perception. Psychophysical optimized color scales are superior to traditional sRGB red-green maps. This leads to an improved explorative assessment of gene expression data.

1 Background

The results of DNA array expression profiling studies are frequently reported in the form of lists of "experiment vs. control" ratios of expression levels. In

⋆ hans.kestler@uni-ulm.de

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 73–84, 2006.

the case of single color setups such as the Affymetrix GeneChip® technology or radioactive hybridizations on nylon membrane arrays, normalized raw values of single or repetitive experiments are often compared to appropriate controls to identify differentially expressed genes on the basis of expression ratios (e.g. "fold change" values generated by the Affymetrix GeneExpress® software). For studies using glass microarrays, many experimental setups involve the simultaneous hybridization of two samples labeled with different fluorescent dyes onto the same array, where one sample is used as a reference to which the actual experimental sample is compared. The ratios of the absolute expression values for each gene are then used as normalized primary data for further analysis.

The use of expression ratios rather than absolute expression values has the advantage that results obtained with different batches of arrays, different array designs or even completely different technologies can be combined for analysis. Furthermore, the "fold change" measure is the biologically most relevant part of the information generated by comparative hybridizations. Nevertheless, by exclusively focusing on expression ratios, important information relating to absolute gene expression values is lost. Depending on the scope of the study, strongly expressed genes may be more attractive (e.g. as therapeutical or diagnostic targets) than weakly expressed ones. Additionally, the quantification of low signal intensities is decreasingly accurate, sometimes resulting in grossly overestimated expression ratios.

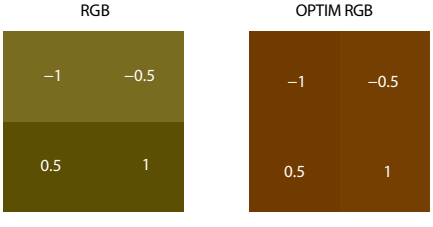


Fig. 1. Uniform versus non-uniform color scales: Four colors in two different color scales visualizing the linear values -1, -0.5, 0.5, and 1 with the total range $[-1, 1]$ are shown. The left diagram shows the standard RGB palette which is used e.g. in the Eisen cluster tool whereas the right diagram shows an optimized RGB palette with near-uniform perceptual distances on computer monitors. It can be seen that the green color in the left diagram is over proportional light and may thus lead to an erroneous interpretation of data.

Visual representations of gene expression ratios are vital for obtaining a general overview of the data sets and are often the first step in generating new hypotheses. The reliability of an observed subset of values is an important factor when assessing data. There is a need for a visual representation which takes these extra information into account.

One possibility to achieve this is to use bivariate color maps e.g. by modulating the hue and the brightness (=value) coordinates in the HSV space in order to represent fold-changes and reliability measures simultaneously. Unfortunately, such color maps are difficult to read since two colorscannot be easily compared

in both variables. Therefore we propose the use of a univariate color map to model one variable in combination with a two dimensional patch grid modeling the other variable by the area of rectangles.

A visualization scheme has to fulfill certain conditions such as the order preservation of the shown values, thus imposing monotonicity constraints on the color scale. Furthermore, the difference of two visualized values x and $x + \delta$ should (at least in a certain neighborhood of x) be nearly proportional to the visual stimulus difference $d(x, x + \delta) \propto \delta$ between the two color representations of the values. In the following, we develop perceptual uniform color scales to provide a most accurate visual representation of data.

The true perception depends on the specific observer, background color, illumination conditions, device properties, distance from the device, object shape, and surrounding objects, which were not taken into account here.

The "standard" red-green palette which is commonly used in microarray studies such as the Eisen cluster/treeview tool [1] is generated by evenly sampling the RGB space. This leads to a perceptually imbalanced scale, i.e. the number of just noticeable differences (JNDs) is not uniformly distributed over the complete palette weighting the green part much stronger than the red part (see figure 1).

Bivariate color maps such as modulating the H and the V channel in the HSV color space allow the representation of two variables in a single color map. One drawback of these bivariate color maps is the difficult readability (compare [2, p. 136]), i.e. when comparing two entities it is not easy to decide which one has the higher ratio and which one has the higher confidence value.

To overcome those problems, we developed a new bivariate visualization scheme which is based on a univariate color map for displaying ratios and a two-dimensional patch grid representing the absolute expression values through patch sizes. A bipolar sequence was chosen since it is important to decide on which side of the zero point (assuming logarithmic ratios) a value lies.

Furthermore, human visual perception suggests the use of red-green color scales [3]. To be in accordance with [1], we chose green for negative values and red for positive values.

2 Methods

The proposed visualization scheme is composed of three parts

1. pre-processing of the gene expression data in order to eliminate unreliable genes and to compute ratios and reliability measures (not further detailed here, see e.g. [4]),
2. a near uniform color scale for visualizing the gene expression ratios,
3. and a two-dimensional grid with rectangular patches coding the reliability values.

2.1 Optimum Color Scales

In the following, we sought color schemes enabling the proportional mapping of values to perceptual stimuli.

The RGB color cube reflects the set of all signals controlling a particular color device (e.g. a color monitor) rather than describing human color perception. The Commission Internationale de l'Eclairage (CIE) proposed psychophysical derived color spaces such as the CIE LUV and CIE LAB space in order to approximate a perceptually uniform color space [5]. The units of the LUV and the LAB space are in *just noticeable differences* (JND's) describing how much two colors differ psychophysically for a standard observer. For the following considerations, we used the LUV space since it is optimized for additive light sources [6, pp.41] such as computer monitors, though any other appropriate color space could be used with our method. Similarly, the CIE LAB space is recommended for modeling reflected light conditions such as print.

```
 1: procedure OPT-SCALE(ĉ1+ ... ĉm+, ĉ1− ... ĉm−, n )
 2:     Δ+ ← d(ĉ1+, ĉm+)
 3:     Δ− ← d(ĉ1−, ĉm−)
 4:     Δ ← min{Δ+, Δ−}/(n − 1)
 5:     c1+ ... cn+ ← OPT(ĉ1+ ... ĉm+, n, Δ)
 6:     c1− ... cn− ← OPT(ĉ1− ... ĉm−, n, Δ)
 7:     return ⟨cn− ... c2−, c1−, c1+ ... cn+⟩
 8: end procedure
 9: procedure OPT(ĉ1 ... ĉm,n,Δ)
10:     c1 ... cn                          ▷ the empty output palette
11:     k ← 2, i ← 1
12:     c1 ← ĉ1
13:     while k ≤ m and i < n do
14:         find j ∈ {k ... m} such that |d(ci, ĉj) − Δ| is minimum
15:         i ← i + 1
16:         ci ← ĉj
17:         k ← j + 1
18:     end while
19:     return ⟨c1 ... cn⟩
20: end procedure
```

Fig. 2. Algorithm OPT-SCALE: This algorithm chooses at most $n \ll m$ colors from a continuous source palette $\langle \hat{c}_1 \dots \hat{c}_m \rangle$ such that all the distances $|d(c_i, c_{i+1}) - \Delta|$ are small for a given $\Delta > 0$ and the color order is not permuted

The LUV space was designed such that the Euclidean distance between two colors c and c' in (L^*, u^*, v^*) coordinates

$$\Delta E^*_{uv} = \sqrt{(\Delta L^*)^2 + (\Delta u^*)^2 + (\Delta v^*)^2}$$

corresponds approximately to the perceptual difference of the stimuli of those two colors. We sought color scales with approximately uniformly distributed points in the LUV space.

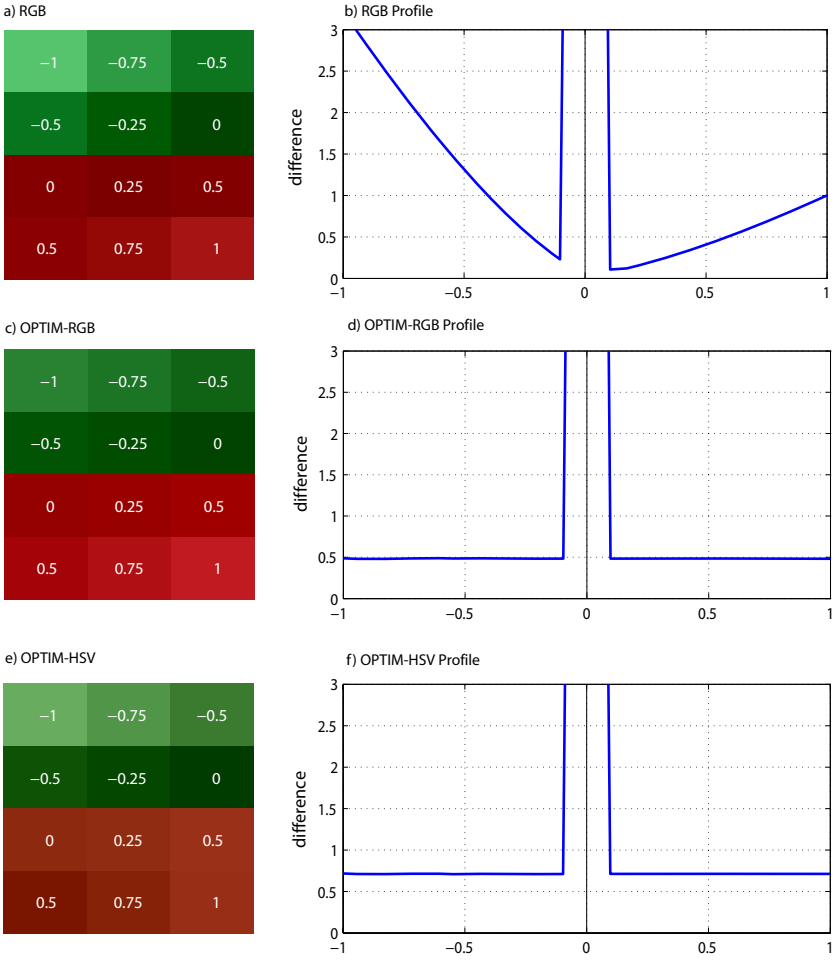


Fig. 3. Perceptual differences of color scales: Diagrams a), c), and e) show examples of the RGB ($\eta = 0$, see text), the OPTIM-RGB ($\eta = 0.1$), and the OPTIM-HSV ($\eta = 0.1$) palette with the corresponding values scaled to the range $[-1, +1]$. The length of the half-palettes were chosen to be $n = 16$. The differences between adjacent horizontal patches are always 0.25 and 0.5 for adjacent vertical patches. It can be seen that the perceptual difference between -1 and -0.75 is much larger than that between -0.75 and 0.5 for the standard RGB scale. Diagrams b), d), and f) show the color difference profiles $n \cdot d(c_i, c_{i+1})$ normalized to the number of colors in the palette for both scales. The RGB palette a) has more (and uneven) JND's on the green part than on the red part. The optimum palettes c), e) show a very uniform difference distribution. Both scales had $n = 33$ values. The optimum palettes were generated by a $r = 128$ times oversampling (see text). For very dark colors the (L^*, u^*, v^*) distances become large (around 0 on the x-axis).

To allow for a widespread usage of our scheme, we used the sRGB color space [7, 8] as a target since sRGB is optimized for computer monitors assuming standardized monitor properties such as the white-point and viewing conditions in order to acquire reproducible results. This was standardized by the International Electrotechnical Commission (IEC) and is referred to as default sRGB IEC 61966-2-1 [9]. In our simulations, the conversion of sRGB colors to the device independent CIE XYZ color space was done using the `sRGB_IEC61966-2-1_withBPC.icc` color profile released by the International Color Consortium (ICC) [8].

The color distance estimation was always performed by transforming a valid sRGB color triplet using the ICC profile into XYZ and finally into (L^*, u^*, v^*) coordinates. Since the ICC transformation process involves table lookups and discretization to 16 bit, a direct sampling in the (L^*, u^*, v^*) space would not necessarily result in a continuous and valid sRGB color scale since the gamut (the set of all colors a device is able to perceive or generate) of the LUV (and XYZ) space - containing all human perceivable colors - is a superset of the sRGB color space. For a good overview of the CIE color spaces see e.g. Hoffmann [10].

The commonly used standard red-green color scale (e.g. [1]) with $n = 2K - 1$ colors is defined by the green channel $g(c_i) = 1 - (i-1)/(K-1)$ for $1 \leq i \leq K$ and $g(c_i) = 0$, $i > K$, the red channel $r(c_{K+1+i}) = i/(K-1)$ for $1 \leq i \leq K-1$ and $r(c_i) = 0$, $i \leq K$, and the blue channel $b(c_i) = 0$ for all i (we define the sRGB coordinates to be in $[0, 1]^3$). So the middle of the scale $(n+1)/2 = K$ is a pure black. Since very low saturated colors are difficult to distinguish [2], we chose to set the saturation to 1.

The visualization concept had the constraints:

1. The absolute value of a ratio is modeled by the brightness. Positive logarithmic ratios are shown in red and negative values are shown in green.
2. There must be a visually perceptible order on the color scale so that it is clear which color corresponds to the greater ratio.
3. The perceptual distances between adjacent colors should be as uniform as possible.

We propose a color search algorithm that was inspired by the linear optimal scale algorithm (LOS) of Levkowitz [6, pp.141] Our algorithm is based on an order preserving sub-sampling of a much larger color sequence $\hat{c}_1 \ldots \hat{c}_m$ fulfilling certain desired perceptually constraints. We chose

1. strictly monotone hue values (H) ranging from black (dark yellow) to green/red (OPTIM-HSV), or constant hue values for each side of the scale (OPTIM-RGB),
2. constant saturation ($s(\hat{c}_i) = 1$ for all i), and
3. strictly increasing brightness values $v(\hat{c}_i) < v(\hat{c}_{i+1})$.

Algorithm OPT-SCALE (see Figure 2) chooses a subset of $n \ll m$ colors $c_1 \ldots c_n$ from the color scale $\hat{c}_i, i = 1 \ldots m$ so that the order criteria are preserved

whereas the perceptual difference between adjacent colors $d(c_i, c_{i+1})$ is optimized to be $\Delta > 0$. If Δ is chosen too large the algorithm may not be able to find enough colors and the output palette could have less than n elements. We chose $\Delta = d(\hat{c}_1, \hat{c}_m)/(n-1)$ to distribute the colors evenly between the first and the last color of the scale. The original color scale $(\hat{c}_i)$ should be a straight line, though Robertson & O'Callaghan [11, p.27] stated that a smooth curving is allowed without affecting the uniformity too strongly in order to exploit the color space more exhaustively.

Algorithm OPT-SCALE optimizes adjacent colors of a scale pair. Since the input scales should be (according to our constraints) not too strongly bent the perceptual differences were approximately $d(c_i, c_{i+k}) \approx k \cdot \Delta$. For greater distances (greater k) in the (L^*, u^*, v^*) space, the perceptual differences become less reliable [11]. Both parts of the scale - black to green and black to red - were interdependently optimized and concatenated since the perceptual bandwidths are different for both parts, which is demonstrated in Figure 3 a).

For the generation of the source color scale, an evenly sampled segment in the HSV space (all HSV and RGB coordinates were defined to be in $[0,1]^3$) was chosen ($h = 0$, $1/6$, $2/6$ correspond to red, yellow, and green, respectively) with saturation $s(\hat{c}_i) = 1$, brightness value $v(\hat{c}_i) = \eta + (1-\eta)(i-1)(m-1)$, and hue $h(\hat{c}_i^g) = 1/6 + 1/6 \cdot (i-1)/(m-1)$ for the green part, and $h(\hat{c}_i^r) = 1/6 - 1/6 \cdot (i-1)/(m-1)$ for the red part of the scale. Since the perceptual difference between black and a pure red/green is unusual high in the CIE LUV space - the algorithm would not find enough colors on the red part of the scale (see above) - we chose a small $\eta = 0.05$ so that both scales are starting at a very dark defined color instead of a true black. Since the perceptual bandwidth of the green part is (according to its distance in the (L^*, u^*, v^*) space) larger than those of the red part, the smaller perceptual bandwidth of the red scale was chosen for both sides $\Delta = d(\hat{c}_1^r, \hat{c}_m^r)/(m-1)$.

The application of algorithm OPT-SCALE (Figure 2) with both scales, lead to the resulting scales $\langle c_1^g \dots c_n^g \rangle$ and $\langle c_1^r \dots c_n^r \rangle$ which were then assembled to the

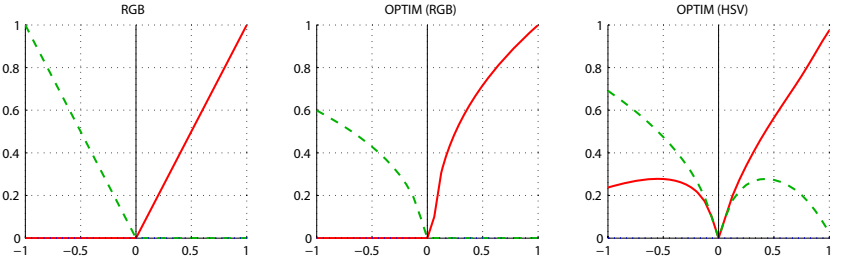


Fig. 4. Proposed color scales: The diagrams show the red channel (straight red line) and the green channel (dashed green line) in RGB coordinates in the range $[0,1]$ for the color scales RGB (left), OPTIM-RGB (middle), and OPTIM-HSV (right). It can be observed that the green part of the scale had to be substantially compressed to get uniform color palettes, in the optimized cases.

final OPTIM-HSV scale $c^t = \langle c^g_n, c^g_{n-1}, \ldots, c^g_1, c_{black}, c^r_1, c^r_2, \ldots, c^r_n \rangle$. Pure black c_{black} is added to the middle of the palette.

The same method was used for a RGB color scale leading to a slightly different scale (OPTIM-RGB). Despite its lower average perceptual distance, compared to the OPTIM-HSV scale, this scale seems to be more intuitive and plausible due to the straightforward ordering of the colors (Figure 3). The $r(c)$ and the $g(c)$ values for the three scales are shown in Figure 4.

2.2 The Visualization Scheme

We propose the use of colored, size varying rectangles to simultaneously represent absolute gene expression values or confidence values and fold changes in a two-dimensional patch grid. The absolute expression values modulate the size of the rectangles (area A), whereas fold changes (ratios) are represented by the rectangle color using one of the previously developed color schemes. We chose black as the background color of the patch grid. The compensation of spatial effects - the perceived color depends on background color, adjacent rectangles, and the exact viewing distance - was not considered in this work. A direct legend comparison would be impossible [11] in this case.

The following three variations enable to observe gene expression fold changes according to different criteria:

(V1) fold changes relative to a fixed sample of interest.
(V2) fold changes among two or more groups of samples. This approach may be useful to quickly assess whether it is possible to discriminate different groups with the data.
(V3) fold changes to the rest of the samples, showing the relative expression value with respect to all other samples. Without pre-information concerning sample groups this approach may give hints if there are relative differences in the expression values or if the samples behave homogeneously.

Let $\mathcal{D} \in \mathbb{R}^{m \times n}$ be a matrix with expression values $d_{ij} \geq 0$ for genes $i = 1 \ldots m$ and samples $j = 1 \ldots n$. Depending on the selected visualization method every gene i in sample j is associated with a corresponding fold change ratio $r_{ij} \in \mathbb{R}$ and an amplitude value $a_{ij} \geq 0$.

The rectangle areas are normalized to the interval $[0, 1]$

$$A_{ij} = \min\{1, \frac{a_{ij}}{\theta_a - a_{min}}\}$$

with $a_{min} = \min_{ij} a_{ij}$, a_{max} similarly, and the cutoff threshold $a_{max} > \theta_a > a_{min}$. The fold change values r_{ij} are normalized to the interval $[-1, 1]$ via the mapping

$$\hat{r}_{ij} = \begin{cases} -r_{ij} / \min_{k,\ell} r_{k\ell} & \text{for } r_{ij} < 0 \\ r_{ij} / \max_{k,\ell} r_{k\ell} & \text{for } r_{ij} \geq 0 \end{cases} \tag{1}$$

thus expanding the positive and the negative fold changes to attain the complete bandwidth. This approach is useful for strongly skewed ratio distributions where

one side of the scale would nearly vanish completely. In this case both sides of the scale must be observed individually and cannot be compared to each other.

The corresponding color is acquired by looking up one of the pre-computed scales with the $\hat{r}_{ij}$ value, so that a ratio of -1 corresponds to the first scale color (green), and a ratio of $+1$ corresponds to the last scale color (red). Intermediate values are linearly interpolated.

Before normalizing the ratios according to equation (1) the dynamics of the r_{ij} is bound by the use of a lower and an upper threshold value determined by the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantile for an $\frac{1}{2} > \alpha > 0$, reflecting the fact that extreme high or low ratios are not reliable. At this step soft thresholds may be used such as e.g. the arctan function.

Limiting the bandwidths of ratios and reliability values is necessary since the dynamics of absolute gene expression values and fold changes exceeds color depth and resolution of computer screens (and those of human vision).

(V1) Single Reference Column: With a pre-defined reference sample (column) $j^* \in \{1 \dots n\}$ the fold changes are defined as

$$r_{ij} = \log_2 \frac{d_{ij}}{d_{ij^*}}$$

for all i, j. The fold changes involving too small values $d_{ij} < \varepsilon$ or $d_{ij^*} < \varepsilon$ for a fixed $\varepsilon > 0$ are discarded and are represented by gray patches.

The reliability value a_{ij} is a function $f(x, y)$ of the value $x = d_{ij}$ (numerator) and the reference value $y = d_{ij^*}$ (denominator). Some proposals for this function are: $f(x, y) = x$ (sample value), $f(x, y) = y$ (reference value), $f(x, y) = \min\{x, y\}$ (pessimistic), etc.

(V2) Class Distinction: Let $\mathbf{c} \in \{1 \dots n_c\}^n$, $n_c \geq 2$ be a vector assigning a group c_j to every sample $j \in \{1 \dots n\}$. The group specific fold change (V2) is then defined as $r_{ij} = f(R_{ij})$ with the set of all log ratios defined by

$$R_{ij} = \left\{ \log_2 \frac{d_{ij}}{d_{ik}} \mid c_k \neq c_j, d_{ik} \geq \varepsilon \right\}$$

and an accumulator function mapping a set of reals to a scalar $f : \mathbb{R}^* \mapsto \mathbb{R}$ which we propose to be the median or the mean value of the set R_{ij}.

(V3) Differential Mode: Visualization (V3) is identical to (V2) by using the group vector $\mathbf{c} \in \mathbb{R}^n$ with $c_j = j$ for $j = 1 \dots n$ thus comparing each column with all other columns. So relative changes of one sample with respect to the rest may be observed. An application of this approach to a pancreas dataset [12] is shown in Figure 5.

The set of all colors an output device can produce is called *gamut*. Obviously, the gamut of a CRT monitor is different from that of a color printer. The printing of gene expression profiles, therefore requires a gamut mapping from the screen representation (to which our scales were optimized) to the printer representation [13, 14].

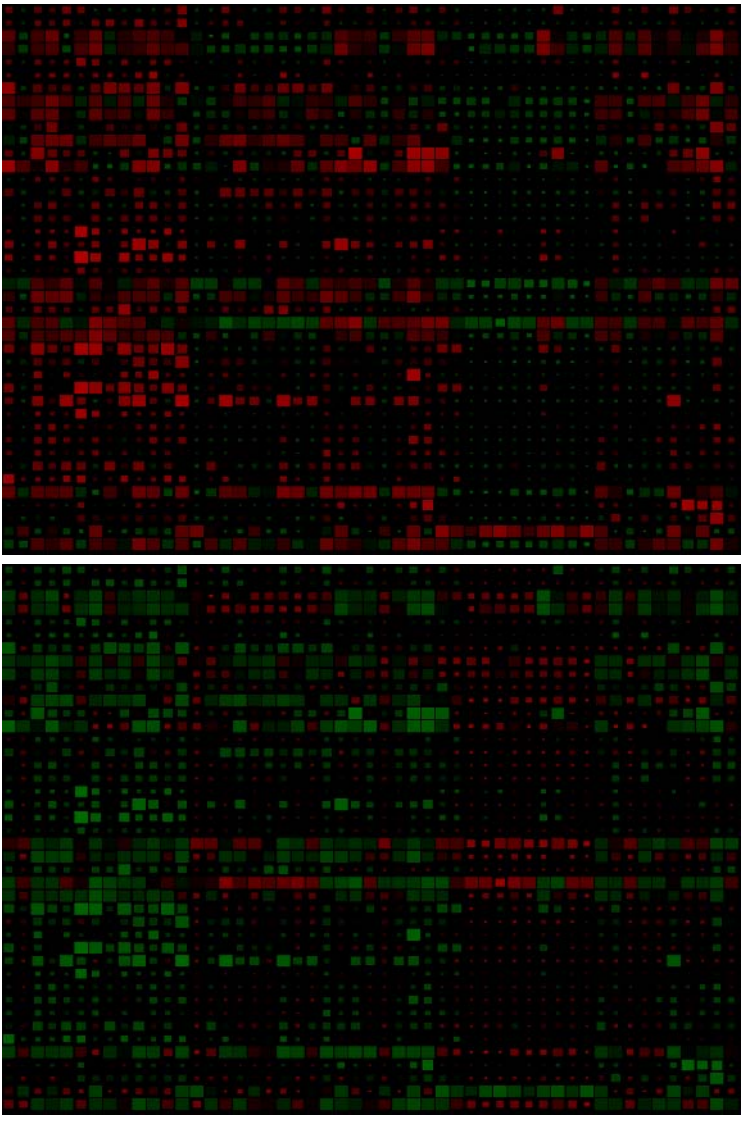


Fig. 5. Ratio visualization (real data): This diagram shows a (clipped) example of visualization approach (V3) applied to a gene expression dataset [12] with 32 positive samples (in the first 32 columns), 19 negative samples (in the last 19 columns), and 231 genes (rows) The ratio cutoff was set to $\alpha = 0.05$ whereas the amplitudes were cut off at the 0.9 quantile. Here, the OPTIM-RGB color scale was used. To illustrate the perceptual similarity (on a computer monitor) a second (lower) panel is generated by exchanging the red and green color scales.

3 Discussion and Conclusion

Since ratios of very low expression values are not reliable (resulting in noise), all genes with uniformly low expression values over all samples should be filtered out in advance to the visualization process (for an overview of gene expression measurement error models, see Strimmer [15]). This can be accomplished by defining a value $\delta > 0$ and counting how many samples of one gene fall below this threshold. If this count exceeds a certain predefined percentage, the gene will be omitted from the analysis.

We have therefore sought to develop a visualization method allowing the combination of expression ratios and absolute expression values for visualization by mapping both values to a two-dimensional patch grid modulating the two orthogonal spaces: area of patches (rectangles) and their corresponding color. We chose this approach since bivariate color maps are difficult to read (compare e.g. [2, p. 136]).

The physiological properties of human vision suggest the use of red-green scales [3, 16] in favor of e.g. yellow-blue scales although the yellow-blue scale is suitable for persons with certain forms of color blindness involving red-green weakness (protanopia and deuteranopia - compare [2, pp. 134]).

We propose two red-green and one blue-yellow color schemes with near optimum perceptual distances for visualizing gene expression ratios. Standard red-green palettes originating from an uniformly (according to the original color space) sampled RGB or HSV color space have unevenly distributed visually perceptible differences, which may lead to biased perception of the data. The first part of this work involves the development of near-uniform color scales, whereas the second part assembles the proposed color schemes into a combined expression value/expression ratio representation showing absolute expression values in combination with fold changes. Absolute values were modeled by the size of rectangles, with areas proportional to the corresponding values and colors depending on the fold changes. The proposed visualization approach may lead to an easier and less error prone assessment of gene expression data.

Acknowledgments

This work is supported by the Stifterverband für die Deutsche Wissenschaft (HAK) and the German Science Foundation, SFB 518, Project C5.

References

1. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. PNAS **95** (1998) 14863–14868
2. Ware, C.: Information Visualization. 2nd edn. Morgan Kaufmann (2004)
3. Spence, I., Efendov, A.: Target detection in scientific visualization. Journal of Exp. Psych.: App. **7** (2001) 13–26
4. Wit, E., McClure, J.: Statistics for Microarrays. Wiley (2004)

5. CIE: Colorimetry, 3rd edition. Technical report, Commission Internationale de l'Eclairage (2004) CIE 15:2004.
6. Levkowitz, H.: Color theory and modeling for computer graphics, visualization, and multimedia applications. Kluwer (1997)
7. Stokes, M., Anderson, M., Chandrasekar, S., Motta, R.: A standard default color space for the internet - srgb. Technical report, W3C - World Wide Web Consortium (1996)
8. ICC: Image technology colour management - architecture, profile format, and data structure. Technical Report ICC.1:2004-10, International Color Consortium (2004)
9. IEC: Multimedia systems and equipment - colour measurement and management - part 2-1: Colour management - default rgb colour space - srgb. Technical report, International Electrotechnical Commission (1999) IEC 61966-2-1.
10. Hoffmann, G.: Cie color space. Technical report, FH Emden, Germany (2005)
11. Robertson, P.K., O'Callaghan, J.F.: The generation of color sequences for univariate and bivariate mapping. IEEE Computer Graphics and Applications **6** (1986) 24–32
12. Fensterer, H., Giehl, K., Buchholz, M., Ellenrieder, V., Buck, A., Kestler, H., Adler, G., Gierschik, P., Gress, T.: Expression profiling of the influence of ras mutants on the tgfb1-induced phenotype of the pancreatic cancer cell line panc-1. Genes Chromosomes Cancer **39** (2004) 224–235
13. Stone, M.C., Cowan, W.B., Beatty, J.C.: Color gamut mapping and the printing of digital color images. ACM Trans. Graph. **7** (1988) 249–292
14. Stone, M.C.: Color printing for computer graphics. In Rogers, D.F., Earnshaw, R.A., eds.: Computer Graphics Techniques - Theory and Practice. Springer (1990) 79–127
15. Strimmer, K.: Modeling gene expression measurement error: a quasi-likelihood approach. BMC Bioinformatics **4** (2003) 10
16. Ware, C.: Color sequences for univariate maps: Theory, experiments and principles. IEEE Comput. Graph. Appl. **8** (1988) 41–49

Combining Speech User Interfaces of Different Applications

Dongyi Song

Department of Informatics, Media Informatics Group
Ludwig-Maximilians-Universität München,
Munich, Germany
Dongyi.song@ifi.lmu.de

Abstract. This paper describes a novel approach to automatically or semi-automatically constructing a multi-application dialogue system based on existing dialogue systems. Nowadays there exist different dialogue systems with general architecture supporting different applications. Yet there is no efficient way to endow the multi-application supported dialogue systems with the corresponding applications. The approach represented in this paper provides an efficient way to integrate different applications into one dialogue system and addresses three issues in multi-application dialogue systems – transparent application switching, task sharing and information sharing by merging the dialogue specifications of different applications into a unified dialogue specification as automatically as possible, which provides necessary domain information for a multi-application supported dialogue system.

1 Introduction

Due to the rapid growing of speech and language processing technologies, speech user interfaces have been developed for a wide variety of applications ranging from information systems to communication systems. To retrieve or update information from various information systems the user has to interact with different speech dialogue systems. Exposing the users to such a complex environment with diverse speech interfaces will result in increased cognitive load and thus bad usability. To enhance the usability a general speech user interface can be constructed. This interface allows the user to access information from different applications simultaneously without awareness of the underlying applications.

For the sake of better understanding of this problem, let's make a simple example. Suppose we have two separate speech user interfaces, one for air ticket reservation and the other for hotel reservation. If a user wants to book an air ticket and afterwards a hotel, he/she calls one speech user interface after the other. A general speech user interface would in this case provide both two functions, so that the user could finish both tasks within the same interface. This general speech user interface has several advantages. First, the user doesn't have to remember or dial different numbers for different services. Second, the common information such as user name, payment information can be shared by both applications. Furthermore, speech user interfaces

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 85 – 96, 2006.

of different information systems and services could be integrated into a single general speech user interface to construct a speech accessing "smart agent". This agent is capable of providing different information about weather, stock, train schedule, etc. and different services such as ticket reservation or hotel reservation.

There exist different multi-application dialogue systems enabling such a general speech user interface. [2, 7, 8] To endow the multi-application dialogue systems with the corresponding applications, there are two different levels for integrating single-applications into a multi-application system – dialogue manager-based integration and dialogue specification-based integration.

Dialogue manager-based integration is suggested in the most multi-domain dialogue systems such as [7, 8]. Different dialogue managers are employed for different applications, and a meta-dialogue manager is applied for managing all dialogue managers. An explicit domain switch is required for switching the appli-cation and there is no interaction between the applications.

As an example for dialogue specification-based integration, an approach for dynamic multi-domain dialogue processing is introduced in [6]. The proposed multi-domain dialogue system focuses on dynamically extending the dialogue systems with new applications. In this system, each domain is specified by a dialogue specification, which is connected in a component named “dialogue specification connection”. The dialogue manager can interpret the content of the dialogue specification connection in runtime and address the right dialogue specification to enable the user to access the desired application. New domains can be easily plugged into the system in runtime by providing the corresponding dialogue specification. However, an explicit domain switch is required because there is always only one dialogue specification active. Further the interoperability between different specifications is not considered in this approach.

Another variant of dialogue specification-based integration is introduced in [2]. This approach brings different applications together into one dialogue system by arranging existing dialogue specifications of each application into an application hierarchy automatically. Based on the description provided by the dialogue specification of each application, the similarity between two applications is calculated. Based on the similarity, the applications are clustered in a binary tree with the most similar applications clustered under the same node in the tree. Given a user utterance, the similarity between the utterance and all applications is computed. If the similarity between the utterance and a single application is beyond a predefined threshold, the desired application is determined. Otherwise the dialogue system navigates along the binary tree with clarification dialogues until the right application is reached. Transparent application switching is enabled by this approach. But interoperability among applications regarding the task sharing problems is not solved in the work yet and is addressed as future work in [2].

Instead of treating each application separately, in this paper I propose a novel approach to merge existing dialogue specifications of different applications automatically or semi-automatically into a unified dialogue specification. By integrating multiple applications into a big unified application in the dialogue specification level, the frame-based dialogue systems supporting sets of contexts [4] such as [1] provide automatically the transparent access to all underlying applications. And based on the analysis of dialogue specifications of different applications, the task

sharing and information sharing issues are addressed in this approach. To minimize the complexity of constructing such a system, the combination tool proposed in this paper supports semi-automatic and in best case automatic generation of the multi-application dialogue system based on existing dialogue systems.

In section 2, I declare the requirements on the dialogue systems and the corresponding dialogue specifications for applying the combination approach. In section 3, I give a formal definition of frame-based dialogue specifications, which conform to the requirements and describe two examples. In section 4, I propose the combination scheme for constructing a general speech user interface for different applications by combing their dialogue specifications. Section 5 introduces the interactive combination tool enabling semi-automatic construction of multi-application dialogue system based on existing dialogue systems. A dialogue example is illustrated in section 6. The section 7 summarizes and discusses open questions for further research.

2 Requirements on Spoken Dialogue Systems and Dialogue Specifications

We do not intent to combine dialogue specifications following different dialogue models, but rather aim to combine dialogue specifications for different applications, which are constructed according the same dialogue model and are thus supported by the same spoken dialogue system. Since most spoken dialogue systems strive to support as many applications as possible, this assumption does not make evident restriction on the combined applications. However, in order to apply the combination scheme represented in this paper, there are some requirements on the adopted dialogue systems and respectively the dialogue specifications.

The adopted spoken dialogue system must be application-independent. In the system, the dialogue management mechanism is transparent to different applications. The system can be regarded as an aggregation of domain-independent dialogue management components and domain-specific knowledge. Porting the spoken dialogue system to a new application means to provide the dialogue specification of it.

The adopted dialogue specifications for defining the application-specific information in the dialogue system must be formal and declarative. An example is VoiceXML [3] or DIANEXML [1]. Because only declarative specifications can be compared and merged, native coding e.g. program codes implementation is not appropriate for these issues.

Among three main classes of dialogue modeling approaches [2], the frame-based dialogue modeling approach is best tailored to our requirements. In a frame-based approach, the algorithm for controlling the interaction between the user and the system is defined in the dialogue manager. The domain information needed for accessing the actual application is specified in a set of frames. Each frame corresponds to a function provided by the application. The necessary information for executing the function is modeled as parameters of the frame. The deployment of a frame-based spoken dialogue system to a specific application requires only the declarative description of that application using a set of frames. In practice, there are

many frame-based dialogue systems indicating the maturity and advantages of the approach. [1, 7]

The combination scheme proposed here is based on the assumption that a frame-based spoken dialogue system is used to support a set of applications, whose dialogue specifications are defined formally and declaratively. To combine exactly this set of applications is the issue addressed in this paper.

3 Frame-Based Dialogue Modeling Approach

In this section, I give a formal definition for frame-based dialogue modeling approaches, which abstracts different implementation variants.

In a frame-based approach an application is modeled as a set of frames.

$$A = \{F_1, F_2, ...F_n\}$$

There are different interpretations of a frame in different approaches. Abstractly, we could specify a frame with the following form:

$$F_i =< ID, FGr, \{PROMPT_1, ..., PROMPT_n\}, \{P_1, ..., P_m\}, Post >$$

$$P_i =< ID, PGr, Infer, \{PPrompt_1, ..., PPrompt_k\} >$$

In short, a frame consists of the following elements:

- *ID* – identity of a frame inside a dialogue specification
- *FGr* – Key grammar (a context free grammar) defining all possible key words identifying a frame. E.g. the frame for "hotel reservation" will be identified by the word "hotel reservation", "book a hotel", "reserve a hotel", etc.
- *PROMPT* – Various prompts, which the system utters with respect to the frame in different situations, e.g. to clarify the user's intention of executing one frame from a set of candidates.
- *P* – Parameters of a frame modeling the necessary information required for executing the frame. Each parameter contains the following elements:
  - *ID* – identity of a parameter inside a dialogue specification
  - *PGr* – A context-free grammar defining all possible natural language user input for the parameter
  - *Infer* – An inference rule for inferring the parameter value based on the dialogue state by the system without asking the user [1]
  - *PPROMPT* – different prompts, which the system utters with respect to the parameter in different situation, e.g. query prompt, confirmation prompt, etc.
- *POST* – Execution of the frame in the backend application and the corresponding information about the execution result

For example, the dialogue specification of an application for "flight reservation (AFR)" consisting of two functions –"flight reservation (FR)" and "check create information (FCI)" would be modeled as $AFR = \{FR, FCI\}$ and the following table gives a detailed specification for each frame and an example for parameter specification.

Table 1. Frame specification of AFR

	FR	FCI
ID	„flight_reservation“	„credit_information“
FGR	{“book an air ticket”, “reserve an air ticket”…}	{“credit card information“, “credit card“ …}
$Prompt_i$	“Do you want to book an air ticket?”	“Do you want to change your credit card information?”
$\{P_1...P_n\}$	{DEPARTURE_CITY, DESTINATION_CITY, DEPARTURE_DATE, DEPARTURE_TIME}	{LAST_NAME, FIRST_NAME}
POST	Reserve the ticket in backend system Prompt “you ticket have been reserved.”	Get the credit card information from backend system and prompt the user about it, i.e. “Your Visa card number is ….” or “ you haven’t provided any credit card information so far. “

Table 2. Parameter example of AFR

	DEPATURE_CITY
ID	„DEPATURE_CITY“
PGr	{„Munich“, „Berlin“, „Frankfurt“…}
Infer	No inference rule
$PPrompt_i$	“Where do you want to fly from?”

Table 3. Frame specification of AHR

	HR	HCI
ID	“hotel_reservation”	“Credit_information”
FGR	{“book a hotel”, “reserve a room”…}	{“credit information”, “ask for credit card information”, …}
$Prompt_i$	“Do you want to book a hotel?”	“Do you want to check your credit card information?”
$\{P_1...P_n\}$	{CITY, DATE, DURATION}	{LAST_NAME, FIRST_NAME}
POST	Reserve the room in the hotel in backend system Prompt “Thank you! The hotel has been reserved. ”	Get the credit card information from backend system and prompt the user about it, i.e. “Your Visa card number is ….”

For the combination purpose, I introduce another modeling example for a hotel reservation application (AHR) consisting of functions "hotel reservation (HR)" and "check credit card information (HCI)". The application is modeled as $AHR = \{HR, HCI\}$ with the following specification for each frame:

4 Combination Scheme

4.1 Basic Principle

The combination approach proposed here is inspired by the idea that different frames of two dialogue specifications can be merged together into one unified dialogue specification:

$$A_1 = \{F_{11}, F_{12}, ..., F_{1n}\} \quad A_2 = \{F_{21}, F_{22}, ..., F_{2n}\}$$

$$A_1 + A_2 = \{F_{11}, F_{12}, ..., F_{1n}, F_{21}, F_{22}, ..., F_{2n}\}$$

An application transparent dialogue system supporting A_1 and A_2 supports the unified dialogue specification naturally. Since a frame-based dialogue system provide transparent access to different functions within one application, after two applications are build into one "big application", the transparent access to functions of both applications can be obviously enabled by the same dialogue system. Most existing multi-application dialogue systems integrate different applications on the application layer, where as the novel approach here merge them into one homogenous application at the function layer.

With this basic principle, based on speech user interfaces of two disjunctive applications we can construct a multi-application dialogue system automatically. Disjunctive applications differ substantially in contents and possible uses. There are no overlaps between them. If each of them is modeled as a set of frames with each frame corresponding to an autonomous function, combining these applications can be carried out by unifying the two sets of frames. For example, home intelligent environment and remote access to information database [5] can be merged all automatically if they are specified according to the formalization introduced in section 3. It is then able to access services of both applications via a single spoken interface simultaneously without any explicit domain switching.

Obviously, this basic idea is best tailored for the disjunctive applications without any functional or semantic overlap. If two applications provide overlapped functions, which become all simultaneously active in the general speech user interface, it turns to be a problem for the dialogue manager to address the right function wished by the user. Also if two applications have some common information such as a user's credit information, such information should be shared by both applications when a general speech user interface for both is provided.

In next two sections, I define these overlaps and propose a method to find them and solve them semi-automatically.

4.2 Functional Overlap

Two applications can provide common functions for the user, and I refer this situation as functional overlap between two applications.

4.2.1 Definition

Two frames are regarded as similar, if they perform the same function/task for the user. Based on the formal definition of a dialogue specification, functional overlap can be defined in the following form:

$$A_1 \text{ and } A_2 \text{ have functional overlap } \leftrightarrow \exists F_i \in A_1, F_j \in A_2 [F_i \approx F_j]$$

Two applications have functional overlap, if there is some frame in A1, which is similar as some frame in A2.

In the examples introduced in section 3, AFR and AHR both provide the same function for “checking credit card information”. The corresponding frames CIF (AFR) and CIH (AHR) are similar and AFR and AHR have functional overlap.

4.2.2 Recognition

Considering all elements of a frame, the most important element, which can indicate the similarity of functions provided by two frames, is the key grammar. In a key grammar, all possible phrases (key phrases) indicating the function of the frame are defined. The key grammar distinguishes a frame from the other. A speech user interface should accept as many key phrases as possible, so that the user has as much flexibility as possible to express his/her intentions. For example, for a frame "flight reservation", there are different possible key phrases such as "book a ticket", "make a reservation", “reserve a flight”, etc. The speech user interface should be able to accept all these expressions, so that the user does not have to remember the "command" for the frame, and the dialogue can remain natural and intelligent. So we could assume all possible natural utterances representing a frame are defined in its key grammars. So we can make the following proposition:

If there is an identical key phrase defined in key grammars of two frames, the frames are functionally identical - similar. Otherwise, they are different. This can be defined by the following form:

$$F_1 =< ID_1, FGr_1, \{PROMPT_{11}, ..., PROMPT_{1n1}\}, \{P_{11}, ..., P_{1n2}\}, Post_1 >$$

$$F_2 =< ID_2, FGr_2, \{PROMPT_{21}, ..., PROMPT_{2n1}\}, \{P_{21}, ..., P_{2n2}\}, Post_2 >$$

$$\exists w[w \in L(FGr_1) \wedge w \in L(FGr_2)] \leftrightarrow F_1 \approx F_2$$

L(G) represents the language defined by the grammar G. An algorithm to compare the key grammars has been designed. But introducing this algorithm would go beyond the scope of this paper.

4.2.3 Solution

To solve the functional overlaps, I propose to merge two similar frames into one frame of the general speech user interface. Because of the fact that it is not relevant for the user which backend application executes the function. He/She only cares the result. If

the user asks for his credit card information, he does not care if the credit information is provided by the "flight reservation system" or the "hotel reservation system" (in case we suppose that the user always uses the same credit card in both systems).

The merged frame is constructed by inheriting dialogue modeling elements from the original frames and determining the right backend application to perform the function.

$$merge(F_1, F_2) =< ID_i, FGr_i, \{PROMPT_{i1},..., PROMPT_{in1}\}, \{P_{i1},..., P_{in2}\}, merge(Post_1, Post_2) >$$

$$i \in \{1,2\}$$

Since two similar frames perform the same function for the user, their dialogue modeling parts are almost identical. So we can take any frame's dialogue modeling part. However, it is not always trivial how the postconditions should be merged together. The function can be performed by one of the applications or by both. For example, the frame FCI in AFR and HCI in AHR in the example, if it is allowed to have different credit cards for "hotel reservation" and "flight reservation", the merged "credit card information" frame should invoke both applications to provide both credit cards information, otherwise it suffices to perform the function by only one application. The right decision is always domain-specific, so there is no generic pattern for merging postconditions. For this issue, I have constructed a combination tool to involve the dialogue designer to make the right merging decision for postconditions. In the next section, this tool is explained in detail.

4.3 Semantic Overlap

Different applications can share common information with each other such as username, password, etc. In such cases, we refer the situation as semantic overlap.

4.3.1 Definition

We say two parameters are similar, if they refer to the same semantic concept and are related to each other. So semantic overlap can be defined in the following form:

A_1 and A_2 has semantic overlap $\leftrightarrow$

$$\exists P_i \in F_n, F_n \in A_1, P_j \in F_m, F_m \in A_2 [P_i \approx P_j \wedge F_n \neq F_m]$$

Two different applications have semantic overlap, if there is some parameter in application A_1, which is similar to some parameter in application A_2. And we do not consider the parameters in similar frames with respect to their similarity.

E.g. the frame FR in application AFR and the frame HR in application AHR have two semantic overlaps - DESTINATION_CITY and CITY, DEPARTURE_DATE and DATE. The destination city of the flight is probably the accommodation city of the hotel. And the departure date of the flight is probably the hotel check in date, if we assume the flight to be a short trip.

4.3.2 Recognition

A parameter defines its range of values with its parameter grammar. If two parameters refer to the same semantic and are related to each other, they must have at least one common value in their value ranges. Otherwise, they would never have the same

value, so there is no chance for information sharing in two applications with respect to these parameters. This can be defined by the following form:

$$P_1 =< ID_1, PGr_1, Infer_1, \{PPrompt_{11}, ..., PPrompt_{1n}\} >$$

$$P_2 =< ID_2, PGr_2, Infer_2, \{PPrompt_{21}, ..., PPrompt_{2n}\} >$$

$$\exists w[w \in L(PGr_1) \wedge w \in L(PGr_2)] \leftrightarrow P_1 \approx P_2$$

Thus if there is any common word defined by the grammars of two parameters from two applications, it is possible for these two applications to share the corresponding information represented by the parameters.

An algorithm to compare the parameter grammars has been designed. But introducing this algorithm would go beyond the scope of this paper. This algorithm can compare parameter grammars and find potential "information shared parameters" automatically. However, not all parameters with common words must share the information with each other. E.g. a date grammar may define a word “Friday” and a last name grammar may define the same word “Friday” as a valid last name. To solve this kind of ambiguity of natural languages, the interactive combination tool confirms each potential “semantic overlap” with the designer.

4.3.3 Solution

To solve the semantic overlap, common semantic information should be shared by different applications. This can be achieved using the inference rule of each parameter. In simple case, the related parameters always have identical value and this can be specified by defining mutual inference in the Infer elements of the parameters. For example, the Infer of CITY in HR will be specified as CITY=ARRIVAL_CITY and the DATE of HR will have an inference rule: DATE = DEPARTURE_DATE.

In practice, the relation could be a little bit complicated. For example, the DATE of HR could be one day later than DEPARTURE_DATE if we consider the inter-continental flight. Such inference rules can not be generated automatically, so the combination tool allows for customized inference rule in case of need.

5 Combination Tool

I have developed a prototype of an interactive combination tool according to the combination scheme introduced in section 3. This tool makes the combination process as automatic as possible and enables the involvement of the designer.

In this section I introduce the work flow of the combination tool. Figure 1 illustrates this combination process.

White boxes refer to the steps performed by the combination tool automatically. Grey boxes indicate steps requiring designer's cooperation. The combination tool works as follows:

1. The combination tool obtains two dialogue specifications D1 and D2 as input.
2. The tool compares D1 and D2, finds the set of overlaps S which has to be handled.
3. If the set S is not empty, go to step 4. If the set S is empty, it means that there is no overlap between two speech applications or the overlaps have all been handled appropriately, then go to step 7.

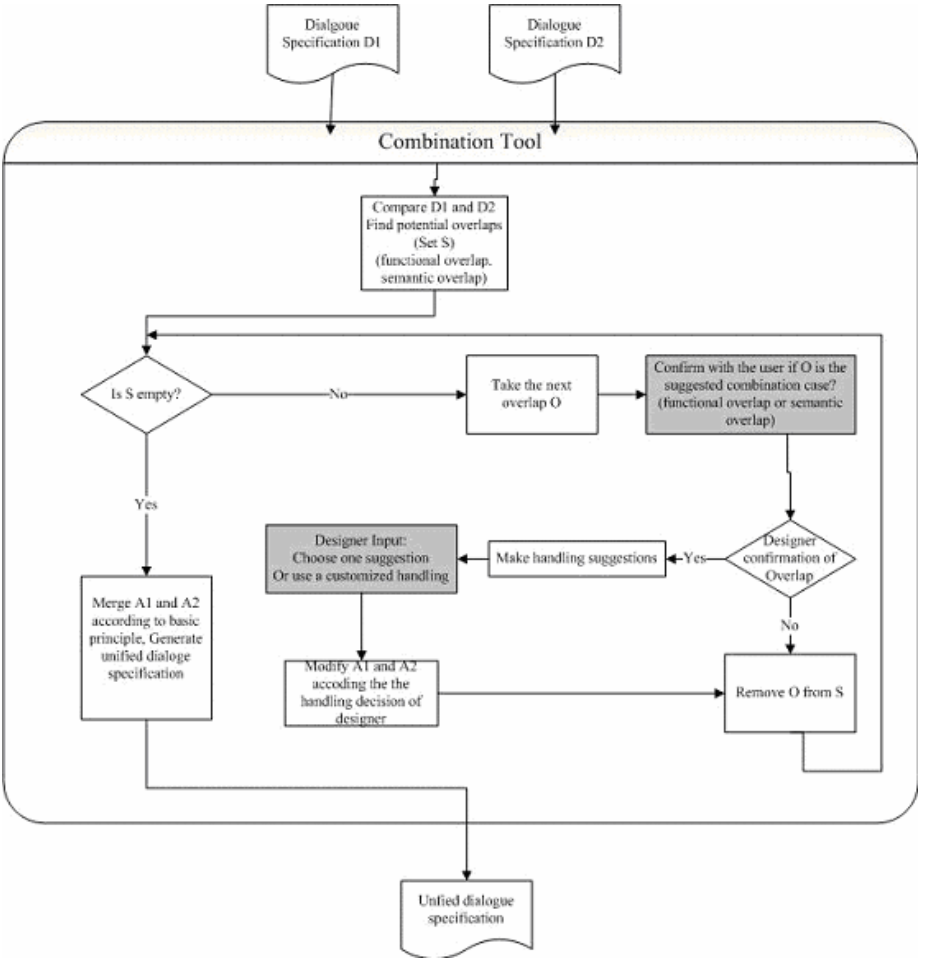


Fig. 1. Work flow of the combination tool

4. The tool takes the next overlap O in the set S and asks the designer for a confirmation if the judged overlap is valid. If the designer declines the judgment, this overlap is removed from set S, and go to step 3. Otherwise the designer confirms the tool's judgment, then go to step 5.
5. The Tool makes several suggestions to handle the overlap O. The designer can choose one suggestion to solve the overlap or input a customized modification.
6. The decision in step 5 will be applied to D1 and D2. The overlap O will be removed from the set S. Go back to step 3.
7. A unified dialogue specification for both applications will be generated according to the basic combination principle, since all overlaps have been handled in the previous steps.

In this process model the tool cooperates with the designer together in order to generate an intelligent speech user interface with acceptable efforts from the designer side.

6 Example

I have combined AFR and AHR with the combination tool. The generated dialogue system supported by the dialogue manager DIANE [1] provides a dialogue as the following example:

S (System): How can I help you?
U (User): I want to fly from Munich to Berlin tomorrow morning at ten o'clock.
S: Flight 731 from Munich to Berlin, starting at ten o'clock on 14th September, Do you want to take this flight?
U: Yes.
S: Your ticket has been reserved.
U: I want to book a hotel.
S: Do you want to reserve a hotel on 14th September in Berlin?
(The system infers the date and city from the previous frame)
U: Yes.
S: The hotel has been booked.
U: Which credit card has been taken actually?
S: Visa card number 12345678.
(FCI and HCI are merged into one and only AFR was chosen to perform the function.)

7 Conclusion and Outlook

The combination scheme proposed in this paper can generate a unified dialogue specification based on declarative specifications of two applications. This general speech user interface enables speech access to both applications. The construction of this multi-application dialogue system is automatic in simple case and is semi-automatic with support of a combination tool, which allows for configuration of a dialogue designer. More speech applications can be combined together by applying the same process iteratively to add other speech applications to the merged application.

With this approach an intelligent general speech user interface can be generated for different applications. This interface improves the rapid provisioning of speech user interfaces through its intelligent nature.

Compared with existing multi-domain or multi-application dialogue system, the novelty of the approach proposed in this paper is the idea of merging different dialogue specifications at the function layer thus enabling transparent access to different applications with a normal frame-based dialogue system. The advantage gained by the approach is an acceptable solution for task and information sharing across applications.

The combination approach has been implemented based on the DIANE speech dialogue system [1] and is fully operational. Case studies with real industrial speech

dialogues from the area of telecommunication system configuration have been carried out.

To extend the current work, an evaluation of the combination tool based on user tests can be carried out for verifying the feasibility of the combination approach. Furthermore, user tests of the generated speech user interface can be carried out for verifying the usability improvement.

Acknowledgements

This work was partly funded by the German Federal Ministry of Education and Research (BMBF) under grant 01IMD01K. The author is solely responsible for the view expressed in this paper. The author wishes to thank Prof. Heinrich Hussmann, Dr. Hans-Ulrich Block and Jürgen Totzke for their kind support during the research and Siemens for funding.

References

1. Block, H.U., Caspari, R. and Schachtl, S.: Callable manuals - access to product documentation via voice. In: Wahlster, W. (ed.) Special Journal Issue Conversational User Interface. it - Information Technology 46 (2004) 6, München, Oldenbourg Wissenschaftsverlag (ISSN 1611-2776), Munich, (2004) 299-304
2. Bui, T., Zwiers, J., Nijholt, A. and Poel, M.: Generic dialogue modeling for multi-application dialogue systems. In 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, July 13 2005, Edinburgh, UK, (2005) pp. 12
3. McGlashan, S., Burnett, D., Carter, J., Tryphonas, S., Ferrans, J., Hunt, A., Lucas, B. and Porter. B.: Voice extensible markup language (voicexml) version 2.0. Technical report, World Wide Web Consortium (W3C), Feb. 2003, http://www.w3.org/TR/voicexml20/
4. McTear, M.: Spoken dialogue technology: enabling the conversational interface. ACM Computing Surveys, Vol. 34(1), (2002) 90-169
5. Neto, Joao p., Mamede, Nuno J., Casaca, Renato and Oliveira, Luis C.: The development of a multi-purpose spoken dialogue system. Proc. EUROSPEECH 2003, (2003)
6. Pakucs, B.: Towards dynamic multi-domain dialogue processing. In Proceedings of EUROSPEECH 2003, Geneva, (2003)
7. Polifroni, J. and Chung, G.: Promoting portability in dialogue management. Proc. ICSLP, Denver, CO, (2002) 2721-2724
8. Seneff, S., Lau, R. and Polifroni, J.: Organization, communication, and control in the GALAXY-II conversational system. Proc. Eurospeech '99, Budapest, (1999) 1271--1274

Learning and Forgetting of Speech Commands in Automotive Environments

Alexander Hof and Eli Hagen

Forschungs- und Innovationszentrum,
BMW Group, Munich,
{eli.hagen, alexander.hof}@bmw.de

Abstract. In this paper we deal with learning and forgetting of speech commands in speech dialogue systems. We discuss two mathematical models on learning and four models on forgetting. Furthermore we describe the experiments used to determine the learning and forgetting curve in our environment. These findings are compared to the theoretical models and based on this we deduce the equations that describe learning and forgetting in our automotive environment most adequately.

1 Introduction

Modern premium class vehicles contain a large number of driver information and driving assistance systems. Therefore the need for enhanced display and control concepts arose. BMW's iDrive is one of these concepts, allowing the driver to choose functions by a visual-haptic interface (see Fig. 1) [Haller, 2003]. Since

Fig. 1. iDrive controller and Central Information Display (CID)

such an interface can according to Vollrath and Totzke [2003] lead to driver distraction, iDrive includes a speech dialogue system (SDS) to overcome this disadvantage. The SDS allows the driver to use a large number of functions via speech commands [Hagen et. al., 2004]. The system offers a help function that can be activated by uttering the keyword 'options'. The options provide help in the form of a list containing speech commands available in the current context. Currently neither the driver's preferences nor his knowledge is taken into consideration.

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 97–106, 2006.

Sample options dialogue.
User: "Phone."
System: "Phone. Say dial name, dial number or name a list."
User: Options.
System: "Options. Say dial followed by a name, for example 'dial Alex', or say dial name, dial number, save number, phone book, quick dialing list, top eight, last eight, accepted calls, missed calls, active calls and or or off."

Our basic concern was to reduce the driver's memory load by reducing irrelevant information. An adaptive help system based upon an individual user model could overcome these disadvantages. In Komatani et. al. [2003] and Libuda and Kraiss [2003], several adaptive components can be included to improve dialogue systems, e.g. user and content adaption, situation adaption and task adaption. In our system we concentrate on user modeling and content adaption.

In this paper we present studies concerning learning and forgetting of speech commands in automotive environments. The results of this study are used to establish a model regarding the driver's knowledge in our SDS domain. This model is used to adapt the content of the optionlists.

2 Learning of Commands

In this section we determine which function most adequately describes learning in our environment. In the literature, two functions that mathematically describe the process of learning complex skills can be found. These laws help to predict the time which is necessary to achieve a task after several trials. One model was suggested by Newell and Rosenbloom [1981] and describes learning with a power law. Recent works on learning theories suggest to discard the power law of practice and thus concentrate on an exponential law [Heathcote et. al., 2002].

$$T = B \cdot N^{-\alpha} \quad \text{(power law)} \tag{1}$$

$$T = B \cdot e^{-\alpha \cdot N} \quad \text{(exponential law)} \tag{2}$$

T represents the time to solve a task, B is the time that is needed at the first trial of a task, N stands for the number of trials and α is the learning rate parameter that is a measure for the learning speed. This parameter α has to be determined empirically. Therefore we conducted memory tests to determine, which of the the two laws best describes the learning curve for our specific environment and depending on the results determine the value of α.

2.1 Test Design for Learning Experiments

The test group consisted of seven persons. The subjects' age ranged from 26 to 43 years. Five of the subjects had no experience with an SDS, two had very little. Novice users were needed because we wanted to observe only novice learning

Table 1. Tasks for learning curve experiments

Task 1	Listen to a radio station with a specific frequency
Task 2	Summary of already used destinations
Task 3	Enter a new destination
Task 4	Start navigation
Task 5	Turn off speech hints
Task 6	3D map
Task 7	Change map scale
Task 8	Avoid highways for route calculation
Task 9	Turn on CD
Task 10	Display the car's fuel consumption

behaviour. The tests lasted about one hour and were conducted in a BMW driving a predefined route with moderate traffic.

Each subject had to learn a given set of ten commands with differing levels of complexity (see table 1). Complexity is measured by the minimal necessary dialogue steps to reach a function. The tasks were not directly named but explained in order not to mention the actual command and thus avoid any influence on the learning process. There was no help allowed except the options function. The subjects received the tasks one by one and had to search for the corresponding speech function in the options. After completion of a task in the testset the next task was presented. The procedure was repeated until all commands had been memorized. For each trial, we measured the time span from SDS activation until the correct speech command was spoken. In order to take the different levels of complexity of a function into account, the time spans were standardized by dividing them through the number of the minimal necessary steps that had to be taken to reach a function.

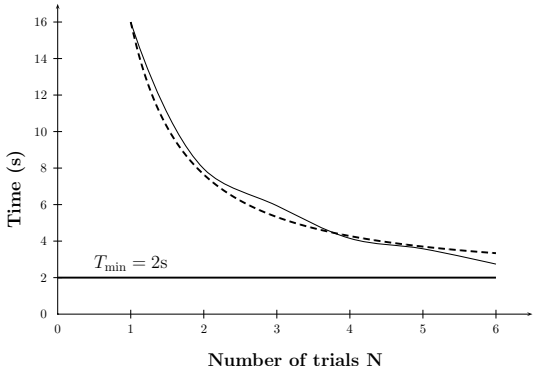
2.2 Results

The resulting learning curve is shown in Fig. 2. In order to determine whether equation (1) or (2) describes this curve more exactly, we used a chi-square goodness-of-fit test [Rasch et. al., 2004]. The more χ^2 tends to zero, the less the observed values (f_o) differ from the estimated values (f_e).

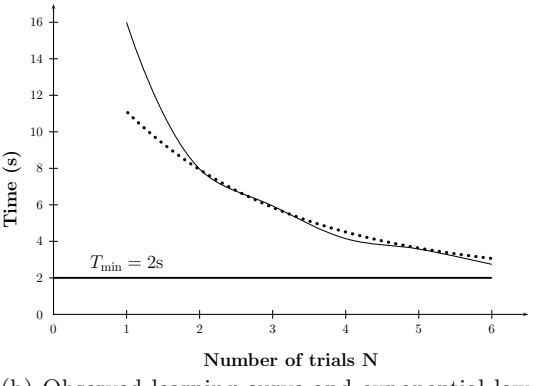
$$\chi^2 = \sum_{i=1}^{k} \frac{(f_o - f_e)^2}{f_e} \tag{3}$$

The power law has a minimum ($\chi^2_{\min} = 0.42$) with a learning rate parameter of $\alpha = 1.31$. The exponential law has its minimum ($\chi^2_{\min} = 2.72$) with $\alpha = 0.41$. This means that the values of the exponential law differ more from the actual value than the power law's values. Therefore we use the power law (see Fig. 2(a)) to describe learning in our environment.

In general we can say that learning takes place very fast in the beginning and with an increasing amount of trials the learning curve flattens and approximates an asymptote. The asymptote at $T_{\min} = 2$s defines the maximum expert level, that means that a certain task can't be completed faster.



(a) Observed learning curve and power law (dashed) with $\alpha = 1.31$



(b) Observed learning curve and exponential law (dotted) with $\alpha = 0.42$

Fig. 2. Learning curves

3 Forgetting of Commands

The second factor influencing our algorithm for the calculation of options is forgetting. If a command was not in use for a longer period of time, we can assume that this command will be forgotten. In this section we determine how long commands are remembered and deduce a function, which most adequately describes the process of forgetting in our environment. In Rubin and Wenzel [1996] 105 mathematical models on forgetting were compared to several previously published retention studies. The results showed that there is no general appliable mathematical model, but a few models fit to a large number of studies. The most adequate models based on a logarithmic function, an exponential function, a power function and a square root function.

$$\mu_{\text{new}} = \mu_{\text{old}} \cdot \ln(t+e)^{-\delta} \quad \text{(logarithmic)} \tag{4}$$

$$\mu_{\text{new}} = \mu_{\text{old}} \cdot e^{-\delta \cdot t} \quad \text{(exponential)} \tag{5}$$

$$\mu_{\text{new}} = \mu_{\text{old}} \cdot (t+\delta)^{-\delta} \quad \text{(power)} \tag{6}$$

$$\mu_{\text{new}} = \mu_{\text{old}} \cdot e^{-\delta \cdot \sqrt{t}} \quad \text{(square root)} \tag{7}$$

The variable μ represents the initial amount of learned items. The period of time is represented through t while δ defines the decline parameter of the forgetting curve. In order to determine the best forgetting curve for SDS interactions, we conducted tests in which the participants' memory skills were monitored.

3.1 Test Design for Forgetting Experiments

The second experiment consisted of two phases, learning and forgetting. In a first step ten subjects learned a set of two function blocks, each consisting of ten speech commands (see table 2). The learning phase took place in a BMW.

Table 2. Tasks for forgetting curve experiments

Function block 1		Function block 2	
Task 1	Start CD player	**Task 11**	Turn on TV
Task 2	Listen to CD, track 5	**Task 12**	Watch TV station 'ARD'
Task 3	Listen to tadio	**Task 13**	Regulate blowers
Task 4	Listen to radio station 'Antenne Bayern'	**Task 14**	Change time settings
Task 5	Listen to radio on frequency 103,0	**Task 15**	Change date settings
Task 6	Change sound options	**Task 16**	Change CID brightness
Task 7	Start navigation system	**Task 17**	Connect with BMW Online
Task 8	Change map scale to 1km	**Task 18**	Use phone
Task 9	Avoid highways for route calculation	**Task 19**	Assistance window
Task 10	Avoid ferries for route calculation	**Task 20**	Turn off the CID

The tasks and the corresponding commands were noted on a handout. The participants had to read the tasks and uttered the speech commands. When all 20 tasks were completed, this step was repeated until all SDS commands could be freely reproduced. These 20 commands built the basis for our retention tests.

Our aim was to determine how fast forgetting took place, so we conducted several memory tests over a time span of 50 days. The tests were conducted in a laboratory environment and should imitate the situation in a car if the driver wants to perform a task (e.g. listen to the radio) via SDS.

$$\begin{aligned} \textit{Intention} &\longrightarrow \textit{Task} \longrightarrow \textit{Command} \longrightarrow \textit{Success} \\ \textit{Icon} &\longrightarrow \textit{Task} \longrightarrow \textit{Command} \longrightarrow \textit{Success} \end{aligned}$$

Because we wanted to avoid any influence on the participant's verbal memory, the intentions were not presented verbally or in written form but as iconic representations (see Fig. 3). Each icon represented an intention and the corresponding speech command had to be spoken. Each subject had to complete a

Fig. 3. Iconic representation of the functions: phone, avoid highways and radio

retention test for every function block after nine and 36 (group A) or 22 and 50 days (group B; see Fig. 4). This method guarantees that each function was only

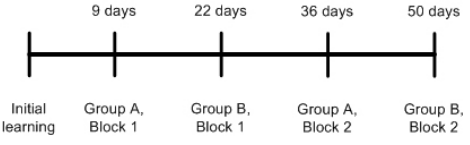


Fig. 4. Test procedure for retention tests

used once and relearning effects could not influence the results. As a measure for forgetting, we used the number of commands recalled correctly after a certain period of time.

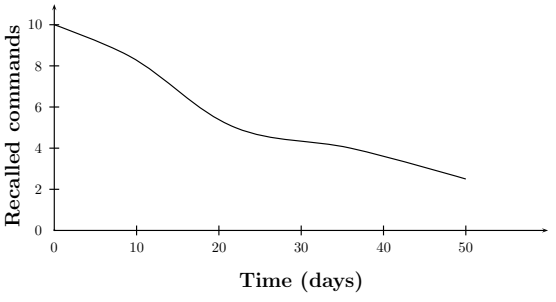
3.2 Results

The observed forgetting curve can be seen in Fig. 5(a). In order to determine whether equation (4), (5), (6) or (7) fits best to our findings, we used the chi-squared goodness-of-fit test (cf. section 2.2). The minima χ^2 for the functions are $\chi^2_{\log} = 2.11$ ($\delta = 0.58$), $\chi^2_{\exp} = 0.12$ ($\delta = 0.027$), $\chi^2_{\text{pow}} = 1.77$ ($\delta = 0.22$) and $\chi^2_{\text{sqrt}} = 0.98$ ($\delta = 0.15$). Because the exponential function (see Fig. 5(b)) delivers the smallest χ^2, we use equation (5) for our further studies.

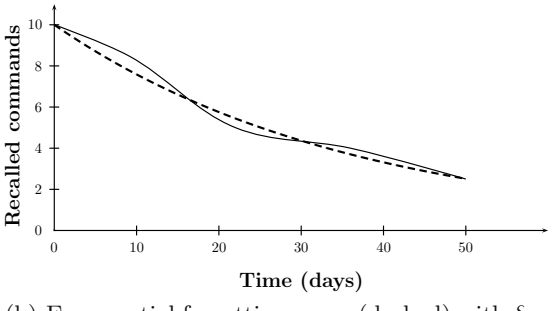
Concerning forgetting in general we can deduce that once the speech commands have been learned, forgetting takes place faster in the beginning. With increasing time, the forgetting curve flattens and with increasing time tends to zero. Our findings show that after 50 days about 75% of the original number of speech commands have been forgotten. Based on the exponential function, we estimate that complete forgetting will take place after approximately 100 days.

4 Providing Adaptive Help

As discussed in previous works, several adaptive components can be included in dialogue systems, e.g. user adaption [Hassel and Hagen, 2005], content adaption,



(a) Empirical determinded forgetting curve



(b) Exponential forgetting curve (dashed) with $\delta = 0.027$

Fig. 5. Forgetting curves

situation adaption and task adaption [Libuda and Kraiss, 2003]. We concentrate on user and content adaption and build a user model that allows us to create an individual profile of the driver's knowledge about certain commands and domains. This profile provides the basis for an adaptive help system which offers only relevant information on certain domains.

4.1 Defining an Expert User

In section 2 we observed that in our environment, the time to learn speech commands follows a power law, depending on the number of trials (N), the duration B of the first interaction and the learning rate parameter (α). If we transform equation (1), we are able to determine the number of trials that are needed to execute a function in a given time T.

$$N = \sqrt[-\alpha]{\frac{T}{B}} \tag{8}$$

If we substitute T with the minimal time $T_{\min}$ an expert needs to execute a function ($T_{\min} = 2$s, cf. section 2.2), we can estimate the number of trials which are necessary for a novice user to become an expert. The only variable is the duration B, which has to be measured for every function at its first usage.

4.2 Knowledge Modeling Algorithm

Our findings from the learning experiments can be used to create an algorithm for the presentation of SDS help. Therefore the function corpus of every context is split into several help layers (see Fig. 6). Each layer contains a maximum

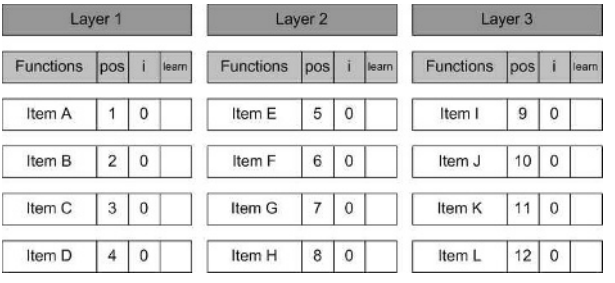


Fig. 6. Exemplary illustration of twelve help items divided into three help layers. Each help item has a number *pos* marking the initial position in the layer, an index i that is used for the monitoring of the usage and a flag *learn* to mark a function as learned.

of 4 ± 2 help items in order to reduce the driver's mental load [Wirth, 2002]. Each item has a number *pos* marking the position within the layers. The initial order in each context is based on our experience with the usage frequency by novice users. The first layer contains help items which either are often used (e.g. dial number) or easy to use. The more complex or infrequent a function is, the more the corresponding help item is put into the latter layers. Every usage of a function is logged by the system and written into an index i.

$$i_{\text{new}} = i_{\text{old}} + 1 \tag{9}$$

The value of N resulting from equation (8) defines a threshold that marks a function as known or unknown (column *learn* in Fig. 9). If a driver uses a function more often than the corresponding threshold ($i > N$), our assumption is that the user is no longer a novice and thus does not need help on this function in the first layer. It can be marked as learned and shifted into the last layer. The other functions move over to the preceding layers (see Fig. 7). If a function is not in use for a longer period of time, the index of this function steadily decreases until the initial value is reached. The decrease itself is based on the results of our forgetting experiments (cf. section 3.2) and described by equation (5). The index for every function increases with every use and decreases as long as the function is not in use. If the index of a help item falls below the corresponding threshold N, the help item is shifted back to its original position (see Fig. 8) and not longer marked as learned.

5 Summary and Future Work

In this paper we presented studies dealing with learning and forgetting of speech commands in an in-car environment. In terms of learning, we compared the

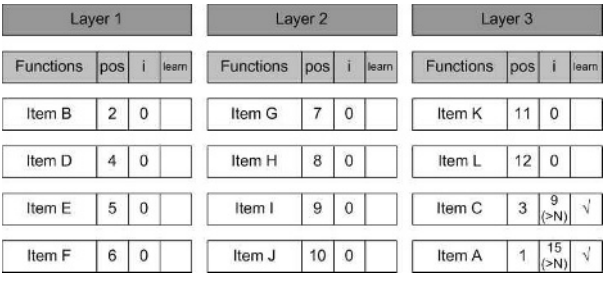


Fig. 7. Items A and C had an initial index of $i = 0$ and were presented in layer 1; after Item A has been used 15 times and Item C nine times, both items are shifted into layer 3 and are marked as learned (assumed that N is in both cases lower than i)

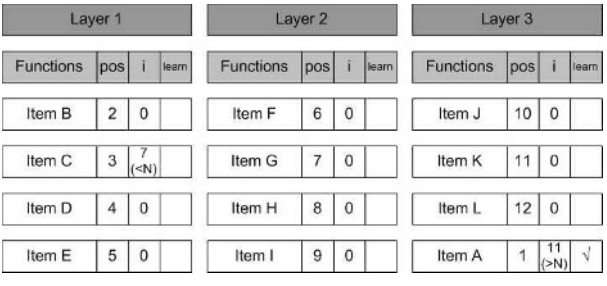


Fig. 8. Items A and C have not been in use for a couple of days, thus the indices for both help items decrease. The index for Item C falls below its threshold N and is shifted back to its original position.

power law of learning and the exponential law of learning as models that are used to describe learning curves. We conducted tests under driving conditions and showed that learning in this case follows the power law of learning. This implies that learning is most effective in the beginning and requires more effort the more it tends towards an expert level.

Concerning forgetting we compared four possible mathematical functions: a power function, an exponential function, a logarithmic function and a square root function. Our retention tests showed that the forgetting curve was described most adequately by the exponential function. Within the observed time span of 50 days about 75% of the initial amount of speech commands have been forgotten.

The test results have been transferred into an algorithm that enables us to specify the driver's knowledge on certain speech commands or domains within the SDS. Based on the learning experiments we are able to deduce a threshold that defines the minimal number of trials that are needed to learn a speech command. The forgetting experiments allow us to draw conclusions on how long this specific knowledge will be remebered. With these key information we established an algorithm for an adaptive help system which will only provide help on speech commands and domains the driver is unfamiliar with.

The help concept we presented in this paper has to be evaluated in usability tests. The main question in this context is if drivers accept and understand the concept of the adaptive presentation of the help items. Furthermore our findings from the tests concerning learning and forgetting have to be validated by a larger number of subjects.

Bibliography

[Hagen et. al. 2004] Hagen, Eli ; Said, Tarek ; Eckert, Jochen: *Spracheingabe im neuen BMW 6er*. ATZ. 2004

[Haller 2003] Haller, Rudolf: The Display and Control Concept iDrive - Quick Access to All Driving and Comfort Functions. In: *ATZ/MTZ Extra (The New BMW 5-Series)* (2003), S. 51–53

[Hassel and Hagen 2005] Hassel, Liza ; Hagen, Eli: Evaluation of a Dialogue System in an Automotive Environment. In: *6th SIGdial Workshop on Discourse and Dialogue*, URL `http://www.isca-speech.org/archive/sigdial6/sgd6_155.pdf`, September 2005, S. 155–165

[Heathcote et. al. 2002] Heathcote, Andrew ; Brown, Scott ; Mewhort, D. J. K.: The Power Law Repealed: The case for an Exponential Law of Practice. In: *Psychonomic Bulletin and Review* 7 (2002), S. 185–207

[Komatani et. al. 2003] Komatani, Kazunori ; Adachi, Fumihiro ; Ueno, Shinichi ; Kawahara, Tatsuya ; Okuno, Hiroshi: Flexible Spoken Dialogue System based on User Models and Dynamic Generation of VoiceXML Scripts. In: *4th SIGdial Workshop on Discourse and Dialogue*, 2003

[Libuda and Kraiss 2003] Libuda, Lars ; Kraiss, Karl-Friedrich: Dialogassistenz im Kraftfahrzeug. In: *45. Fachausschusssitzung Anthropotechnik der DGLR: Entscheidungsunterstützung für die Fahrzeug- und Prozessführung*, Oktober 2003, S. 255–270

[Newell and Rosenbloom 1981] Newell, Allen ; Rosenbloom, Paul: Mechanisms of skill acquisition and the law of practice. In: Anderson, J. R. (Ed.): *Cognitive skills and their acquisition*. Hillsdale, NJ : Erlbaum, 1981

[Rasch et. al. 2004] Rasch, Björn ; Friese, Malte ; Hofmann, Wilhelm ; Naumann, Ewald: *Quantitative Methoden*. Springer, 2004

[Rubin and Wenzel 1996] Rubin, David ; Wenzel, Amy: One Hundred Years of Forgetting: A Quantitative Description of Retention. In: *Psychological Review* 103 (1996), Nr. 4, S. 734–760

[Vollrath and Totzke 2003] Vollrath, Mark ; Totzke, Ingo: Möglichkeiten der Nutzung unterschiedlicher Ressourcen für die Fahrer-Fahrzeug-Interaktion. In: *Der Fahrer im 21. Jahrhundert*. Düsseldorf : VDI-Verlag, 2003

[Wirth 2002] Wirth, Thomas: *Die magische Zahl 7 und die Gedächtnisspanne*. 2002. – URL `http://www.kommdesign.de/texte/gedaechtnisspanne.htm`

Help Strategies for Speech Dialogue Systems in Automotive Environments

Alexander Hof and Eli Hagen

Forschungs- und Innovationszentrum,
BMW Group, Munich,
{eli.hagen, alexander.hof}@bmw.de

Abstract. In this paper we discuss advanced help concepts for speech dialogues. Based on current research results in the field of human-machine-interfaces, we describe two advanced help concepts based on hierarchical structuring of help dialogues. Furthermore we explain the test design for our usabilty experiments and present the methods and measures we used to collect our test data. Finally we report the results from our usability tests and discuss our findings.

1 Introduction

A growing number of premium class vehicles include advanced human-machine-interfaces in order to provide access to driver information and assistance systems. Besides a visual-haptic interface, some interfaces include a speech dialogue system (SDS), e.g. BMW's iDrive. The iDrive help is activated by uttering the speech command 'options'. The options provide context specific help in the form of a static list, containing available speech commands in the current context.

Sample options dialogue for the current system.
User: "Phone."
System: "Phone. Say dial name, dial number or name a list."
User: Options.
System: "Options. Say dial followed by a name, for example 'dial Alex', or say dial name, dial number, save number, phone book, quick dialing list, top eight, last eight, accepted calls, missed calls, active calls and or or off."

Neither the users' preferences are taken into account nor are the options aggregated in a certain logic or dialogue. Findings in the area of human-machine-interaction (HMI) show that in dual task situations the driver's mental load is minimized, if the information is presented in a deep hierarchy [Totzke u. a., 2004].

In this paper we describe two concepts we have developed in order to present the options in a separate dialogue. Both concepts are based on a hierarchical

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 107–116, 2006.

structure. In section 2.1 we describe a help concept based on semantic aggregation. In section 2.2 we describe a second concept where the aggregation depends on the frequency of usage. Finally we describe the usability tests we conducted in order to compare the two prototypes (cf. section 3). The results will show if the findings concerning hierarchies in speech dialogues are conform to the findings in the field of graphical user interfaces.

2 Features of the Enhanced Help Concepts

In Nielsen [1994] and Shneiderman [1997], we find several guidelines for the design of human-machine-interfaces in general. Interface guidelines for speech dialogues are treated in Bernsen u. a. [1998]. We took these guidelines into account and developed two prototypes of an enhanced help system. Both prototypes have in common that the help item are not presented as an unordered list. The help items are ordered in a hierarchy. The domain we used for the tests of the concepts was set to the phone menu.

2.1 Prototype 1: Semantic Aggregation of Help Items

One possible approach is the aggregation of help items in semantic groups. All speech commands with a similar meaning are ordered in a hierarchy (see Fig. 1) and put into clusters. In the first level, basic phone functions (e.g. dial number) are presented. Additionally the drivers can get help on the related lists (e.g. re-dial list, missed calls, etc.). If the drivers choose the lists, they can get an overview of the lists or of the related functions which can be performed within the lists. The two basic categories are navigational functions (e.g. next entry, previous entry) and managing functions (e.g. delete entry, read all entries). One

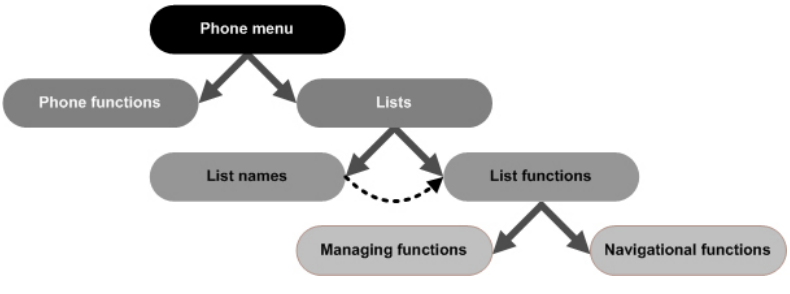


Fig. 1. Help dialogue structure for prototype 1 in the phone context

advantage of this help structure is the reduction of the short-term memory load by clustering not more than 4±2 speech commands per prompt [Wirth, 2002]. Besides that, some clusters can be reused in different contexts (e.g. N 4.1, basic navigation in any kind of lists), which supports the establishment of a mental model about the available functions of the SDS. Thus the driver will be able

to build transfer knowledge for various contexts [Shneiderman, 1997]. Furthermore the repeated usage of identical clusters contributes to a consistent dialogue [CEN, 1996].

Sample options dialogue for prototype 1.
User: "Phone."
System: "Phone. Say dial name, dial number or name a list."
User: "Options."
System: "Phone options. Say dial name or dial number. If you want information about the phone, say phone functions. To get help on the lists, say lists."
User: "Lists."
System: "Lists. Choose between list names oder list functions."
User: "List names."
System: "List names. Re-dial, top 8, adressbook, missed calls or accepted calls and current calls. Choose a list or say managing functions or navigation functions."

2.2 Prototype 2: Aggregation of Help Items Depending on Usage

Our second approach employs a hierarchical structures as well, but the aggregation of help items follows another principle. Speech commands are aggregated in clusters according to the frequency of use (see Fig. 2). The order of the commands is based on our experience with the usage frequency by novice users. If the options are activated, the drivers gets information on the most important functions (e.g. dial number, dial name, re-dial list, etc.). They can access further commands that are not so frequently used by uttering "more" and get in a secondary help level (e.g. incoming calls, accepted calls, etc.). In both levels the drivers get a hint that they can access additional functions for the presented lists. These functions are according to section 2.1 navigational functions or managing functions.

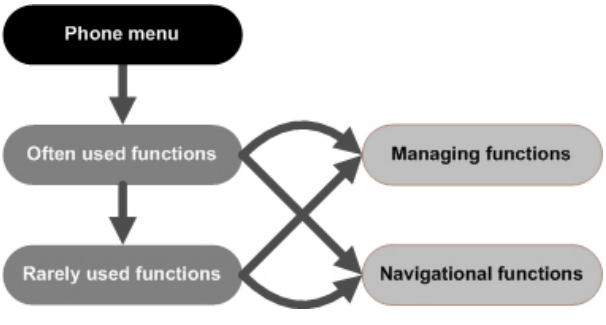


Fig. 2. Help dialogue structure for prototype 2 in the phone context

Sample options dialogue for prototype 2.
User: "Phone."
System: "Phone. Say dial name, dial number or name a list."
User: "Options."
System: "Phone options. Say dial name or dial number, adressbook, redial list or missed calls. To ask for additional information on the lists, say navigation functions or managing functions. To hear more available commands, say more."
User: "Managing functions."
System: "Managing functions. Say show list, read list or delete list. To choose an entry, say its name or its number."

3 Test Design

In order to compare the help concepts, a BMW 7 series was equipped with the two prototype systems. Each prototype was evaluated by a different test group (between-subject-design) [Rubin, 1994]. In order to get useful results, the size for each of the two test groups was according to Nielsen [1994] set to four subjects. Each subject was an experienced driver, having a driving licence for at least four years and kilometres travelled per year of at least 10,000 km. No subject had any experience with the usage of iDrive.

The tests were conducted in real traffic on a predefined route. The route was 15 km long and included several traffic conditions from heavy city traffic to light suburban traffic. In order to minimize traffic variations, each test was conducted from 9:00 to 10:30 or from 13:00 to 14:30. The tests themselves consisted of three parts (see table 1). During the driving phase, the subjects were asked to perform twelve representative tasks (see table 2) using the help function. The help function was introduced to all participants before the tests began.

The task instructions were embedded into longer descriptions of certain situations, e.g. "You took a rest in a restaurant. You have returned to your car and want to know if someone tried to call you while you have been in the restaurant.". Since there was not given any direct hint on the corresponding voice command, the users had to find these commands in the help system on their own.

Table 1. Test phases

Part	Description	Duration
1	Introduction into the car functions and iDrive, short description of the test and the usage of the help function	15 min
2	Driving phase, working on tests tasks	45 min
3	Questionnaire	30 min

Table 2. Test tasks

Task 1:	Navigate to the menu 'commuinication'
Task 2:	Navigate to the phone
Task 3:	Call Irina Adam from the adressbook
Task 4:	Dial the number 123456
Task 5:	Find the redialling list
Task 6:	Browse the redialing list
Task 7:	Getting read the redialing list by the system
Task 8:	Show list of missed calls
Task 9:	Show list of short messages
Task 10:	Browse short messages list
Task 11:	Call the sender of a short message
Task 12:	Speak and Save a note

4 Evaluation Method

For the evaluation of the two prototypes we conducted a qualitative and a quantitative analysis. According to Rubin [1994], preference and performance measures can be employed. Preference measures represent the subjects' individual rating, while performance measures depend on observed data.

4.1 Preference Measures

We collected the preference measures with a questionnaire. We used a total number of 26 questions (see appendix 6), which contained questions about the subjects' knowledge concerning the usage of computers, the stress caused by the driving situation and the test, the usability of the options and the quality of the prompts.

The rating scale ranged from 1 (positive rating) to 6 (negative rating). We used the median to determine average values for each question.

4.2 Performance Measures

As performance measures we used the efficiency for each task (see formula 1) [Nielsen, 1994]. The efficiency equals 100% if no more interactions are made than necessary and tends to 0% the more interactions are made than necessary.

$$\text{Efficiency} = \frac{\text{Necessary number of interactions}}{\text{Observed number of interactions}} \cdot 100 \quad (1)$$

As a second performance measure we use the number of mistakes during a task. According to Reason [1990] we classified the mistakes in three categories: M1 (mistakes due to lack of understanding), M2 (mistakes caused by a lack of attention, speech recognition (ASR) failures or forgetting of necessary interactions) and M3 (mistakes due to other reasons, e.g. software bugs). The most interesting mistakes are those from category M1. All mistakes in this category result from incorrect interactions due to conceptual weaknesses. So these mistakes are

directly related to the help system and are in contrast to M2 and M3 not influenced by external delimitting factors like speech recognition failures or software bugs.

5 Test Results

5.1 Resulting Preference Data

The analysis of the questionnaire (see appendix 6) delivered very similar ratings for both prototypes (see Fig. 3). Thus we focused on the questions that showed

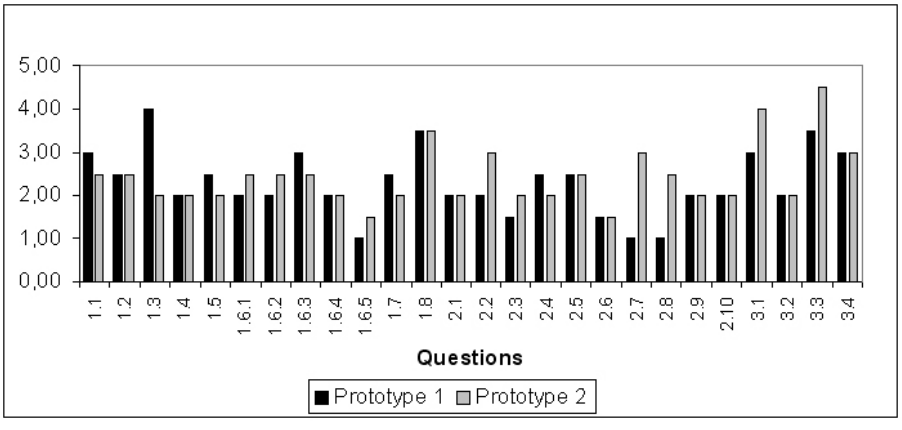


Fig. 3. Results of the questionnaire

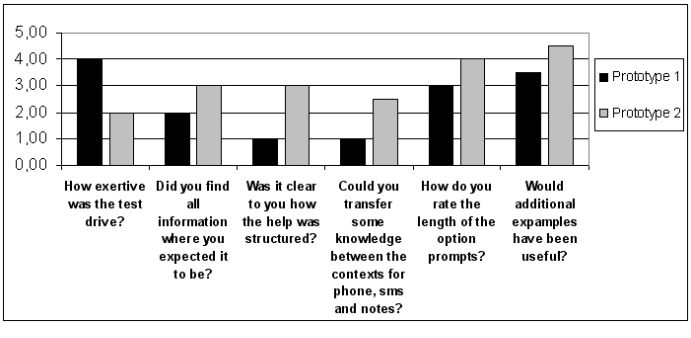


Fig. 4. Results of the questionnaire showing different median values for the concepts

a greater difference of the median for both prototypes (see Fig. 4). Prototype 1 is rated better than protoype 2 in the fields concerning the structure of the options (where informationis expected to be found, structure of the options,

establishment of transfer knowledge). Additionally, the lenght of the prompts was rated better for prototype 1. The prompts for the semantic cluster concept are due to deeper structure shorter than the prompts for prototype 2 that is based on the frequency of usage. The only aspect where prototype 2 receives better results is the stress during the test situation. If this effect is based on different traffic situations, individual differences or is related to the conceptual structure of the options can not be clearly deduced from the data. Concerning the preference data, we can derive that the concept of semantic clustering contributes to a more consistent dialogue than the concept based on the frequency of usage.

5.2 Resulting Performance Data

We monitored the number of interactions that the users needed to perform a certain task (see Fig. 5) since this information is needed for further calculations of the efficiency. The analysis of the efficiency can be done in two ways, either

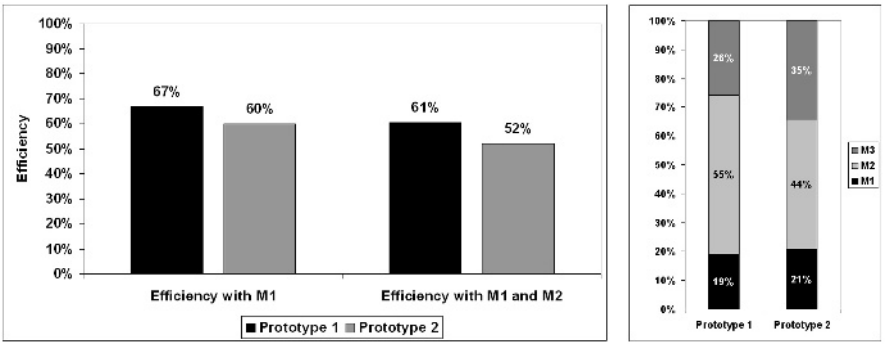
		Prototype 1	Prototype 2
Task No.	No. of necessary Interactions	Performed Interactions	Performed Interactions
1	3	6	4,5
2	2	31,75	22
3	4	7,25	12,5
4	5	32,5	20
5	4	27,25	14,25
6	6	7,25	11
7	5	6	6
8.1	5	5,5	4,5
8.2	4	16	16,25
9	3	5	3,5
10	3	8,5	5,5
11	4	16,5	15,75
12	5	8,75	9

Fig. 5. Number of necessary interactions per task and average number of performed interactions per task (including M1, M2 and M3 mistakes)

including M2 mistakes or omitting M2 mistakes. The most interesting mistakes are M1 mistakes, because these mistakes reveal conceptual weaknesses. We can not draw direct conclusions on the quality of the concepts from M2 mistakes because these mistakes are stronger related to technical or traffic conditions. But since ASR failures or timeout situations could possibly indicate conceptual weaknesses as well, M2 mistakes are also taken into consideration. Since M3 mistakes only include system errors, they are not taken into consideration at all for the calculation of the efficiency.

The analysis for both prototypes showed that prototype 1 achieved slightly better results concerning the efficiency than prototype 2 (see Fig. 6(a)). This

could be related to the deeper hierarchy and thus would coincide with the findings of Totzke u. a. [2004] concerning higher performance with deeper hierarchies in dual task situations. The analysis of the mistakes M1 (see Fig. 6(b)) showed



(a) Overall efficiency

(b) Distribution of M1, M2 and M3 mistakes by percentage

Fig. 6. Performance data

that the distribution of mistakes by percentage for prototype 1 has nearly the same value (19%) as for prototype 2 (21%). In terms of M1 mistakes both concepts seem to behave very similar.

The analysis of M2 mistakes reveals that the semantic clustering concept causes 11% more mistakes than the concept based on the frequency of usage. This could be related to longer dialogues for prototype 1, because the hierarchy is deeper for the semantic clustering concept and therefore implies more dialogue steps and more possible sources of error for the user.

6 Summary

In this paper we presented two prototypes for advanced help systems for speech dialogues in automotive environments. We developed two concepts, both based on a hierarchical structure. Prototype 1 employs a semantic aggregation, whereas prototype 2 is based on an aggregation depending of the frequency of usage.

We conducted usability tests with (in terms of speech dialogues) novice users in real traffic conditions in order to evaluate the two prototypes. We collected preference and performance data and analysed the findings.

Regarding the preference data, prototype 1 delivered the better results. The adavantage of the semantic concept is its clear hierarchy that supports the establishment of a mental model of the help structure. The performance data showed that the distribution of M1 mistakes reveals no significant difference between both prototypes. Concerning M2 mistakes (speech recognition failures), prototype 1 caused more mistakes, but this seems to be related to a higher number

of dialogue steps that are necessary to navigate through the help dialogue. The efficiency was 7% (9% inluding M2 mistakes) higher for the prototype based on semantic structuring.

Based on these findings, we conclude that a deeper hierarchical structure contributes to a more efficient help dialogue. Furthermore our findings support the thesis that deeper menu structures are more useful in dual task situations than broad structures. Thus we assume that the findings of Totzke u. a. [2004] in the field of grapical interfaces are valid for voice user interfaces as well.

Bibliography

[Bernsen u. a. 1998] Bernsen, Nils O. ; Dybkjær, Hans ; Dybkjær, Laila: *Designing Interactive Speech Systems: From First Ideas to User Testing.* London : Springer-Verlag, 1998

[CEN 1996] CEN: *Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten, Teil10: Grundsätze der Dialoggestaltung (DIN EN 9241-10).* 1996

[Nielsen 1994] Nielsen, Jakob: *Usbility Engineering.* Academic Press, 1994

[Reason 1990] Reason, James: *Human Error.* Cambridge University Press, 1990

[Rubin 1994] Rubin, Jeffrey: *Handbook of Usability: How to Plan, Design and Conduct Effective Tests.* Toronto : John Wiley & Sons, Inc., 1994

[Shneiderman 1997] Shneiderman, Ben: *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* 3. Ausgabe. Reading, Massachusetts : Addison-Wesley, 1997

[Totzke u. a. 2004] Totzke, Ingo ; Rauch, Nadja ; Krüger, Hans-Peter: Kompetenzerwerb und Struktur von Menüsystemen im Fahrzeug: "Breiter ist besser?". In: *Entwerfen und Gestalten. 5. Berliner Werkstatt für Mensch-Maschine-Systeme* Bd. 18, 2004, S. 226–249

[Wirth 2002] Wirth, Thomas: *Die magische Zahl 7 und die Gedächtnisspanne.* 2002. – URL `http://www.kommdesign.de/texte/gedaechtnisspanne.htm`

A Questionnaire

Number	Question
1.1	How do you rate the difficulty of the test tasks?
1.2	How do you rate the difficulty of the driving task?
1.3	How exertive was the test drive?
1.4	Did the test drive cause stress?
1.5	How well did you accomplish the test drive?
1.6.1	Did you feel well during the test drive?
1.6.2	Did you feel save during the test drive?
1.6.3	Did you feel calm during the test drive?
1.6.4	Did you feel relaxed during the test drive?
1.6.5	Have you been very concentrated during the test?
1.7	Did the driving task detain you to in dealing with the additional test tasks?
1.8	How did the test task influence your driving performance?
2.1	How well did you perform with the speech options?
2.2	Did you find all information where you expected it to be?
2.3	Did you get an overview over the available speech commands?
2.4	Did you always know your position within the dialogue?
2.5	How well could you access the relevant information?
2.6	How well did the help function support you in dealing with test tasks?
2.7	Was it clear to you how the help was structured?
2.8	Could you transfer some knowledge between the contexts for phone, sms and notes?
2.9	How do you rate the learnability of the options?
2.10	How do you rate the comprehensibility of the options?
3.1	How do you rate the length of the option prompts?
3.2	Was the phrasing of the prompts unambiguous?
3.3	Would additional expamples have been useful?
3.4	Would an additional graphical support have been useful?

Information Fusion for Visual Reference Resolution in Dynamic Situated Dialogue

Geert-Jan M. Kruijff[1], John D. Kelleher[2], and Nick Hawes[3]

[1] Language Technology Lab, DFKI GmbH
gj@dfki.de
http://www.dfki.de/~gj
[2] Dublin Institute of Technology
John.Kelleher@comp.dit.ie
www.comp.dit.ie/jkelleher
[3] School of Computer Science, University of Birmingham
n.a.hawes@cs.bham.ac.uk
http://www.cs.bham.ac.uk/~nah

Abstract. Human-Robot Interaction (HRI) invariably involves dialogue about objects in the environment in which the agents are situated. The paper focuses on the issue of resolving discourse references to such visual objects. The paper addresses the problem using strategies for *intra-modal fusion* (identifying that different occurrences concern the same object), and *inter-modal fusion*, (relating object references across different modalities). Core to these strategies are sensori-motoric coordination, and ontology-based mediation between content in different modalities. The approach has been fully implemented, and is illustrated with several working examples.

1 Introduction

The context of this work is the development of dialog systems for human-robot collaboration. The framework presented in this paper addresses a particular aspect of situated dialog, namely reference resolution. Reference resolution in situated dialog is a particular instance of the anchoring problem [Coradeschi and Saffiotti, 2003]: how can an artificial system create and maintain correspondences between the symbols and sensor data that refer to the same physical object?

In a dialog, human participants expect their partner to construct and maintain a model of the evolving linguistic context. Each referring expression used in the dialog introduces a representation into the semantics of its utterance. This representation must be bound to an element in the context model in order for the utterance's semantics to be fully resolved. Referring expressions that access a representation in the context are called *anaphoric*. In a *situated* dialog, human participants expect their partner to not only construct and maintain a model of the linguistic discourse, but also to have full perceptual knowledge of the environment. This introduces a form of reference, called *exophoric* reference. Exophoric references denote objects that have entered the dialog context through a non-linguistic modality (such as vision), but have not been previously evoked into the context. Consequently, for a robot to participate in a situated dialog,

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 117–128, 2006.

the framework it uses for reference resolution must support the integration of different forms of perceptual knowledge with the context models it constructs through dialog. The importance of anaphoric and exphoric references to situated discourse is evidenced by the frequency with which they occur. For example, the two most common cases of definite descriptions in the TRAINS corpus on situated dialogue were anaphoric and exophoric definites [Poesio, 1994].

The dynamics of environmental interaction make exophoric reference resolution a genuine problem. The movement of objects and agents in the environment may result in changes to how objects are perceived, and thus how these objects may be referred to. For example, you may be talking to a robot about an orange juice carton, and then rotate it so the robot now faces the side of the carton. Regardless of this change, the robot should understand that it is still the same orange juice carton which you talked about earlier. To manage environmental change, we must consider how different sightings of an object may be identified as being one and the same object. Furthermore, we should ensure that this identification enables the robot to construct a more complete understanding of the object. This will allow us to address the uncertainty and partial coverage of the robot's perception. For example, the robot should be able to combine the perceptual information it gets from the front and side sightings of the same orange juice carton to construct a more complete representation of it.

In this paper, we present an approach that addresses the problem of resolving references in situated dialog. The approach uses a combination of two fusion strategies. To establish whether different object occurrences in a single modality concern one and the same object, we use an *intra-modal fusion* strategy. The results of this strategy are *equivalence classes* which store different occurrences of the same object within a single modality, and describe an object at a conceptual level. To establish relations between the equivalence classes that have been established in different modalities (i.e. establishing cross-modal bindings) we use an *inter-modal fusion* strategy. For this, we use ontology-based mediation, exploiting the conceptual character of equivalence classes. Inter-modal fusion provides the basis for exophoric reference resolution, and can help the further completion and disambiguation of content within modalities. The resulting approach has been fully implemented, and used on a Pioneer PeopleBot mobile robot in HRI scenarios dealing with human-augmented mapping, and visual object learning. We will discuss several working examples.

Contributions. This paper presents a novel approach to the resolution of contextual references to visual objects in dynamic environments. We base reference resolution on establishing relations between equivalence classes, rather than individual instances. This makes it possible to handle changes in sightings of objects, without losing the conceptual integrity of the object. Maintaining this integrity means we can refer to the object as still "the same" as before, even though the situation has changed.

Overview. §2 highlights some of the factors that affect situated reference resolution and reviews previous approaches. In §3 we describe our approach, and in §4 we describe how the approach has been implemented. Following this, in §5, we provide various worked examples to illustrate the implementation. The paper finishes with conclusions and future work.

2 Data and Previous Work

In §1 we introduced the concepts of anaphoric and exophoric references and noted how the dynamics of the environment and the agents in the environment can make the resolution of these types of reference difficult. In this section we discuss one particular difficulty, namely the temporal aspect of situated discourse reference induced by these dynamics, and how it affects reference resolution. We also review some of the previous approaches to reference resolution against this background.

The types of referring expressions we are concerned with denote physical objects that persist in time and space. Some of the properties of these objects are preserved across time while others change. One consequence of this environmental variability is that connecting object representations across different modalities must take this temporal dimension, and its attendant variability, into account. [Coradeschi and Saffiotti, 2003] observe that we cannot model this binding as a one-shot process. However, previous approaches to intra-modal fusion, for example [Loutfi et al., 2005, Gurevych et al., 2003] and [Alexandersson and Becker, 2003] focus on fusing only the most recent perceptual representation of an object with the representation of a linguistic referent that denotes the object. For example, [Loutfi et al., 2005] propose a *track* function that ensures that the cross-modal binding points to "the most recent and adequate representation of the object". [Gurevych et al., 2003] use the *overlay* operator of [Alexandersson and Becker, 2003] which computes "the maximally compatible combination of new information and old information". Adopting such an approach results in the loss of information relating to the previous perceptual states of objects. If a robot does not maintain knowledge about the previous state of an object related to the discourse, it will not be able to answer questions relating to that object's prior state. The following example dialog highlights this problem:[1]

1. **scene:** A ball is placed to the right of a box.

H1 "*This*$_i$ is *a box*$_i$"
R1 "OK"
H2 "*This*$_j$ is *a ball*$_j$."
R2 "OK"

2. **scene change:** The ball is moved to the left of the box.

H3 "Where is *the ball*$_j$?"
R3 "*It*$_j$ is to the left of *the box*$_i$."
H4 "Where was *it*$_j$ when you first saw *it*$_j$?"

In our framework, the resolution of an anaphoric reference, for example the pronominal references in R3 and H4, represents one type of intra-modal fusion. When a representation is first evoked in a dialog, an equivalence class is created to hold it. Subsequent linguistic references to this representation are then all fused into the same (linguistic)

[1] In this example the indices i and j indicate that all the referring expressions marked by a particular index refer to the same physical entity, and that the representations of these references are intra-modally bound within the linguistic context model.

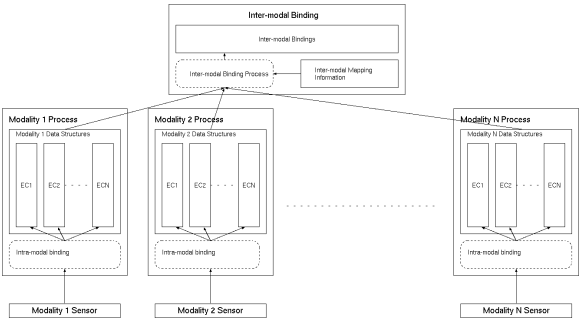


Fig. 1. Binding processes

equivalence class to denote that they are all about the same representation. The details of how we do this are presented in §3.1.

Our framework supports resolving exophoric references, for example the deictic references in H1 and H2, through inter-modal fusion. The reference to the object from the dialog is first entered into a linguistic equivalence class. It is then fused inter-modally (i.e. across modalities), with an equivalence class from another modality that represents the referent in that modality (e.g. the visual equivalence class that stores multiple sightings of the object being spoken about). We present the details of this in §3.2.

Given the above context, unless the robot retains information regarding the original location of the ball, it will not be able to answer the question posed in H4. This interaction points to the necessity of maintaining a perceptual history, similar to a discourse history. Indeed, the requirement for such a history has been noted elsewhere, see [Byron, 2003]. Finally, as noted in §1 the partial coverage and uncertainty inherent in robotic perception provides an extra layer of difficulty to reference resolution. An interesting effect of maintaining such a perceptual history is that it provides a mechanism for dealing with noise in real world sensor data. For example, it is possible that a vision sensor may not detect either the ball or the box at the time question H3 is posed to the robot, even though they are in the robot's visual field. However, if the robot maintains a history of an object's perceptual states (including where it appears in the world), it can use the last known location of the object to construct an answer to the question.

3 Approach

In the previous sections we briefly described our notion of an equivalence class, and indicated how these are created through intra-modal fusion. We then discussed how equivalence classes are associated through a process of inter-modal fusion to support references across modalities (e.g. the resolution of exophoric references). Figure 1 gives an illustration of these processes, which we discuss in more detail below.

3.1 The Formation of Equivalence Classes Through Intra-modal Fusion

An *equivalence class* (EC) represents an object at two levels: a concept-level (i.e. intensional) characterization of the properties of an object, and an instance-level

(i.e. extensional) characterization of different sightings of the object. By separating these two levels of representation we can deal with the issue of change, and information completion. While sightings track change, the conceptual level provides a constant (though extensible) representation of the identity of the object.

We establish equivalence classes for objects within individual modalities, through *intra-modal fusion*. Intra-modal fusion represents a level of modality-specific information fusion. The input to an intra-modal binding process is formed by sensor events. These events are sets of entity descriptions from the input modality. Each of these structures contain information related to a particular sensed entity. For example, a sensor event from a visual sensor would consist of a set of structures each describing the visual properties of an entity in a scene.

The function of the intra-modal binding process is to bind each structure in a sensor event to an equivalence class that represents the entity the sensor data describes. For example, in visual fusion we use visual categorization to establish whether we are still looking at the same kind of object, based on aspects of visual appearance. Sensorimotoric coordination, i.e. the alignment between coordinate systems in perceptual and motoric modalities, subsequently enables us to determine whether the spatial positioning of an object has remained the same despite movement of the robot, or manipulation of the object. In our dialogue model, we use anaphoric binding to relate different mentions of (or references to) a discourse object to an equivalence class for that object.

Currently, we adopt a conservative approach to intra-modal binding, by trying to minimise the number of equivalence classes that are created within each modality. As the examples above illustrate, the particular process used for intra-modal binding is specific to the type of data being processed. Regardless of modality specific details, we propose a general strategy for intra-modality binding. The steps are as follows:

1. For each structure in a sensor event, retrieve the set of equivalence classes whose previously bound structures do not conflict with the input structure.
2. If the number of retrieved ECs $== 0$, create a new EC and bind the structure to it.
3. If the number of retrieved ECs $== 1$, bind the structure to the retrieved EC.
4. If the number of retrieved ECs > 1 trigger a conflict resolution strategy.

Step 4 of the above description introduces the notion of *conflict resolution*. The conflict occurs because the input sensor structure could possibly be bound to more than one EC. The purpose of conflict resolution to to determine which, if any, of these equivalence classes is the appropriate one. This issue is pervasive in any binding process where there is the possibility of ambiguity.

In the implementation of the architecture (described in §4), we resolve conflicts by binding the input structure to the EC with the largest amount of agreement between the sensor data in the structure and the sensor data in the structures already bound to the EC. In situations where this overlap heuristic still does not provide a single EC to bind to we can use a simple first-come first-served approach, or have the robot initiate a clarification dialogue as described in [Kruijff et al., 2006a]. Conflict resolution is one of the primary focuses for our future work.

Another point of note relating to the intra-modal binding process is that this process may trigger the inter-modal binding process. In the next section we describe inter-modal binding in more detail, and how it interacts with intra-modal binding.

3.2 Reference Resolution Through Inter-modal Fusion

The purpose of inter-modal fusion is to establish relations between equivalence classes across different modalities. Figure 1 shows the three components involved in inter-modal binding: inter-modal mapping information, the actual inter-modal binding process, and the resulting inter-modal bindings. Inter-modal mapping information addresses the problem of establishing a relation between the *content* in different modalities. For this, we exploit the conceptual characterization that an equivalence class provides for an object. We use ontology-based mediation (cf. [Wache et al., 2001] and [Gurevych et al., 2003]) to provide an ontological mapping between (modality-specific) conceptual systems to establish whether we can relate the content of the equivalence classes.

The inter-modal mapping information informs the inter-modal binding process about any correspondences between possible sensor inputs in different modalities. For example, this inter-modal mapping information helps us to bind the linguistic token *red* (specified conceptually as a *color*) to a cluster of visual sensor readings around the value *rgb(255,0,0)* (which we can also conceptualize as a color).

The inter-modal bindings component contains a set of structures that represent the inter-modal bindings that have been previously created by the binding process. Each structure is referenced by an id (e.g. b_1). This binding id is a unique identifier for the binding, and is used for indexing. Each binding structure contains at most one equivalence class from each modality, and at least one equivalence class in total. All equivalence classes bound together in this way can be assumed to refer to the same conceptual entity. The existence of an inter-modal binding structure with only one bound equivalence class indicates that there is an equivalence class in a particular modality that has not been bound to a representation from another modality. This may occur if, for example, the agent has seen an object that has not yet been talked about.

The inter-modal binding process is triggered by the intra-modal binding process. This happens when a new EC is created, and when an addition to an EC extends the range of sensor data associated with it (i.e. when a new attribute is used to help distinguish the additional entity from other entities, rather than when an existing attribute changes in value). The latter case is important as the addition of new information to an EC may cause a conflict between the extended EC and one or more of the ECs it may be inter-modally bound to.

When a new equivalence class EC_m is created in modality m, the inter-modal binding process must execute the following steps:

1. Retrieve the set of inter-modal binding structures that (1) do not have an EC from m bound to them, and (2) are already bound to ECs that can be aligned with EC_m. Whether alignment is possible is determined by applying the inter-modal binding information.
2. If the number of retrieved inter-modal binding structures $== 0$, create a new inter-modal binding structure containing EC_m.
3. If the number of retrieved inter-modal binding structures $== 1$, bind EC_m to the returned inter-modal binding structure.
4. If the number of retrieved inter-modal binding structures > 1, trigger a conflict resolution strategy.

As with the intra-modal binding algorithm, the issue of ambiguity introduces the need for conflict resolution. We adopt a similar resolution strategy here as was outlined for intra-modal binding: the candidate binding structures are ordered by the degree of possible alignment between the ECs bound to them and the input EC. Again, in situations where this overlap heuristic does not provide a single inter-modal binding structure, we use a simple first-come first-served approach or initiate a clarification dialogue.

4 Implementation

We have implemented the approach of §3 in a distributed architecture which integrates different sensorimotoric and cognitive modalities. The architecture enables a robot to move about in an indoor environment, and have a dialogue with a human about visual and spatial aspects of the situation. We have used this system in scenarios for human-augmented mapping and simple visual object manipulation, using a Pioneer PeopleBot.

Figure 2 shows the relevant aspects of the architecture. We have subsystems for communication, spatial localization & mapping, and visual processing. We use a BDI-based process to mediate between the different subsystems. We use beliefs to provide a common ground between different modalities, rather than being a layer on top of the different modalities. Beliefs thus provide a means for cross-modal information fusion, in its minimal form by co-indexing references to information in individual modalities [Gurevych et al., 2003]. Below we describe the communication subsystem and the visual subsystem in greater detail.

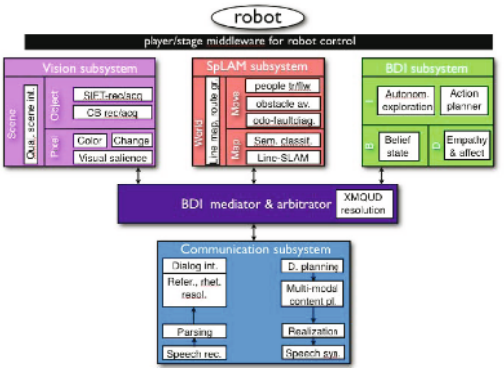


Fig. 2. The Implemented Architecture

The communication subsystem consists of several components for the analysis and production of natural language. It has been implemented as a distributed architecture using the Open Agent Architecture [Cheyer and Martin, 2001], following the idea of agent-based models for multi-modal dialogue systems [Allen et al., 2000].

For analysis we use the Sphinx4 speech recognition engine[2] with a domain-specific JSAPI speech grammar. The string-based output of Sphinx4 is then parsed with OpenCCG[3]. OpenCCG employs a combinatory categorial grammar [Baldridge and Kruijff, 2003] to yield a representation of the linguistic meaning that the string (i.e. the utterance) represents. We represent linguistic meaning as an ontologically rich, relational structure in a description logic-like formalism

[2] http://cmusphinx.sourceforge.net/sphinx4/
[3] http://openccg.sf.net

[Baldridge and Kruijff, 2002]. In structural dialogue analysis we relate the linguistic meaning of an utterance to the current dialogue context, in terms of how it rhetorically and referentially relates to preceding utterances. This yields an updated model of the situated dialogue context [Asher and Lascarides, 2003, Bos et al., 2003].

In the communication subsystem we use dialogue planning to produce flexible, contextually appropriate interaction. Based on a need to communicate, arising from the current dialogue flow or from another modality, the dialogue planner establishes a communicative goal. In turn, we plan the content to express this communicative goal, possibly in a multi-modal way using non-verbal (pose, head moves) and verbal means. In these planning steps, we can inquire the models of the situated context (e.g. dialogue context, visually scene) to ensure that the content we plan is contextually appropriate [Kruijff, 2005]. We realize verbal content using the OpenCCG realizer, which generates a string for the utterance, and then synthesize this string using text-to-speech[4].

In the vision subsystem, we have implemented visual scene understanding based on three cues: identity, color, and size of objects in the scene.[5] We use SIFT (Scale Invariant Feature Transform) features [Lowe, 2004] to recognize object identity. Each SIFT feature is a vector representing a particular arrangement of pixels centered on a particular point. When learning an object, SIFT features are extracted from the object (segmented by a fixed-size bounding box) and stored in a database along with a description of the object. During object recognition, SIFT features are again extracted from the image, and these are compared to the features associated with the previously learned objects. If the number of feature matches for an object is over a given threshold, the affine transformation of the object is estimated based on affinities between matched features. We obtain the pose of the object by applying this affine transformation to the model's segmentation mask. The robot calculates the color histogram over the segmented region of the IHS color space. The peak of smoothed histogram indicates the color. The vision subsystem consists of several CORBA[6] servers. We use an OAA agent to serve as a mediator between the communication subsystem and the vision subsystem.

5 Examples

In this section we provide various working examples to illustrate the implementation of our approach (§3) within our larger architecture for human-robot interaction (§4).

Figure 3 gives a flow diagram for processing simple dialogue describing a new visual object and some of its properties. We analyze the utterance "This is a box" in terms of its *linguistic meaning* and a (complex) characterization of its *dialogue act*. We model linguistic meaning as a relational structure over ontologically sorted content [Baldridge and Kruijff, 2002], and determine dialogue acts from the content and mood of the utterance. For "This is a box" this yields an assertion stating that an observed

[4] http://mary.dfki.de

[5] The vision subsystem was primarily implemented by Gregor Berginc (University of Ljubljana) and Bastian Leibe (TU Darmstadt).

[6] http://www.corba.org/

endurant (physical object) is an instance of a given type (a box). We inform BDI mediation of the linguistic meaning, its dialogue act, and the ECs for its discourse referents. Based on this, BDI mediation then decides to trigger a learning event in vision.

In vision we use a bounding box-method to determine the region of interest in the image for which we should learn a SIFT-based model. We create a visual referent id for the resulting model, and store this id in a new visual EC for the object. We provide the EC with a structural description of the object ("box") based on what was said [Kruijff et al., 2006b]. We then return the identifiers of the sighting and its visual EC to BDI mediation. BDI mediation creates a belief in which the dialogue and visual ECs are connected, and informs the communication subsystem that a visual model has been successfully acquired for the robot to provide feedback.

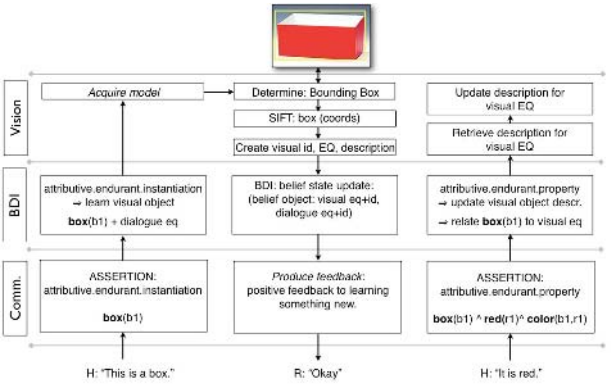


Fig. 3. Incremental update

When we next say, "It is red", reference resolution links the "it" to the discourse referent for "box". We provide the linguistic meaning (with the resolved reference) to the BDI mediator, as before, with a characterization that it expresses an assertion attributing a property to an object. Based on the dialogue act, BDI mediation retrieves the visual object EC to which the discourse referent for "box" corresponds, and informs the vision subsystem to update its description for the visual EC with the property "red".

The mechanisms for incrementally updating structural descriptions of visual object ECs thus rely on our ability to use the identifiers of ECs for co-indexing across utterances, relating the content we attribute to an object over the course of a dialogue. The same mechanism we can use in relating structural descriptions across different visual object models. For example, assume we say "This is an orange juice carton," rotate the carton and then say "This is its side." We can resolve the possessive pronoun to refer to the earlier mentioned carton. This provides the basis on which we can relate the (newly created) visual EC for "side" to the earlier created visual EC for "orange juice carton." The possessive expresses we have a *part-of* relation between "side" and the visual object that corresponds to the discourse referent to which we have resolved the pronoun. Using this information, we subsequently create a structural description for "side" which expresses this part-of relation between side, and the EC for "orange juice carton."

An interesting challenge is presented by "This is the side": We do not have a possessive pronoun, only the definite determiner. In this case, we use reasoning over

hierarchical models to establish that a *side* is a part of an *object*. The definite determiner leads us to check whether we can consider it a part of a visually salient object. In the current scene, the carton is salient, and it is a type of object which has sides; hence "the side" is most likely the side of the carton, and we can again use co-indexation across structural descriptions for visual ECs to establish a relation.

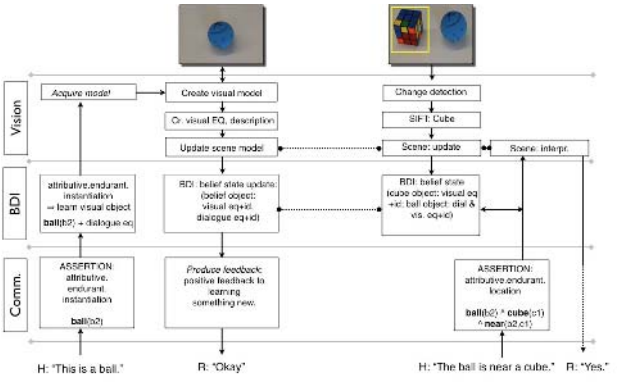


Fig. 4. Changing scenes

Figure 4 illustrates how we can use ECs to deal with changing scenes. After the visual system has learned a model for a new object ("ball"), we reorganize the scene, placing the ball next to a (known) cube. In this new scene, the original position of the ball has changed. We store the new scene in a ordered scene *history*. Now we say "The ball is near a cube", asserting the position of a (known) object. To interpret this assertion, we first need to resolve the new sighting of a ball object as an instance of the visual EC which corresponds to the discourse EC for "ball". Provided we can do so, we can then evaluate that the asserted spatial relation holds [Kelleher and Kruijff, 2005], and respond.

Finally, Figure 5 illustrates how we deal with references to previous scenes. As previously discussed, scene changes are inherent to dynamic environments but provide a problem for current approaches to vision/language-fusion (§2). To address this problem, we again use the ECs to relate different sightings of the same object, and provide history over the scenes in which these different sightings occurred.

When the human asks "Where was the ball at first?", the communication subsystem analyzes the utterance and determines its meaning as expressing a question about the location of an object. The meaning also indicates that the location is temporally circumscribed, and that we are after the location of the ball. BDI mediation uses the dialogue EC for "ball" to determine for which visual EC we should retrieve a scene in which we had a sighting. It then queries the scene history for a past scene – specifically, for the first scene as indicated by the temporal modifier "at first".

Based on the retrieved scene, BDI mediation establishes the dialogue ECs for the visual objects. This information is then provided to the communication subsystem, together with the scene, so that we can generate a spatial-locative expression to describe the past scene [Kelleher and Kruijff, 2005]: "The ball was to the left of the cube."

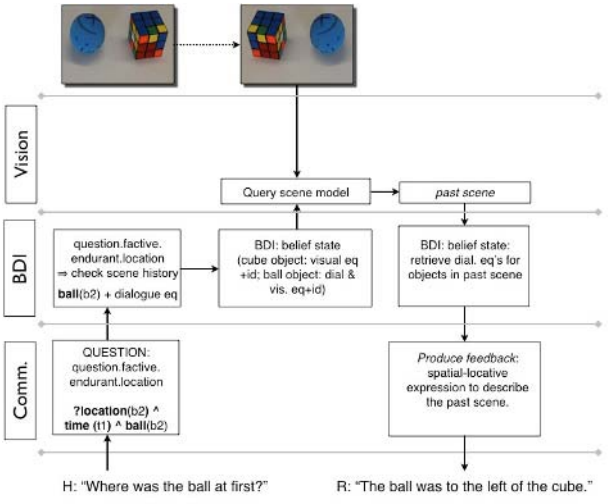


Fig. 5. Previous scenes

6 Conclusions

We have presented an approach to reference resolution for situated human-robot dialog. The dynamics of the environment and the noise inherent in robot perception makes this task genuinely difficult. The framework distinguishes between intra- and inter-modal fusion. Intra-modal fusion establishes whether different percepts from a single modality concern one and the same object. We introduced the notion of an equivalence class to represent the set of percepts from a single modality that relate to the same object. Inter-modal fusion relates the use of object references across different modalities, e.g. the resolution of an exophoric linguistic reference against an object in the robots perceptual field. Within the framework, inter-modal fusion results in the binding of equivalence classes from different modalities. A key element of the inter-modal fusion process is the use of ontology-based mediation to provide a mapping between conceptual systems to establish whether we can relate percepts from different modalities. One of the main advantages of this framework is that it provides a mechanism for dealing with the temporal dimension of situated reference. In contrast with previous approaches, the inter-modal binding process does not restrict linguistic reference to the current perceptual state. Rather, due to the fact that equivalence classes retain all the prior percepts relating to an object, a linguistic reference can leverage any of the prior perceptual states of the object.

References

[Alexandersson and Becker, 2003] Alexandersson, J. and Becker, T. (2003). The formal foundations underlying overlay. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*, pages pp. 22–36, Tilbugy, The Netherlands.

[Allen et al., 2000] Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., and Stent, A. (2000). An architecture for a generic dialogue shell. *Journal of Natural Language Engineering*, 6(3):1–16.

[Asher and Lascarides, 2003] Asher, N. and Lascarides, A. (2003). *Logics Of Conversation*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, United Kingdom.

[Baldridge and Kruijff, 2002] Baldridge, J. and Kruijff, G.-J. M. (2002). Coupling CCG and hybrid logic dependency semantics. In *Proceedings of ACL 2002*, Philadelphia, Pennsylvania.

[Baldridge and Kruijff, 2003] Baldridge, J. and Kruijff, G.-J. M. (2003). Multi-modal combinatory categorial grammar. In *Proceedings of EACL 2003*, Budapest, Hungary.

[Bos et al., 2003] Bos, J., Klein, E., and Oka, T. (2003). Meaningful conversation with a mobile robot. In *Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Li nguistics (EACL'03)*, Budapest, Hungary.

[Byron, 2003] Byron, D. (2003). Understanding referring expressions in situated language some challenges for real-world agents. In *Proceedings of the First International Workshop on Language Understanding and Agents for the Real World*.

[Cheyer and Martin, 2001] Cheyer, A. and Martin, D. (2001). The open agent architecture. *Journal of Autonomous Agents and Multi-Agent Systems*, 4(1):143–148.

[Coradeschi and Saffiotti, 2003] Coradeschi, S. and Saffiotti, A. (2003). An introduction to the anchoring problem. *Robotics and Autonomous Systems*, 43(2-3):85–96.

[Gurevych et al., 2003] Gurevych, I., Porzel, R., Slinko, E., Pfleger, N., Alexandersson, J., and Merten, S. (2003). Less is more: Using a single knowledge representation in dialogue systems. In *Proceedings of the HLT-NAACL Workshop on Text Meaning*, pages pp. 14–21, Edmonton, Canada.

[Kelleher and Kruijff, 2005] Kelleher, J. D. and Kruijff, G.-J. M. (2005). A context-dependent model of proximity in physically situated environments. In *Proceedings of the ACL-SIGSEM workshop The Linguistic Dimension of Prepositions*, Colchester, England.

[Kruijff et al., 2006a] Kruijff, G., Kelleher, J., Berginc, G., and Leonardis, A. ((Under Review) 2006a). Structural descriptions in human-assisted robot visual learning. In *Human Robot Interaction*, Salt Lake City, Utah.

[Kruijff, 2005] Kruijff, G.-J. M. (2005). Contextually appropriate utterance planning for CCG. In *Proc. of the 9th European Workshop on Natural Language Generation*, Abderdeen, Scotland.

[Kruijff et al., 2006b] Kruijff, G.-J. M., Kelleher, J., Berginc, G., and Leonardis, A. (2006b). Structural descriptions in human-assisted robot visual learning. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction (HRI'06)*, Salt Lake City, UT.

[Loutfi et al., 2005] Loutfi, A., Coradeschi, S., and Saffiotti, A. (2005). Maintaining coherent perceptual information using anchoring. In *Proc. of the 19th IJCAI Conf.*, Edinburgh, UK. Online at http://www.aass.oru.se/˜asaffio/.

[Lowe, 2004] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. In *Int. Jnl. Computer Vision*, pages 91–110.

[Poesio, 1994] Poesio, M. (1994). *Discourse Interpretation and the Scope of Operators*. Ph.d. dissertation, University of Rochester.

[Wache et al., 2001] Wache, H., Vögele, T., Visser, U., and G. Schuster, H. S., Neumann, H., and Hübner, S. (2001). Ontology-based integration of information - a survey of existing approaches. In *Proceedings of IJCAI 2001 Workshop "Ontologies and Information S haring"*, Seattle WA.

Speech and 2D Deictic Gesture Reference to Virtual Scenes

Niels Ole Bernsen

Natural Interactive Systems Laboratory, Campusvej 55, DK-5230 Odense, Denmark
nob@nis.sdu.dk
http://www.nis.sdu.dk

Abstract. Humans make ample use of deictic gesture and spoken reference in referring to perceived phenomena in the spatial environment, such as visible objects, sound sources, tactile objects, or even sources of smell and taste. Multimodal and natural interactive systems developers are beginning to face the challenges involved in making systems correctly interpret user input belonging to this general class of multimodal references. This paper addresses a first fragment of the general problem, i.e., spoken and/or 2D on-screen deictic gesture reference to graphics output scenes. The approach is to confront existing sketchy theory with new data and generalise the results to what may be a more comprehensive understanding of the problem.

1 Introduction

Speech and deictic (pointing, delimiting, etc.) gesture input is known as an excellent multimodal input combination for interacting with many different kinds of application. The reason is that, by itself, unimodal speech is often poor at providing unambiguous reference to spatial objects, unambiguously specifying spatial manipulations for the system to do, etc. [Bernsen 2002, cf. Bolt 1980]. With today's rapidly evolving multimodal technologies, we will soon be able to provide robust camera-captured 3D deictic gesture input into virtual reality scenes. Still, spontaneous *2D* deictic gesture and spontaneous spoken input addressing 3D virtual reality scenes remains the state of the art. This may be viewed as a good thing from the point of view of scientific methodology. It induces us to first develop applicable theory for the speech and 2D deictic gesture input case before attempting to generalise the theory two-fold, i.e., to the general case of handling spontaneous speech and spontaneous (i) *3D* deictic gesture input which, moreover, (ii) not only refers to visually perceived scenes but also to aurally perceived 3D sound sources, tactilely perceived objects, olfactorily perceived sources of smell, and gustatorily perceived sources of taste.

As for the background of this paper, the author and colleagues received a sense of the complexity of the problem when developing speech/gesture input fusion for a domain-oriented (or non-task-oriented) system enabling English user conversation with 3D-embodied fairytale author Hans Christian Andersen [Martin et al. 2006, to appear]. In the course of conversation, users may refer, using spontaneous speech and/or 2D tactile-screen deictic gesture, to objects in Andersen's study which he might tell stories about. Figure 1 shows Andersen in front of some seven such

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 129 – 140, 2006.

objects, i.e., the six pictures on the wall and his pen on the writing desk. In the absence of applicable theory for semantic-level speech/gesture input fusion, our approach to input fusion in the Andersen system was a series of cautious design decisions which favoured confidence in gesture input over confidence in spoken input. Whilst largely successful, it was clear to us from the start that those decisions do not scale to applications in which the spoken input contents are typically richer than being mostly a semantically and pragmatically redundant reflection of the gesture input contents.

Fig. 1. Hans Christian Andersen in his study surrounded by gesturable objects

Even if not directly *applicable* to the speech/gesture problem in the Andersen system, there is, in fact, *relevant* theory around. In particular, [Landragin 2006, to appear] proposes a sketch of how to handle, theoretically as well as algorithmically, the speech and 2D tactile-screen deictic references to virtual objects. The paper is based primarily on a small corpus collected under controlled circumstances and on partial algorithmic implementation in two task-oriented system research projects. As Landragin points out, his solution sketch is potentially limited by the small and in other ways limited corpus he has had the opportunity to analyse.

The question to be addressed below can now be stated in simple terms. If we test Landragin's approach with data from a rather different corpus, what happens? Will the theory sketch stand or must it be significantly expanded to capture the full dimensionality of the problem? In the latter case, we will probably have to analyse additional corpora and do more conceptual homework before arriving at stable theory on which to base processing of spontaneous speech and 2D deictic gesture references. The data analysed and discussed in this paper is from the user test of the second Andersen system prototype.

2 The General Problem Approached

The general problem may be characterised as follows. (1) users may use speech and/or 2D deictic gesture (henceforth: gesture) to refer to virtual objects as part of dialogue or conversation. (2) The system must be able to interpret, and appropriately respond to, whichever referential input is provided by the users, no matter which task, domain, or interactive application the user is addressing. (3) We need a formal model of the problem that can reliably serve as a basis for algorithm development. It should be noted that the problem is not about speech/gesture input fusion *per se*, it is about spoken *and/or* 2D deictic gesture reference to virtual objects. Multimodal fusion is an issue only when the user speaks and gestures more or less simultaneously. Even in a universe of discourse in which this form of multimodal reference is possible, we often make unambiguous reference to scenes and objects using speech-only or gesture-only.

2.1 Landragin's Approach

A slightly extended version of Landragin's model is shown below. Referential speech/ gesture input may be represented as the quadruplet:

/ referring mode / grammatical category of referring expression / deictic gesture yes/no / other information / (1)

Referring mode is a pragmatic descriptor of the referential act made, such as *indicate a particular referent*. This descriptor is modality-independent in Landragin's model due to some tacit assumptions made. *Grammatical category* is the grammatical category of the spoken referring expression, such as *indefinite noun phrase*. *Deictic gesture yes/no* marks if the referential communicative act did or did not include deictic gesture. *Other information* has been added to Landragin's model. It enables us to add information required for input interpretation, such as that the input is anaphoric or elliptic. In addition, the model includes the notion of a *reference domain,* i.e., a subset of scene objects with something in common, such as forming a spatial group or being similar in shape, size, or perceptual salience. Humans introduce reference domains to simplify reference interpretation by sub-dividing the visible scene into sub-domains within which it is easier to disambiguate spoken or speech/gesture reference to particular entities. The nature of human perception must be taken into account because, e.g., the relative *salience* of visually perceived objects and the phenomenon of *perceptual grouping* are important underlying mechanisms for resolving referential ambiguity in conversation involving speech and 2D gesture.

The main corpus on which the model is based is a corpus of 98 data points from a Wizard of Oz exercise in which subjects had to manipulate, i.e., identify and request system actions onto, virtual objects, such as triangles and squares [Wolf et al. 1998].

Table 1 shows the result of Landragin's corpus analysis. Six *referring modes* were found. These referential actions are to: (1) introduce a new referent by creating it -*new-ref*; (2) extract any referent -*ext-any-ref*- or (3) extract a particular referent -*ext-par-ref*- from an already delimited reference domain; (4) indicate a particular referent that is, or has been, focused by, e.g., gesture or prior spoken reference -*ind-par-ref*; (5) indicate a particular reference domain, using gesture to focus on a particular object in the domain -*ind-par-dom*; (6) referring to a generic entity, e.g., 'triangles' -*gen-ref*.

Table 1. Landragin's corpus analysis. Dem is demonstrative, NP is noun phrase. P is pronoun.

Referring mode	**Grammatical category**	**Gesture**	**Example**	**Other information**
new-ref	Indefinite NP	**no**	Create **a square**	Kataphor
ext-any-ref	Indefinite NP	**no**	Delete **a square**	Anaphor
ext-par-ref	Definite NP	yes	**The square**	
	Definite NP	**no**	**The square, The square** to the left	Anaphor or contextually obvious
	Dem NP	**no**	Delete **this square**	Anaphor
	Dem P	**no**	**This one**	Anaphor
ind-par-ref	Indefinite NP	yes	Delete **a square**	
	Definite NP	yes	**The square**	
	Definite NP	**no**	**The square**	Anaphor
	Dem NP	yes	**This square**	
	Dem NP	**no**	**This square**	Anaphor
	Personal P	**no**	Delete **it**	Anaphor
	Dem P	yes	**This one**	
ind-par-dom	Definite NP	yes	**The squares** [pointing to one of them]	
	Definite NP	**no**	**The group**	Anaphor
	Dem NP	yes	**These squares** [poin ting to one of them]	
	Dem NP	**no**	**This group**	Anaphor
	Personal P	**no**	Delete **them**	Anaphor
gen-ref	Indefinite NP	**no**	**A square** has four sides	
	Definite NP	**no**	**The square** has four sides	
	Dem NP	yes	**These forms** [pointing to one of them]	
	Dem NP	**no**	**These forms**	Anaphor
	Dem NP	yes	**This form**	
	Personal P	**no**	I have added a red square because **they** are eye-catching	Anaphor
	Dem P	**no**	**These ones** are eye-catching	Anaphor

The *grammatical categories* expressing the referring modes in the corpus are five: *indefinite noun phrases*, *definite noun phrases*, *demonstrative noun phrases*, *demonstrative pronouns*, and *personal pronouns*, see Table 1 for examples. The table also shows if deictic gesture accompanies spoken reference and adds other information.

2.2 Potential Need for Generalisation

How representative is the universe of referential discourse of Table 1 of the full complexity of the speech/2D gesture reference problem? Drawing upon Modality Theory [Bernsen 2002], the corpus has the following properties: (1) *static* graphics output domain in which the user perceives the *entire* collection of objects; (2) *mere output objects*, i.e., the geometric shapes in the corpus do not themselves represent information; (3) *2D objects* which do not allow for occlusion, objects resting on larger objects, etc.; (4) *simple and easy-to-label objects,* such as triangles and circles of different colour. In addition, (5) users' spoken input seems to be *simple and partially controlled* language. Users appear to speak explicitly at all times, always saying, e.g., "this triangle" rather than "this", i.e., using a demonstrative noun phrase when a pure demonstrative might suffice. What the users do in addition to object reference is to command simple changes, such as 'select', 'create' or 'delete'; (6) there are *no gesture-only object references*. Users never seem to shortcut by just pointing to objects.

Clearly, there are lots of applications which could benefit from multimodal speech/gesture input reference and which do not share the limitations just listed. It is thus a real possibility that, were we to analyse a corpus from an application domain with different general characteristics, we might find a different pattern of referential quadruplets, forcing extension to the quadruplet formal model *in spe*.

Let us use the method of generalisation-by-negation to broaden the scope of a general theory of speech/2D gesture reference, following the numbering above and introducing a couple of definitions (DEFn). *DEF1*: let us call what the user sees and what is being referred to, a visual or graphical *scene*. Note that the scene itself, and not just the objects *in* the scene, may have properties and be referred to. (1) The scene may be *static or dynamic*. It may include objects which move, change or act, and the scene itself may shift dynamically, e.g., because the user changes the virtual camera angle or the scene itself shifts beyond the user's control. *DEF2*: some scenes may be described as *scene worlds*, i.e., as the sum total of all possible scenes in the application, past, present and future. Thus, users might refer to past-but-not-present *scene world snapshots*, their objects and properties, and to future and expected but not-yet-perceived snapshots. (2) Scene objects may be either *mere objects*, such as a triangle, or *representational objects* which represent information, such as an image showing several objects. Users may correctly refer to this image using both singular and plural pronouns, as in "What is this [pointing]?" vs. "Who are they [pointing]?" [Martin et al. 2006, to appear]. (3) Scene worlds may be *2D or 3D*. 3D introduces new referentially relevant aspects, such as the backside of objects or objects viewed from above, occluded by other objects, or resting on larger objects. (4) Contrary to simple geometric shapes, *complex real-world-like objects* do not necessarily have a single standard label and are easier to mislabel when referring to them. (5) The general case of spoken input is uninstructed *spontaneous speech*. A general solution to our problem must be able to handle any kind of spoken reference to scenes, objects and properties,

irrespective of linguistic complexity, speech acts performed, etc. (6) As for 2D deictic gesture, it must be assumed that users will sometimes make *gesture-only reference*. This can be done correctly and successfully, as we shall see, because application-specific deixis sometimes, at least, comes with implicit or explicit semantics and pragmatics.

There may be other dimensions along which we should generalise in order to describe the full scope of a theory of speech/2D gesture reference. However, the above generalisations demonstrate that the universe of referential speech-2D deictic gesture discourse is far larger than the one addressed in Landragin's main corpus. In fact, those generalisations might provide an approximate target for a general theory.

Still, these are general arguments. We need to analyse new corpora in which users refer to scenes, objects and properties in sectors of the universe of referential speech-2D deictic gesture discourse other than those addressed by Landragin's model in order to identify new specific types of referential discourse compared to those in Table 1.

3 A Different Corpus

In this section we present and analyse a corpus of English speech and 2D deictic gesture which represents a rather different fragment of the universe of referential discourse compared to Landragin's main corpus. This new corpus reflects interaction with an application having all the properties generated in Section 2.2, including: a dynamic scene world; representational objects; 3D scenes and objects; complex photo-realistic, real-world-like objects with multiple properties; spontaneous conversational speech input; and gesture-only reference.

The corpus was produced as a follow-up to the user test of the second Hans Christian Andersen (HCA) system prototype made in February 2005. By contrast with the larger user test which involved Danish kids speaking English, the corpus to be discussed here was recorded with four native English speaking children, two girls and two boys, aged between 10 and 13 years. The native English test had two test conditions. In the first condition, the users were (i) instructed in how to use the keyboard for changing virtual camera angle and making HCA move when in non-autonomous mode, the tactile screen, and the microphone headset; and (ii) coached in spontaneous, free-style conversation with HCA, such as in re-phrasing input if not understood rather than just repeating the input. For the second condition which lasted 20-25 minutes per subject, the users were provided with a handout which proposed a series of 11 issues they could try to address in conversation at their leisure and in random order. We shall be looking at the second-condition corpus consisting of four conversations with HCA. All module interactions were logged and the spoken user input was recorded and transcribed. To correct for gesture recognition and interpretation errors relative to what the users actually did, the gesture log data was augmented with data from the two-camera video recordings of the user-system interactions. All HCA development and test corpora are available at NISLab's website [www.nis.sdu.dk] and are described more comprehensively in [Bernsen et al. 2006, to appear].

3.1 Corpus Annotation

With one major exception, all speech and/or 2D deictic gesture references to scene worlds and scenes, scene properties, and scene objects and their properties have been annotated. The exception is the many spoken input utterances in which the *sole* scene referent is HCA but in which no further reference is made to visual scene contents. Rather, reference is made to abstract discourse entities. Thus, e.g., an input utterance referring to "your hair" is annotated since HCA's hair is visible but, e.g., reference to "your fairytales" is not annotated because HCA's work is an abstract discourse object. Similarly, "you are ugly" is included since 'ugly' refers to visible properties of HCA, whereas, e.g., "you are cool" is not annotated. By implication, speech-only input which refers anaphorically to abstract discourse entities, such as the volunteered user comment "That is a sad story", is not annotated. If there is gesture, any accompanying speech is annotated for any reference.

Ignoring annotation beyond the scope of this paper, the corpus was annotated as follows: (1) Find the next data point to be annotated given the criteria stated at the start of this section. (2) If the data point includes spoken reference, identify the referring word or phrase and its grammatical category. If the category is in Landragin's coding scheme, mark this. If not, create new quadruplet. (3) Assign one of Landragin's referring modes to the spoken reference. Create new referring mode if referring mode is not in Landragin's coding scheme. (4) Mark if gesture accompanies spoken input. (5) Mark other relevant information.

The results of annotating the 78 data points in the corpus are shown in Table 2

Table 2. Results of annotating the native English Andersen corpus. Dem is demonstrative, NP is noun phrase. P is pronoun, ref is reference, S is speech.

Referring mode	Grammatical category	Gesture	Example	Other information
1. ind-par-ref	Dem NP	yes	Have you written **these books**?	
2.	Pure Dem	yes	What is **this**? **This**?	S redundant
3. gen-ref	Indefinite NP	yes	Did you always write with **a feather**?	
4.	Indefinite NP	yes	Do you like **trains** [pointing to picture of locomotive]	Indirect ref
5.	Indefinite NP	yes	Do you **read** a lot? [pointing to stack of books]	Indirect ref, ellipsis
6. scene-world-ref	Pure Dem	yes	Is **this** were you live? [circling gesture]	Metonym
7.	Pure Dem	**no**	Is **this** where you live?	Metonym
8. unique ref	Definite NP	**no**	**Your nose** is very big	Exophor

Table 2. (*continued*)

9.	Personal P	**no**	**You** are very old	Exophor
10. gesture-only-ref	N/A	yes	[pointing to anything]	N/A
11. indep-S-gesture-ref	Definite NP	yes	I do not like **your hair** [pointing to featherpen]	Exophor
12.	Personal P	yes	Have **you** been baptized? [pointing to featherpen]	Exophor

3.2 Annotation Analysis

Table 2 shows 12 quadruplets, only one of which, i.e., Quadruplet 1 */ ind-par-ref / Dem NP / yes / - /* is also present in Landragin's main corpus, cf. Table 1 Let us look at the 11 others in order, using Qn for Quadruplet (n).

Q2, the pure demonstrative *this* or, rarely, *that*, used in combined speech-gesture reference, occurred more frequently than any other quadruplet in the corpus with 38 or close to 50% of all data points. Q2 is always part of an explicit or implicit interrogative, such as "What is this?" We consider "What is this?" strictly *redundant* relative to the accompanying gesture, so let's explain the claim made earlier that deictic gesture itself can have a particular semantic and pragmatic meaning which may be relative to the application. HCA encourages the user to point to objects which he can tell stories about. So, when this happens, the meaning of the pointing *includes* the meaning of the spoken question "What is this?" and its variations in number or in elliptical expression. Note that the pointing gesture has its own pragmatic meaning in addition to the abstract spoken meaning pattern it includes. If the object referred to is not visually salient or in other ways prominent in the scene, the spoken "What is this?" may not succeed in uniquely referring to the object, whereas well-formed pointing to the object will always do that. So, the redundancy is asymmetric.

Q3, Q4 and Q5 are all *gen-ref* quadruplets in which the NP refers to *a kind* of object rather than to the particular object referred to in the accompanying gesture. From the point of view of input interpretation, Q4 and Q5 are of particular interest. In Q4 the user points to a picture, i.e., an object which itself represents information, of a *locomotive* but asks about *trains*. We call this *indirect reference* and the system must figure out, somehow, that the user is right in talking about trains here rather than declaring the user wrong or asking what the user means. In Q5 we also have a form of indirect reference, i.e., the ellipsis "Do you read [books] a lot?" or "Do you read a lot [of books]?" The system must figure this out.

Q6 and Q7 belong to our first new referring mode, i.e., *scene-world-ref*. In Q6, the user makes a circling gesture on the screen and asks about the entire scene world. What the user actually sees on the screen at the time is part of HCA's study but the user refers to the study as a whole or possibly to HCA's entire apartment. Since the question is not understood by HCA, the user repeats the question without accompanying gesture (Q7) and gets the correct response from HCA. We have marked the user's pure demonstrative reference and encircling gesture as *metonymic* because the communicative intent is to refer to the whole by referring to part of it.

Q8 and Q9 belong to a second new referring mode, i.e., *unique-ref*. Without using gesture, the user succeeds in uniquely referring to on-screen objects using a definite NP and a personal pronoun, respectively. In reference theory, these grammatical forms would normally be anaphoric references which presuppose that their referents, i.e., HCA's nose and his physical person, respectively, have been introduced already. This is why there are so many anaphoric quadruplets in Landragin's main corpus, cf. Table 1 However, explicit prior referent introduction is not necessary in *exophoric* reference by which we successfully refer to prominent, or otherwise unique, objects, perceptually or otherwise, with no prior introduction. For instance, we don't need to introduce "the sun" as referent prior to saying, e.g., "The sun is hot today". In fact, Table 1 illustrates another reference phenomenon in addition the anaphor, i.e., *kataphoric* reference, in which we "refer ahead" to a referent that will be introduced later, as is, indeed, the case for an object which will only come into existence as an effect of the user's command "Create a square".

Q10 is a third new, *gesture-only-ref* referring mode which shows, furthermore, that not all referring modes are modality-independent (cf. Section 2.1). Q10 works perfectly well in the application context, we argue, because gesture already has a well-defined semantic-pragmatic meaning, i.e., what we would render linguistically as, e.g., "What is this?" This is why the system has no problem interpreting the user's intended meaning and also why, had the user said at the same time "What is this?", this spoken utterance would have been redundant relative to the deictic gesture (cf. above).

Finally, the fourth new referring mode, *indep-speech-gesture-ref*, highlights another limitation of Landragin's corpus, i.e., that speech and gesture are always complementary in that corpus. Both *complementarity*, i.e., that speech and gesture both contribute necessary and non-redundant parts of the user's intended meaning, and redundancy (cf. above), imply that speech and gesture are *semantically related and consistent*. In real life speech/gesture interaction, however, speech-gesture *inconsistency* is bound to occur from time to time and so is speech and gesture which is not semantically related but, rather, *independent* from each other. Although our small native English corpus does not have a case of the former phenomenon, Q11 and Q12 are cases of the latter.

4 Corpus Comparison

Section 2.2 presented a series of conceptual arguments why the interactive task and the application domain with which Landragin's main corpus was generated must be considered severely limited in many respects compared to the full potential universe of application of speech and 2D gesture. In the course of the argument, we introduced a series of notions of aspects of potential scenes which might be addressed through speech and 2D deictic gesture and which did not (appear to) apply to the scene and the interaction with it in the application used for collecting Landragin's main corpus. The notions were: *dynamic scenes, scene worlds, representational scene objects, 3D scenes and objects, complex, real-world-like objects, rich spontaneous speech,* and *gesture-only reference*. The implication was the hypothesis that a speech/2D deictic gesture corpus collected with a different form of interaction and a different application domain, might include quadruplets different from those listed in Table 1.

The scene world of the HCA system is characterised by all the notions introduced in Section 2.2. And the analysis of the native English HCA corpus in Section 3.2 confirms the hypothesis that a corpus of this nature would exhibit very different referential phenomena from those reported in Table 1. In fact, the confirmation is massive to the extent that only a *single* quadruplet among the 25 quadruplets in Table 1 was found in the HCA native English corpus. Even if [Landragin 2006, to appear] were right in claiming that this one, i.e., Quadruplet 1 in Table 2, */ ind-par-ref / Dem NP / yes / - /*, is the most common way of referring to scene objects using speech and 2D deictic gesture – the HCA data does not bear this out since pure demonstrative reference was done five times more frequently than reference using a demonstrative NP - it seems highly likely that we still have more to learn about the varieties of speech and 2D deictic gesture references. It would seem näive to claim that the *union* of Tables .1 and 2 comes close to constituting the total number of speech and 2D gesture reference phenomena. Rather, the user-HCA conversation provides a first taste of unconstrained speech/2D deictic gesture interaction and that's it. To better understand why, let us revisit the corpus analyses in Tables 1 and 2.

4.1 Absent from HCA Conversation but Present in Landragin

In comparing the corpus analyses, we propose to focus on the quadruplet values *referring mode*, *gesture yes/no* and *other information*, considering *grammatical category* a more detailed quadruplet value to be worked out when we have a better grasp of the reference problem as a whole.

Among the six referring modes in Table 1, four were not found in the native English HCA corpus, i.e., *new-ref, ext-any-ref, ext-par-ref*, and *ind-par-dom* (see Section 2.1 for definitions). Among these, *new-ref* (e.g., "Create a square") and *ext-any-ref* (e.g., "Delete a square"), might be viewed as tied to a particular family of application. Also *ind-par-dom* in which a particular reference domain is indicated through a gesture that focuses on a particular object in the domain, might be in this category given the many-look-alike-objects-character of Landragin's main application domains. The absence of *ext-par-ref* which is used to extract a particular reference from a reference domain activated through gesture, verbal reference, description, previous references to objects in the domain, or visual salience – is more surprising.

However, it is easy to imagine slight extensions to the HCA system which would enable the occurrence of *new-ref* and *ext-any-ref*. All we need is an HCA who can act in certain ways when encouraged by the user. For *ind-par-dom* and *ext-par-ref*, we do not even need system modification. All we need for *ind-par-dom* to occur is a user who says, e.g., "Tell me about the pictures" [pointing to a single picture in Figure 1]. This just did not happen in our data. Similarly, all we need for *ext-par-ref* is a user who says, e.g., "Let us talk about the pictures above your desk" followed by "tell me about the one on the left". It may be concluded that the absence from the HCA corpus of some of the referring modes in Table 1 in no way implies that those referring modes are strongly tied to a particular family of applications.

A note on the scope of a theory of speech/2D deictic gesture reference concerns the Table 1 *gen-ref* cases of using an indefinite or definite noun phrase without accompanying gesture in descriptions of triangles-in-general. It is not obvious that these cases refer to scene aspects at all even though the scene objects include, e.g., triangles.

4.2 Present in HCA Conversation but Absent in Landragin

Section 3.2 describes four new referring modes, i.e., *scene-world-ref, unique-ref, gesture-only ref*, and *indep-speech-gesture-ref*. Of these, *scene world ref* enables reference to the scene world as a whole. Since the scene world is by definition not visible as a whole at any one time, pure demonstrative reference and deictic gesture-only reference to it must necessarily be metonymic whereas non-metonymic linguistic reference can be easily done. The referring mode *unique-ref* is exophoric reference to unique scene objects and properties. Exophoric reference does *not* depend on visual salience as discussed by Landragin but may be done to non-salient objects as long as these are unique in the scene world. *Gesture-only-ref* represents a much-needed theory extension acknowledging gesture-only reference. Finally, *indep-speech-gesture-ref* reflects another necessary theory extension which takes into account that more or less simultaneous input speech and deictic gesture may be semantically and pragmatically independent. In addition, we found some new *gen-ref* quadruplets, including some involving indirect reference and elliptical reference. Taking the two corpora together, we have found anaphoric, kataphoric and exophoric reference. And, having found speech-gesture complementarity, redundancy and independence, it is easy to predict that a new corpus could show speech-gesture inconsistency – an 11th referring mode.

5 Conclusion

If we make the common assumption that the sign of mature theory, such as a theory of speech and/or 2D deictic gesture reference to scene aspects, is its ability to predict and explain the large majority of data within its scope, the inevitable conclusion is that the data we have looked at is only the tip of the iceberg. Statistical convergence between the categories of the theory and the phenomena present in the data corpora used for its development would seem a long way off. We can forget about general algorithms for speech and/or 2D deictic gesture interpretation for the time being.

What is needed is, first of all, new speech and 2D deictic gesture corpora whose analysis can add more concepts for a general theory than those presented in Landragin and in this paper. Secondly, it is necessary to take a new look at the structure of a theory which could incorporate the results. A possible top-level organising principle is the distinction between speech-only reference, gesture-only reference and more or less simultaneous speech-gesture reference. We also need a more thorough analysis of the inventory of scene worlds than made above. We have had a glimpse of the fact that 2D deictic gesture reference is inherently complex, so, we should also look at, e.g., what is the significance of different deictic gesture shapes, such as points, circles, semi-circles, straight and curved lines, crosses involving several separate contacts with the screen, doodles, etc.; are there relevant differences between tactile-screen deictic 2D gesture and, e.g., mouse deictic gesture; what is the significance of the temporal aspects of speech/deictic gesture reference; and what is the semantics and pragmatics of 2D deictic gesture for different application families? We also need, at some point, a thorough theoretical comparison with, and possible multimodal extension of, linguistic reference theory [Kamp and Reyle 1993], or rather, perhaps, subsumption of both under a unified theory of multimodal reference, considering purely linguistic reference as a special case.

Finally, we need to take a system development point of view of it all. Like humans, the system must evaluate the input before planning its response. For instance, it must evaluate the truth of a definite NP like "This triangle [pointing]". What if "this" object which is being both pointed to and classified as a triangle is *not* a triangle?

Acknowledgement

The HCA system was built in the NICE (Natural Interactive Conversation for Edutainment) project supported by the EU's Human Language Technologies programme. The support is gratefully acknowledged. I would like to thank Svend Kiilerich for his video-based annotation correction of the HCA native English corpus.

References

1. Bernsen, N. O.: Multimodality in Language and Speech Systems - from Theory to Design Support Tool. In: Granström, B., House, D., and Karlsson, I. (eds.): Multimodality in Language and Speech Systems. Kluwer Academic Publishers, Dordrecht (2002) 93-148
2. Bernsen, N.O., Dybkjær, L., Kiilerich, S.: H.C. Andersen Conversation Corpus. In: Proceedings of the Language Resources and Evaluation Conference, Genoa, May 2006 (to appear)
3. Bolt, R.A.: Put-That-There: Voice and Gesture at the Graphics Interface. Computer Graphics 14(3) (1980) 262-270
4. Kamp, H., Reyle, U.: From Discourse to Logic. Kluwer Academic Publishers, Dordrecht (1993)
5. Landragin, F.: Visual Perception, Language and Gesture: A Model for their Understanding in Multimodal Dialogue Systems. Special Issue of on Multimodal Interaction, Signal Processing (2006, to appear)
6. Martin, J.C., Buisine, S., Pitel, G., Bernsen, N. O.: Fusion of Children's Speech and 2D Gestures when Conversing with 3D Characters. Special Issue of on Multimodal Interaction, Signal Processing (2006, to appear)
7. Wolf, A., De Angeli, Romary, L.: Acting on a Visual World: The Role of Perception in Multimodal HCI. In: Proceedings of the AAAI'98 Workshop: Representations for Multi-Modal Human-Computer Interaction. Madison, Wisconsin (1998)

Combining Modality Theory and Context Models

Andreas Ratzka

Universität Regensburg, Lehrstuhl für Informationswissenschaft, 93040 Regensburg, Germany
Andreas.Ratzka@sprachlit.uni-regensburg.de
http://www-iw.uni-regensburg.de

Abstract. This paper outlines a research plan with the purpose of combining model-based methodology and multimodal interaction. This work picks up frameworks such as modality theory, TYCOON and CARE and correlates them to approaches for *context of use modelling* such as the interaction constraints model and the unifying reference framework for multi-target user interfaces. This research shall result in methodological design support for multimodal interaction. The resulting framework will consist of methodological design support, such as a design pattern language for multimodal interaction and a set of model-based notational elements.

1 Introduction

This research aims at contributing to model-based support for developing multimodal systems. Traditional model-based approaches are focusing on separation and mapping between task specification, abstract interaction and concrete interaction with the purpose of helping developers to design device independent applications running on different desktop platforms and, for instance, mobile devices. Recent approaches include *context of use models*, such as user models, situation models, platform-models etc. (see for instance [8] or [9]).

A first, but central step in this research is to connect approaches for context of use modelling to theories of multimodal interaction. This is necessary because up to now, model-based approaches and modality description frameworks are existing in parallel and attempts to connect them are rare. Two of those are outlined next.

The authors of [15] present the output presentation tool MOST which uses a rule-based approach to perform context-appropriate modality selection.

The work presented in [5] and [6] covers context based mappings of application requirements to interaction styles based on a collection of modality claims. Due to the high complexity of this domain, the author's research group has given up the aim of generating a comprehensive "grammar" of modality combinations. One fruitful outcome of their research is a decision support tool, giving advice on the use of speech in interactive systems [5].

The research presented in this paper assumes that there is a great need in methods and design support for multimodal interaction. Attempts to generate hard-wired mappings through highly formal notations often fail because of the lack of formalisation

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 141 – 151, 2006.

and externalisation of design knowledge. Due to the dynamic development of interaction technologies a complete formalisation of design aspects seems to be unachievable. Therefore semiformal approaches, such as design patterns, will gain in importance.

The outcome of our work will be offering design support to developers of multimodal interactive systems. This design support may consist of design patterns, guiding the developer in the use and allocation of input and output modalities as well as notations for reusable and consistent interaction specification.

This paper presents ongoing research. The results presented here are based on literature review and derived conclusions are drawn upon introspection and still require empirical validation.

The following section describes modality theory and our extensions made in order to cover aspects, such as channel properties and interaction granularity, which can be mapped to from context of use factors depicted later on.

2 Modality Properties

Several approaches, such as TYCOON [12], CARE [10] and others [13], [18] have introduced terminologies for describing aspects of modality combination. As a first step in the development of multimodal systems, we have to consider which modalities are *equivalently* exchangeable with respect to certain subtasks and scenarios and which ones are specialized. This notion of equivalence vs. specialization [12] or equivalence vs. assignment [10] can be applied to interaction modalities in general but also to single modality properties, such as the ones covered by modality theory [3], [4].

The first subsection describes modality theory developed by Bernsen [3], [4]. In the following subsections we introduce new dimensions of description which are necessary for mapping of modality properties and aspects of the task and context of use. As modality theory focuses on output modalities, it does not cover data input. In order to overcome this shortcoming, we attempt to introduce further modality properties for data input.

2.1 Bernsen's Modality Theory

Bernsen introduced a comprehensive taxonomy of output modalities [3], [4]. Output modalities can be analysed as linguistic versus non linguistic ones, as analogue ones or non analogue ones, arbitrary or non arbitrary modalities, static or dynamic ones. Finally, modalities can be classified as acoustic, haptic or graphic ones, concerning the perception channel.

Vernier and Nigay added the analysis dimensions deformed vs. non deformed, local vs. global and precise vs. vague. They organized these dimensions into articulatory syntactic and semantic levels [18].

In the following subsections we describe our refinements to this modality description framework. The articulatory level, consisting of channel and the opposition pair static vs. dynamic, can be described by more general properties such as chromatic resolution, spatial resolution and spatial selectivity. These can be more easily mapped to task requirements and context factors than channel names. Furthermore, we are introducing additional modality properties for data input.

2.2 Refining the Articulatory Level for Data Output

The articulatory level comprises the physical modality aspects of articulation, perception and communication channel. As for the articulatory level, modality theory has introduced the information channel (graphics, acoustics, haptics) and the opposition pair static vs. dynamic as modality properties. For our purposes, we will refine the modality description at this level to achieve a more detailed characterisation of perceptual and articulatory abilities of humans as well as situational aspects. Therefore, we introduce the channel properties chromatic resolution, spatial resolution and spatial selectivity. Chromatic resolution covers stimulus diversity such as colour (graphics) or timbre (acoustics) and determines together with the spatial resolution the coding capacity of a channel.

Table 1. Channel Properties at the Articulatory Level

Channel	Acoustics	Graphics	Haptics
Chromatic resolution	Very High	High	Low
Spatial resolution	Low	High	Very High
Spatial selectivity	Ambient	Directed	Touch

The chromatic resolution of acoustic stimulus is very high (diversity of timbres, and pitch) but the spatial one is relatively low. Complex codes impose therefore the use of acoustic *dynamic* modalities. Acoustic stimulus can be perceived at high and near distances with low spatial selectivity (value: ambient). Everyone around can listen to spoken conversations. Therefore, in public surroundings spoken interaction is not desirable, because the people around might feel annoyed. Furthermore private information should not be revealed to the public.

The visual channel has high spatial and chromatic (colours, brightness etc.) resolution. At the same time visual stimulus can be perceived even at high distances, but only if there is a direct visual connection from the person to the emitting device.

The diversity of discrete haptic stimulus is relatively low. At the other hand the haptic channel is characterised by very high spatial resolution. The resulting coding capacity is, however, exploited only by sign systems for special user groups, such as the Braille script for blind. Moreover, haptic stimulus can be perceived only on direct touch and therefore does not allow for hands free interaction.

Haptic and graphic modalities are characterised by high spatial selectivity (touch or directed). In this case abstract device properties affecting the interaction area should be considered as well. Such abstract device properties contain position, size, distance to the user, and the degree, to which these properties are adjustable to current interaction needs.

2.3 Refining Modality Theory for Data Input

Articulatory Level. For characterising input modalities, we should not only consider the channel, via which the information is being conveyed, but also the aspect of signal

source, that is, which limbs are involved in the user *action*. We can distinguish *vocal*, *mimic*, *manual* and *pedal* actions. In most of the cases, vocal actions are interpreted acoustically by the device, but lip readers are interpreting speech related lip movements visually (see for instance [2]). Facial expressions and manual gestures can be interpreted visually. Manual and pedal actions are usually interpreted mechanically, that is the user activates a key or a switch mechanically. Both aspects, the *articulatory signal source* and the *channel* are to be considered for the selection of appropriate input modalities. Therefore both aspects will be covered by our modelling framework.

Syntactic Level. The syntactic level describes aspects of the external constitution and composition of a sign. In terms of interaction modalities, this comprises the opposition pairs linguistic/non-linguistic and deformed/non-deformed [18]. We further add the aspect of interaction granularity.

For data input, it is necessary to consider the input granularity of the interaction. Multi-token utterances are coarse grained. That means that the user can input a lot of tokens in one step. In contrast directed dialogue consists of single-token utterances: The user is forced to input token after token. Hence, directed dialogue is a finer grained dialogue strategy. Directed dialogue implies higher interaction robustness for novice users as well as in unfavourable environmental situations, but might be annoying for expert users who prefer using multi-token shortcuts.

Spelling is an even finer grained input modality, making use of the double segmentation of linguistic signs (see [11]). The same holds for typing in contrast to gesture selection. Spelling and typing are the only ways of defining new tokens, whereas for selecting a token from a limited set, spelling and typing are equivalent to more coarse grained modalities, such as gesture-based selection. At the same time, transitions to finer grained modalities are successfully employed in error handling and prevention strategies (see for instance [16], [17]).

As this *syntactic* aspect of granularity plays an enormous role in interaction design, our framework covers them as modality properties.

3 Frameworks for Context of Use Analysis

In order to determine which modalities should be used equivalently and which ones in a specialized way – according to task and situation, one has to generate models of the task to be performed and the context of use.

Calvary et al. distinguish domain models (comprising task and concept models), context models (consisting of user, platform and environment models), and adaptation models (comprising evolution and transition models) [9].

Bürgy presents a comprehensive interaction constraints model for mobile wearable industrial applications which analyses the interaction according to the target user, device, environment, application and task [8].

Calvary presents a survey of models used in so called *multi target applications*, but gives no details about context of use modelling or the mapping of context factors to modality properties. Bürgy's work contains a very detailed analysis of context of use properties but disregards multimodal interaction.

In the following subsections we describe our framework for *context of use modelling*. Our work shall close the gap in current research which has been depicted above. Our framework consists of models of the task, the situation, the user and the device. In section 4 we show how these models can be mapped onto the modality properties defined in section 2.

3.1 Task Model

Our framework describes task models at the contextual, semantic and discourse levels.

The contextual level (or task context) covers the combination of different subtasks within the interactive system as well as beyond this one (e.g. using a navigation system while driving). This is comparable to the so called task type in Bürgy's interaction constraints model [8], which determines the combination of primary and secondary tasks.

The semantic level describes which kind of data has to be exchanged between system and user. Software for picture editing has to display analogous graphic data and to receive at the same time analogous gestures such as drawing a line. A conversion of these analogous domain data is hardly possible nor recommended. Non analogous dialog acts, such as selecting a tool from the palette can be performed by non analogous modalities, such as WIMP-like pointing gestures or speech-based commands as has been presented successfully in [14].

The discourse level covers aspects of topic changes and information priority. For system output in response to user input we can anticipate user attention. For system output which is not motivated by discourse context it is important to consider the information priority. System acts such as notifying the user about an important event should be allocated to an attention catching modality e.g. speech. If, in contrast to this, the notification is not urgent, the user should not get disrupted.

3.2 Situation Model

The situation covers activity-related, articulatory and social aspects determining modality choice.

The activity-related aspect of the situation (or usage context) is comparable to the task context. It considers activities the user is performing while interacting with an application. In contrary to the task context, these activities (e.g. using a mobile messaging system while driving) are not motivated by the task itself.

The articulatory aspect of the situation covers environmental conditions affecting the interaction channel physically, such as noise and lighting.

The social aspect covers environmental conditions affecting the desirability of certain interaction modalities. If the user is situated in a public environment, it is not desirable to transmit private information such as PINs or email messages via voice. Furthermore, surrounding people might feel annoyed by a spoken man machine conversation.

3.3 User Model

The user model covers articulatory and perceptual skills (articulatory aspect) as well as user expertise (lexical aspect).

More experienced users prefer more efficient but complex dialogue acts, whereas for novices a simple step by step strategy might be more appropriate.

3.4 Device Model

The choice of the device itself is determined by a trade-off between costs and the above described task related and task independent aspects. The device restricts modality use to the currently supported interaction channels. Furthermore, physical properties determine modality combination and integration of primary and secondary tasks.

4 Dependencies Between Modality Properties and Context of Use

This section illustrates dependencies between modality properties (see chapter 2) and context of use aspects (see chapter 3). A preliminary overview of these dependencies is given in Table 2 (see below).

Discourse properties of the task determine modality choice at the articulatory level. This means, if a task consists in notifying the user of an unanticipated event, the user might not pay attention to system output. If the event is of high priority, the system should use an output modality with ambient spatial selectivity. At the same time we have to consider the contextual aspect of the task as well as the activity-related aspect of the situation. If the user is performing a primary task while interacting with the device, it is important to know how much attention has to be devoted to this primary task and whether disruptive system output such as speech or a distracting display might affect security aspects. There is no general solution for this conflict of objectives.

The semantic aspect of the task covers the data which is being processed and exchanged by the user and the application. This aspect affects modality usage at all levels. Painting programs imply the use of analogous modalities, as pictures rank among analogous modalities. The syntactic level is being affected as well, as the data determines the kind of interaction languages to be used. In order to show how the task influences the syntactic aspect of the modalities, we take an address book application as an example. For creating new contacts we have to use finer grained input modalities such as spelling and typing than for simply selecting existing entries. The semantic aspect of the task determines the articulatory aspect of the modality as well. Drawing applications usually involve graphical output modalities and haptic input for typical tasks such as drawing polygons etc.

The contextual aspect of the task as well as the activity-related aspect of the situation describe which activities the user is performing. The contextual aspect of the tasks covers activities motivated by the task itself (such as driving a car while using a navigation system), whereas the activity-related aspect of the situation lacks this motivation. In both cases one can anticipate the attention resources of the user, the free interaction channels, direction of gaze, which is important for selecting modalities at the articulatory level. If we know the direction of gaze, we can judge whether a visual message at a certain position can be perceived by the user. Therefore, recent car navigation systems display route instructions in a head-up display. Otherwise a more ambient modality (such as speech) has to be used.

The social aspect of the situation covers the question whether there are surrounding people who might be annoyed by acoustic system output or who should not know about the private data exchanged between user and system. In this case ambient modalities such as speech should be avoided.

The articulatory aspect means factors such as lighting and noise level. High noise level discourages the use of speech recognition, because acoustic signals cannot be transmitted reliably. At the same time it encourages the use of finer grained input modalities (syntactic level), such as directed dialogue instead of natural language or even spelling as methods for error recovery and prevention.

Articulatory aspects of the user model cover perceptual and motoric impairments, which determine modality selection at the articulatory level. The lexical aspect, that is user expertise, determines modality selection at the syntactic level – that is the choice of interaction language (direct manipulation, command language etc.) and dialogue strategy (directed dialogue versus natural language).

Table 2 exemplifies how the task model of our framework might motivate modality selection. Fields marked with a plus are indicating that the column category encourages the use of the related row category. Fields marked with a minus on the contrary are indicating dissuasion. It should be noted, that the categories in this matrix are

Table 2. Influence of the task model on modality properties

			Task						
			discourse aspect		semantic aspect			contextual aspect	
			high priority notification	low priority information	analogous graphics	linguistic token creation	linguistic token selection	safety critical visual attention requirements	no attention requirements
Modality	semantic aspect	analogous			+	–			
		digital			–	+	+		
	syntactic aspect	linguistic			–	+	+		
		non-linguistic			+	–	–		
		multi token				–	+		
		single token				–	+		
		spelling				+			
	articulatory aspect	acoustic	+	–	–			+	–
		graphic		+	+			–	+
		haptic			+				
		vocal							
		manual							
		ambient	+	–				+	–
		directed		+				–	+
		touch							

parts of a preliminary set and might be even more diversified. Most of the fields are empty, because it is difficult to decide for each aspect independently without consideration of other factors.

Alternative representations, such as rules or pattern languages [1], [7], [19] (see appendix) might be more convenient for covering the influence of more complex factor combinations.

Design patterns are usually based on well established designs, design experience and good practice. Multimodal interactive systems are relatively new and lack dissemination, so that they apparently do not have a basis for design patterns. Nevertheless, design patterns can serve in this domain not only for communicating design recommendations based on modality theory and context of use modelling but also for documentation of design discussions, such that design consistency and reuse of requirement specifications can be achieved more easily. The following lines point out, how design patterns for multimodal interactions might look like.

Our pattern *non-disruptive notification* (table 3) points out how the discourse aspect as well as the contextual aspect of the task and the activity-related aspect of the situation influence modality selection at the articulatory level.

Table 3. Design Pattern *Non-Disruptive Notification*

Non-Disruptive Notification	
Problem	An event occurs, but the user's attention has to be directed to an important, potentially safety critical activity. The user wants to decide himself when to retrieve new information.
Use when	You are developing an application which involves user notification about events. The application scenario (either the task or the situation) involves safety critical activities or requires high concentration.
Solution	Use output modalities of high spatial selectivity, such as graphics or (for blind users) haptics. The information should be displayed at a consistent place. If there might be a lot of information to be notified about, the user should be given a standardised command for retrieving it. The presence of new information should be indicated at a consistent place on the display.

5 Conclusion and Future Work

In this paper we have extended modality theory by introducing new modality properties at the syntactic level. Furthermore, we have refined modality properties at the articulatory level.

We have introduced the channel properties chromatic resolution, spatial resolution and spatial selectivity. For input modalities, our framework considers beyond the interaction channel the action *source* (vocal, mimic, manual, pedal). Finally we have extended the syntactical level by including aspects of dialogue strategy as interaction granularity.

At the same time, we have introduced models which should be taken into account for modality selection and allocation at several levels. These are task model, situation model, user model and device model. Section 4 points out the mapping of these

models to modality properties. The appendix shows design patterns for multimodal interaction resulting from these mappings.

Our next step will be to verify and refine our theory. Therefore we are modelling a multimodal organizer. We have chosen this scenario because it requires sufficiently complex data input (for instance the input of a series of message recipients) and covers a broad range of contexts of use including mobile computing.

Drawing upon the insights of this we will create a design pattern language [1], [7], [19] (see appendix) as well as model-based notational elements for multimodal interaction.

References

1. Alexander, C., Ishikawa, S., Silverstein, M., Jacobson, M., Fiksdahl-King, I., Angel, S.: A Pattern Language. Oxford University Press (1997)
2. Benoît et al. Audio-visual and Multimodal Speech Systems. Audio-visual and Multimodal Speech Systems. Handbook of Standards and Resources for Spoken Language Systems - Supplement Volume (2000)
3. Bernsen, N. O.: Modality Theory: Supporting Multimodal Interface Design. Proc. ERCIM (1993)
4. Bernsen, N. O.: A toolbox of output modalities. Representing output information in multimodal interfaces. WPCS-95-10. Centre for Cognitive Science, Roskilde University (1995)
5. Bernsen, N. O.: Bernsen, N. O.: Towards a tool for predicting speech functionality. Speech Communication 23, (1997) 181-210
6. Bernsen, N. O.: Multimodality in language and speech systems - from theory to design support tool. In Granström, B., House, D., and Karlsson, I. (Eds.): Multimodality in Language and Speech Systems. Dordrecht: Kluwer Academic Publishers (2002) 93-148
7. Borchers, J. O.: A Pattern Approach to Interaction Design. AI & Society Journal of Human-Centred Systems and Machine Intelligence 15(4): 359–376. Springer-Verlag (December 2001)
8. Bürgy, C.: An Interaction Constraints Model for Mobile and Wearable Computer-Aided Engineering Systems in Industrial Applications. Doctoral Dissertation, University of Pittsburgh, Pennsylvania, USA (2002)
9. Calvary, G., Coutaz, J., Thevenin, D., Limbourg, Q., Bouillon, L., Vanderdonckt, J.: A unifying reference framework for multi-target user interfaces. Interacting with Computers 15(3): 289-308 (2003)
10. Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J., Young, R. M.: Four Easy Pieces for Assessing the Usability of Multimodal Interaction: The CARE Properties. Proc. Interact '95, Chapman & Hall, London, (1995) 115-120.
11. Eco, U.: Einführung in die Semiotik. Fink, München (1994)
12. Martin, J.-C.: Towards “intelligent” cooperation between modalities. The example of a system enabling multimodal interaction with a map. Proc. IJCAI-97 Workshop on Intelligent Multimodal Systems, Nagoya, Japan (1997)
13. Nigay, L., Coutaz, J.: A Design Space For Multimodal Systems: Concurrent Processing and Date Fusion. Proc. INTERCHI’93, ACM Press, NY, USA (1993)
14. Hiyoshi, M. and Shimazu, H.: Drawing pictures with natural language and direct manipulation. Proc. Coling-94, Kyoto, Japan (1994)
15. Rousseau C., Bellik Y. and Vernier F.: Multimodal Output Specification / Simulation Platform. Proc. ICMI 2005, Trento, Italy (2005)

16. Suhm, B., Myers, B., Waibel, A.: Multimodal error correction for speech user interfaces. ACM Trans. Comput.-Hum. Interact. 8, 1 (Mar. 2001), 60-98
17. Tan, Y. K., Sherkat, N., Allen, T.: Error recovery in a blended style eye gaze and speech interface. In Proc. ICMI '03. ACM Press, New York, NY, 196-202 (2003)
18. Vernier, F., Nigay, L.: A Framework for the Combination and Characterization of Output Modalities. DSV-IS (2000) 35-50
19. Van Welie, M., and van der Weer, G. C.: Pattern Languages in Interaction Design: Structure and Organization. Proc. Interact'03 (2003)

Appendix: Multimodal Design Patterns

This appendix shows five design patterns for multimodal interaction which are based on modality theory. In contrast to two-dimensional matrices, design patterns allow multidimensional mappings. In addition design patterns allow the formulation of design recommendations at different levels of abstraction. Our patterns are structured according to [19]. Future work will augment our pattern collection.

Design Pattern 1. Non-Disruptive Notification

Non-Disruptive Notification	
Problem	An event occurs, but the user's attention has to be directed to an important, potentially safety critical activity. The user wants to decide himself when to retrieve new information.
Use when	You are developing an application which involves user notification about events. The application scenario (either the task or the situation) involves safety critical activities or requires high concentration.
Solution	Use output modalities of high spatial selectivity, such as graphics or (for blind users) haptics. The information should be displayed at a consistent place. If there might be a lot of information to be notified about, the user should be given a standardised command for retrieving it. The presence of new information should be indicated at a consistent place on the display.

Design Pattern 2. Alert

Alert	
Problem	An important event occurs, but the user's attention is not directed towards system output.
Use when	You are developing an application which involves user notification about important events. The user should get informed without delay. A disruption of the user in his primary tasks has no severe implications.
Solution	Use ambient modalities in order to notify the user. When the user's direction of gaze or hand position can be anticipated because of the primary task, visual and haptic notification techniques might be appropriate as well.

The design pattern *alert* could be described in more detail. Depending on the context of use, the alert should be performed by an appropriate modality. We could cover these factors by a more detailed description in the field *solution* or, as exemplified on

the following page, by drawing a pattern hierarchy (see also [7]). Therefore, the specialized patterns *visual*, *haptic* and *acoustic alert* contain the additional field *extension of* which relates them to the generalized pattern *alert*.

Design Pattern 3. Visual Alert

Visual Alert	
Problem	An important event occurs, but the user's attention is not directed towards system output.
Use when	You are developing an application which involves user notification about important events. The application scenario (either the task or the situation) allows the prediction of the user's direction of gaze. The use of ambient modalities such as sound is not appropriate because of task or situational aspects.
Solution	Use head-up displaying techniques in order to notify the user visually about the event.
Extension of	Alert

Design Pattern 4. Haptic Alert

Haptic Alert	
Problem	An important event occurs, but the user's attention is not directed towards system output.
Use when	You are developing an application which involves user notification about important events. The application scenario allows to predict, that the user touches a operator device with his hands.
Solution	Use haptic signals such as vibration in order to notify the user haptically about the event.
Extension of	Alert

Design Pattern 5. Acoustic Alert

Acoustic Alert	
Problem	An important event occurs, but the user's attention is not directed towards system output.
Use when	You are developing an application which involves user notification about important events. The application scenario does not allow the prediction of the user's direction of gaze nor the position of the user's hands. The use of ambient modalities such as sound is not discouraged by situational or task-related aspects.
Solution	Use sound or speech output in order to notify the user acoustically about the event.
Extension of	Alert

Visual Interaction in Natural Human-Machine Dialogue

Joseph Machrouh and Franck Panaget

France Telecom R&D, TECH/EASY labs
2, avenue Pierre Marzin - BP 50702
22307, Lannion Cedex, France
{joseph.machrouh, franck.panaget}@francetelecom.com

Abstract. In this article, we describe a visual component able to detect and track a human face in video streaming. This component is integrated into an embodied conversational agent. Depending on the presence or absence of a user in front of the camera and the orientation of his head, the system begins, continues, resumes or closes the interaction. Several constraints have been taken into account: a simple webcam, a low error rate and a minimum computing time that permits the whole system to run on a simple pc.

1 Introduction

Embodied Conversational Agents (ECA) bring unique opportunities for delivering users new kinds of interaction with computer systems. They adopt some properties of human face-to-face communication [Cassell et al., 1999]. Our ECA Nestor [Pelé et al., 2003] comprises a continuous speech recognizer, a text-to-speech synthesizer, an avatar control module, and a dialoguing rational agent [Sadek et al., 1997] having both multimodal fusion and multimodal fission components. Multimodal fusion merges users' inputs from different media (spoken language, mouse click, text, ...). Multimodal fission divides the agent's reactions into outputs on appropriate media (text, speech, images, body movements, ...). One of the applications using Nestor is PlanResto [Pelé et al., 2003], a restaurant web guide. It permits users to look for a restaurant in Paris by specifying location, culinary speciality or price.

The user interacts with the system in natural language (spoken or written), or through mouse clicks. Nestor answers requests in various ways, it may use any appropriate combination of gestures, text, speech or images (maps or photos). Our goal is to improve Nestor's communicating abilities by adding a computer vision component permitting it to recognize (some elements of) non-verbal communication.

Turk [Turk, 2004] highlights several functionalities that a visual module integrated into an interactive system should offer. In this paper, we focus on two of them:

– face detection and location : How many people are in the scene and where are they?

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 152–163, 2006.

– head and face tracking : Where is the user's head, and what is the specific position and orientation of the face

The organization of this paper is as follows. Section 2 provides details on the visual system architecture. In section 3 we present the experimental results. Finally, a brief review of vision contribution in interactions is described in section 4, followed by a conclusion where we expose some prospects for our work.

2 Visual System Architecture

The vision module that are intended to be integrated in a human-computer dialogue process must be fast and efficient. They must be able to track in real time yet not absorb a major share of computational resources: other tasks must be able to run while the vision module is being used. We will develop an architecture centered on a supervising process similar to SERVP [Crowley and Bedrune, 1994]. This architecture comprises several specialized modules (see figure 1): vision supervisor module, eye detection module and face detection module.

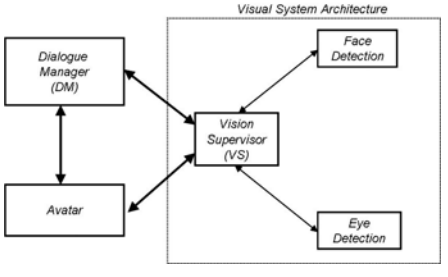


Fig. 1. Visual system architecture

2.1 Supervisor Module

Mainly, the Vision Supervisor (VS) manages various specialized vision modules according to the dialogue context, environment and computational cost.

To ensure a strong knowledge about the environment, VS updates a database containing information about the number of people standing in front of the camera, their position and movements, information about the illumination and a lot of other clues (head orientation and gaze direction, color histograms of clothes and background regions). This information can be considered as a memory which is used to modulate the future processing. For example, if a user leaves the field of the camera or when his face is not detected by the face detection module, and if this user reappears after a few frames, VS considers that is the same person. So the dialogue manager continues the dialogue with him.

VS also considers the dialogue context to coordinate the different modules. Indeed, when the system asks a yes-no question, the supervisor will activate the eye detection module to detect a head shaking, whereas if the system describes the list

of the restaurants, then the supervisor chooses the gesture detection one. VS, in relation with the dialogue manager, defines the task to be processed according to the communication context. It directs the various visual processing modules and merges the information they transmit. It interacts with the other interface components to harmonise their behaviour (for instance, the avatar may follow the user with its eyes). VS also deals with luminosity variations (see below).

2.2 Face Detection Module

To ensure a real time application, it's crucial to process as few operations as possible. It is nevertheless essential to maintain a strong knowledge about people's characteristics. So, face detection must be active on each frame to supply people location. Consequently to reduce computational cost, Face detection uses two different approaches for face detection: the first one is based on a convolutionnal neural network named "Convolutional Face Finder" (CFF) [Garcia and Delakis, 2004] and the other one is based on skin colour segmentation (see figure 2).

Skin colour is not robust enough in light changes and camera noises. When there is a change in luminosity, VS runs the CFF algorithm to reinitialize the feature extraction in order to extract a new value and to permit the skin colour algorithm to track the faces.

When the system uses only the skin colour algorithm, it does not detect the presence of the other users who have just appeared in front of the camera. To resolve this problem, CFF is executed at the request of the dialogue manager, when the user turns his head or when it spends time to answer.

CFF. CFF permits to locate the presence of someone in front of the camera. It uses a neural-based face detection scheme to precisely locate multiple faces of minimum size 20x20 pixels and variable appearance in complex real world images. A good detection rate of 90.3% with 8 false positives have been reported on the CMU test set.

Even if this algorithm offers great accuracy, it is inefficient when heads are rotated more than ±30 degrees in image plane and turned more than ±60 degrees. Moreover, the processing time increases with the number of faces detected in the image. But, in our context of natural human-computer dialogue, it is exactly when a face is detected that the other components of a dialogue system require CPU.

Skin colour regions. Most existing face detection systems use histogram color for segmentation [Hsu et al., 2002]. The skin color model can be used for face localization [Cai and Goshtasby, 1999] [Kovac et al., 2003], tracking [Bradski, 1998] and hand localization [Ahmad, 1995].

The main difference between those systems is the choice of colorimetric space. There are HSV [Hsu et al., 2002], I1I2I3 [Menezes et al., 2003], TSL [Tomaz et al., 2003], YIQ [He et al., 2003] and the most used one is YCrCb [Chen et al., 2003] [Foresti et al., 2003].

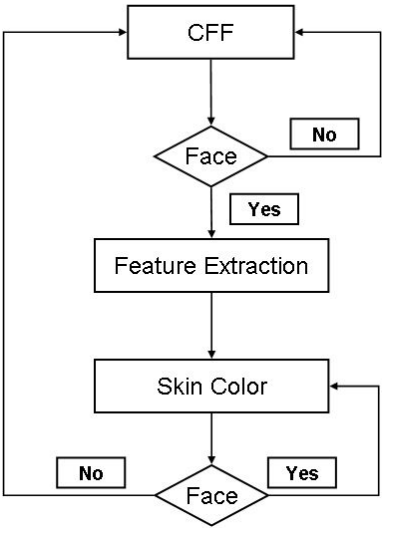


Fig. 2. Architecture of face detection module

In [Chai and Ngan, 1999], the authors determined Cb and Cr range to detect a skin color ($R_{Cb} = [77, 127]$, $R_{Cr} = [133, 173]$). In addition to its low computational cost, this method permits to track a face whatever its orientation but it alos detects skin colour regions such a face, hand, arm. Moreover, it does not result in robust systems.

CFF/Skin colour regions. In our system we use the two alogrithms. CFF is used to detect a face in frontal position and to allow the skin colour algorithm to locate that face in the following images. This method makes it possible to detect a face whatever its orientation in a minimum time interval in spite of luminosity variations. We choose YCrCb colorimetric space [Machrouh et al., 2006] to perform skin colour segmentation.

The result of CFF is a rectangle around the face. Using a simple histogram of this area, which enables us to extract all shades of the detected face's colour, is not efficient enough for two reasons. Firstly, this rectangle contains eyes, eyebrows, hair, and glasses. Secondly, there might exist shade variations due to the camera's noise. We thus propose, on the one hand, to select a sub-area of the rectangle, where the probability of having skin coloured pixels is higher, and on the other hand, to represent skin colour distribution by a two-dimensional Gaussian law.

To avoid possible noise from non-skin coloured pixels, we use a priori knowledge of a human head's proportion and form to determine the sub-area E to pick up skin colour samples (see figure 3a).

For the initialization of our model, we choose to represent skin colour distribution by a two-dimensional Gaussian law of parameters the mean μ and the covariance matrix Σ of all the normalized pixel components c in the area E.

Fig. 3. (a) Skin colour is initialized in the area E, (b) face detection with CFF+skin colour

$$p(c/skin) = \frac{1}{2\pi.\left|\sum\right|^{\frac{1}{2}}}.e^{-\frac{1}{2}(c-\mu)^T \sum^{-1}(c-\mu)} \tag{1}$$

$$\mu = \frac{1}{M}.\sum_{i=1}^{M} c_i \quad and \quad \sum = \begin{bmatrix} \sigma_{CrCr} & \sigma_{CrCb} \\ \sigma_{CrCb} & \sigma_{CbCb} \end{bmatrix} \tag{2}$$

Where M is the number of pixels and $c_i = \begin{pmatrix} Cr_i \\ Cb_i \end{pmatrix}$ is the colour vector of pixel i (Cr_i and Cb_i represent the Cr and Cb components of pixel i in YCrCb format) and $\sigma_{xx} = \frac{\sum_{i=1}^{M} x_i^2}{M} - \mu_x^2$ and $\sigma_{xy} = \frac{\sum_{i=1}^{M} x_i.y_i}{M} - \mu_x\mu_y$

With this method, only one scan of the area is necessary.

For the face tracking, we consider a pixel to be skin colored if:

$$p(c/skin) = \frac{1}{2\pi.\left|\sum\right|^{\frac{1}{2}}}.e^{-\frac{1}{2}(c-\mu)^T \sum^{-1}(c-\mu)} \geq \lambda \tag{3}$$

which in effect performs Mahalanobis' distance minimization.

Once the skin colour filtering is performed, we determine clusters of connected pixels (connected component analysis). Clusters and holes of area less than 0.5% of the frame area are respectively discarded and filled so that only a small number of clusters are considered for further analysis.

We then extract the characteristics from the clusters (surface, perimeter, compactness, average, variance...) to detect faces (see figure 3b).

2.3 Eye Detection

In human-machine interaction, eye detection is the first step toward evaluation of head orientation and gaze direction [Feng and Yuen, 2001] [Kumar et al., 2002]. Our aim in eye detection is to recognize some communication gestures such as head nods and orientation.

Our approach is as follows: locating the face using the face detection module, estimating the rough position of the eyes and improving eye localisation using the eye detection module, which operates a processing sequence based on eye region colorimetric specificities. In YCrCb space, the chrominance (Cr, Cb) and luminance (Y) information can be exploited to extract eye region. According to our experiment in many face databases, the area around the eyes has specific colorimetric values. Cb values are higher than Cr ones [Hsu et al., 2002]. Concerning Y values, this area contains both high and low values.

The goal of the process is to accentuate the brighter and darker pixels of the eyes, initially through the chrominance (Cr Cb) and through the luminance (Y) as shown in figure 4.

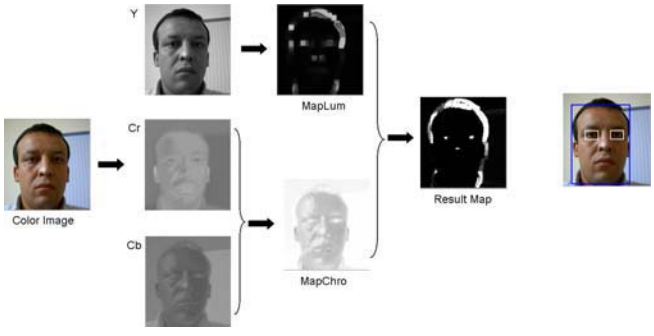


Fig. 4. EyeMap construction procedure

First, we will try to emphasize eye brightness through chrominance. We note that around eye region we have $Cb > Cr$. This implies $Cr - Cb < 0$, so neg(Cr-Cb) will have saturated values (255) around the eyes.

$$MapChro = neg(C_r - C_b) \tag{4}$$

where $neg(x)$ is the negative of x (i.e. 255-x). And in a second time, we process through luminance Y: we dilate to propagate the high values and erode for the low ones. The division result will have high values around the eye region.

$$MapLum = \frac{Dil(Y)}{Ero(Y)} \tag{5}$$

The result map, obtained by the AND operation of the two resulting maps MapChro and MapLum, shows isolated clusters at eyes location. A simple connected component analysis based on pixel connectivity (already performed in Face Detection) is sufficient to determine clusters (or components). Then, we consider the head position, inter-reticular distance and eyes characteristics (compactness, shape) in order to choose among the different clusters to identify eyes.

3 Results

We have evaluated our system in video streaming and two databases image series:

– the video streaming consists of 15 video files representing television news, each one containing 5700 frames.

– the first database, the head pose database [Gourier et al., 2004], consists of 15 sets of images. Each set contains two series of 93 images of the same person at different poses. There are 15 people in the database, wearing glasses or not and having various skin colours. The pose or head orientation is determined by 2 angles (h,v), which vary from -90 degrees to +90 degrees.
– The second database is a set of images collected on the World Wide Web (4868 images), called www database. These colour images have been taken under varying lighting conditions and with complex backgrounds. Furthermore, these images contain multiple faces with variations in colour, position, scale, orientation, 3D pose and facial expression. This database was sorted according to face pixel size into 5 subsets.

3.1 Face Detection Results

In the first test, we compared the results of CFF only with the results of our CFF/skin colour algorithm. In order to evaluate the improvement of the tracking error rate, we compare the two algorithms on the head pose database. The left part of figure 5 shows that, as mentionned in section 2.2, CFF alone cannot detect faces in all positions whereas the CFF/skin colour permits to detect the faces in many positions (figure 5 right).

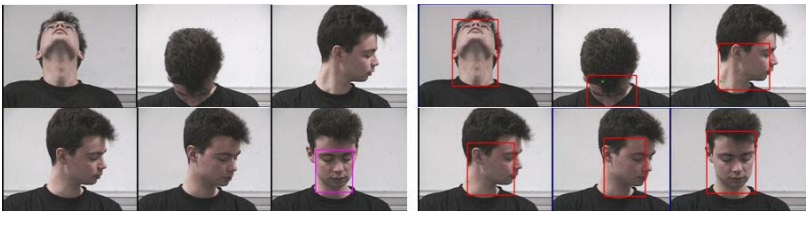


Fig. 5. Left: the six images shows the face detection by CFF. Right: CFF/Skin colour can detect face in multiple positions.

More precisely on this database, CFF detected on average 48% of the faces. On the other hand, 98% of the faces were correctly detected when using both algorithms (see figure 6).

Our face detection algorithm cannot detect a user when his face is not in frontal position in the first image (since it is CFF that initialises the face detection). But, in a large part of human-machine dialogue except in vehicle environment, the user usually has his face in front of the camera when he begins a dialogue with the system.

The second experiment consists in computing the processing time of both algorithms in a video streaming. Figure 7 shows the detection result in video sequence.

Another result is the evolution of the CPU consumption. Figure 8 shows the time processes of CFF alone and CFF + Skin colour algorithm in a image video

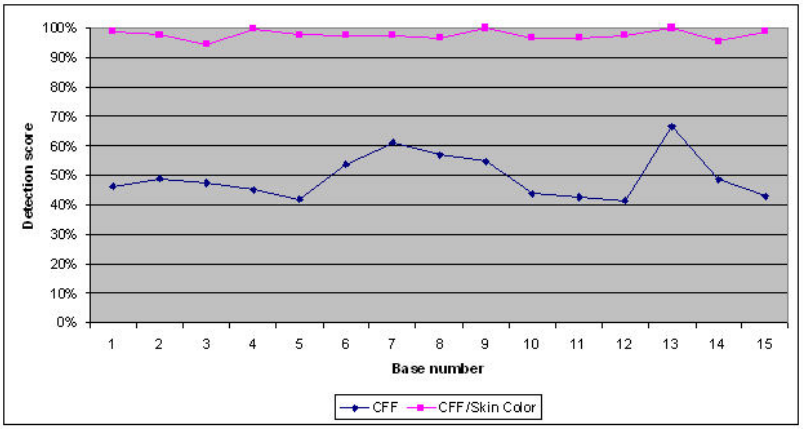


Fig. 6. Face detection score rate

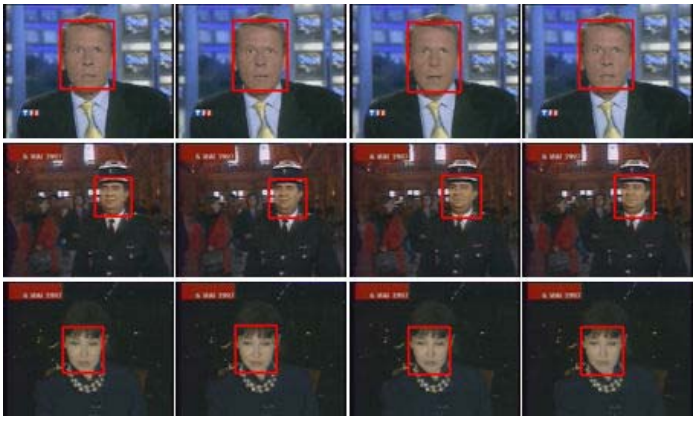


Fig. 7. Face detection in a video streaming

sequence. When no face is present in the image, CFF needs 100 ms to respond (The skin colour algorithm is not running). When a face is present, CFF needs 150 ms to detect and track it whereas the skin colour algorithm only need 10 ms. This corresponds to our expectations since it is precisely when a person is present that the other system components must run.

3.2 Eye Detection Results

The second test consists in applying the face detection module and the eye detection module in the www database. This test is applying on a database contains different sizes images of women and men. Table 1 shows the total detection rate on both image databases. We can see that the rate detection is better when the pixel size face is larger.

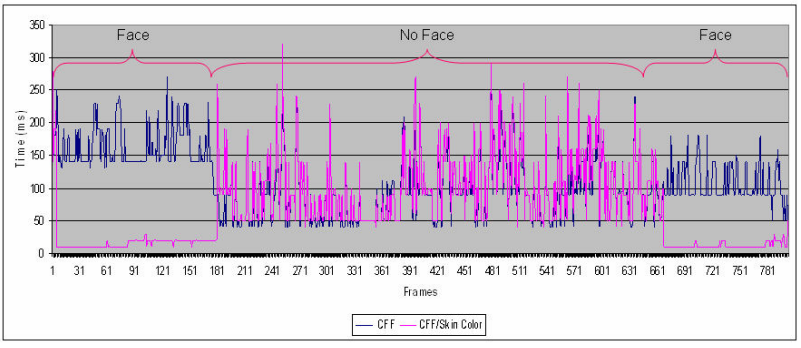


Fig. 8. CPU consumption

Table 1. Eye detection results on www database. DR: Detection Rate FP: False Positives).

Face size	39×38	80×91	147×166	179×204	205×250
DR	84.12%	87.04%	93.55%	93.75%	94.87%
FP	3.44%	4.01%	3.54%	3.51%	3.94%

4 Contribution of Vision in Interaction

In combining visual information with data obtained from speech recognition component, human-machine interaction is significantly improved. Currently, visual details permit Nestor to build a coherent dialogue by taking into

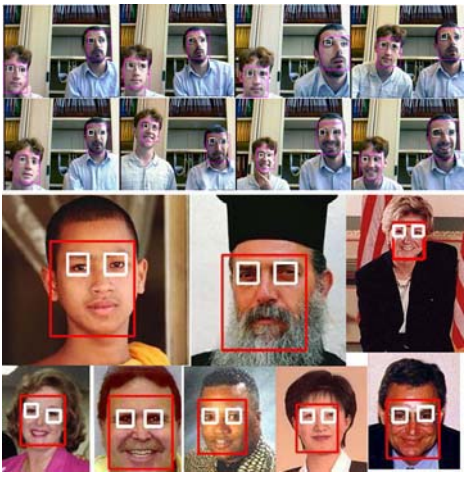


Fig. 9. Face detection and Eye tracking result on a subset of the images database

Fig. 10. Application of our system in an Embodied Conversational Agent

consideration the visual context. For instance, Nestor engages the dialogue by greeting a user who just appeared in front of the camera.

Nestor : *"Hello. My name is Nestor. I can help you to find a restaurant in Paris."*

User : *"I want a restaurant near Montparnasse train station"*

Nestor : *"There are more than 100 restaurants near Montparnasse train station. You can choose a culinary speciality, for example a French or an Italian restaurant"*

if the user is still looking at the screen, the system proposes other culinary specialities

Nestor : *"There is also a Chinese or Japanese restaurant"*

if the user is looking somewhere else, the system suspends the dialogue until the user is looking at the screen.

Another example of visual context improving interaction is that considering head orientation may change how the dialogue is carried out. For instance, if the user looks at someone else when speaking, the system might deduce that those words are not directed to it. If the user turns his head and goes away, the system will detect it and might then say:

Nestor :*"I can see that you are leaving, I hope that you are satisfied with the information I gave you. Bye"*

User's head movements can also inform Nestor when signs of approval or disapproval are made. The dialog will go on without the spoken response of the user.

Nestor :*"Would you like more information about this restaurant or would you like..."*

If the user nods, Nestor will react like this:

Nestor :*"This restaurant is located..."*

5 Conclusion and Future Work

In this article, we described the architecture of the visual system integrated in our embodied conversational agent Nestor. Considering the visual context

significantly improves the interaction. Despite the fact that image processing requires many resources, the improvement justifies the computational time surplus.

Nevertheless, we have seen that we can save some resources without hurting performance, in managing the processes with the Vision Supervisor. Our system runs in real-time on a basic personal computer (all tests were performed on a Pentium 4 2.2 Ghz).

This architecture has been tested by several people in front of the camera, and also on image databases. First results are satisfying. But, vision supervisor (including specialized vision modules) has to be tested on video databases.

In the future, this system will be able to detect the user's gaze direction. The current system can detect gestures but cannot yet recognize them. A learning phase for everyday communication gestures will start soon, following previous work in our laboratory [Marcel and Bernier, 1999].

References

[Ahmad, 1995] Ahmad, S. (1995). A usable real-time 3d hand tracker. In *proceeding of the 28th Asilomar Conference on Signals, Systems and Computers*, pages 1257–1261, Pacific Grove, CA, USA.

[Bradski, 1998] Bradski, G. R. (1998). Computer vision face tracking for use in a perceptual user interface. In *proceeding of IEEE Workshop on Applications of Computer Vision*, pages 214–219, Princeton, NJ, USA.

[Cai and Goshtasby, 1999] Cai, J. and Goshtasby, A. (1999). Detecting human faces in color images. *Image Vision Computing*, 18:63–75.

[Cassell et al., 1999] Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H., and Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *CHI'99: Proceedings of the SIGGHI Conference on Human Factors in Computing Systems*, pages 520–527, Pittsburgh, Pennsylvania, USA.

[Chai and Ngan, 1999] Chai, D. and Ngan, K. (1999). Face segmentation using skin-color map in videophone applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(4):551–564.

[Chen et al., 2003] Chen, M., Chi, M., Hsu, C., and Chen, J. (2003). Roi video coding based on h.263+ with robust skin-color detection technique. *IEEE Transactions on Consumer Electronics*, 49(3):724–730.

[Crowley and Bedrune, 1994] Crowley, J. L. and Bedrune, J. M. (1994). Integration and control of reactive visual process. In *Proceeding of the 3rd European Conference on Computer Vision (ECCV 94)*, Stockholm, sweden.

[Feng and Yuen, 2001] Feng, G. C. and Yuen, P. (2001). Multi-cues eye detection on gray intensity image. *Pattern Recogntion*, 34(5):1033–1046.

[Foresti et al., 2003] Foresti, G., Micheloni, C., Snidaro, L., and Marchiol, C. (2003). Face detection for visual surveillance. In *Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP03)*, Mantova, Italy.

[Garcia and Delakis, 2004] Garcia, C. and Delakis, M. (2004). Convolution face finder: A neural architecture for fast and robust face detection. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 26(11):1408–1423.

[Gourier et al., 2004] Gourier, N., Hall, D., and Crowley, J. L. (2004). Estimating face orientation from robust detection of salient facial features. In *Proceeding of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures*, Cambridge, UK.

[He et al., 2003] He, X., Liu, Z., and Zhou, J. (2003). Real-time human face detection in color image. In *Proceedings of the Second lnternational Conference on Machine Learning and Cybernetics*, Xi'an, China.

[Hsu et al., 2002] Hsu, R. L., Abdel-Mottaleb, M., and Jain, A. K. (2002). Face detection in color images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(5):696–706.

[Kovac et al., 2003] Kovac, J., Peer, P., and Solina, F. (2003). Human skin colour clustering for face detection. In Zajc, B., editor, *EUROCON 2003 - International Conference on Computer as a Tool*, Ljubljana, Slovenia.

[Kumar et al., 2002] Kumar, R. T., Raja, S. K., and Ramakrishnan, A. G. (2002). Eye detection using color cues and projection functions. In *Proceeding of International Conference on Image Processing*, volume 3, pages 337–340, Rochester, NY, USA.

[Machrouh et al., 2006] Machrouh, J., Panaget, F., Bretier, P., and Garcia, C. (2006). Face and eyes detection to improve natural human-computer dialogue. In *Proceeding of the second IEEE-EURASIP International Symposium on Control, Communications, and Signal Processing*, Marrakech, Morocco.

[Marcel and Bernier, 1999] Marcel, S. and Bernier, O. (1999). Hand posture recognition in a body-face centred space. In *Gesture-Based Communication in Human-Computer Interaction: International Gesture Workshop GW'99.*, volume 1739, pages 97–100. Lecture Notes in Computer Science.

[Menezes et al., 2003] Menezes, P., Brethes, L., Lerasle, F., Dans, P., and Dias, J. (2003). Visual tracking of silhouettes for human-robot interaction. In *Proceeding of International Conference on Advanced Robotics (ICAR01)*, volume 2, pages 971–976, Coimbra, Portugal.

[Pelé et al., 2003] Pelé, D., Breton, G., Panaget, F., and Loyson, S. (2003). Let's find a restaurant with nestor: A 3d embodied conversational agent on the web. In *Proceeding of AAMAS Workshop on embodied conversational characters as individual*, Australia.

[Sadek et al., 1997] Sadek, D., Bretier, P., and Panaget, F. (1997). Artimis: Natural dialogue meets rational agency. In *Proceeding of the 15th International Joint Conference on Artificial Intelligence (IJCAI'97)*, pages 1030–1035, Nagoya, Japan.

[Tomaz et al., 2003] Tomaz, F., Candeias, T., and Shahbazkia, H. (2003). Improved automatic skin detection in color images. In Sun, C., Talbot, H., Ourselin, S., and Adriaansen, T., editors, *Proceeding of VIIth Digital Computing: Techniques and Applications*, pages 419–427, Sydney, Australia.

[Turk, 2004] Turk, M. (2004). Computer vision in the interface. *Communications of the ACM*, 47(1):60–67.

Multimodal Sensing, Interpretation and Copying of Movements by a Virtual Agent

Elisabetta Bevacqua[2], Amaryllis Raouzaiou[1], Christopher Peters[2], George Caridakis[1], Kostas Karpouzis[1], Catherine Pelachaud[2], and Maurizio Mancini[2]

[1] Image, Video and Multimedia Systems Laboratory, National Technical University of Athens, Greece
http://www.image.ntua.gr
[2] LINC, IUT de Montreuil, Université de Paris 8
http://www.univ-paris8.fr

Abstract. We present a scenario whereby an agent senses, interprets and copies a range of facial and gesture expression from a person in the real-world. Input is obtained via a video camera and processed initially using computer vision techniques. It is then processed further in a framework for agent perception, planning and behaviour generation in order to perceive, interpret and copy a number of gestures and facial expressions corresponding to those made by the human. By *perceive*, we mean that the copied behaviour may not be an exact duplicate of the behaviour made by the human and sensed by the agent, but may rather be based on some level of interpretation of the behaviour. Thus, the copied behaviour may be altered and need not share all of the characteristics of the original made by the human.

1 Introduction

The ability for an agent to provide feedback to a user is an important means for signalling to the world that they are animate, engaged and interested. Feedback influences the plausibility of an agent's behaviour with respect to a human viewer and enhances the communicative experience.

In this paper, we present a scenario whereby an agent senses, interprets and copies a range of facial and gesture expression from a person in the real-world. Input is obtained via a video camera and processed initially using computer vision techniques. It is then processed further in a framework for agent perception, planning and behaviour generation in order to perceive, interpret and copy a number of gestures and facial expressions corresponding to those made by the human. By *perceive*, we mean that the copied behaviour may not be an exact duplicate of the behaviour made by the human and sensed by the agent, but may rather be based on some level of interpretation of the behaviour [1]. Thus, the copied behaviour may be altered and need not share all of the characteristics of the original made by the human.

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 164–174, 2006.

Of particular interest is that a subset of the framework has already been used in conjunction with synthetic vision to implement conversation initiation behaviours between agents in a virtual environment based on their goals and perception of the attention shown by others in them through gaze [2]. In this paper, we also describe how the same framework may be used with real world input. This is an important feature of our work, as one of our long term objectives is to endeavor towards an agent framework that allows agents to interact in a seamless manner with both real and virtual environments in order to further investigate the inherent interactional differences between such environments.

2 State of the Art

There is a long history of interest in the problem of recognising emotion from facial expressions, and extensive studies on face perception during the last twenty years [3]. Ekman and Friesen elaborated a scheme to annotate facial expressions named Facial Action Coding System (FACS) [4] to manually describe facial expressions, using still images of, usually extreme, facial expressions. In the nineties, automatic facial expression analysis research gained much interest mainly thanks to progress in the related fields such as image processing (face detection, tracking and recognition) and the increasing availability of relatively cheap computational power. Head pose and especially facial expression having a very important role in the Human Computer Interaction (HCI), many researchers tackle the problem of facial expression analysis [5], [6] and head movements [7], [8]. Regarding feature-based techniques, Donato et al [9] tested different features for recognizing facial Action Units (AUs) and inferring the facial expression in the frame. Analysis of the emotional expression of a human face requires a number of pre-processing steps which attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, such as characteristic points in these regions, or model facial gestures using anatomic information about the face.

The detection and interpretation of hand gestures has become an important part of HCI in recent years [10]. The HCI interpretation of gestures requires that dynamic and/or static configurations of the human hand, arm, and even other parts of the human body, be measurable by the machine. First attempts to address this problem resulted in mechanical devices that directly measure hand and/or arm joint angles and spatial position. The so-called *glove-based* devices best represent this type of approach. Analysing hand gestures is a comprehensive task involving motion modeling, motion analysis, pattern recognition, machine learning, and even psycholinguistic studies. The first phase of the recognition task is choosing a model of the gesture. Among the important problems involved in the analysis are those of hand localization, hand tracking, and selection of suitable image features. The computation of model parameters is followed by gesture recognition.Hand localization is locating hand regions in image

sequences. Skin color offers an effective and efficient way to fulfill this goal. An interesting approach of gesture analysis research [11] treats a hand gesture as a two- or three dimensional signal that is communicated via hand movement from the part of the user; as a result, the whole analysis process merely tries to locate and track that movement, so as to recreate it on an avatar or translate it to specific, predefined input interface, e.g. raising hands to draw attention or indicate presence in a virtual classroom. There are many systems available for synthesising the animation of a virtual agent. Badler's research group developed EMOTE (Expressive MOTion Engine [12]), a parameterized model that procedurally modifies the affective quality of 3D character's gestures and postures motion. From EMOTE the same research group derived FacEMOTE [13], a method for facial animation synthesis that altered pre-computed expressions by setting a small set of high level parameters taken from Laban Parameters. Wachsmuth's group [14] described a virtual agent capable of imitating natural gestures performed by a human using captured data. Imitation is conducted on two levels: when mimicking, the agent extracts and reproduces the essential form features of the stroke which is the most important gesture phase; the second level is a meaning-based imitation level that extracts the semantic content of gestures in order to re-express them with different movements.

3 General Framework

The present work takes place in the context of our general framework (Figure 1) that is adaptable to a wide range of scenarios. The framework consists of a number of interconnected modules. At the input stage, data may be obtained from either the real world, through visual sensors, or from a virtual environment through a synthetic vision sensor.

Visual input is processed by computer vision [15] (see Section 4.1) or synthetic vision techniques [2], as appropriate, and stored in a short-term sensory storage. This acts as a temporary buffer and contains a large amount of raw data for short periods of time. Elaboration of this data involves symbolic and semantic processing, high-level representation and long-term planning processes. Moreover, it implies an interpretation of the viewed expression (e.g. FAPs → anger), which may be modulated by the agent (e.g. display an angrier expression) and generated in a way that is unique to the agent (anger → another set of FAPs). The generation module [16, 17] synthesises the final desired agent behaviours (Section 4.2). In this paper we present a system of an ECA able to perceive facial and gesture expressions performed by a real user. In order to demonstrate such a capability we present, in Section 5, a simple scenario where the ECA perceives and reproduces the user's movements by using a generation module. That is, in our system, the resulting animation is not a pure copy of the perceived data. In the future, we aim to use this capability to implement a more complex decisional model: by *decisional model*, we refer to a model capable of deciding which movement the ECA will perform, in accordance with the current user's behaviour.

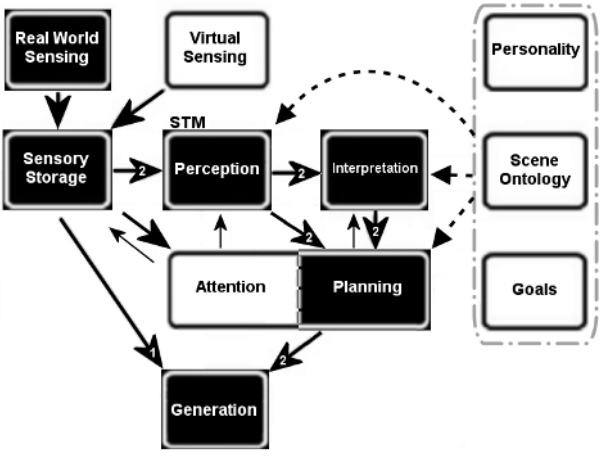


Fig. 1. The general framework that embeds the current scenario. Large arrows indicate the direction of information flow, small arrows denote control signals, while arrows with dashed lines denote information availability from modules associated with long term memory. Modules with a white background are not applicable to the scenario described in this paper.

4 Description of Expressivity Model

The expressivity of behaviors, that is the way behaviors are executed, is an integral part of the communication process. Several researchers ([18], [19], [20], [21], [22]) have investigated human motion characteristics and encoded them into dimensions. In particular Wallbott and Sherer have conducted perceptual studies that show that human beings are able to perceive and recognise a set of these dimensions [19].

Some authors refer to body motion using dual categories such as slow/fast, small/expansive, weak/energetic, small/large, unpleasant/pleasant. To model expressivity, in our framework we use 6 parameters, derived from perceptual studies conducted by [19], each represented by dual category:

- *Overall Activation*: quantity of movements in a timespan
- *Spatial Extent*: amplitude of movements (e.g., amount of space taken up by body or of emotion arousal)
- *Temporal Extent*: duration of movements (e.g., quick vs sustained actions)
- *Fluidity*: smoothness and continuity of overall movement (e.g., smooth vs jerky)
- *Power*: dynamic properties of the movement (e.g., weak vs strong)
- *Repetition*: tendency to rhythmic repeats of specific movement.

Evaluation studies conducted on our model show that spatial and temporal dimensions are easily recognised, whereas fluidity and power are more difficult to interpret. Repetition of a gesture has been often mistaken as being a single complex gesture rather than the repetition of a simple gesture [23].

4.1 Analysis

Facial analysis includes a number of processing steps which attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, such as characteristic points in these regions, or model facial gestures using anatomic information about the face. Although FAPs [24] provide all the necessary elements for MPEG-4 compatible animation, we cannot use them for the analysis of expressions from video scenes, due to the absence of a clear quantitative definition framework. In order to measure FAPs in real image sequences, we have to define a mapping between them and the movement of specific feature points (FPs), which correspond to salient points on the human face [25]. The proposed facial analysis subsystem can detect facial expressions in good lighting conditions. Additionally, the face should not be in an angle omitting characteristic facial features like eye or lip corner. The proposed facial feature extraction scheme is based on a hierarchical, robust scheme, where soft a priori assumptions are made on the pose of the face or the general location of the features in it. Gradual revelation of information concerning the face is supported under the scope of optimisation in each step of the hierarchical scheme, producing a posteriori knowledge about it and leading to a step-by-step visualisation of the features in search. Face detection is performed first through detection of skin segments or blobs, merging them based on the probability of their belonging to a facial area, and identification of the most salient skin color blob or segment. Primary facial features, such as eyes, mouth and nose, are treated as major discontinuities on the segmented, arbitrarily rotated face. Following face detection, morphological operations, erosions and dilations, taking into account symmetries, are used to define the most probable blobs within the facial area to include the eyes and the mouth. Searching through gradient filters over the eyes and between the eyes and mouth provide estimates of the eyebrow and nose positions. Based on the detected facial feature positions, feature points are computed and evaluated.

The next step of the system is the tracking of head and hand. The input image sequences of the gesture analysis subsystem are real videos captured at an acted session. The gestures that our subsystem can detect should be distinguishable in the 2-D frame we have at our disposal, e.g. a hand moving towards the camera-in the vertical plane cannot be detected. Several approaches have been reviewed for a gesture analysis module. The major factors taken under consideration are computational cost and robustness, resulting in an accurate near real-time skin detection and tracking module. The general process involves the creation of moving skin masks, namely skin color areas that are tracked between subsequent frames. By tracking the centroid of those skin masks, an estimate of the user's movements is produced. A priori knowledge concerning the human body and the circumstances when filming the gestures was incorporated into the module indicating the different body parts (head, right hand, left hand). For each frame (Figure 2, top left) a skin color probability matrix is computed by calculating the joint probability of the Cr/Cb image values (Figure 2, top center). The skin color mask is then obtained from the skin probability matrix using

thresholding (Figure 2, top right). Possible moving areas are found by thresholding the pixels difference between the current frame and the next, resulting in the possible-motion mask (Figure 2, bottom center). This mask does not contain information about the direction or the magnitude of the movement, but is only indicative of the motion and is used to accelerate the algorithm by concentrating tracking only in moving image areas. Both color and motion masks contain a large number of small objects due to the presence of noise and objects with color similar to the skin. To overcome this, morphological filtering is employed on both masks to remove small objects. All described morphological operations are carried out with a disk-structuring element with a radius of 1% of the image width. The distance transform of the color mask is first calculated (Figure 2, bottom right) and only objects above the desired size are retained. These objects are used as markers for the morphological reconstruction of the initial color mask. The color mask is then closed to provide better centroid calculation. For the next frame, a new moving skin mask is created, and a one-to-one object correspondence is performed. Object correspondence between two frames is performed on the color mask and is based on object centroid distance for objects of similar (at least 50%) area. In the case of hand object merging and splitting, e.g., in the case of clapping, we establish a new matching of the left-most candidate object to the user's right hand and the right-most object to the left hand. The tracking algorithm is responsible for classifying the skin regions in the image sequence of the examined gesture based on the skin regions extracted from the described method.

As far as expressivity dimensions are concerned, they have been designed for communicative behaviours only [17], [26]. Each dimension acts differently for

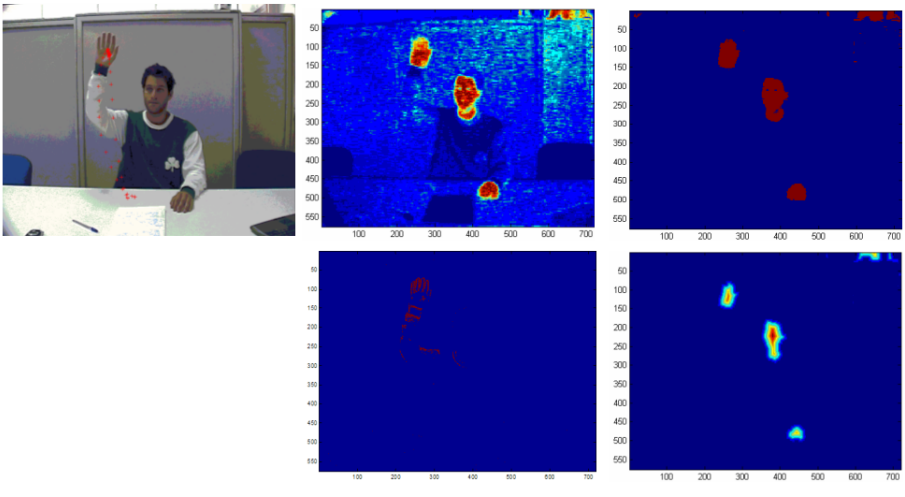


Fig. 2. Top left: example of video frame. Top center: Cr/Cb image. Top right: skin color mask. Bottom center: possible-motion mask. Bottom right: distance transform of the color mask.

each modality. For an arm gesture, expressivity works at the level of the phases of the gesture: for example the preparation phase, the stroke, the hold as well as on the way two gestures are co-articulated [27, 20]. We consider the six dimensions of expressivity as defined in Section 4. Overall activation is considered as the quantity of movement during a conversational turn. In our case it is computed as the sum of the motion vectors' norm (as shown in formula 1). Spatial extent is modeled by expanding or condensing the entire space in front of the human that is used for gesturing and is calculated as the maximum Euclidean distance of the position of the two hands (see formula 2).

$$OA = \sum_{i=0}^{n} \mid \boldsymbol{r}(i) \mid + \mid \boldsymbol{l}(i) \mid . \tag{1}$$

$$SE = \max(\mid d(r(i) - l(i)) \mid). \tag{2}$$

The average spatial extent is also calculated for normalisation reasons. The temporal parameter of the gesture determines the speed of the arm movement of a gesture's meaning carrying stroke phase and also signifies the duration of movements (e.g., quick versus sustained actions). Fluidity differentiates smooth/graceful from sudden/jerky ones. This concept seeks to capture the continuity between movements, as such, it seems appropriate to modify the continuity of the arms' trajectory paths as well as the acceleration and deceleration of the limbs. To extract this feature from the input image sequences we calculate the sum of the variance of the norms of the motion vectors. Finally, the power is identical to the first derivative of the motion vectors calculated in the first steps.

4.2 Synthesis

Table 1 shows the effect that each expressivity parameter has on the production of head movements, facial espressions and gestures. The *Spatial Extent* (SPC) parameter modulates the amplitude of the movement of arms, wrists (involved in the animation of a gesture), head and eyebrows (involved in the animation of a facial expression); it influences how wide or narrow their displacement will be. For example let us consider the eyebrows raising in the expression of surprise: if the value of the *Spatial Extent* parameter is very high the final position of the eyebrows will be very high (i.e. the eyebrows move under a strong of muscular contraction). The *Temporal Extent* (TMP) parameter shortens or lengthens the motion of the preparation and retraction phases of the gesture as well as the onset and offset duration for facial expression. It speeds up or slows down the rising/lowering of the eyebrows. The animation of the agent is generated by defining *key frames* and computing the interpolation curves passing through these frames using TCB-Splines. The *Fluidity* (FLT) and *Power* (PWR) parameters act on the interpolation curves. *Fluidity* increases/reduces the continuity of the curves allowing the system to generate more of less smooth animations. Let us consider its effect on the head: if the value of the *Fluidity* parameter is very low the resulting curve of the head movement will appear as generated through linear interpolation resulting as jerky movement. *Power* introduces a

gesture/expression overshooting, that is a little lapse of time in which the body part involved by the gesture reaches a point in space further than the final one. For example the frown displayed in the expression of anger will be stronger for a short period of time, and then the eyebrows will backtrack to reach the final position. The last parameter, *Repetition* (REP) increases the number of stroke of gestures to obtain repetition of the gestures themselves in the final animation. Let us consider the gesture "wrists going up and down in front of the body with open hands and palms up", a high value of the *Repetition* parameter will increase the number of the up and down movements. On the other hand this parameter decreases the time period of head nods and head shakes to obtain more nods and shakes in the same lapse of time.

The synthesis module is MPEG-4 compatible. Facial expressions are described by FAPs value. Gestures are computed through the interpolation of a sequence of static positions defined by shoulder and arm rotation (arm position), hand shape (chosen in a set of predefined shapes) and palm orientation [28].

Table 1. Effects of Expressivity parameters over head, facial expression and gesture

	HEAD	FACIAL EXPRESSION	GESTURE
SPC	wider/narrower movement	increased/decreased muscular contraction	wider/narrower movement
TMP	shorter/longer movement speed	shorter/longer onset and offset	shorter/longer speed of preparation and retraction phases
FLT	increases/reduces continuity of head movement	increases/reduces continuity of muscular contraction	increases/reduces continuity between consecutive gestures
PWR	higher/shorter head overshooting	higher/shorter muscular contraction overshooting	more/less stroke acceleration
REP	more/less number of nods and shakes	not implemented yet	more/less number of repetitions of the same stroke

5 Application Scenario and Implementation

In section 3 we described the general framework of the system we are aiming at that is able to analyse a real scene and generate the animation of a virtual agent. In this section we present a scenario that is a partial implementation of this framework. Currently our system is able to extract data from the real world, process it and generate the animation of a virtual agent. Either synthesized gesture or facial expression are modulated by the intrinsic expressivity parameters extracted from the actor's performance. Figure 3 is a sub-set of the architecture shown in Figure 1. It describes the architecture required by our current application scenario. The modules of this diagram correspond to the modules in black

in the original architecture. The input coming from the real world is a predefined action performed by an actor. The action consists of a gesture accompanied by a facial expression. Both, the description of the gesture and of the facial expression are explicitly requested to the actor and previously described to him in natural language (for example the actor is asked "to wave his right hand in front of the camera while showing a happy face"). The *Perception* module analyses the resulting video extracting the expressivity parameters of the gesture (see Section 4) and the displacement of facial parts that is used to derive the FAPs values corresponding to the expression performed. The FAPs values and the Expressivity parameters are sent to the *Interpretation* module. If the facial expression corresponds to one of the prototypical facial expression of emotions, this module is able to derive its symbolic name (emotion label) from the FAPs values received in input; if not the FAPs values are used. Instead, the symbolic name of the gesture is sent manually because the *Interpretation* module is not able to extract the gesture shape from the data yet. Finally, how the gesture and the facial expression will be displayed by the virtual agent is decided by the *Planning* module that could compute a modulation either of the expressivity parameters or of the emotion. Then the animation, consisting of variation of FAPs and BAPs values during time, is calculated through the Face and the

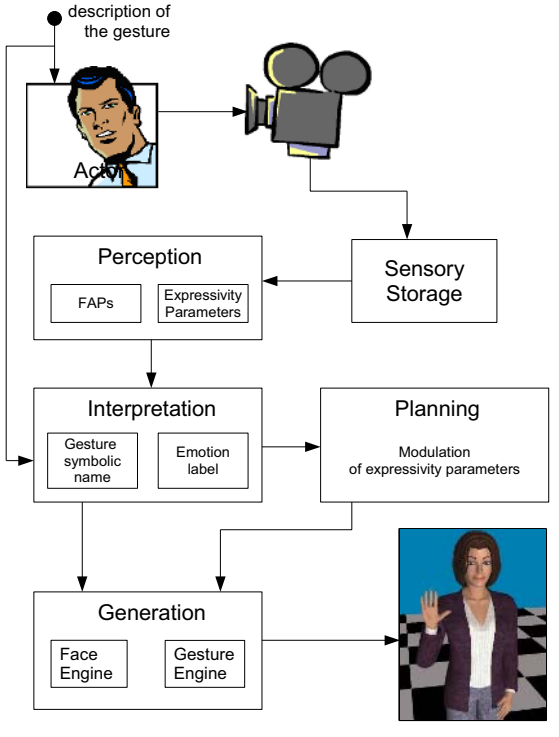


Fig. 3. Diagram of the proposed implementation

Gesture Engine and displayed by the virtual agent. The system does not work in real-time yet, but we aim to develop real-time capabilities in the near future. We also intend to evaluate our system through perceptual tests in order to estimate the correctness of movements.

6 Conclusions

We have presented our general framework consisting of a number of interconnected modules and a sample scenario whereby an agent senses, interprets and copies a range of facial and gesture expression from a person in the real-world. The animation of the agent is based on different types of data: raw parameters values, emotion labels, expressivity parameters, and symbolic gestures specification. To do so the system is able to perceive and interpret gestural and facial expressions made by an actor.

A very interesting extension to the framework would be the addition of visual attention capabilities. As seen in the design of Figure 1, attention may be used to select certain information in the sensory storage, perception or interpretation stages for access to further stages of processing, as well as modulating planning and for some behaviour generation, such as orienting an agent's gaze. An attention system, applicable to both real and virtual environments, in a unified framework, is an interesting prospect. Finally, we also aim to use the analysis-synthesis loop as a learning phase to refine the synthesis model of expressivity and of behaviour.

References

1. Martin, J.C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M., Pelachaud, C.: Levels of representation in the annotation of emotion for the specification of expressivity in ecas. In: International Working Conference on Intelligent Virtual Agents, Kos, Greece (2005) 405–417
2. Peters, C.: Direction of attention perception for conversation initiation in virtual environments. In: International Working Conference on Intelligent Virtual Agents, Kos, Greece (2005) 215–228
3. Scherer, K., Ekman, P.: Approaches to Emotion. Lawrence Erlbaum Associates (1984)
4. Ekman, P., Friesen, W.: The Facial Action Coding System, Consulting Psychologists Press. San Francisco, CA (1978)
5. Tian, Y., Kanade, T., Cohn, J.: Facial expression analysis. In: Handbook of face recognition, S.Z. Li and A.K. Jain, ed.,. (2003)
6. Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Recognizing facial expression: machine learning and application to spontaneous behavior. In: Computer Vision and Pattern Recognition. Volume 2., Computer Vision and Pattern Recognition, CVPR 2005 (2005) 568–573
7. Morency, L.P., Sidner, C., Lee, C., Darrell, T.: Contextual recognition of head gestures. In: Proceedings of ICMI'05, Trento, Italy (2005)
8. Gratch, J., Marsella, S., Maatman, M., Gratch, J., Marsella, S.: Natural behavior of a listening agent. In: Proceedings of IVA05, Kos, Greece (2005)

9. Donato, G., Bartlett, M., Hager, J., Ekman, P., Sejnowski, T.: Classifying facial actions. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume 21. (1999)
10. Wu, Y., Huang, T.: Hand modeling, analysis, and recognition for vision-based human computer interaction. In: IEEE Signal Processing Magazine. Volume 18. (2001) 51–60
11. Wexelblat, A.: An approach to natural gesture in virtual environments. In: ACM Transactions on Computer-Human Interaction. Volume 2. (1995) 179–200
12. Chi, D., Costa, M., Zhao, L., Badler, N.: The EMOTE model for effort and shape. In: ACM SIGGRAPH '00, New Orleans, LA (2000) 173–182
13. Byun, M., Badler, N.: FacEMOTE: Qualitative parametric modifiers for facial animations. In: Symposium on Computer Animation, San Antonio, TX (2002)
14. Kopp, S., Sowa, T., Wachsmuth, I.: Imitation games with an artificial agent: From mimicking to understanding shape-related iconic gestures. In: Gesture Workshop. (2003) 436–447
15. Rapantzikos, K., Avrithis, Y.: An enhanced spatiotemporal visual attention model for sports video analysis. In: International Workshop on content-based Multimedia indexing (CBMI), Riga, Latvia (2005)
16. Pelachaud, C., Bilvi, M.: Computational model of believable conversational agents. In Huget, M.P., ed.: Communication in Multiagent Systems. Volume 2650 of Lecture Notes in Computer Science. Springer-Verlag (2003) 300–317
17. Hartmann, B., Mancini, M., Pelachaud, C.: Implementing expressive gesture synthesis for embodied conversational agents. In: Gesture Workshop, Vannes (2005)
18. Johansson, G.: Visual perception of biological motion adn a model for its analysis. Perception and Psychophysics **14** (1973) 201–211
19. Wallbott, H.G., Scherer, K.R.: Cues and channels in emotion recognition. Journal of Personality and Social Psychology **51** (1986) 690–699
20. Gallaher, P.E.: Individual differences in nonverbal behavior: Dimensions of style. Journal of Personality and Social Psychology **63** (1992) 133–145
21. Ball, G., Breese, J.: Emotion and personality in a conversational agent. In J. Cassell, J. Sullivan, S.P., Churchill, E., eds.: Embodied Conversational Characters. MITpress, Cambridge, MA (2000)
22. Pollick, F.E.: The features people use to recognize human movement style. In Camurri, A., Volpe, G., eds.: Gesture-Based Communication in Human-Computer Interaction - GW 2003. Number 2915 in LNAI. Springer (2004) 10–19
23. Hartmann, B., Mancini, M., Buisine, S., Pelachaud, C.: Design and evaluation of expressive gesture synthesis for embodied conversational agents. In: Third International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), Utretch (2005)
24. Tekalp, A., Ostermann, J.: Face and 2-d mesh animation in MPEG-4. In: Signal Processing: Image Communication. Volume 15. (2000) 387–421
25. Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K., Kollias, S.: Parameterized facial expression synthesis based on mpeg-4. In: EURASIP Journal on Applied Signal Processing. Volume 2002., Hindawi Publishing Corporation (2002) 1021–1038
26. Wallbott, H.G.: Bodily expression of emotion. In: European Journal of Social Psychology. Volume 28. (1998) 879–896
27. Harrigan, J.A.: Listener's body movements and speaking turns. In: Communication Research. Volume 12. (1985) 233–250
28. Hartmann, B., Mancini, M., Pelachaud, C.: Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis. In: Computer Animation'02, Geneva, Switzerland, IEEE Computer Society Press (2002)

Perception of Dynamic Facial Expressions of Emotion

Holger Hoffmann, Harald C. Traue, Franziska Bachmayr, and Henrik Kessler

University Clinic for Psychosomatic Medicine and Psychotherapy, Ulm
{holger.hoffmann, harald.traue, franziska.bachmayr}@uni-ulm.de,
henrik.kessler@uniklinik-ulm.de

Abstract. In order to assess subjects' ability to recognize facially expressed emotions it is more realistic to present dynamic instead of static facial expressions. So far, no time windows for the optimal presentation for that kind of stimuli have been reported. We presented dynamic displays where the face evolves from a neutral to an emotional expression to normal subjects. This study measured the optimal velocities in which facial expressions were perceived as being natural. Subjects (N=46) viewed morphed sequences with facial emotions and could adjust the velocity until satisfied with the natural appearance. Velocities for each emotion are reported. Emotions differed significantly in their optimal velocities.

1 Introduction

Presentation of facial displays of emotion is an important method in emotion perception studies in various basic and applied sciences. However, one of the great drawbacks of facial emotion perception research is that in almost all studies typically only *static* displays of facial emotions were used. Using *dynamic* stimuli, where the face evolves from neutral to an emotional expression naturally over time, seems to be a worthwhile new approach for various reasons. First, static images do clearly not represent real-life facial expressions. Presentation of film sequences, where the emotion evolves gradually in a face would be a much more natural approach. Second, recent neuroimaging [6] and computer modelling research has proposed different brain areas for the recognition of dynamic versus static faces, inducing additional interest in the use of dynamic stimuli to untangle underlying recognition mechanisms. And third, there is empirical data showing that subjects recognize emotions better when displayed dynamically as opposed to static or multi-static pictures [1]. Although dynamics in facial emotion presentation have been recognized to be important, actual information on temporal properties of facial expressions is scarce.

This study explored the perception of dynamic emotions. We presented subjects morphed sequences of actors portraying six different facial emotions (anger, fear, happiness, sadness, disgust and surprise). The expression evolves from neutral to emotional with varying durations (.2 to 3.0 seconds). Subjects had to adjust the duration of the presented emotional sequence until they considered the sequence most realistic. The aim was to obtain an optimal duration for each emotion which is perceived as most natural to produce dynamic displays with emotion-specific film speeds.

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 175 – 178, 2006.

2 Methods

2.1 Subjects

Forty-six healthy subjects were tested who gave written consent to participate in our study. The group (N=46) aged 19 to 61 years (mean: 22.1 years, SD: 8.2 years), 35 (76 %) of them were female.

2.2 Stimuli

Image sets (neutral & emotional) of 8 different expressions for each emotion were selected for this study (4 male & 4 female; each half caucasian and japanese actors), resulting in 48 different actors portraying the emotions. Pictures used were taken from the JACFEE set (Japanese and Caucasian Facial Expressions of Emotion, [5]), where actors portray one of six basic emotions (anger, disgust, fear, happiness, sadness, surprise). The set also contains one neutral picture for each actor. Various studies have shown the reliability and validity of the JACFEE set in displaying the intended emotions [2].

In our study we produced emotional sequences where the expression changes from neutral to emotional using both pictures from the JACFEE as end-points and synthesized morphed images between them. Sequences were generated using various speeds with 5 to 75 frames (by an increment of 5 frames) using a frame rate of 25 frames s^{-1}. This results in 15 video clips for each actor with film durations between .2 and 3.0s (in steps of 200ms). Morph sequences were generated using a self-developed morphing tool, called FEMT (Facial Expression Morphing Tool, [4]). This software uses common morphing algorithms to synthesize emotional sequences. For optimization in producing facial morphs, additional techniques (e.g. multiple layers) have been implemented.

2.3 Procedure

Subjects were tested individually using a self-developed presentation software, taking approximately 15 minutes to complete. Subjects saw 48 emotional sequences consecutively showing the development from a neutral to an emotional face. They were instructed to adjust the velocity of each sequence as long as necessary until the sequence was perceived as being most natural or realistic respectively. This was done by repeatedly pressing “faster/slower” buttons (in steps of +/- 200/600ms). The actual emotion sequence was embedded in a prior 1000ms neutral face and a 300ms full-blown emotional face afterwards. Both, the velocity of each first sequence and the order of all sequences were randomized. The emotion name was displayed at the top of the screen for each presented sequence. Furthermore, subjects had the possibility to repeat the current sequence as often as required.

Table 1. Descriptive statistics for each emotion. Mean values show optimized temporal windows for the presentation of emotional sequences in milliseconds.

Emotion	Minimum [ms]	Maximum [ms]	Mean [ms]	SD [ms]
Anger	300	1775	**907**	367
Disgust	400	1925	**939**	330
Fear	250	1425	**715**	332
Happiness	325	1375	**826**	306
Sadness	350	2000	**1198**	400
Surprise	225	1425	**572**	295

3 Results

Table 1 shows descriptive statistics for each emotion. It is important to notice that different emotions had their peaks at different display durations, giving first indication that "natural" time frames could be emotion-specific. Multiple t-tests showed significant differences in time windows between emotions with the exception of "anger - disgust". The respective means show the velocity that was considered to be realistic for each emotion (figure 1a). Fear and surprise were best rated at around 550-700ms and show a small standard deviation of about 300ms. Happiness, disgust and anger seem to be considered natural when displayed for 850-950ms. Finally sadness is perceived as natural well above 1000ms and shows the widest standard deviation of 400ms. Figure 1b shows histograms containing preferred temporal windows for each emotion. For this figure we used the total of 368 single values (46 subjects × 8 expressions for each emotion).

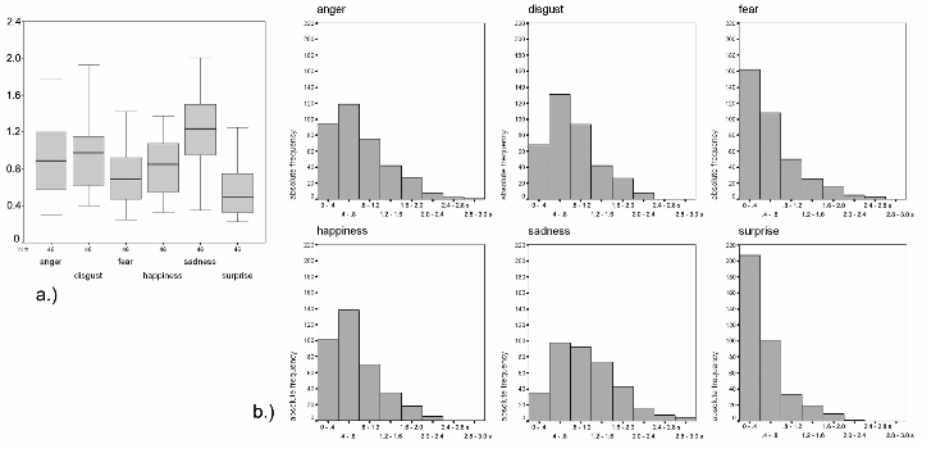


Fig. 1. a.) Boxplot of chosen velocities for every emotion. The y axis shows the according time in seconds. b.) Distribution of absolute frequencies of chosen velocities (N= 368 single films).

4 Discussion

This is the first study to present durations for which facially expressed basic emotions have to be displayed in order to be perceived as being natural by subjects. Fear and surprise seem to be the "rigidly short" emotions with display times of 550-700ms and narrow standard deviation. On the other extreme lies sadness as "long and variable" (1200ms) with a relatively wide standard deviation. Disgust, happiness and anger are "in between" with a mean duration of 850-950ms.

Interestingly, there is no single optimal duration across emotions, but results support the assumption of emotion-specific display durations. The question whether temporal characteristics are something inherent to the emotion itself or depend on context (and/or so-called display rules) remains to be answered in further studies.

Our results are mainly in line with findings from two different studies varying display durations of facial emotion sequences. Sato [7] found that surprise was considered natural when presented at a fast speed (255ms) and sadness at a slow speed (1020ms).

Kamachi [3] found that *recognition* of dynamic facial emotions depends on the duration of the film shown, with sadness being better recognized at slow speed (3400ms), happiness and surprise at fast speed (200ms) and anger in between (900ms). Kamachi's results indicate that subjects tend to associate certain emotions with specific display durations where recognition is best. Kamachi's and our results underline the notion that temporal characteristics of facial expressions are inherent to an emotion in such a way that they facilitate recognition.

References

1. Ambadar, Z., Schooler, J. W., & Cohn, J. F. (2005). Deciphering the enigmatic face: the importance of facial dynamics in interpreting subtle facial expressions. Psychol Sci, 16(5), 403-410.
2. Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., & Ton, V. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability Data and Cross-National Differences. Journal of Nonverbal Behavior, 21, 3-21.
3. Kamachi, M., Bruce, V., Mukaida, S., Gyoba, J., Yoshikawa, S., & Akamatsu, S. (2001). Dynamic properties influence the perception of facial expressions. Perception, 30(7), 875-887.
4. Kessler, H., Hoffmann, H., Bayerl, P., Neumann, H., Basic, A., Deighton, R. M., Traue, H. C. (2005). Measuring emotion recognition with computer morphing: New methods for research and clinical practice. Nervenheilkunde, 24, 611-614.
5. Matsumoto, D., & Ekman, P. (1988). Japanese and Caucasian Facial Expressions of Emotion (JACFEE) and Neutral Faces (JACNeuF). [Slides]: Dr. Paul Ekman, Department of Psychiatry, University of California, San Francisco, 401 Parnassus, San Francisco, CA 94143-0984.
6. Sato, W., Kochiyama, T., Yoshikawa, S., Naito, E., & Matsumura, M. (2004). Enhanced neural activity in response to dynamic facial expressions of emotion: an fMRI study. Brain Res Cogn Brain Res, 20(1), 81-91.
7. Sato, W. Yoshikawa, S. (2004) The dynamic aspects of emotional facial expressions. Cognition And Emotion, 18(5), 701-710

Multi-level Face Tracking for Estimating Human Head Orientation in Video Sequences

Tobias Bausch, Pierre Bayerl, and Heiko Neumann*

Universität Ulm, Neuroinformatik, 89069 Ulm, Germany
tobias.bausch@web.de, {pierre.bayerl, heiko.neumann}@uni-ulm.de

Abstract. We propose a hierarchical scheme of tracking facial regions in video sequences. The hierarchy uses the face structure, facial regions and their components, such as eyes and mouth, to achieve improved robustness against structural deformations and the temporal loss of image components due to, e.g., self-occlusion. The temporal deformation of facial eye regions is mapped to estimate the head orientation around the yaw axis. The performance of the algorithm is demonstrated for free head motions.

1 Introduction

Approaches to face image analysis can be roughly subdivided into tasks for recognition of identity and the interpretation of changeable aspects. In the latter, several computational approaches have been suggested for the detection of head position and orientation ([5][7]), eye gaze [8] and facial expression ([2][3]). However, most investigations consider static images as input or if they process temporal sequences the analysis is often based on individual snapshots. We propose to extract motion information from sequences to track faces. The estimated deformations of facial structures during motion are then used to measure the head orientation of the observed human face.

2 Multi-level Face Tracking and Head Pose Estimation

Our approach takes video sequences as input to continuously track a human face and its facial regions. We employ a region-based registration algorithm for correspondence finding and subsequent estimation of head orientations from image sequences.

2.1 Region-Based Facial Motion Tracking

The key problem to motion computation is to find corresponding matches in subsequent image frames. Utilizing small regions or localized features for correspondence finding renders algorithms insensitive against distortions, but makes them sensitive to noise and leaving ambiguities due to multiple possible interpretations. Using large

* Part of this work has been supported by a grant from the Ministry of Science, Research and the Arts of Baden-Württemberg (Az: 23-7532.24-13-19/1) to H.N.

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 179 – 182, 2006.

regions, on the other hand, leads to more robust detection and unambiguous interpretation, but often suffers from distortions due to the varying appearances of object shapes through rotations, e.g., leading to perspective foreshortening.

To track faces over time, we assume an approximately flat facial region such that deformations in frame-to-frame image registration can be accounted for by an affine warping transformation, $warp(\vec{x};\vec{p})$, with $\vec{p} = (p_1, p_2, ..., p_6)^T$ ([4]). The frame-to-frame image registration is computed using a quadratic grey level error measure using the inverse compositional image alignment method ([1])

$$\min_{\Delta\vec{p}} E(\Delta\vec{p}) = \min_{\Delta\vec{p}} \sum_{\vec{x}\in N} \left[I_1(warp(\vec{x};\Delta\vec{p})) - I_2(\vec{x})\right]^2 \quad (1)$$

as a computationally more efficient alternative to the Lucas-Kanade tracker ([6]).

2.2 Facial Component Matching in Temporal Sequences

Region-based matching of surface regions over longer temporal sequences renders the problem of deterioration of match quality due to changes in surface appearance. As a consequence, the matching must be supplied with a proper consistency measure for the error measure E. We investigated several methods for frame-wise segment tracking. Among others, we utilized frame-to-frame matching, $M_{f\text{-}f}$, in which a previously selected region (rendered by a circumscribing polygon) is warped into the next frame. Alternatively, a model-matching approach, M_m, was employed in which the current image frame I_t is compared against a constant model I_{model} (often the first image in the sequence). The former approach has deficits due to instabilities by accumulating propagated errors, while the model matching approach suffers from long-term deterioration of the reference from current image frames when time proceeds. To utilize the best features, both approaches were combined into $M_{f\text{-}f\rightarrow m}$.

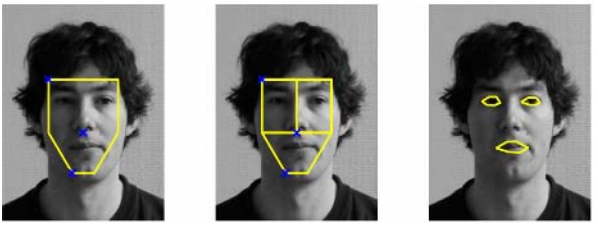


Fig. 1. Hierarchical multi-level feature tracking scheme. The polygons need to be initialized once in the first image frame using a set of control points. Three levels are employed that analyze central face area, L_{face}, facial sub-regions, L_{reg}, and facial components, L_{comp}.

In order to take advantage of more global as well as localized information, we propose a hierarchical tracking at three levels namely full face, facial regions and facial components, L_{face}, L_{reg} and L_{comp} (see Fig. 1). The method must be initialized in the first image by manual selection of individual feature points of a polygon frame. All levels employ the model matching-approach M_m utilizing different motion predictions for different levels. The predicted motion warp at L_{face} is computed by a weighted average of motions of its components derived at the next lower level in the hierarchy.

The weighting of the individual contributions is inverse proportional to the error measure of the region-based motion for the particular region. At the region level, L_{reg}, the $M_{f\text{-}f\rightarrow m}$ method is applied using the frame-to-frame motion estimate without initial prediction, and then employs the model matching for similarity measure. At the component level, L_{comp}, the motion estimates for both eyes and the mouth region are predicted by results from the region levels, using model matching to find the best fit.

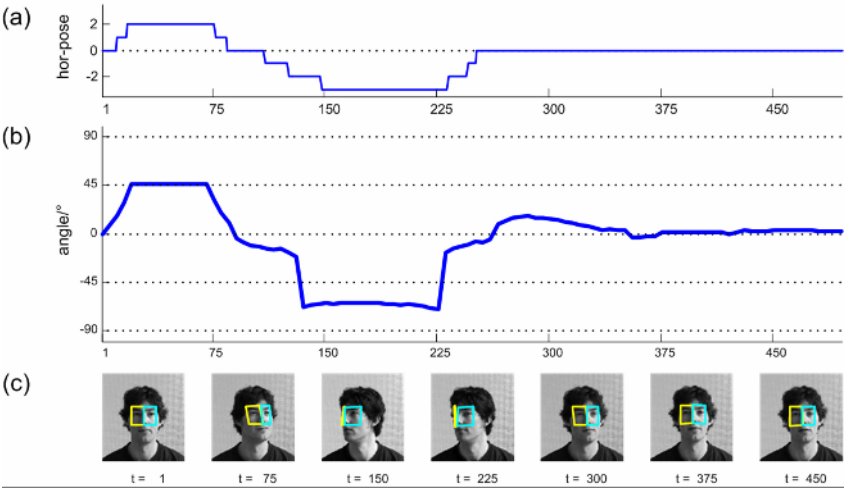


Fig. 2. Estimating head rotation. (a) Manually labeled head rotations in 7 categories ranging from -3 to +3 corresponding to extremely leftwards, leftwards, nearly centered, centered, ..., extremely rightwards. (b) The estimated angle deduced from the spatial extent of automatically detected and tracked facial sub-regions (see text). The estimated head pose roughly corresponds to the manually labeled data. (c) Individual frames from the input sequence and the corresponding detected sub-regions around the eyes utilized for pose estimation.

During an extended tracking period a facial region may get lost temporarily. We employ a statistical measure to detect whether a patch disappeared to resume tracking. The criterion is based on the observation that different facial sub-regions are approximately coherent. A facial loss is detected when translational components along the x- and the y-axis of different facial regions at one time step show high variances, particularly, if $var(t_x)\cdot var(t_y) >$ threshold. In such cases the face is searched in an extended area at L_{face} and iterated until the error E drops below a preset level.

2.3 Estimating Head Rotation

With the hierarchical multi-level approach it is possible to track moving faces for long temporal sequences and to estimate the head pose for visual communication. Our approach assumes a dominant horizontal facial rotation (around the yaw axis) utilizing a geometric model of the head as a cylindrical object with circular base and orthographic projection for simplicity. The method investigates the two juxtaposed facial sub-regions around the eyes which increase or decrease in their horizontal extension due to the head rotation. The change in the size of the sub-region that has better

visibility is mapped using a non-linear function (derived from the geometric model) to yield possible solutions for the rotation angle. The other sub-region is utilized to select one of those solutions as the final estimate.

3 Results

A comparison of the different tracking methods demonstrates the desired stability and quality (as measured by the mean distance of detected polygons) for the combined matching approach, $M_{f\text{-}f\rightarrow m}$ (results not shown). Thereupon the hierarchical scheme was employed to track faces demonstrating the robustness against intermediate loss of facial sub-regions. The analysis of the upper left and right sub-regions (eye regions) and their mapping to head orientation angles is shown in Fig. 2. The results demonstrate that the yaw axis rotation is robustly estimated over long video sequences.

4 Summary and Further Work

We propose an approach for coupled hierarchical facial component tracking based on an affine image registration method. The hierarchical multi-level approach achieves robust tracking of faces and their sub-components even in cases when some elements are temporarily lost. This algorithm is augmented by a non-linear mapping to estimate the orientation angle of the moving head. Extensions of the approach could be employed along different directions. The current initialization of the facial region can be automatized through a face detection approach. The incorporation of localized features (at a fourth level L_{feat}), such as, e.g., eye corners, may help to further improve tracking performance. The prediction of translational motion components might be improved using, e.g., a Kalman filter approach on region centers.

References

1. Baker, S., Matthews, I.: Lucas-Kanade 20 years on: A unifying framework: Part1. Carnegie-Mellon Univ., Robotics Institute, CMU-RI-TR-02-16 (2002)
2. Black, M.J., Yacoob, Y.: Recognizing facial expressions in image sequences using local parametrized models of image motion. Int.'l J. of Computer Vision 25 (1997) 23–48
3. Essa, I.A., Pentland, A.P.: Facial expression recognition using a dynamic model and motion energy. In: Proc. 5th Int.'l Conf. on Computer Vision, ICCV'95 (1995) 360–367
4. Gee, A., Cipolla, R.: Determining the gaze of faces in images. Image and Vision Comp. 30 (1994) 639–647
5. Krüger, N., Pötzsch, M., von der Malsburg, C. Determination of face position and pose with a learned representation based on labeled graphs. Im. Vis. Comput. (1997) 15(8) 665–673.
6. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proc. 7th Int.'l Conf. on Artif. Intell., IJCAI'81 (1981) 674–679
7. Rae, R., Ritter, H.J.: Recognition of human head orientation based on artificial neural networks. IEEE Trans. on Neural Networks 9 (1998) 257–265
8. Tan, K.-H., Kriegman, D.J., Ahuja, N.: Appearance-based eye gaze estimation. In: Proc. 6th IEEE Workshop on Applications of Computer Vision, WACV'02 (2002) 191–195

The Effect of Prosodic Features on the Interpretation of Synthesised Backchannels

Åsa Wallers, Jens Edlund, and Gabriel Skantze

Department for Speech Music and Hearing, KTH
Lindstedtsv. 24, 100 44 Stockholm, Sweden
{wallers, edlund, gabriel}@speech.kth.se

Abstract. A study of the interpretation of prosodic features in backchannels (Swedish /a/ and /m/) produced by speech synthesis is presented. The study is part of work-in-progress towards endowing conversational spoken dialogue systems with the ability to produce and use backchannels and other feedback.

1 Introduction

Spoken dialogue among humans is an intricate and fine-tuned process which puts high demands on the participants' ability to perceive and produce inputs and outputs according to the flow of the dialogue, as well as to the context. In a conversation, the participants take turns talking, and the speaker transition is for the most part a very smooth interaction with little speech overlap [1].

Interaction control in spoken dialogue systems is an active area of research. We are becoming increasingly good at dealing with online analysis of human speech and great efforts have been spent to make systems give properly timed feedback. Many researchers working with the development of spoken dialogue systems have shown interest in prosodic features when trying to make the system handle the turns properly in the conversation (e.g. [2, 3]).

As our research systems become more human-like and better at timing their responses, other shortcomings become more apparent. In human-human dialogue, feedback and back-channels make up a significant part of the interaction, and a spoken dialogue system that is to be deemed responsive and human-like needs similar capabilities. We have made preliminary user studies indicating that backchannels have a great effect on how a conversation proceeds, and similar observations are described in more detail by Riccardi and Gorin [4].

A problem that has to be overcome in order to achieve system backchannels is that the interpretation of feedback backchannels, such as *ah* and *m*, may depend on their prosody. In the type of unrestricted conversations we are aiming at in our research systems (e.g. Waxholm, August, AdApt [5], and currently Higgins [6]) the demands on flexible output generation makes canned speech difficult to use. Instead, we aim to include prosodic variation in synthesised feedback and backchannels.

2 Method

Previously we looked at the effect of prosody on one-word elliptical feedback [7]. Here, we attempt to the same with the Swedish one-syllable back-channels *m* and *a*.

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 183 – 187, 2006.

These are commonly used in Swedish: in Swedish MapTask dialogues [8], we found that *a* and *m* made up 34% and 15% of the backchannels, respectively.

The stimuli consisted of monosyllabic renditions of /a/ and /m/ synthesised using an experimental version of LUKAS diphone Swedish male MBROLA voice implemented as a plug-in to the WaveSurfer speech tool. Although disyllabic /mm/ and /aa/ also occur frequently in Swedish, we used the monosyllabic versions only, partly to constrain the dimensionality of the experiment and partly in an attempt to make the experiment consistent with [7].

For each of the two test words the parameters peak POSITION, peak HEIGHT, and DURATION were manipulated. Three peak positions were obtained by time-shifting the focal accent peaks in intervals of 100 ms comprising EARLY, MID and LATE peaks. The LOW and HIGH peak height was set to 130 and 160 Hz. The durations SHORT and LONG were set to 450 and 650 ms. Combination of the two backchannels and the three properties gave a total of 24 different stimuli, schematically represented in Fig. 1.

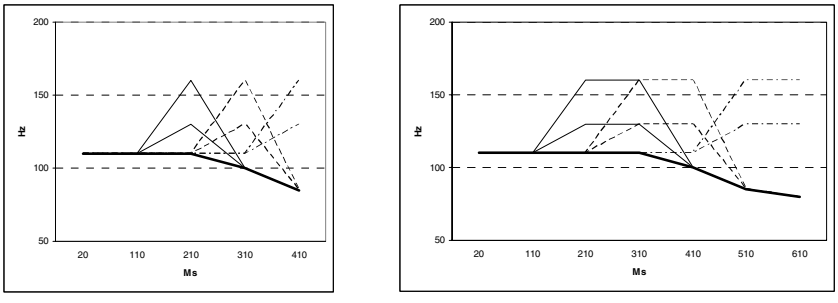


Fig. 1. The prosodic properties of the short and long stimuli, respectively

A preliminary listening test made it clear that the results found in [7] (i.e. a mapping between peak position and level of grounding) would not be reproduced on these stimuli. Subjects' comments showed a wide range in interpretations of the stimuli, so we resorted to a two-step exploratory approach.

Table 1. List of backchannel interpretations

Interpretation (in English translation)	Abbreviation
Good, you are in the right place.	RIGHT PLACE
Oh, NOW I understand where you are.	OH!
Really? That was unexpected.	UNEXPECTED
Oh, you are in the wrong place.	WRONG PLACE
Okay, but I need more information.	CONTINUE

In a first experiment, five listeners were subjected to dialogue fragments consisting of a human speaker uttering "On my left I have a X house…", where X was one of the colours red, yellow, and blue, followed by one of the system backchannels. This was repeated for each stimuli and colour, and after each fragment, the listener

was asked to write down a free interpretation of the system's response. These interpretations were then manually summarised and condensed into the five paraphrases found in Table 1.

The second step was a perceptual test where the eight participants were asked to listen to the backchannels in the context, and then chose the one of the five paraphrases they felt best represented the meaning of the backchannel. Each stimulus was played three times and the order was randomised.

3 Results

In general, the results show that the variation in prosodic features does effect the interpretation of the backchannels significantly, and in general, the choice of stimulus (/m/ or /a/) had a greater effect than anticipated, with /a/ tilting the interpretations heavily towards OH! and /m/ showing a preference for CONTINUE (Fig. 2). Unfortunately it is difficult to draw any general conclusions about the individual effect of each parameter – POSITION, HEIGHT, and DURATION – from the material. The exploratory experiment design makes it difficult to test for significance, and as this is work-in-progress, we will limit the presentation to raw numbers.

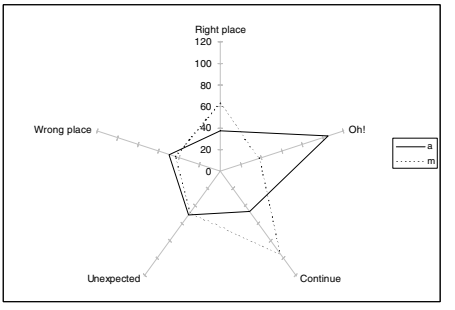
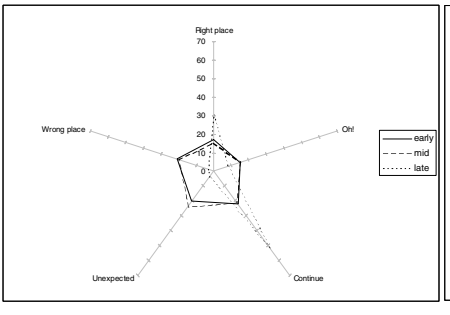
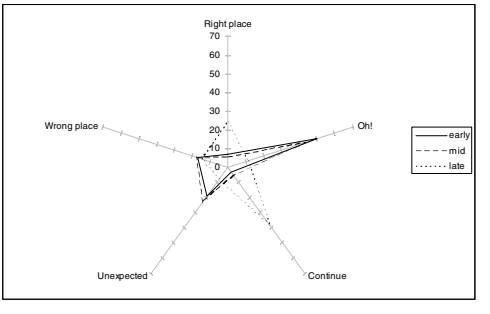


Fig. 2. Distribution of votes for /a/ and /m/





Figs. 3 and 4. Distribution of votes for different setting of PEAK for /m/ and /a/,respectively

The feature that shows the greatest difference in numbers is peak POSITION. EARLY and MID position give similar results both for /a/ and /m/, but LATE leads to a different interpretation. In the case of /m/, the interpretation of EARLY/MID POSITION stimuli is very vague, with close to equal distribution amongst the five interpretations (Fig. 3). For /a/, the same features leads to a bias towards the OH! interpretation (Fig. 4). For /m/ and /a/ alike, a LATE peak position shifts the bias heavily towards the more neutral

CONTINUE interpretation (Figs. 3 and 4). The HEIGHT and DURATION parameters show less clear influence on the distribution of interpretations.

The stimuli reaching the highest consensus amongst the subjects are /a/ LONG EARLY HIGH peak and /a/ LONG MID HIGH peak, where the OH! Interpretation obtained 75% and 71% of the votes, respectively. For /m/, LONG and SHORT LATE LOW peak yielded the highest consensus, with 63% and 58% of the votes, respectively, for CONTINUE. Finally, for both /m/ and /a/ RIGHTPLACE obtained 46% of the votes in the SHORT LATE HIGH peak setting. The full results are available in detail in [9].

4 Discussion and Future Work

The preliminary work described here has taught us valuable lessons:

- it is indeed possible to produce synthesized backchannels with variable prosody that is perceived and interpreted in a consistent manner by human subjects
- our previous findings on the interpretation of prosodic patterns applied to synthesised monosyllabic one word clarification ellipses ([7]) are not directly applicable on synthesised monosyllabic backchannels
- peak POSITION may effect the interpretation of monosyllabic backchannels more than DURATION and HEIGHT

As the study is context dependent, it needs expansions in several ways to test its relevance to a wider domain. Presently we intend to examine if the backchannels are sufficiently salient to signal system reactions in a spoken dialogue system, as well as to study how system bachchannels are perceived in general is such a setting.

Finally, we made several observations that beckon further investigation. For example, the interpretation categories derived from the open interpretations may be grouped into RIGHTPLACE, WRONGPLACE and CONTINUE on the one hand, and OH! and UNEXPECTED on the other. The first group combine interpretations suggesting that nothing unexpected has occurred from the point of view of the producer of the backchannel, whereas the second group combines interpretations that include an element of surprise.

Acknowledgements

This research was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations and was supported by the EU project CHIL (IP506909).

References

1. Levinson, S. (1983). Pragmatics. Cambridge: Cambridge University Press.
2. Edlund, J & Heldner, M (2005). Exploring Prosody in Interaction Control. Phonetica, 62(2-4).
3. Ferrer, L., Shriberg, E., Stolcke, A (2002).: Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody in human-computer dialog. Proc. Int. Conf. spoken Lang. Processing, Denver, pp. 2061–2064.

4. Riccardi, G. & Gorin, A.L. (2000): Spoken Language Adaptation over Time and State in a Natural Spoken Dialog System. IEEE Trans. on Speech and Audio.
5. Gustafson, J. (2002): Developing Multimodal Spoken Dialogue Systems. Empirical Studies of Spoken Human-Computer Interaction. TRITA-TMH 2002:8, ISSN 1104-5787.
6. Edlund J, Skantze G & Carlson R (2004): Higgins - a spoken dialogue system for investigating error handling techniques. In Proceedings of the International Conference on Spoken Language Processing (ICSLP) 2004 (pp. 229-231). Jeju, Korea
7. Edlund, J, House, D, & Skantze, G (2005): The Effects of Prosodic Features on the Interpretation of Clarification Ellipses. In Proceedings of Interspeech 2005, Lisbon, Portugal
8. Helgason, P. (2002). Preaspiration in the Nordic languages: synchronic and diachronic aspects. Doctoral dissertation, Stockholm University, Stockholm.
9. Wallers, Å. (2005): Minor Sounds with Major Impact. Master thesis, KTH, Stockholm.

Unsupervised Learning of Spatio-temporal Primitives of Emotional Gait

Lars Omlor and Martin A. Giese

Laboratory for Action Representation and Learning/Department of Cognitive Neurology,
Hertie Institute for Clinical Brain Research, University of Tübingen, Germany
http://www.uni-tuebingen.de/uni/knv/arl/index.html

Abstract. Experimental and computational studies suggest that complex motor behavior is based on simpler spatio-temporal primitives. This has been demonstrated by application of dimensionality reduction techniques to signals from electrophysiological and EMG recordings during execution of limb movements. However, the existence of such primitives on the level of kinematics, i.e. the joint trajectories of complex human full-body movements remains less explored. Known blind source separation techniques, e.g. PCA and ICA, tend to extract relatively large numbers of components or source signals from such trajectories that are typically difficult to interpret. For the analysis of emotional human gait patterns, we present a new method for blind source separation that is based on a nonlinear generative model with additional time delays. The resulting model is able to approximate high-dimensional movement trajectories very accurately with very few source components. Combining this method with sparse regression, we identified spatio-temporal primitives for the encoding of individual emotions in gait. We verified that these primitives match features that are important for the perception of emotions from gait in psychophysical studies. This suggests the existence of emotion-specific movement primitives that might be useful for the simulation of emotional behavior in technical applications.

1 Introduction

Human full-body movements are characterized by a large number of degrees of freedom. This makes the accurate synthesis of human trajectories for applications in computer graphics and robotics a challenging problem. The analysis of motor behavior suggests the existence of simple basis components, or spatio-temporal primitives, that form building blocks for the realization of more complex motor behavior [1]. Since such basic components cannot be directly observed, several studies have tried to identify spatio-temporal primitives by application of unsupervised learning techniques, like PCA or ICA [2], to data from electrophysiological and EMG recordings (e.g.[3] [4]). Similarly, one can try to apply such methods directly to joint angle trajectories. However, the complexfull-body

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 188–192, 2006.

movements typically result in the extraction of a relatively large number of basic components or source signals that are difficult to control and to interpret (e.g. [5]). In our study we tried to learn movement primitives of emotional gait from joint angle trajectories. We developed a new technique for blind source separation, which is based on a non-linear generative model that, opposed to normal PCA and ICA, can model time delays between source components and the individual joint angles. Opposed to the other existing algorithms for blind source separation with delays [6, 7], our method scales up to large problems, allows dimensionality reduction, and requires no additional sparseness assumptions. It provides a much better approximation of gait data with few basic components than other existing methods.

By approximating gait trajectories by superpositions of the extracted component signals and applying a regression algorithm, which extracts important features by sparsification, we were able to determine the most important spatio-temporal components for the expression of different emotions in gait. To validate the results from our kinematic analysis, we compared the extracted components with results from psychophysical perception experiments that have identified features that are critical for the perception of emotion from gait. We found good correspondence between these features and the components derived from the joint trajectories.

Trajectory data. Using a VICON motion capture system with 7 cameras, we recorded the gait trajectories from thirteen lay actors executing walking with four basic emotional styles (happy, angry, sad and fear), and normal walking. Each trajectory was executed three times by each actor. Approximating the marker trajectories with a hierarchical kinematic body model (skeleton) with 17 joints, we computed joint angle trajectories. Rotations between adjacent segments were described as Euler angles, defining flexion, abduction and rotations about the connecting joint. Data for the unsupervised learning procedure included only the flexion angles of the hip, knee, elbow, shoulder and the clavicle, since these showed the most reproducible variation.

Blind source separation. We first applied 3 established methods for the estimation of the source signals to our trajectory data: PCA, fast ICA and bayesian ICA [8] with a positivity constraint for the elements of the mixing matrix . These methods required at least 5 sources for a reconstruction of the original trajectories, that explained at least 90 % of the variation of the data.

We then performed separate ICAs for the individual joints, resulting in separate source variables for each individual joint. By computing the cross-correlation functions between different sources, we found that the sources of different joints are often astonishingly similar, and differ only by additional time delays. This has motivated us to develop a new source separation algorithm that appropriately models such delays.

Signifying by x_i the i-th trajectory and by s_j the j-th unknown source signal, the data is modeled by the following *nonlinear* generative model:

$$x_i(t) = \sum_{j=1}^{n} \alpha_{ij} s_j(t - \tau_{ij}) \tag{1}$$

The model is specified by the linear mixing coefficients α_{ij} and the time delays τ_{ij} between source signals and trajectory components. The problem of blind source separation with time delays has only been rarely been treated in the literature (e.g. [6, 7, 9]), and the existing solutions were not applicable because they either require positive signals or were not suitable for dimensionality reduction.

An efficient algorithm for the solution of this problem, which scales up for high-dimensional problems, was obtained by representing the signals in time-frequency domain using the Wigner-Ville transform, that is defined by

$$Wf(x,\omega) := \int \mathrm{E}\left\{ f(x + \frac{t}{2})\overline{f(x - \frac{t}{2})} \right\} e^{-2\pi i \omega t} dt \tag{2}$$

Applying this integral transformation to equation (1) one obtains:

$$\begin{aligned} Wx_i(\eta,\omega) &= \int \mathrm{E}\Big\{ \sum_{j,k=1}^{n} \alpha_{ij}\overline{\alpha_{ik}} s_j(\eta + \frac{t}{2} - \tau_{ij})\overline{s_k}(\eta - \frac{t}{2} - \tau_{ik}) \Big\} e^{-2\pi i \omega t} dt \\ &= \sum_{j}^{n} |\alpha|_{ij}^2 W s_j. \end{aligned} \tag{3}$$

The last equality sign above is due to the (approximate) independence of the sources. With the additional assumption that the data coincides with the mean of its distribution ($x_j \approx E(x_j)$) one can compute the first and the zero's order moment from equation (3) resulting in the two equations:

$$|\mathcal{F}x_i|^2(\omega) = \sum_{j}^{n} |\alpha|_{ij}^2 |\mathcal{F}s_j|^2(\omega) \tag{4}$$

$$|\mathcal{F}x_i(\omega)|^2 \cdot \frac{\partial}{\partial\omega}\arg\{\mathcal{F}x_i\} = \sum_{j}^{n} |\alpha|_{ij}^2 \cdot |\mathcal{F}s_i|^2 \cdot \left[\frac{\partial}{\partial\omega}\arg\{\mathcal{F}s_j\} + \tau_{ij} \right] \tag{5}$$

Here $\mathcal{F}$ denotes the Fourier transform. From these equations the unknowns can be estimated. To recover the unknown sources s_j, mixing coefficients α_{ij} and time delays τ_{ij} we used the following two step algorithm:

1. First, equation (4) is solved using non-negative ICA [8]. (This step could also be realized exploiting non-negative matrix factorization.)
2. Iteration of the following two steps:
   (a) Equation (5) is solved numerically for $\frac{\partial}{\partial\omega}\arg\{\mathcal{F}s_j\}$, and by integration $\mathcal{F}s_j$ is obtained with initialization $\tau_{ij} = 0$.
   (b) The mixing matrix and the delays are obtained by solving the following optimization problem (with $\mathbf{S}(\boldsymbol{\tau_j}) = (s_k(t_i - \tau_{jk}))_{i,k}$, $\mathbf{A} = (\alpha_{ij})_{ij}$):

$$[\widehat{\boldsymbol{\tau}_j}, \widehat{\mathbf{A}}] = \mathrm{argmin}_{[\boldsymbol{\tau_j}, \mathbf{A}]} \| x_j - \mathbf{A} \cdot \mathbf{S}(\boldsymbol{\tau_j}) \| \tag{6}$$

   This minimization is accomplished following [10], assuming uncorrelatedness for the sources and independence of the time delays.

To construct a mapping between the linear weights $\mathbf{A}$ and the emotional expression we considered the following multi-linear regression model

$$\mathbf{a}_j \approx \mathbf{a}_0 + \mathbf{C} \cdot \mathbf{e}_j \tag{7}$$

where $\mathbf{a}_0$ is a vector with the weights for neutral walking, and $\mathbf{a}_j$ the weight vector for emotion j. $\mathbf{e}_j$ is the j-th unit vector. The columns of the matrix $\mathbf{C}$ encode the deviations in weight space between emotion j and neutral walking. To obtain sparsified solutions for this matrix, we solved the regression problem by minimizing the following cost function (with $\gamma > 0$) with quadratic programming:

$$E(\mathbf{C}) = \sum_j \|\mathbf{a}_j - \mathbf{a}_0 + \mathbf{C} \cdot \mathbf{e}_j\|^2 + \gamma \sum_{ij} |C_{ij}| \tag{8}$$

2 Results

Figure 1 presents the approximation accuracy (explained variance of the whole data set) as a function of the number of extracted sources for 5 different blind source separation methods: PCA, fast ICA, probabilistic ICA with a positivity constraint for the elements of the mixing matrix, and our new method with and without a positivity constraint for the weights α_{ij}. The new algorithm reaches an accuracy of 97% with only 3 source signals (even in presence of a positivity constraint), while the other methods require at least 6 sources to achieve the same level of accuracy.

For additional validation, we animated an avatar with the approximated trajectories using 3 estimated sources. While the animations using components learned with the new algorithm are almost indistinguishable from animations with the original motion capture data, while animations generated with 3 sources extracted by the other methods contain many artifacts.

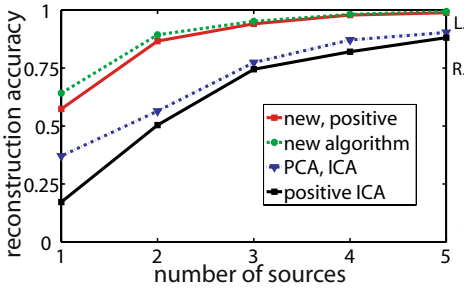


Fig. 1. Comparison of different blind source separation algorithms. Explained variance is shown for different numbers of extracted sources.

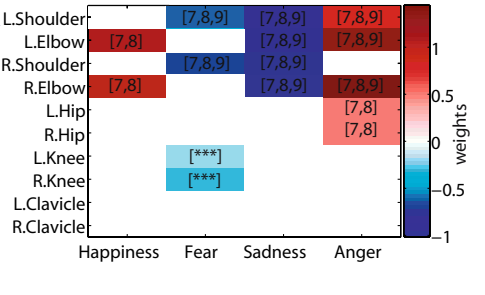


Fig. 2. Elements of the weight matrix $\mathbf{C}$, encoding emotion-specific deviations from neutral walking, for different degrees of freedom. Numbers indicate references describing psychophysical experiments that have reported the same critical components for visual emotion recognition.

To verify whether the extracted spatio-temporal components are biologically meaningful, we compared the non-zero elements of the sparsified regression matrix **C** with results from psychophysical experiments on the perception of emotional gaits. These experiments show that perception of emotions depends on specific changes (of the joint angle amplitudes) of individual degrees of freedom relative to the pattern of neutral walking. As illustrated in Figure 2, we found almost perfect consistency between the weights of the matrix **C** (color coded in the figure) and these features. An example is an increased step length for angry walking, or decreased arm movements for sad walking. The numbers in Figure 2 indicate the features and references of behavioral recognition studies that reported consistent features.

3 Conclusions

The proposed new algorithm accomplishes highly accurate approximation of gait trajectories with very few extracted source components. Selective modulation of the extracted primitives allows to simulate different emotional styles, and the required modulation reflects specific changes in selected joints that are consistent with features that are important for the visual perception of emotional gaits. This supports the biological relevance of the extracted sources and emotion-specific kinematic components. Future work will try to exploit such learned primitives as basis for character animation with high degrees of realism.

References

[1] Flash, T., Hochner, B.: Motor primitives in vertebrates and invertebrates. Curr Opin Neurobiol. **15(6)** (2005) 660–666
[2] Cichocki, A., Amari, S.: Adaptive blind signal and image processing. John Wiley, Chichester (2002.)
[3] Ivanenko, Y., Poppele, R., Lacquaniti, F.: Five basic muscle activation patterns account for muscle activity during human locomotion. J Physiol. **556(Pt 1)** (2004) 267–282
[4] d'Avella, A., Bizzi, E.: Shared and specific muscle synergies in natural motor behaviors. Proc Natl Acad Sci U S A **102(8)** (2005) 3076–3081
[5] Safonova, A., Hodgins, J., Pollard, N.: Synthesizing physically realistic human motion in low-dimensional, behavior-specifc spaces. ACM Trans. Graph. **23(3)** (2004) 514–521
[6] Bofill, P.: Underdetermined blind separation of delayed sound sources in the frequency domain. Neurocomputing **Vol. 55** (2003.) 627–641
[7] Yeredor, A.: Time-delay estimation in mixtures. Acoustics, Speech, and Signal Processing **5** (2003) 237–240
[8] Højen-Sørensen, P., Winther, O., Hansen, L.: Mean field approaches to independent component analysis. Neural Computation **14** (2002) 889–918
[9] Torkkola, K.: Blind separation of delayed sources based on information maximization. ICASSP'96 (1996) 3509–3512
[10] Swindelhurst, A.: Time delay and spatial signature estimation using known asynchronous signals. IEEE Trans. on Sig. Proc. **ASSP-33,no. 6** (1998) 1461–1470

Talking with HIGGINS: Research Challenges in a Spoken Dialogue System

Gabriel Skantze, Jens Edlund, and Rolf Carlson

Department for Speech Music and Hearing, KTH
Lindstedtsv. 24, 100 44 Stockholm, Sweden
{gabriel, edlund, rolf}@speech.kth.se

Abstract. This paper presents the current status of the research in the Higgins project and provides background for a demonstration of the spoken dialogue system implemented within the project. The project represents the latest development in the ongoing dialogue systems research at KTH. The practical goal of the project is to build collaborative conversational dialogue systems in which research issues such as error handling techniques can be tested empirically.

1 Introduction

This paper presents the current status of the research in the HIGGINS project and the spoken dialogue system implemented within the project. The project represents the latest development in the ongoing dialogue systems research at KTH (for an overview, see [1]). The practical goal of the project is to build collaborative conversational dialogue systems in which research issues such as error handling techniques can be tested empirically.

The initial HIGGINS domain – pedestrian city navigation and guiding – is similar to the now classic MapTask [2] domain and to a number of guide systems such as REAL [3]. In this domain, a user gives the system a destination and the system guides the user verbally. For simulation purposes, a 3D model of a city is used (see fig. 1). The system does not have access to the users' positions, but must rely on their descriptions of their surroundings. Since the user is moving, the system must continually update its

Fig. 1. The 3D simulation that is used for user tests

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 193 – 196, 2006.

model of the user's position and provide new, possibly amended instructions until the destination is reached. The surroundings the user and system talk about contain complex landmarks and relations that are challenging to interpret and represent semantically. Compared to simpler domains, the users' utterances (e.g. natural language descriptions of the surrounding environment) tend to be longer, more disfluent and filled with pauses. The domain is implemented in Swedish.

The second major domain to be implemented in HIGGINS was the KTH Connector [4] – a telephony based personal secretary whose task it is to mediate communication between callers and (potentially occupied) callees[1]. Again, the users and the system can reason about complex concepts that are challenging to model – notably relations in time. The KTH Connector is implemented in English.

Finally, several toy domains have been implemented to highlight and test particular HIGGINS features, notably a voice controlled chess board and the language training game *Is it blue?* [5].

2 The HIGGINS Spoken Dialogue System

The HIGGINS spoken dialogue system has a distributed architecture with modules communicating over sockets. Each module has a clearly defined input and output, and can be implemented in any language, running on any platform. All messages and resources are encoded in XML.

The complex semantics used in the HIGGINS domains call for deep semantic structures, and a main focus of the project to date has been developing and testing robust and "error aware" modules for interpretation: the semantic interpreter PICKERING [6] and the discourse modeller GALATEA [7], both implemented in Oz[2]. PICKERING is designed to work with continuous incremental input from a probabilistic speech recogniser. It allows insertions and non-agreement inside phrases, and combines partial results to return a limited list of semantically distinct solutions. The semantic structures that PICKERING produces are represented as rooted unordered trees of semantic concepts. Nodes in the tree may represent for example attribute-value pairs, objects, relations and properties.

GALATEA provides the next step in the interpretation. Whereas PICKERING builds a model of the semantics of an utterance, but does not consider context outside the utterance, GALATEA takes the communicative acts that PICKERING finds in an utterance and does a context aware interpretation of them, resulting in a dynamic model of the discourse.

In HIGGINS, the traditional aspects of dialogue management are not implemented in a single module. Instead, the processing is divided so that GALATEA models the discourse, and the discourse model is then sent to an action manager, which consults the discourse model and a domain database to make decisions and generate communicative system acts. These acts are sent back to GALATEA, as well as to a generator. Thus, GALATEA models communicative acts both from the user and the system.

---

[1] The KTH CONNECTOR is part of the EU-funded project CHIL – a project investigating automatic tracking and support of human interactions.

[2] http://www.mozart-oz.com

3 Research Challenges

The HIGGINS domains are chosen to give rise to a number of research challenges. Here, we will discuss a few of our current foci: concept-level error handling, interaction control, and how well the HIGGINS techniques adapt to other domains.

Concept-Level Error Handling
A challenging research issue is how to handle errors (due to imperfect speech recognition) for individual concepts in the deep semantic structures, using for example elliptical clarification requests, in order to make the dialogue natural and efficient. PICKERING can assign confidence scores to individual concepts in the semantic result, based on the word confidence scores from the speech recogniser. GALATEA may then store these scores in the discourse model as a measure of *concept grounding status*. Dialogue (1) illustrates different ways of using this information:

(1) **U** I have a red building on my left
Sa Red?
Sb Red, hm. How many storeys does it have?
Sc How many storeys does the red building have?

If the grounding status for a concept is low, the system may use an elliptical clarification request (as in Sa), display understanding (as in Sb) or modify the way it refers to objects (as in Sc), in order to "boost" this grounding status. Since GALATEA handles anaphora and ellipsis resolution, it can accurately update the grounding status for the concepts that are involved.

As elliptical clarification requests lack syntax, the prosodic realisation becomes more important. An example is "red" in Sa and Sb, which should be pronounced differently. We are currently performing studies of how prosodic features affect the interpretation of synthesised fragments in such situations [8, 9].

Another error handling method is *late error detection,* i.e. to find possible errors in the discourse at later stages in the dialogue. If, for example, the system finds that there is no possible location for the user given its current beliefs, it may try to remove concepts with low grounding status.

As seen above, the discourse model allows for different error handling strategies. Currently, we are collecting data of users talking to the system. Based on analysis of this data, we investigate methods for choosing strategy.

Interaction Control
Another interesting challenge is how the system should best handle interaction control. Dialogue (2) shows a typical problem in spoken dialogue systems, where the voice activity detection in the automatic speech recogniser has erroneously told the system that the speaker is done talking.

(2) **U** to my left I see a /SIL/ yellow building
S what do you see to your left

Silence detection alone, as used by most systems today, is not sufficient to deal with this kind of situation. Methods involving both syntactic and semantic completeness

(e.g. [6]) as well as prosody [10] have been shown to improve the situation, and we are currently investigating their use in HIGGINS.

Generalisability
An important question is to what extent the techniques developed within the HIGGINS project apply to other domains. One of the reasons for dividing dialogue management into a discourse modeller and an action manager is that it allows the discourse modeller to be fairly generic (GALATEA is simply configured using XML), while the action manager is highly domain dependent. The action planner may have to be reimplemented for each new domain, but this is facilitated by the facts that it can be implemented in any programming language, and that much of the work typically done by a dialogue manager (e.g. ellipsis and anaphora resolution) is already dealt with by GALATEA.

As mentioned above, there are currently several domains implemented in HIGGINS. More domains will be tested in an attempt to find out how general the methods are, and how easy (or difficult) they are to use.

Acknowledgements

This research was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations and was supported by the EU project CHIL (IP506909).

References

1. Gustafson, J. (2002): Developing Multimodal Spoken Dialogue Systems. Empirical Studies of Spoken Human-Computer Interaction. TRITA-TMH 2002:8, ISSN 1104-5787.
2. Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., & Weinert, R (1991): The HCRC Map Task corpus. Language and Speech 34(4) 351-366
3. Baus, J., Kray, C., Krüger, A., & Wahlster, V (1991): A resource-adaptive mobile navigation system. In proc. of the International Workshop on Information Presentation and Natural Multimodal Dialog.
4. Edlund, G. & Hjalmarsson, A. (2004): Applications of Distributed Dialogue Systems: the KTH Connector. In proc. of the ISCA Tutorial and Research Workshop on Applied Spoken Language Interaction in Distributed Environments (ASIDE 2005) Aalborg, Denmark
5. Hjalmarsson, A.. & Wik, P. (2005): Is it blue?, Term paper, Course in NLP, GSLT, Sweden
6. Skantze, G. & Edlund, J. (2004). Robust interpretation in the Higgins spoken dialogue system. In proc. of ITRW on Robustness Issues in Conversational Interaction 2004.
7. Skantze, G: Galatea – a discourse modeller supporting concept-level error handling in spoken dialogue systems. In proc. of SigDial 2005
8. Edlund, J., House, D., & Skantze, G (2005): The effects of prosodic features on the interpretation of clarification ellipses. In proc. of Interspeech 2005, Lisbon, Portugal
9. Wallers, Å., Edlund, J., & Skantze (under review): Small sounds of great importance. Submitted to Perception and Interactive Technologies (PIT06), Kloster Irsee, Germany
10. Edlund, J & Heldner, M (2005). Exploring Prosody in Interaction Control. Phonetica, 62(2-4).

Location-Based Interaction with Children for Edutainment

Matthias Rehm, Elisabeth André, Bettina Conradi, Stephan Hammer, Malte Iversen, Eva Lösch, Torsten Pajonk, and Katharina Stamm

Multimedia Concepts and Applications, University of Augsburg, Germany
{rehm, andre}@informatik.uni-augsburg.de
http://www.interactive-multimedia.de

Abstract. Our mixed-reality installation features two cooperating characters that integrate multiple users by location-based tracking into the interaction, allowing for dynamic storylines.

1 Motivation

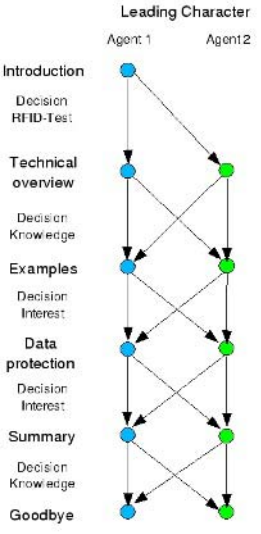
A typical situation for an edutainment scenario is an exhibition or museum event where a group of users can acquire knowledge in a playful way. In most cases, this experience is limited to the traditional on-screen edutainment game with a maximum of two players. We present an edutainment installation that combines robustness and simplicity with creativity and fun allowing a group experience for a large number of users without confronting them explicitely with computer equipment or restraining them to the seat in front of a computer screen.

Interacting with multiple users at the same time is a challenge at the level of the interface. Speech or gesture recognition techniques work fairly robust with a single user in a limited domain and under controlled conditions (e.g., [Hill et al., 2003]; [Latoschik, 2005]; [Nakano et al., 2003]). Multithreaded conversations with different users at the same time are beyond the scope of the current recognition systems. Thus, interactions with multiple users have to restrain themselves either to well-focused, strictly turn-based interactions (e.g., [Rehm and André, 2005]) or they have to come up with new ideas of interactions. In this paper, we present an application that renounces visible input devices but nevertheless allows a group of users to interact with two collaborating agents in an immersive and dynamic experience.

The second challenge is introduced by focusing on a still atypical user group. Generally, a computer user is an adult, in most user studies, e.g., a student. There are only some preliminary investigations on how children as a user group differ from adults in their interests and interaction habits. In the Victec project which features an interactive narrative approach against bullying behavior, [Hall et al., 2004] show that there are significant differences between adults and children in regard to their interests in the presented characters and the storyline. To capture the interaction habits of children with a hand-free interactive system, [Höysniemi and Hämäläinen, 2004] conducted a Wizard-of-Oz experiment interested in different ways to depict the action of swimming. Contrary to the

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 197–200, 2006.

expectations, children exhibited surprising creativity in this task and came up with 17 different ways to indicate the action. [Paiva et al., 2003] specifically targeted children as a user group and developed a tangible interface in the form of a sensor-equipped puppet. Manipulating the puppet allows for influencing the emotional state of the characters on screen and thus their behavior.



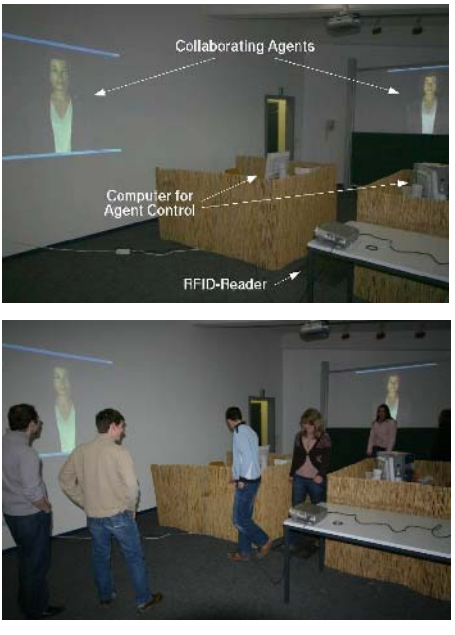


Fig. 1. Left: An overview of the dynamic eventflow between two agents which is based on user decisions. Right: The upper image shows the installation with the areas for the two agents. In the lower image the voting process at one decision point is shown.

2 System Features

In our installation, two agents collaborate in explaining the Radio Frequency IDentification (RFID) technology to a group of users. The two agents have different personalities with different interest. One agent, e.g., is more interested in the possibilities presented by the technology whereas the other one considers problems concerning data protection as crucial. During the performance, the users influence the storyline by a majority voting process making use of the same technique that is explained to them.

2.1 Invisible Input

Targeting groups of children as users, we searched for a possibility to abandon visible and explicit input devices as well as a direct confrontation with computer equipment. The metaphor used is that of taking the side of others and thus giving support to them. At each decision point, the users can support one of the agents. Consequently, the agent which gets the most support by the children

will get the floor and carries on with the presentation. The users movements are tracked by using RFID. RFID allows for reading the information on a chip from a distance. Users are equipped with passive RFID-tags which have a unique identification number. The reading devices are placed between the two locations of the Greta agents (Fig. 1 (right)). Thus, when a user changes place and decides to support the other agent, her RFID-tag is registered while crossing over to the other agent's area.

2.2 Dynamic Eventflow

To ensure the users interest even if she visits the installation more than once, the order of events is not fixed but established by a mix of user reactions, temporal constraints as well as randomization. Thus, a dynamic storyline enfolds during the demonstration. The story line is represented by finite state machines which are defined in XML similar to SceneMaker [Gebhard and Klesen, 2005]. The XML-file is parsed at the beginning to build the corresponding finite-state machine, allowing for unlimited revisions and re-designs of the application. A typical storyline consists of a greeting, followed by a short general introduction of the technique, elaborations on user specified topics interspersed with small knowledge tests. Figure 1 (left) gives an overview of the event flow (not the actual finite state machines).

2.3 Cooperative Characters

The two characters are spatially distributed and occupy different areas of a room. By positioning the two agents in different areas, users can actively support an agent by moving towards it. Creating two agents with different personalities makes the decision process of the users crucial to the unfolding storyline. The topics discussed as well as the focus on these topics depends on the agents personalities. Thus, the group members influence their experience by their decision for supporting one or the other agent. Figure 1 (left) shows that for each topic the role of the leading character is defined. The leading character is the one which is supported by the users. This character dwells on the current topic from its perspective. The other character is not rendered speechless during this time but comments on the other agent's performance bringing its own perspective on the topic into play.

The agents choose their turns from a library containing around 200 different dialogue acts. In one of the dialogues, e.g., agent one gives an account of a party where an RFID-transponder was injected subcutaniously for registering orders. The second agent claims the floor by listing the drinks the other agent had during the last two weeks, thus directing attention to the sensitive topic of data protection.

3 Concluding Remarks

By employing RFID as input device for group interactions we created a location-based edutainment installation that is targeted at children as a user group. By

moving towards the agents, users can support their preferred character in the application which makes the group interaction both simple and robust. A formal evaluation of the system is pending and will be conducted in April with around 120 schoolgirls during the German "GirlsDay 2006" event.[1]

References

[Gebhard and Klesen, 2005] Gebhard, P. and Klesen, M. (2005). Using Real Objects to Communicate with Virtual Characters. In Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P., and Rist, T., editors, *Intelligent Virtual Agents (IVA'05)*, pages 99–110. Springer, Berlin.

[Hall et al., 2004] Hall, L., Woods, S., Sobral, D., Paiva, A., Dautenhahn, K., Wolke, D., and Newall, L. (2004). Designing empathic agents: Adults versus kids. In Lester, J. C., Vicari, R. M., and Paraguacu, F., editors, *Intelligent Tutoring Systems*, pages 604–613. Springer, Berlin.

[Hill et al., 2003] Hill, R. W., Gratch, J., Marsella, S., Rickel, J., Swartout, W., and Traum, D. (2003). Virtual humans in the mission rehearsal exercise system. *KI – Künstliche Intelligenz*, (4):5–10.

[Höysniemi and Hämäläinen, 2004] Höysniemi, J. and Hämäläinen, P. (2004). Describing children's intuitive movements in a perceptive adventure game. In Martin, J.-C., Os, E. D., Kühnlein, P., Boves, L., Paggio, P., and Catizone, R., editors, *Multimodal Corpora: Models Of Human Behaviour For The Specification And Evaluation Of Multimodal Input And Output Interfaces*, pages 21–24.

[Latoschik, 2005] Latoschik, M. E. (2005). A User Interface Framework for Multimodal VR Interactions. In *Proceedings of the IEEE seventh International Conference on Multimodal Interfaces (ICMI)*.

[Nakano et al., 2003] Nakano, Y. I., Reinstein, G., Stocky, T., and Cassell, J. (2003). Towards a Model of Face-to-face Grounding. In *Proceedings of the Association for Computational Linguistics*, Sapporo, Japan.

[Paiva et al., 2003] Paiva, A., Costa, M., Chaves, R., Piedade, M., ao, D. M., Sobral, D., Höök, K., Andersson, G., and Bullock, A. (2003). Sentoy: an affective sympathetic interface. *International Journal of Human-Computer Studies*, 59(1–2):227–235.

[Rehm and André, 2005] Rehm, M. and André, E. (2005). Catch me if you can — Exploring lying agents in social settings. In *Proceedings of AAMAS 2005*, pages 937–944.

[1] http://www.girls-day.de

An Immersive Game - Augsburg Cityrun

Klaus Dorfmueller-Ulhaas, Dennis Erdmann, Oliver Gerl, Nicolas Schulz, Volker Wiendl, and Elisabeth André

Augsburg University, Eichleitnerstr. 30, 86159 Augsburg, Germany
dorfmueller-ulhaas@informatik.uni-augsburg.de
http://mm-werkstatt.informatik.uni-augsburg.de

Abstract. We present a platform for creating immersive 3D games including a new interface for navigating through virtual scenes. One innovative part of this application is a crowd simulation with an emergent behaviour of virtual characters. While the users has to move very quickly through a crowd of nearly two hundred virtual characters he is supported by a precise, fast operating, and unobtrusive navigation interface.

1 Introduction

A novel interaction technology is shown by Augsburg cityrun. While the user moves in front of a 3D projection screen, he is wearing shutter-glasses enabling a 3D impression (compare Fig. 1). Virtual characters autonomously move through

Fig. 1. *Left:* The user controls his navigation by bending his body. Wireless tracking is performed with infrared cameras to enable an unobtrusive interface. *Right:* Crowd simulation is done by enabling emergent behaviour of each character.

the inner city of Augsburg. The user‘s task in this game is to catch a specific pedestrian, but to avoid hitting others in a crowd within a close time limit. The player controls the continuation of his journey only by moving his upper body. An optical tracking system recognizes markers attached to the player's shutter-glasses and is thus able to capture the player's movements. The impression of immersion is achieved by combining the optical tracking system with the navigation control and a special 3D projection system (see Fig. 2) that adapts

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 201–204, 2006.

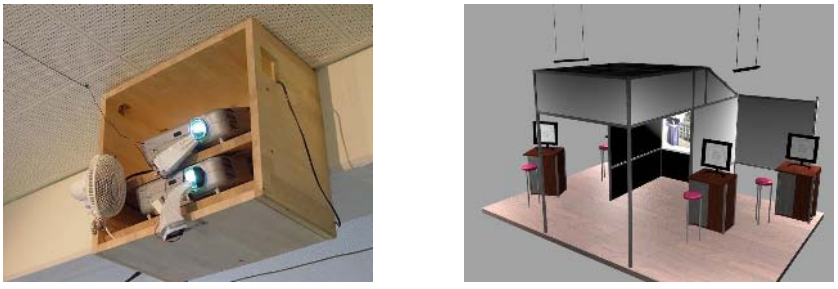


Fig. 2. *Left:* The shuttered stereo projection system. *Right:* The virtual model of the exhibit at the CeBit 2006 in Hannover.

itself dynamically to the viewpoint of the user in front of the projection screen. This feeling is enhanced by 3D sound effects. The complete gaming system is distributed among two independent computers. The first computer tracks the user's position by optical tracking. The data of the the player's eye positions is sent via UDP sockets to the game system. The receiver is responsible for doing the entire game processing, including sound and stereoscopic graphics rendering.

2 Crowd Simulation

In order to make the game more vivid and to avoid a blank and lifeless city, we populated the scene with lifelike and convincingly acting characters. For this reason a crowd simulation module has been implemented. The agents in this system are represented as particles similar to Reynolds famous boids [Reynolds, 1987]. Collision avoidance between characters is achieved by calculating physically based repulsion forces between particles, which are accumulated to a steering force determining the proximate position. [Heigeas et al., 2003] propose a three zone model for setting the strength of the repulsion forces. The resulting force function can be adjusted by the physical parameters stiffness and viscosity. We have adopted the model, but calculate these parameters automatically. In the process a continuous function has been created providing smoother movements of our agents. The realtime rendering of multiple animated characters is still a serious challenge. Many crowd visualization systems are using some sort of impostors [Aubel et al., 2000], [Dobbyn et al., 2005] to reduce the polygon count. Impostors are prerendered images that show a character with important animation keyframes from different camera angles. Usually they have a flat appearance, are very static, and can consume much texture memory. Since we are low on memory due to the large amount of texture data in the city scene we opted for a real 3D representation of characters in combination with skeletal animation. The skinning is done on the GPU with vertex shaders in order to get the best performance and keep the CPU free for other tasks. Another promising optimization that fits perfectly for a city scene is occlusion culling [Sekulic, 2004]. This GPU-accelerated technique helped us to determine which

characters are hidden behind buildings and hence don't have to be transmitted to the rendering pipeline.

3 User Navigation and Perspective Correction

Fast navigation through a crowd requires a very sensitive and accurate as well as unobtrusive and intuitive input system. There are systems using no markers for user tracking such as a system described in [Wren et al., 1997]. However, our application is highly dynamic and thus requires a navigation method that cannot be warranted by pure natural feature trackers. We therefore employ an optical stereoscopic tracking system that works similar to the approach described in [Dorfmüller, 1999]. The interaction between the virtual player's movement in the virtual world and the real user's position is done by a simple but efficient approach. We define a coordinate system in the real user's interaction area. The origin of this coordinate system is defined during the external camera calibration process. In case, the user stands close to the origin no movements in the virtual space are produced. Otherwise, the position of the user is converted into a polar coordinate system. The outcome of this procedure is an angle to one of the coordinate axes and the distance to the origin as shown in Fig. 3. The angle is used for the direction of movements and the distance constrols velocity. In case of a desktop VR system that generally uses a small display, it is assumed that the user is observing the scene close to an axis perpendicular to the center of the monitor screen. However, this assumption does not hold when viewing virtual objects on huge screens e.g. in CAVEs [Cruz-Neira et al., 1993]. Within our application the user is not restricted in his movements. In fact, he is motivated to perform large movements in front of the screen as depicted in Fig. 3. Therefore, a perspective correction is necessary and has been done by implementing an asymmetric viewing frustum. Furthermore, the scene becomes alive through three-dimensional sound effects emitted in relation to the user's position together with an asymmetric viewing perspective.

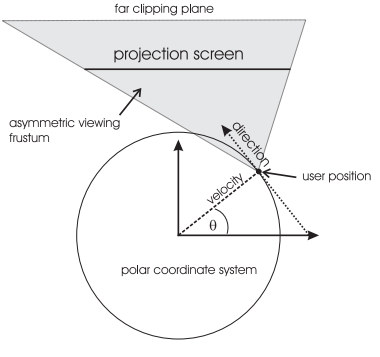


Fig. 3. User position given in polar coordinates controls navigation. The user position requires a perspective correction with an asymmetric viewing frustum.

4 Concluding Remarks

We described a system that consists of a 3D sound system, a 3D projection screen with shutter-glasses and an optical tracking system that transmits positional changes to the 3D world. Due to the optical tracking system, no cables are required to capture user movements, but only the markers attached to the glasses. We presented the system at the Open Lab event where more than one hundred people interacted with it. Even though the users had no experience with 3D tracker interfaces they were able to intuitively interact with the game without feeling hindered by obtrusive cables. The funny body poses of the players during the navigation and the witty comments of the virtual passers-by additionally contributed to the joy of playing. We hope to collect more user data when the complete system is demonstrated at CeBit 2006 in Hannover (see Fig. 2). The interface control in combination with real-time visualization and animation of virtual characters is so far unknown in game applications. Through the skilful combination of an optical tracking system with the navigation control and the projection system, a unique 3D experience is achieved.

References

[Aubel et al., 2000] Aubel, A., Boulic, R., and Thalmann, D. (2000). Real-time display of virtual humans: Levels of detail and impostors.

[Cruz-Neira et al., 1993] Cruz-Neira, C., Sardin, D., and Defanti, T. (1993). Surround-screen projection-based virtual reality: The design and implementation of the cave. In *SIGGRAPH 93 conference proceedings*, pages 135–142, Onahium. ACM Press.

[Dobbyn et al., 2005] Dobbyn, S., Hamill, J., O'Conor, K., and O'Sullivan, C. (2005). Geopostors: a real-time geometry / impostor crowd rendering system. In *SI3D '05: Proceedings of the 2005 symposium on Interactive 3D graphics and games*, pages 95–102, New York, NY, USA. ACM Press.

[Dorfmüller, 1999] Dorfmüller, K. (1999). An optical tracking system for vr/ar-applications. In Gervautz, M., Hildebrand, A., and Schmalstieg, D., editors, *Virtual Environment'99*, pages 33–42. Springer Verlag, Wien, New York.

[Heigeas et al., 2003] Heigeas, L., Luciani, A., Thollot, J., and Castagné, N. (2003). A physically-based particle model of emergent crowd behaviors. In *Graphicon*.

[Reynolds, 1987] Reynolds, C. W. (1987). Flocks, herds and schools: A distributed behavioral model. In *SIGGRAPH '87: Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 25–34, New York, NY, USA. ACM Press.

[Sekulic, 2004] Sekulic, D. (2004). Efficient occlusion culling. In Fernando, R., editor, *Gpu Gems: Programming Techniques, Tips and Tricks for Real-Time Graphics*. Addison Wesley Pub Co Inc.

[Wren et al., 1997] Wren, C. R., Azarbayejani, A., Darrell, T., and Pentland, A. (1997). Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785.

Gaze-Contingent Spatio-temporal Filtering in a Head-Mounted Display

Michael Dorr, Martin Böhme, Thomas Martinetz, and Erhardt Barth

Institute for Neuro- and Bioinformatics
University of Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany
{dorr, boehme, martinetz, barth}@inb.uni-luebeck.de
http://www.inb.uni-luebeck.de

1 Introduction

The spatio-temporal characteristics of the human visual system vary widely across the visual field. Recently, we have developed a display capable of simulating arbitrary visual fields on high-resolution natural videos in real time by means of a gaze-contingent spatio-temporal filtering [1]. While such a system can also be a useful tool for psychophysical research, our main motivation is to develop gaze-guidance techniques. Because the message an image sequence conveys depends on the exact pattern of eye movements an observer makes, we propose that in future information and communication systems, images will be augmented with a recommendation of where to look, of how to view them. Ultimately, we want to incorporate gaze guidance technology into mobile applications; such technology, integrated into a head-mounted display (HMD), could use computer vision techniques to enhance human visual performance.

In our demonstration, we will show a first implementation of such a device in the form of a system that implements our gaze-contingent spatio-temporal filtering algorithm in an HMD with video-see-through. Subjects will be able to walk around, seeing their natural visual environment inside the HMD. We will demonstrate that we then can manipulate what the subjects see in real time.

2 Gaze-Contingent Spatio-temporal Filtering

Visual acuity is highest only in the very centre of the human retina, the fovea, and drops off sharply towards the periphery. This variable-resolution effect has been simulated on image sequences in a gaze-contingent manner, e.g. [2].

Such "foveation" has typically been used to reduce the bandwidth required for video transmission or to improve the perceived video quality at a certain bandwidth [3]. Another application has been the simulation of visual field defects, e.g. to educate students or relatives of patients suffering from such defects. For an overview of applications of gaze-contingent displays in general, see e.g. [4].

Our work, however, focuses on the guidance of eye movements [5], and so we are interested in the effect that foveation has on the observer's eye movements. Based on the observation that movement or change in the visual periphery is a

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 205–207, 2006.

Fig. 1. Example of our spatio-temporal filtering algorithm. The effects in the middle and the right pictures can be combined in an arbitrary manner. Gaze position is indicated by the little white square to the left (below the white sail). Left: Still shot of the original movie. Middle: Spatial blur that is increased towards the visual periphery. Right: Temporal blur. Note the disappearance of the little girl to the right.

strong cue for eye movements, we have extended the techniques used by [2] such that they also filter in the temporal domain.

In the spatial case, the video sequence is filtered using an arbitrary two-dimensional map that specifies the desired spatial resolution at each pixel location relative to the direction of gaze. This is achieved by blending between levels of a multi-resolution pyramid computed for each frame in the video. The extension to the temporal domain is conceptually simple, but computationally challenging, because ultimately, each video frame that is displayed is a weighted sum of more than 250 high-resolution video frames surrounding it.

In experiments with our gaze-contingent system, we have indeed been able to show that eye movements are influenced by a temporal foveation. Furthermore, if the strength of the effect is kept below a certain threshold, it is not detected by the observer [6].

3 Head-Mounted Display

Our HMD is a custom-made model manufactured by a company that also sells off-the-shelf head-mounted displays (Trivisio), with eye-tracking capabilities fitted in collaboration with the SensoMotoric Instruments GmbH. For each eye, the

Fig. 2. Picture of our head-mounted display. Note the two scene cameras facing forward; two gaze cameras are integrated into the case.

HMD has a separate VGA input connected to a microdisplay with 800x600 pixels spatial and 60 Hz temporal resolution. Because our work focuses on the effect of visual stimulation in the periphery, the optics were chosen so that the field of view spans a wide 50 degrees of visual angle diagonally (40x30 degrees). Furthermore, four cameras are attached to the display. Two digital scene cameras (640x480 pixels spatial, up to 48 Hz temporal resolution) face forward, covering the field of view of the display. They can be connected to a computer via USB to enable a so-called "video see-through" mode, where the image sequence recorded by the cameras is immediately fed back to the displays. This setup allows us to perform arbitrary manipulations on the visual input in a highly natural environment. Two further cameras are placed in-between the microdisplays and the user's eyes and record the eye movements. Their analogue outputs are digitized using a framegrabber box connected to a computer via firewire and they allow for gaze tracking at 50 Hz with an accuracy of better than 1 degree.

Acknowledgments

Research has been supported by the German Ministry of Education and Research (BMBF) under grant number 01IBC01B with acronym ModKog. We thank SensoMotoric Instruments GmbH, Teltow, Germany, for their eye-tracking support.

References

1. Böhme, M., Dorr, M., Martinetz, T., Barth, E.: Gaze-contingent temporal filtering of video. In: Eye Tracking Research and Applications (ETRA). (2006) (in press).
2. Perry, J.S., Geisler, W.S.: Gaze-contingent real-time simulation of arbitrary visual fields. In Rogowitz, B.E., Pappas, T.N., eds.: Human Vision and Electronic Imaging: Proceedings of SPIE, San Jose, CA. Volume 4662. (2002) 57–69
3. Geisler, W.S., Perry, J.S.: A real-time foveated multiresolution system for low-bandwidth video communication. In Rogowitz, B., Pappas, T., eds.: Human Vision and Electronic Imaging: SPIE Proceedings. (1998) 294–305
4. Duchowski, A.T., Cournia, N., Murphy, H.: Gaze-contingent displays: A review. CyberPsychology & Behavior **7** (2004) 621–634
5. Barth, E., Dorr, M., Böhme, M., Gegenfurtner, K.R., Martinetz, T.: Guiding eye movements for better communication and augmented vision. In: Perception and Interactive Technologies. (same volume)
6. Dorr, M., Böhme, M., Martinetz, T., Barth, E.: Visibility of temporal blur on a gaze-contingent display. In: APGV 2005 ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization. (2005) 33–36

A Single-Camera Remote Eye Tracker

André Meyer, Martin Böhme, Thomas Martinetz, and Erhardt Barth

Institute for Neuro- and Bioinformatics, University of Lübeck
Ratzeburger Allee 160, D-23538 Lübeck, Germany
{meyera, boehme, martinetz, barth}@inb.uni-luebeck.de
http://www.inb.uni-luebeck.de

1 Introduction

Many eye-tracking systems either require the user to keep their head still or involve cameras or other equipment mounted on the user's head. While acceptable for research applications, these limitations make the systems unsatisfactory for prolonged use in interactive applications. Since the goal of our work is to use eye trackers for improved visual communication through gaze guidance [1, 2] and for Augmentative and Alternative Communication (AAC) [3], we are interested in less invasive eye tracking techniques.

In recent years, a number of so-called "remote" eye-tracking systems have been described in the literature (see e.g. [4] for a review). These systems do not require any equipment to be mounted on the user and allow the user to move their head freely within certain limits. The most accurate remote eye tracking systems that have been described in the literature to date use multiple cameras and achieve an accuracy of 0.5 to 1.0 degrees [5, 6, 7, 8, 9]. We need this level of accuracy for our applications but would ideally like to achieve it using a single fixed wide-field-of-view camera for reasons of cost and complexity. That single-camera systems can achieve an accuracy in the range of 0.5 to 1.0 degrees is demonstrated by a commercial system [10], but no implementation details have been published, and we are not aware of any similar system being described in the literature.

In this paper, we will describe a single-camera remote eye tracking system we are working on that is designed to achieve an accuracy in the range of 0.5 to 1.0 degrees.

The hardware of the eye tracker consists of the following components: (i) A single calibrated high-resolution camera (1280x1024 pixels), (ii) two infrared light-emitting diodes (LEDs) to either side of the camera that illuminate the face and generate corneal reflexes (CRs) on the surface of the cornea, and (iii) a display located above the camera and the LEDs (see Fig. 1).

The eye tracking software consists of two main components: The image processing algorithms that are used to extract the location of the pupils and corneal reflexes from the image and the gaze estimation algorithm that is used to calculate the position the user is fixating on the screen.

These two components were developed independently and have recently been integrated into a running system, which will be shown at the workshop. No accuracy measurements have been made yet on the complete system, but tests

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 208–211, 2006.

Fig. 1. Physical setup of the eye tracking hardware, consisting of a single high-resolution camera below the display and two infrared LEDs to either side of the camera

on simulated data show the gaze estimation algorithm can achieve an accuracy of one degree or better.

2 Algorithms

Due to space constraints, we can only give a brief outline of the image processing and gaze estimation algorithms but will demonstrate the functionality of the system at the workshop.

The algorithm used to extract the locations of the pupils and CRs from the camera image is based on the Starburst algorithm [11], which was modified to fit the needs of the remote eye tracking setting. An open source implementation of this algorithm is available under the name "openEyes", but we chose to reimplement the algorithm for better integration within our existing computer vision and eye tracking framework.

In general terms, the pupil and CR location extraction algorithm proceeds as follows: The locations of the CRs are extracted by applying a difference of Gaussians and searching for maxima; the approximate pupil centre is determined as the darkest pixel in the vicinity of the CRs; contour points are identified on rays shot from the centre of the pupil as well as on secondary rays shot from these primary contour points; and an ellipse is fitted to the these contour points.

The gaze estimation algorithm is based on a physical model of the eye. Figure 2 shows the eye model, which models the following components of the eye:

- The surface of the cornea. This is modelled as a spherical surface with a centre of curvature CC and a curvature radius of r_{cornea}. The corneal surface plays a role in two effects that are relevant for eye tracking: First, the corneal reflexes are generated by reflections of the infrared LEDs at the corneal surface; and second, the image of the pupil that we observe through the cornea is distorted by refraction at the corneal surface.

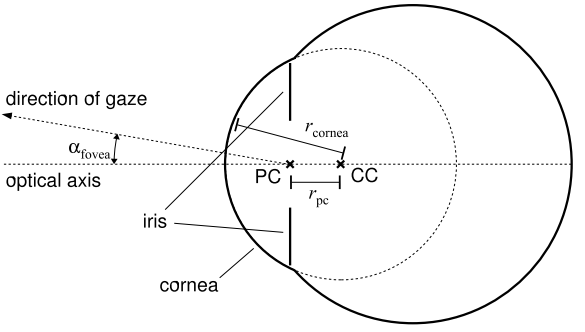


Fig. 2. Eye model used for gaze estimation. PC: Pupil centre. CC: Cornea centre. The model contains three user-dependent parameters α_{fovea}, r_{cornea} and r_{pc} (see text for an explanation).

- The pupil centre, referred to as PC. This is the point located at the centre of the pupil disc (we make the idealized assumption that the pupil is perfectly circular). The pupil centre PC is a certain distance from the centre of corneal curvature CC, and we call this distance r_{pc}.
- The angular offset between the optical axis of the eye and the direction of gaze (referred to as α_{fovea}), which is caused by the fact that the fovea does not lie on the optical axis but is offset temporally and slightly upwards (at the moment, we only model the horizontal component of this offset).

Given the position and orientation of an eye relative to the camera, the eye model predicts where the pupil and the CRs should be observed in the camera image. The inverse problem can also be solved, allowing us to determine the direction of gaze from the pupil and CR position. The values of the model parameters for a particular user are determined by asking the user to fixate a series of calibration points and finding the set of parameter values that best explain the observations.

3 Results

The gaze estimation component was tested individually on simulated test data. To mimic the effects of camera noise, a certain amount of random error was added to the measurements of pupil centre and CR position.

Table 1 shows the eye model parameters that were estimated for varying amounts of camera error, along with the resulting gaze estimation error. Assuming a maximum measurement error of half a pixel, which we believe our image processing algorithms can achieve, we obtain a gaze estimation error of about one degree.

To date, no accuracy measurements have been made on the complete system; we plan to present these results at the workshop.

Table 1. Results of testing gaze estimation algorithm on simulated data

camera error	r_{cornea}	r_{pc}	α_{fovea}	mean gaze error
0.0 px	7.93 mm	4.52 mm	5.98 °	0.22 °
0.1 px	7.53 mm	4.28 mm	6.16 °	0.22 °
0.3 px	7.53 mm	4.26 mm	6.15 °	0.59 °
0.5 px	7.34 mm	4.12 mm	6.20 °	1.04 °
1.0 px	6.94 mm	3.84 mm	6.36 °	2.22 °
2.0 px	6.04 mm	3.24 mm	7.07 °	5.03 °
ground truth	7.98 mm	4.44 mm	6.00 °	

Acknowledgments

Research was supported by the German Ministry of Education and Research (BMBF) under grant number 01IBC01 with acronym *ModKog*.

References

1. Itap: Information technology for active perception website (2002) `http://www.inb.uni-luebeck.de/Itap/`.
2. Barth, E., Dorr, M., Böhme, M., Gegenfurtner, K.R., Martinetz, T.: Guiding eye movements for better communication and augmented vision. (same volume)
3. COGAIN: (Network of excellence on communication by gaze interaction) `http://www.cogain.org`.
4. Morimoto, C.H., Mimica, M.R.M.: Eye gaze tracking techniques for interactive applications. Computer Vision and Image Understanding **98** (2005) 4–24
5. Yoo, D.H., Chung, M.J.: A novel non-instrusive eye gaze estimation using cross-ratio under large head motion. Computer Vision and Image Understanding **98** (2005) 25–51
6. Shih, S.W., Wu, Y.T., Liu, J.: A calibration-free gaze tracking technique. In: Proceedings of the 15th International Conference on Pattern Recognition. (2000) 201–204
7. Ohno, T., Mukawa, N.: A free-head, simple calibration, gaze tracking system that enables gaze-based interaction. In: Eye Tracking Research and Applications (ETRA). (2004) 115–122
8. Brolly, X.L.C., Mulligan, J.B.: Implicit calibration of a remote gaze tracker. In: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW '04). Volume 8. (2004) 134
9. Beymer, D., Flickner, M.: Eye gaze tracking using an active stereo head. In: Proceedings of Computer Vision and Pattern Recognition (CVPR). Volume 2. (2003) 451–458
10. Tobii: (Tobii 1750 eye tracker, Tobii Technology AB, Stockholm, Sweden) `http://www.tobii.se`.
11. Li, D., Winfield, D., Parkhurst, D.J.: Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In: Proceedings of the IEEE Vision for Human-Computer Interaction Workshop at CVPR. (2005) 1–8

Miniature 3D TOF Camera for Real-Time Imaging

Thierry Oggier, Felix Lustenberger, and Nicolas Blanc

CSEM SA, Zurich Center, 8048 Zurich, Switzerland
thierry.oggier@csem.ch
http://www.csem.ch

Abstract. In the past, measuring the scene in all three dimensions has been either very expensive, slow or extremely computationally intensive. The latest progresses in the field of microtechnologies enable the breakthrough for time-of-flight (TOF) based distance-measuring devices. This paper describes the basic principle of the TOF measurements and a first specific implementation in a state-of-the-art 3D-camera "SwissRanger SR-3000" [T. Oggier et al., 2005]. Acquired image sequences will be presented as well.

1 Introduction

The visionary dream of providing operators the capability of interacting without using any tactile input devices or other actuators seems to be a rather futuristic scenario. Most approaches so far have failed due to unusable real-world representation by the sensing devices. However, thanks to the latest progresses achieved in the field of optical distance measurements, state-of-the-art 3D-camera systems seem to fill this gap perfectly.

2 Principle of the TOF Distance Measurement

In theory, each TOF system contains essentially two components: a light source and a detector unit [T. Oggier et al., 2004] . The light source synchronously emits an intensity-modulated wave. This wave propagates from the TOF camera system to the scene and is reflected by the scene back to the sensor where the sensor captures its time of flight. By knowing the time for the light to travel to the object and back to sensor, the distance of the target can be deduced.

In practical implementations, imaging TOF systems make use of low-frequency modulation techniques instead of applying pure pulse methods that require extremely high bandwidth. E.g. in homodyne systems, the modulation frequency of the light source is sinusoidally modulated, synchronous to the demodulating imager. The image sensor is capable of making a demodulation of the optical field that is reflected from the objects in the scene back to the camera. Thus, the image sensor contains an array of lock-in pixels that can derive the phase information in each pixel. In the general implementation, the phase information is obtained by sampling four times the incoming signal within one modulation period.

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 212–216, 2006.

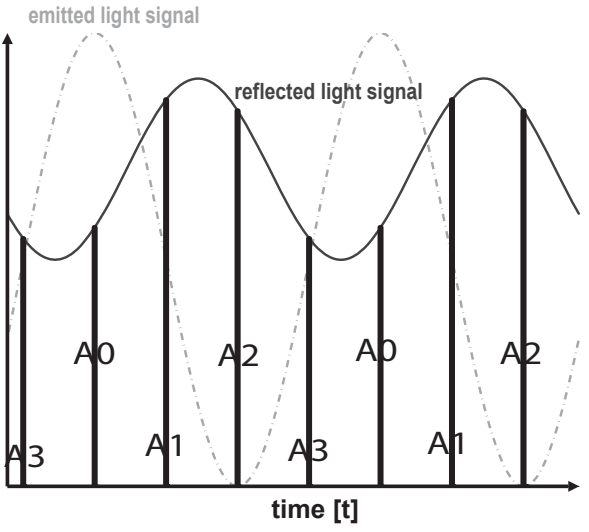


Fig. 1. Illustration of the four sampling process

Figure 1 illustrates the emitted and the detected sine waves. The four samples A_0, A_1, A_2 and A_3 allow a non-ambiguous and complete determination of the incoming sine wave. Finally, the phase difference is directly related to the target's distance.

3 State-of-the-Art 3D Camera SwissRanger SR-3000

The latest version in the SwissRanger 3D camera family is the so-called SR-3000 camera [SwissRanger, 2006]. The camera has been manufactured with the goal of serving different applications and industries, varying from human-machine interaction, safety and security to mobile robotics applications.

The pixel resolution of the SR-3000 camera is QCIF (174 × 144 pixels). For each pixel, the camera outputs the coordinates x, y, z and the intensity i. All pixels can inherently suppress background light illumination. The SR-3000 camera is modularly set up. Each printed circuit board (PCB) has its specific tasks and can therefore easily be replaced and adjusted to the customer's specification. E.g. to bottom board is dedicated to the interface. In the standard SR-3000 camera version, this board supports a USB2.0 interface. However, it can easily be replaced by other interfaces such as Ethernet, FireWire or CameraLink. A photograph of the SR-3000 camera is shown in Figure 2.

The SR-3000 camera's body size is $50 \times 67 \times 42.3\,\mathrm{mm}^3$ and it weighs 162 g. The acquisition speed is mainly limited by the reflected amount of light. For short ranges of 1 to 2 meters, the frame rates reaches up to 30 complete frames per second. For acquiring ranges of a few meters, usually frame rates of 20 frames per second can be achieved.

Fig. 2. Photograph of the SR-3000 3D-camera

The distance resolution achieved with the SR-3000 camera is dependant on the frame rate, e.g. exposure time, the target's reflectivity and distance. Table 1 shows some corner measurement and the corresponding depth resolution.

Table 1. Overview of corner measurements

Distance resolution [a]				
Distance [m] [b]	0.3	1	2	3
Frame Rate [Hz]	29	20	15	12
Distance Resolution [mm]	2.5	6	13	22

[a] Acquisition at room temperature; target with 90% reflectivity, 20 MHz modulation frequency
[b] At full image frame

4 Application Examples

4.1 SR-3000 for Head Tracking

A first use of the SR-3000 shown is head tracking. In the illustrated example, full frame acquisitions are performed at 30 frames per second. Due to the highly accurate and robust distance information, facial characteristics such as the nose can easily be determined and located. Figure 3 sketches a captured face in different views.

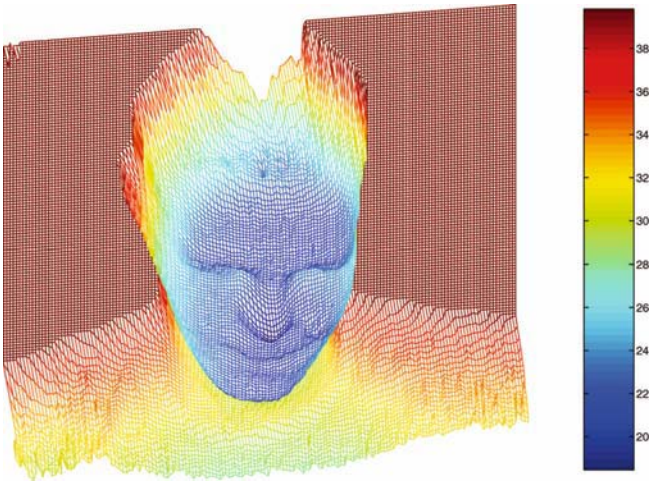


Fig. 3. Head tracking example

4.2 SR-3000 for People Tracking

The real-time distance information allows an easy tracking of people. Such features can be used e.g. for interactive gaming, service robots controlling [N. Blanc et al., 2004] and even for surveillance and safety applications. For people tracking, usually ranges of a few meters are required. Figure 4 illustrates a typical example of the three-dimensional scene acquired by the SR-3000 camera for people tracking. The full frame rate for these example measurements is at about 20 frames per second.

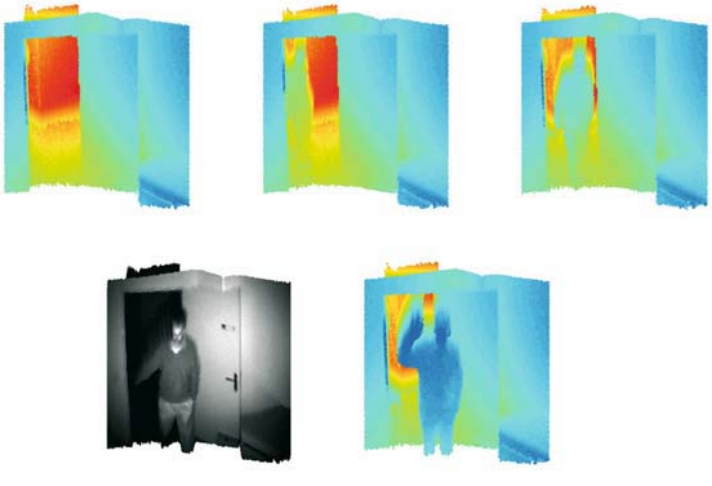


Fig. 4. Sample sequence illustrating color-coded 3D-scene for e.g people tracking

5 Summary and Conclusion

The state-of-the-art 3D camera SR-3000 enables completely new approaches for the development of peripheral devices. The availability of the third dimension of the operator renders the tracking task much easier while at the same time keeps the required computational cost at a minimum level.

References

[N. Blanc et al., 2004] N. Blanc et al. (2004). Miniaturzed smart camera for 3d-imaging in real time. *Sensors, Proc. IEEE*, 1:471–474.

[SwissRanger, 2006] SwissRanger (2006). http://www.swissranger.ch.

[T. Oggier et al., 2004] T. Oggier et al. (2004). An all-solid-state optical range camera for 3d real-time imaging with sub-centimetre depth resolution (swissranger). *Proc. SPIE*, 5249:534–545.

[T. Oggier et al., 2005] T. Oggier et al. (2005). Swissranger sr3000 and first experiences based on miniaturized 3d-tof cameras. *ETH 1st RIM Days*.

Author Index

André, Elisabeth 40, 53, 197, 201

Bachmayr, Franziska 175
Barth, Erhardt 1, 205, 208
Bausch, Tobias 179
Bayerl, Pierre 9, 65, 179
Bee, Nikolaus 40
Bernsen, Niels Ole 129
Bevacqua, Elisabetta 164
Blanc, Nicolas 212
Böhme, Martin 1, 205, 208
Buchholz, Malte 73

Caridakis, George 164
Carlson, Rolf 193
Conradi, Bettina 197

Dorfmueller-Ulhaas, Klaus 201
Dorr, Michael 1, 205

Edlund, Jens 183, 193
Erdmann, Dennis 201

Fürstenau, Norbert 20

Gegenfurtner, Karl 1
Gerl, Oliver 201
Giese, Martin A. 188
Gress, Thomas M. 73

Hagen, Eli 97, 107
Hammer, Stephan 197
Hawes, Nick 117
Hof, Alexander 97, 107
Hoffmann, Holger 175

Ishizuka, Mitsuru 40
Iversen, Malte 197

Karpouzis, Kostas 164
Kelleher, John D. 117
Kessler, Henrik 175
Kestler, Hans A. 73
Kim, Jonghwa 53
Kolesnik, Marina 32
Kruijff, Geert-Jan M. 117

Layher, Georg 9
Lösch, Eva 197
Lustenberger, Felix 212

Machrouh, Joseph 152
Mahler, Thorsten 65
Mancini, Maurizio 164
Martinetz, Thomas 1, 205, 208
Meyer, André 208
Müller, Andre 73

Nakasone, Arturo 40
Neumann, Heiko 9, 65, 179

Oberhoff, Daniel 32
Oggier, Thierry 212
Omlor, Lars 188

Pajonk, Torsten 197
Palm, Günther 73
Panaget, Franck 152
Pelachaud, Catherine 164
Peters, Christopher 164
Prendinger, Helmut 40

Raouzaiou, Amaryllis 164
Ratzka, Andreas 141
Rehm, Matthias 197

Schulz, Nicolas 201
Skantze, Gabriel 183, 193
Song, Dongyi 85
Stamm, Katharina 197
Stynen, Andy 32

Traue, Harald C. 175

Wallers, Åsa 183
Weber, Michael 65
Weidenbacher, Ulrich 9
Wiendl, Volker 201