Proceedings of GW2007 - 7th International Workshop on Gesture in Human-Computer Interaction and Simulation 2007 – POSTER SESSION

Editors
Miguel Sales Dias - Microsoft & ADETTI/ISCTE, Portugal
Ricardo Jota - IST/Technical University of Lisbon, Portugal

Editorial Production
Ana Rita Leitão, ADETTI, Portugal

Online Proceedings:
http://www.adetti.pt/

Logo and Cover Credits
Ana Rita Leitão – ADETTI

Local Organiser
ADETTI - Associação para o Desenvolvimento das Telecomunicações e Técnicas de Informática
Avenida das Forças Armadas, Edifício ISCTE
1600-082 Lisboa, PORTUGAL
Tel: +351 21 7826480 Fax: +351 21 7826488
www.adetti.pt
(Printed copies can be ordered from ADETTI)

Sponsors

ADETTI - Associação para o Desenvolvimento das Telecomunicações e Técnicas de Informática, Grupo Português de Computação Gráfica, ISCTE – Instituto Superior de Ciências do Trabalho e da Empresa, Fundação para a Ciência e Tecnologia, Associação de Turismo de Lisboa, Springer

ISBN: 978-972-8862-06-0
May 2007

Foreword

The International Gesture Workshop is an interdisciplinary event where researchers working on human gesture-based communication present and exchange ideas and advanced research currently in progress, on gesture across multi-disciplinary scientific disciplines. This workshop encompasses all fundamental aspects of gestural studies in the field of Human-Computer Interaction and Simulation, including all multifaceted issues of modelling, analysis and synthesis of human gesture, encompassing hand and body gestures and facial expressions. A focus of these events is a shared interest in using gesture in the context of sign language analysis, understanding and synthesis. Another stream of interest is the user centric approach of considering gesture in multimodal human-computer interaction, in the framework of the integration of such interaction into the natural environment of users. In addition to welcoming submission of work by established researchers, it is the tradition of the GW series of workshops to encourage submission of student work at various stages of completion, enabling a broader dissemination of finished or on-going novel work, and the exchange of experiences in a multi-disciplinary environment.

Submissions include papers, posters and demonstrations.

GW2007 is the 7th European Gesture Workshop in the GW series initiated in 1996. Since this event, the Gesture Workshops have been held roughly every second year, with fully reviewed post-proceedings typically published by Springer-Verlag.

In GW2007 53 contributions were received, of which were accepted 15 full papers,16 short papers and 10 as posters and demos.

Two brilliant key-note speakers honoured the event with their presentations: Dr. Andrew Wilson, member of the Adaptive Systems and Interaction group at Microsoft Research, and Prof. Joaquim Jorge, Associate Professor of Computer Science at Instituto Superior Técnico (IST/UTL), the School of Engineering of the Technical University of Lisboa, Portugal

Miguel Sales Dias
Workshop Chair

Workshop Chair
Miguel Sales Dias - Microsoft & ADETTI/ISCTE, Portugal

Poster and Demo Chair
Ricardo Jota - IST/Technical University of Lisbon, Portugal

Local Organising Committee
Ana Rita Leitão - ADETTI, Portugal
Ricardo Jota - IST/Technical University of Lisbon, Portugal
Rafael Bastos - ADETTI, Portugal

International Programme Committee
Annelis Braffort - LIMSI, Orsay, France
António Augusto Sousa - INESC Porto/FEUP
Antonio Camurri-InfoMus Lab, DIST, University of Genova, Italy
Bencie Woll - Sign Language and Deaf Studies / UCL DCAL Research Centre, UK
Christian Vogler-Washington University, USA
Daniel Arfib - CNRS-LMA, Marseille, France
David McNeill - Center for Gesture and Speech Research, university of Chicago, USA
Gildas Ménier - VALORIA, University of Bretagne Sud, France
Gualterio Volpe-InfoMus Lab, DIST, University of Genova, Italy
Héloir Alexis -VALORIA, University of Bretagne Sud, France
Hermann Ney - Aachen University, Germany
Ipke Wachsmuth - Bielefeld University, Germany
Isabel Hub Faria - Faculdade de Letras da Uni. de Lisboa, Portugal
Jean-François Kamp - VALORIA, University of Bretagne Sud, France
Joaquim Jorge - Instituto Superior Técnico, Portugal
Joaquim Madeira - DETI-IEETA/UA, Portugal
Jorge Salvador Marques - IST/ISR, Portugal
José Manuel Rebordão - INETI & FCUL, Portugal
Leonel Valbom- UM, Portugal
Manuel João Fonseca-IMMI – INESC-ID; Portugal
Manuel Próspero dos Santos - FCT/UNL, Portugal
Marcelo Wanderley - McGill University, Canada
Marianne Gullberg - Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
Miguel Sales Dias - Microsoft & ADETTI/ISCTE, Portugal
Nicolas Courty - VALORIA, University of Bretagne Sud, France
Nuno Correia - DI/FCT/UNL, Portugal
Peter Wittenburg - Max-Planck Institute of Psycholinguistics, Nijmegen, The Netherlands
Philippe Gorce - LESP, Université de Toulon et du Var, France
Pierre-François Marteau - VALORIA, University of Bretagne Sud, France
Richard Kennaway - Norwich, United Kingdom
Ronan Boulic - Virtual Reality Lab, EPFL, Switzerland
Seong-Whan Lee - Korea University, Korea
Sylvie Gibet - Valoria, Université de Bretagne Sud, France
Teresa Chambel - FC/UL, Portugal
Timo Sowa - Bielefeld University , Germany
Winand Dittrich - University of Hertfordshire, United Kingdom
Ying Wu - Northwestern University, Evanston, USA

Poster Session

Poster Session

A VR-based virtual hand system for experiments on perception of manual gestures

Sebastian Ullrich[1], Jakob T. Valvoda[1], Marc Wolter[1],
Gisela Fehrmann[2], Isa Werth[2], Ludwig Jäger[2], and Torsten Kuhlen[1]

[1] Virtual Reality Group, RWTH Aachen University
[2] SFB/FK 427 “Media and Cultural Communication”, University of Cologne
s.ullrich@rz.rwth-aachen.de

1 Introduction

This paper contributes an integrated and flexible pipeline approach to sign language processing in Virtual Reality (VR), that allows for interactive experimental evaluations with high ecological validity. Initial steps deal with real-time tracking and processing of manual gestures. Hand shape and motion data is rendered in virtual environments with varying spatial and representational configurations. Besides flexibility, the most important aspect is the seamless integration within a VR-based neuropsychological experiment software.

2 Flexible pipeline approach

We propose a system architecture that covers all prerequisites to allow interactive experimental evaluations of sign language in virtual environments. With a single system, reusability and flexibility is limited. Therefore, our approach is to decompose the workflow into a pipeline with reusable components: data input, model, representation and experimental platform.

The first component is an arbitrary hardware or software source producing **data input** (e. g., modeling, motion capturing, synthesized gestures, kinematic algorithms). This includes automatic or manual post-processing and mapping of the data to a model. The **model** consists of a complete description of a time-varying hierarchical structure based on human anatomy. It combines the data from an arbitrary number of data inputs to provide a consistent and complete

Fig. 1. Representations in VR for intuitive interaction and experiments.

model. For the **representation** to a human user the structural model must be visualized. Several techniques can be chosen to vary the degree of realism or to change the appearance. Thus, the same structural model can be visualized differently, e. g., as puristic dots, as stick-figure or as realistic human model. The **experimental platform** adds the semantic context for sign language experiments. Here, the representations are integrated into interactive virtual environments according to a user-defined experimental design. This includes the specification of a chain of events structured into sessions, blocks, and trials, of analysis parameters and the interaction possibilities of the human user. The last step of the workflow is the **user**, who is able to interact with the experimental platform (e. g., reacting to stimuli) or with the data inputs directly.

For our implementation we have utilized VRZula [1] and ReactorMan [2].

3 Results

The pipeline and components have been successfully utilized in different setups, e. g., to build a motion database for perception experiments. Real-time visualizations have been displayed during recording with a tracked data glove, which allowed to find errors and adjust calibration instantly. In a post-processing session the recorded gestures have been edited and cut. Inverse kinematics has been used to animate the upper limbs from shoulder to the hand. The signing space was adjusted interactively by moving the hand position of the looping gestures whilst the arm automatically aligned accordingly. Over 100 items from German Sign Language have been recorded, categorized, adjusted and stored in a database. In addition, 100 non-signs have been conceived and created. The VR-setup allows to conduct the experiments with stereoscopic projection with proper depth perception and contributes to the validity of the experiments.

4 Conclusion & Future Work

This paper presents a system for processing of sign language for experimental evaluation in virtual environments. A pipeline has been proposed to reach flexibility in terms of data input during acquisition of signs. It allows for multimodal and immersive representation of the signs and adds an integrated platform for controlled experiments. The system has proven its capability in first studies and enables a large variety of experiments in VR that will contribute to the understanding of sign language processing.

References

1. Valvoda, J.T., Kuhlen, T., Bischof, C.: Interactive Virtual Humanoids for Virtual Environments. In: Eurographics Symposium on Virtual Environments. (2006) 9–12
2. Valvoda, J.T., Kuhlen, T., Wolter, M., et al.: NeuroMan: A Comprehensive Software System for Neuropsychological Experiments. CyberPsychology and Behaviour **8**(4) (2005) 366–367

Towards a Video Corpus for Signer-Independent Continuous Sign Language Recognition

Ulrich von Agris and Karl-Friedrich Kraiss

Institute of Man-Machine Interaction, RWTH Aachen University, Germany
{vonagris,kraiss}@mmi.rwth-aachen.de

Research in the field of continuous sign language recognition has not yet addressed the problem of interpersonal variance in signing. Applied to signer-independent tasks, current recognition systems show poor performance as their training bases upon corpora with an insufficient number of signers. In contrast to speech recognition, there is no benchmark which meets the requirements for signer-independent recognition.

Although signer-independence is an essential precondition for future applications such as translation systems, only little investigations have been made in this field so far. This unexplored gap is the subject of a current research project which aims for achieving signer-independence in continuous sign language recognition. For this purpose a new video corpus containing articulations of a large number of signers will be recorded.

System Overview Following sign language recognition system constitutes the basis for our ongoing research work. A detailed description is given in [1,2]. The system utilizes a single video camera for data aquisition to ensure user-friendliness. In order to cover all aspects of sign languages sophisticated algorithms were developed that robustly extract manual and facial features, even in uncontrolled environments. The classification stage is based on hidden Markov models and allows recognition of continuous sign language. For statistical modeling of reference models each sign is represented either as a whole or as a composition of smaller subunits – similar to phonemes in spoken languages.

Related Work Similar to the early days of speech recognition, most papers focus on the recognition of isolated signs. Only a few recognition systems were reported that can process continuous signing. Here most research was done within the signer-dependent domain, i.e. every user is required to train the system himself before being able to use it. Altogether only three corpora [3,4] reported in literature comprise sentences articulated by more than one signer. However, these databases are of limited use since they do not sufficiently cover interpersonal variance. Either the training population is to small or the corpus includes too many signs that occur only once or twice in the whole dataset.

Video Corpus In the following some details about the new video corpus are presented. The corpus' content was already specified, but recordings are still in progress and will be finished within the next months. After the project the whole database will be made available for interested researchers in order to establish the first benchmark for signer-independent continuous sign language recognition.

The vocabulary comprises 450 signs in German Sign Language (DGS) representing eight different word types such as nouns, verbs, adjectives and numbers. Those signs were selected which meet the following criteria: They should occur most frequently in everyday conversation and should not be dividable into smaller signs. In certain respects these signs can be regarded as some kind of basic signs. For the selection several books and visual media commonly used for learning DGS were evaluated.

All 450 signs are different with regard to their manual parameters. However, similar to other sign languages, many of them can be combined with several facial expressions leading to a multitude of different meanings. For example, the signs TECHNIK (ENGINEERING) and POLITIK (POLICY) are identical with respect to gesturing and can only be distinguished by their non-manual parameters. Therefore 226 additional signs with a different facial expression were selected and integrated into the database.

Furthermore, some of the 450 basic signs can be concatenated for creating new signs with a different meaning. For example, the sign ZAHNARZT (DENTIST) is composed of the two basic signs ZAHN (TOOTH) and ARZT (PHYSICIAN). According to this concept 124 composed signs were collected and integrated as well.

Altogether 800 different meanings can be expressed with the selected vocabulary of 450 basic signs. For continuous recognition overall 780 sentences were constructed. All sentences are meaningful and grammatically well-formed. There are no constraints regarding a specific sentence structure. Each sentence ranges from two to eleven signs in length. No intentional pauses are placed between signs within a sentence, but the sentences themselves are seperated. The annotation follows the specifications of the Aachener Glossenumschrift, which was developed by the Deaf Sign Language Research Team (DESIRE) at the RWTH Aachen University.

For modeling interpersonal variance in articulation each sentence will be performed by several signers. The number of signers must be chosen in such a way that variability is sufficiently represented within the corpus. Influencing factors on the articulation have to be explored and taken into consideration during the casting period. For the moment we will start recording with 20 native signers of different sexes and ages. Therefore a total of 15.600 articulated sentences will be stored in the new database.

Since we use a vision-based approach for sign language recognition the corpus will be recorded on video. In order to facilitate feature extraction recordings are conducted under laboratory conditions, i.e. controlled environment with diffuse lighting and a unicolored blue background. All video clips are recorded on hard disk using an image resolution of 780×580 pixels at 30 fps. This high spatial resolution ensures reliable extraction of manual and facial features from the same input image.

References

1. Kraiss, K.F., ed.: Advanced Man-Machine Interaction. Springer (2006)
2. von Agris, U., Zieren, J., Canzler, U., Bauer, B., Kraiss, K.F.: Recent developments in visual sign language recognition. Springer Journal on Universal Access in the Information Society, "Emerging Technologies for Deaf Accessibility in the Information Society" (to appear, 2007)
3. Fang, G., Gao, W., Chen, X., Wang, C., Ma, J.: Signer-independent continuous sign language recognition based on srn/hmm. In: International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction, Springer (2002) 76–85
4. Zahedi, M., Dreuw, P., Rybach, D., Deselaers, T., Ney, H.: Continuous sign language recognition - approaches from speech recognition and available data resources. In: Second Workshop on the Representation and Processing of Sign Languages. (2006)

An Analytic Approach for Optimal Hand Gestures

Helman Stern, Juan Wachs, and Yael Edan

Department of Industrial Engineering and Management,
Ben-Gurion University of the Negev, Beer Sheva, 84105, Israel
{helman, juan, yael)@bgu.ac.il

Hand gestures to control systems require accurate recognition, high learnability, usability, ergonomic design and comfort. Unfortunately, most gesture interfaces are designed with the technical consideration of recognition accuracy as the central focus. In this research we consider hand gestures that are unencumbered, i.e.; can be captured with camera vision devices. The selection of a set of hand gestures that consider both recognition accuracy as well as the ease of learning, lack of stress, cognitively natural and ease of implementation is still an open research question.

An optimal hand gesture vocabulary, GV, is defined as a set of gesture-command pairs, such that it will minimize the time τ for a given user (or users) to perform a task, (or collection of tasks). One of the major problems is how to select the best subset of gestures from a very large number of possible gesture configurations. Another problem is that task completion time, as a function of GV, has no known analytical form. We thus, propose three different performance measures as proxies: intuitiveness Z_1(GV), comfort Z_2(GV) and recognition accuracy Z_3(GV). Maximizing each of the measures over the set of all feasible GVs defines a multiobjective optimization problem (MOP). For the MOP, a set of Pareto solutions can be used to aid the decision maker to select the GV according to his own desires. Because finding the solutions the MOP requires a large amount of computation time, an analytical methodology is proposed in which the MOP is relaxed to a dual priority objective problem. Here recognition accuracy is considered of prime importance, while the human performance objectives are secondary. The optimal GV methodology architecture is comprised of four modules (Fig. 1). In Module 1 human psycho-physiological input factors are determined. In Module 2 gesture subsets, subject to machine gesture recognition accuracy are determined. Module 3 constitutes a command - gesture matching procedure, and Module 4 finds the Pareto solutions.

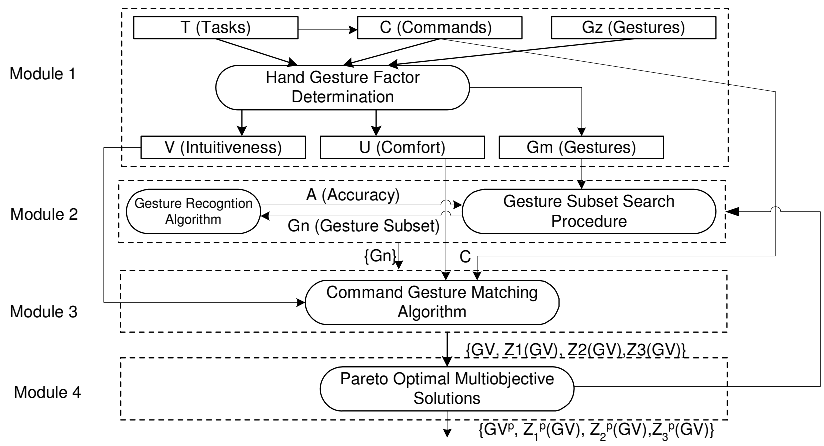

Fig. 1. Architecture of optimal hand gesture vocabulary

The task set T, the large gesture master set G_z and a set of commands C of size n are the input parameters to the first Module 1. The objectives of Module 1 are to find a user intuitiveness (direct and complementary) matrix, a comfort matrix based on command transitions and fatigue measures, and to reduce the large set of gestures, to the master set G_m. For Module 2, the necessary inputs are the reduced master set of gestures G_m, and a recognition algorithm to determine the gesture recognition accuracy, A. Five real life instances per gesture type, performed by each of eight subjects, were used in this Module. This module employs an iterative search procedure to find a set of feasible gesture subsets, satisfying a given pre specified accuracy level. For this, a meta- heuristic search called "confusion matrix derived solution" (CMD) is proposed. The method is initiated by finding the accuracy of the gesture master set G_m and its confusion matrix $\mathcal{C}_m$ .From $\mathcal{C}_m$ recognition accuracies are extracted that are associated with gesture subsets, G_n of size n (m>n). The CMD procedure finds a set of gesture subsets G_n that exhibit accuracies above some tolerable level A_{min} . In Module 3, n by n matching in the form of a quadratic binary integer assignment problem (QAP), is solved by simulated annealing in which the human factor measures are maximized. The resulting gesture-command assignment constitutes the gesture vocabulary, GV. To determine the feasibility of the approach, a robotic arm control task using hand gestures is used. The CMD algorithm found five solutions (gesture subsets), that satisfied a minimum acceptable accuracy of A_{min}=98.33%. Each of the gesture subsets were matched to commands using the QAP to obtain a set of related GV solutions. This was done by scaling the intuitiveness and the comfort measures by weights ranging from 0 to 10, in steps of 1, such that $w_1+w_2=10$) that reflected the importance of each factor on the solution. For each pair of weights (w_1,w_2) and a gesture subset G_n, a GV solution is obtained with its associated values Z_2 and Z_3. These values are plotted in Fig. 5 to show the intuitiveness vs. comfort trade offers for each of the five G_n solutions.

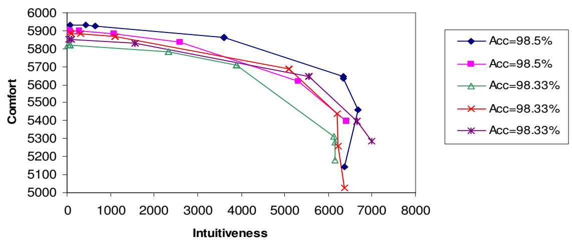

Fig. 2. Intuitiveness vs. comfort families (5 curves)

Thirteen Pareto points (non dominated solutions) were found from all the solutions in Fig. 2. The decision maker can now select a solution from the Pareto set according to his/her own preferences. Fig.3 shows one of the Pareto solutions selected by considering accuracy, intuitiveness, and comfort as the 1st, 2nd and 3rd priorities, respectively. One can see the solution contains many complementary gesture-command pairings. For example, the left and right commands are represented by wrist flips. Also, the command closing and opening the gripper are represented by closing and opening the fist. Sixteen subjects performed a human evaluation of the results obtained by the CMD method. The results showed that highly intuitive, comfortable and easily recognizable gestures resulted in shorter task completion times at the .5% level of significance (p=0.0059).

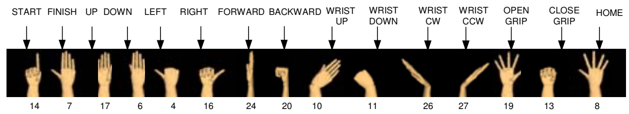

Fig. 3. A GV selected by the decision maker from the Pareto solutions

Analysis of Emotional Gestures from Videos for the Generation of Expressive Behaviour in an ECA

Ginevra Castellano[1], Maurizio Mancini[2]

[1] InfoMus Lab, DIST - University of Genova
Viale Causa 13, I-16145, Genova, Italy
Ginevra.Castellano@unige.it
[2] LINC, IUT de Montreuil, University of Paris8
m.mancini@iut.univ-paris8.fr

In human-computer interaction the ability for systems to understand users' behaviour and to respond with appropriate feedback is an important requirement for generating an affective interaction. Virtual agent systems represent a powerful human-computer interface, as they can embody characteristics that a human may identify with and they may therefore interact with the user in an affective manner. Several systems have been proposed in which virtual agents provide visual feedback or response by analysing some characteristics of the users' behaviour [1] [2] [3].

In this paper we focus on modelling a bi-directional communication between an embodied conversational agent and humans based on the non-verbal channel of the communication involving movement and gesture. Specifically, our research considers movement and gesture expressivity as a key element both in understanding and responding to users' behaviour. We present a system capable of acquiring input from a video camera, processing information related to the expressivity of human movement and generating expressive copying behaviour. Our system is based on the integration of two different software platforms: EyesWeb (www.eyesweb.org) [4] for video tracking and analysis of human movement, and the Greta embodied conversational agent (ECA) for behaviour generation [5]. Figure 1 shows an overview of the system architecture.

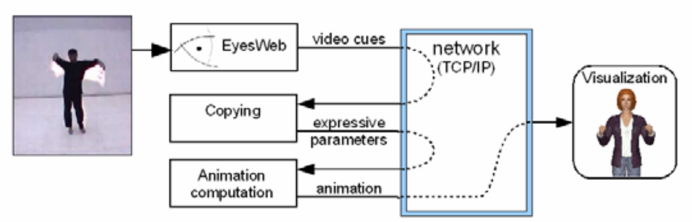

Fig.1. Overview of the system architecture modules. (1) *EyesWeb:* this module performs the automatic extraction of motion cues (e.g., amplitude, fluidity, velocity and acceleration of movement). (2) *Copying*: motion cues are mapped into agent's expressivity parameters. (3) *Animation computation* and *Visualization*: these modules, given the animation data in input, create a graphical representation of a virtual human.

In this scenario, we model a bi-directional communication based on real-time analysis of movement and gesture expressivity and generation of expressive copying behaviour: for each gesture performed by a human, the agent responds with a gesture that exhibits the same movement qualities. We defined a mapping between the expressive cues that we analyse in humans and the correspondent expressive parameters of the agent (Figure 2).

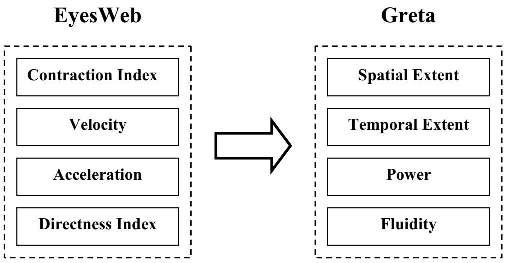

Fig.2. Mapping between EyesWeb and Greta expressive parameters.

To design and test our system we analysed gestures in affective videos to generate an expressive behaviour in the Greta agent. We analysed an extract of videos from the GEMEP (GEneva Multimodal Emotion Portrayals) corpus, a corpus of acted emotional expressions collected by the Geneva Emotion Research Group. We extracted motor indicators conveying information about the expressivity in the stroke. Results show that we defined successful indicators to drive the synthesis process: the synthesised gestures reproduce the expressivity of the real ones. We also investigated which gesture phase to consider as the analysis unit: we found that, since in the Greta agent the expressivity acts on the stroke, working on the stroke may increase the effectiveness of the synthesised gesture to reproduce the expressivity of the real one.

Acknowledgements. The research work has been realised in the framework of the EU-IST Project HUMAINE (Human-Machine Interaction Network on Emotion), a Network of Excellence (NoE) in the EU 6th Framework Programme (2004-2007). We thank Tanja Baenziger and Klaus Scherer for the concession of videos of the GEMEP corpus. We also thank Catherine Pelachaud and Antonio Camurri for their advice.

References

1. Kopp, S., Sowa, T. and Wachsmuth, I.: Imitation games with an artificial agent: From mimicking to understanding shape-related iconic gestures, Gesture Workshop, 2003, pp. 436–447.
2. Reidsma, D., Nijholt, A., Poppe, R.W., Rienks R.J., and Hondorp, G.H.W.: Virtual Rap Dancer: Invitation to Dance, in Conference on Human Factors in Computing Systems, April 2006, G. Olson and R. Jeffries (eds), ACM, Montréal, Canada, pp. 263-266, 2006.
3. Mancini, M., Castellano, G., Bevacqua, E. and Peters, C.: Copying behaviour of espressive motion, in Proc. of the the Mirage Conference 2007.
4. Camurri, A., Coletta, P., Massari, A., Mazzarino, B., Peri, M., Ricchetti, M., Ricci, A. and Volpe, G.:Toward real-time multimodal processing: EyesWeb 4.0, in Proc. AISB 2004 Convention: Motion, Emotion and Cognition, Leeds, UK, March 2004.
5. Pelachaud, C. Multimodal expressive embodied conversational agents. In MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia, pages 683 {689, New York, NY, USA, 2005). ACM Press.

Doctor-Computer Interface using Gestures

Juan Wachs[1], Helman Stern[1], Yael Edan[1],
Michael Gillam[2], Craig Feied[2], Mark Smith[2] and Jon Handler[2]

[1]Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Be'er-Sheva, Israel, 84105,
{juan, helman, yael}@bgu.ac.il
[2]Institute for Medical Informatics, Washington Hospital Center, 110 Irving Street, NW, Washington, DC, 20010,
{feied,smith,handler,gillam}@medstar.net

1 System Overview

The motivation of this system is the sterile control for the interaction between doctors/surgeons and medical imaging systems. A web-camera placed above a screen (Fig. 1) captures a sequence of images of the hand. The hand is tracked by a tracking module which segments the hand from the background using color and motion cues. This is followed by black/white thresholding and various morphological image processing operations. The location of the hand in each image is represented by the 2D coordinates of its centroid. The system is implemented through a two level architecture. The lower level, named Gestix, provides tracking and recognition functions, while the higher level, named Gibson, manages the user interface.

(a) (b)

Fig. 1. Gesture capture system

The Gibson image browser is 3D visualization medical tool that enables examination of images, such as CT scans and X-rays. To interface the Gestix gesture recognition routines with the Gibson system, information such as, the centroid of the hand (used for image navigation controls), its size representing zoom, and orientation for rotation angle, are used to enable screen operations in the Gibson GUI. Both the gesture interface and the Gibson image browser are embedded in ActiveX controls which are communicated using messages and windows events.

2 Hand Tracking and Operation Modes

Gesture operations are initiated by a calibration procedure in which a skin color model of the users hand is constructed. Control between dynamic gestures used for browsing through images and pose gestures (used for rotation and zoom) are affected by mode switch gestures. Superimposed over the image is a rectangular frame. The area inside the frame is called the "neutral area". Movements of the hand across the boundary of the rectangle constitute directional browser commands. When a doctor decides to perform a specific operation on a medical image, he/she places the hand in the 'neutral area' momentarily, and an attention window event is called. The spatio-temporal information and other attributes of the posture are sent to a "mode detector" to determine whether a zoom or rotation pose gesture is presented, see Fig 2. Interaction is designed in this way because the doctor will often have his hands in the 'neutral area' without intending to control the Gibson data browser. Only when a flick gesture is moved towards one of the four quadrants (left, right, up, down), is the image browser moved in the direction of the flick, see Fig 3.

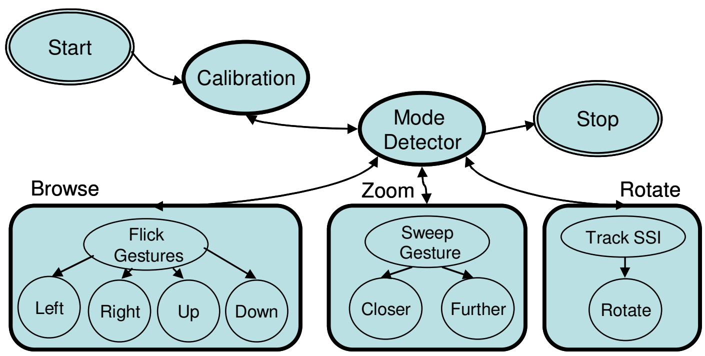

Fig. 2. State machine for the gesture-based medical browser

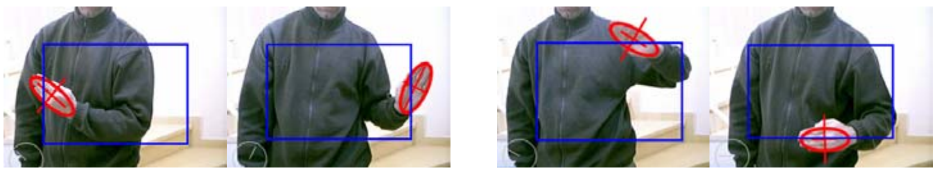

Fig 3. Four quadrants mapped to browsing operations

3 System Implementation

The software requirements of Gestix are Windows XP with Visual Studio 6.0 installed. Hardware requirements include a regular USB webcam connected to a Laptop with an Intel Pentium processor (1.8 GHz) and 500 Mb RAM. Artificial fluorescent light is recommended.

Enhancing a Sign Language Translation System with Vision-Based Features

Philippe Dreuw, Daniel Stein, and Hermann Ney

Lehrstuhl für Informatik 6– Computer Science Department,
RWTH Aachen University – D-52056 Aachen, Germany
{surname}@cs.rwth-aachen.de

Introduction. For automatic sign language translation, one of the main problems is the usage of spatial information and its proper representation and translation, e.g. the handling of spatial reference points in the signing space. Such locations are encoded at static points in signing space as spatial references for motion events.

We present a new approach starting from a large vocabulary speech recognition system which is able to recognize sentences of continuous sign language speaker independently [4]. The manual features obtained from the tracking, are passed to the statistical machine translation system to improve its accuracy. On a publicly available benchmark database, we achieve a competitive recognition performance and can similarly improve the translation performance by integrating the tracking features.

Tracking System. Relevant body parts such as the head and the hands have to be found for feature extraction, but most systems can only produce candidate regions. To extract features which describe manual components of a sign, the dominant hand is tracked in each image sequence. Therefore, a robust tracking algorithm is required as the signing hand frequently moves in front of the face, may temporarily disappear, or cross the other hand. Our head and hand tracking framework is based on the work of [1]. These hand features, which are usually used within the recognition framework, can also be used within the translation framework, in order to improve the translation error rate, too.

Translation System. We use a statistical machine translation system to automatically transfer the meaning of a source language sentence into a target language sentence [2].

Following the notation convention, we denote the source language with J words as $f_1^J = f_1 \dots f_J$, a target language sentence as $e_1^I = e_1 \dots e_I$ and their correspondence as the a-posteriori probability $\Pr(e_1^I|f_1^J)$. Our baseline system maximizes the translation probability directly using a log-linear model:

$$p(e_1^I|f_1^J) = \frac{\exp\left(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum_{\tilde{e}_1^I} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(\tilde{e}_1^I, f_1^J)\right)}, \quad (1)$$

with a set of different features h_m, scaling factors λ_m and the denominator a normalization factor that can be ignored in the maximization process. We choose the λ_m by optimizing an MT performance measure on a development corpus using the downhill simplex algorithm. For a complete overview of the sign language

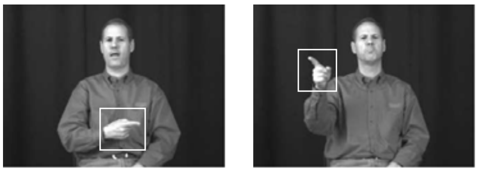

Fig. 1. Sample frames for pointing near and far used in the translation.

translation system, see [3]. The tracking positions of the dominant-hand were clustered and their mean calculated. Then, for deictic signs, the nearest cluster according to the Euclidean distance was added as additional word information for the translation model.

Experimental Results. The RWTH-Boston-Hands database[1] for the evaluation of hand tracking methods in sign language recognition systems has been prepared. It consists of a subset of the RWTH-Boston-104 videos [4]. The positions of both hands have been annotated manually for 1119 frames in 15 videos. We achieve a 2.30% tracking error rate for a 20×20 search window [1].

In the translation, the incorporation of the tracking data for the deixis words helped the translation system to discriminate between deixis as distinctive article, locative or discourse entity reference function. For example, the sentence `JOHN GIVE WOMAN IX COAT` might be translated into *John gives the woman the coat* or *John gives the woman over there the coat* depending on the nature of the pointing gesture `IX`. Using the tracking data, the translation error rate in preliminary experiments (40 test sentences) improves from 44.2% Levenshtein word error rate to 42.5 %, and from 42.5% position independent word error rate to 40.3%.

Summary & Conclusions. We presented a vision-based approach to continuous ALSR and a statistical machine translation approach for ASLT. The results suggest that hand tracking information is an important feature for sign language translation, especially for grammatically complex sentences where discourse entities and deixis occur a lot in signing space.

References

1. P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. Tracking Using Dynamic Programming for Appearance-Based Sign Language Recognition. In *7th Intl. Conference on Automatic Face and Gesture Recognition*, IEEE, Southampton, pages 293–298, April 2006.
2. A. Mauser, R. Zens, E. Matusov, S. Hasan, and H. Ney. The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation. In *IWSLT*, Kyoto, Japan, pages 103–110, November 2006. Best paper award.
3. D. Stein, J. Bungeroth, and H. Ney. Morpho-Syntax Based Statistical Methods for Sign Language Translation. In *11th Annual conference of the European Association for Machine Translation*, Oslo, Norway, pages 169–177, June 2006.
4. M. Zahedi, P. Dreuw, D. Rybach, T. Deselaers, and H. Ney. Using Geometric Features to Improve Continuous Appearance-based Sign Language Recognition. In *17th BMVC*, Edinburgh, UK, pages 1019–1028, September 2006.

[1] `http://www-i6.informatik.rwth-aachen.de/~dreuw/database.html`

Towards a Gesture Repertoire for Cooperative Interaction with Large Displays

Wim Fikkert and Paul van der Vet

Human Media Interaction*, University of Twente
PO Box 217, 7500 AE Enschede, The Netherlands
{f.w.fikkert,p.e.vandervet}@ewi.utwente.nl
http://hmi.ewi.utwente.nl

Future, 'smart' environments are emerging rapidly: in meeting rooms, in public areas and in the home. They are fraught with new, computational technologies that demand easy access. Human-computer interaction (HCI) develops 'natural' interfaces that analyse, react to and are based upon natural human behaviour [1]. We study a key natural interaction modality, manual gesture input, with large displays in smart environments for face-to-face single- and multi-user settings.

Gesture input has traditionally been based on a repertoire of predefined, single-user gestures. Such a set of predefined gestures can easily become idiosyncratic, which does not match the—intended—natural interaction. We describe preliminary work on a series of user-centred experiments that lead to a natural gesture repertoire for HCI in both single-user and multi-user situations.

A Real World Context: The Life Sciences

The life sciences are the setting for our research. The *e*-BioLab, under development by the Microarray Department headed by Timo Breit at the University of Amsterdam, is used by life scientists who are involved in 'omics' experiments [2]. In multidisciplinary teams, these users carry out omics experiments and seek biological meaning in the resulting huge, complex datasets. This results in dynamic, diverse sequences of interaction tasks. The *e*-BioLab continues to evolve towards a finished facility, partly based on experiences in everyday, real-life use.

Currently, the large display (3x2m) in the *e*-BioLab is controlled by an operator. His role is to follow the team's discussion and to fulfil visualization requests for the team. These requests can be implicit, in which case the operator will act on his experience, and explicit, coming directly from the users. Our observations so far show that explicit requests vary from visualizing particular data, to highlighting a correlation between data views, to applying data manipulation.

Capturing User Behaviour during Interactions

Interaction tasks can be classified as either navigation, selection or manipulation. Clearly, these tasks can be subdivided further. We propose three consecutive

* This work is part of the BioRange program carried out by the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI)

user studies that provide more insight in naturally occurring behaviour when performing interaction tasks with a large display. This behaviour will be new because the interaction is novel itself, especially with human behaviour as input.

First, we observe microarray analysis and interpretation in *e*-BioLab meetings. An extensive task analysis of those meetings is currently being performed. Multiple audiovisual streams are recorded in the *e*-BioLab, combined with frame-based logging of its large display. Human behaviour in this footage is annotated with an emphasis on users' gestures towards both each other and the display. For annotation, we use an expanded notation for sign language [3] that includes the gesture's target, for example, in deictic messages. The resulting repository of gestures describes human behaviour, with an emphasis on gestures, that were clearly used to say something about, or have something to do with, the display.

Second, we perform a so-called Wizard of OZ experiment: users have the feeling that they are operating the display themselves through both explicit commands and gestures. The operator responds to these commands depending on their context, using our instructions. These observations are crucial for our goal. For natural single-user gesture interaction, gestures have to be both sufficiently recognisable and distinct from each other. For multi-user situations, there is an supplementary requirement that gestures made towards team members can be distinguished from gestures intended to operate the display.

Third, we introduce an automated gesture recognition system. Ideally, such a system is based on unobtrusive detection. Its aim is to polish our previous findings and to arrive at a robust, stable system. The current state of the art in such systems makes it unclear if a somewhat obtrusive solution must suffice [1].

Discussion

We have described a method to arrive at a natural, multi-user gesture repertoire for HCI with large displays. Even without the third experiment, we will have constructed a significant gesture repertoire that life scientists can use to operate a large display. Porting our gesture repertoire to other user communities requires further investigation. However, we believe that it can be generalised to empirical scientists in general. They carry out tasks that are highly constrained by both their task environment and their explorative research approach.

References

1. Pantic, M., Pentland, A., Nijholt, A., Huang, T.: Human computing and machine understanding of human behavior: A survey. In: ICMI '06: Proceedings of the 8th int. conf. on Multimodal interfaces. Volume 8., ACM Press (2006) 239–248
2. Rauwerda, H., Roos, M., Hertzberger, B., Breit, T.: The promise of a virtual lab in drug discovery. Drug Discovery Today **11**(5–6) (2006) 228–236
3. Kipp, M.: Gesture Generation by Imitation - From Human Behavior to Computer Character Animation. PhD thesis, Saarland University, Saarbruecken, Germany, Boca Raton, Florida (2004)

Gestures to control sound/music objects in a three-dimensional environment

Leonel Valbom[1], Adérito Marcos[1]

[1] Department of Information Systems, University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal
{lvalbom, marcos,}@dsi.uminho.pt

Immersive Musical Instrument

During our research we developed a prototype based on a new idea of musical instrument that expands the concepts of traditional musical elements and allows the integration of spatial dimension using 3D music/sound objects into the musical environment, using physical, visual and sound immersion.

The created immersive musical instrument, which we named as WAVE, integrates virtual reality technologies in order to make possible the visual immersion, the visualization of three-dimensional graphic objects and the spatial displacements of sound sources. The system does not restrict the performance of different musical styles or types of music because it can handle traditional and new music/sound elements and adds the capability of the use of the spatial dimension with sound in composition or performance.

In terms of visualization, the developed immersive musical instrument is based on the "polyvalent spatial sound object" represented on the screen by a modelled 3D graphic object with sound properties that creates in the user the illusion of being immersed in an interactive visual and sound environment.

The visualization display shows to the user few spherical objects that represent long audio files or pre-mixed music distinguished from the parallelepiped objects. These last objects are geometric representations of fingertips based on the keys of Jacob Duringer' Monolithic Two-Dimensional Keyboard to facilitate horizontal and vertical fingering techniques.

The graphic objects, spherical and parallelepiped, have built in actions and they behave as sound objects, as defined by Michel Chion [1], the composer or performer can organize a vast amount of sound/musical material relying more on perception than in physical models.

The interactive gesture is constructed based on two wands that are virtually positioned on the hands of the user in order to point, select or grab objects on the screen.

Controlling gestures

There are now significant works about how to extract significant information from gestures captured by several types of devices, especially three-dimensional tracking systems. In our research, we created a model of immersive musical instrument that is controlled by gestures. However, our work was focused in a certain type of gestures that allow for a facilitated move of sound objects in the space, control sound trajectories, select, and trigger actions and events in three-dimensional environment.

When performing musical instruments, the motor system, proprioception, vision and auditory sense controlled by central nervous system involve complex cognitive operations sometimes with redundancy in terms of management in several parts of cortex in cooperation with other structures like basal ganglia and cerebellum which control the muscle activity. Therefore, it is important to keep the simplicity of gestures. We used in the prototype the same gestures that we use to move objects in our daily life (macro gestures),

combined with small and precise gestures that we use to act in particular functions built-in the daily life objects, like the gestures of the fingers to manipulate real objects (micro gestures).
The micro gestures have small amplitude; they are related to the fingers and allow fast movements in small spaces. These kinds of gestures are the only way to perform music with the virtuosism required in some musical works. Macro gestures have median or big amplitude, they are related to the arms movements and they represent the natural way to achieve far objects in a space or move these objects.

Gesture Interface

WAVE uses two sensors and actuators to support interaction with the system. Using the Cadoz terminology [2], the actuators aim to answer to the excitation gestures like trigger notes, sounds and functions built-in objects and the sensors react to the modification gestures, like modification of trajectories, musical/sound.

The results of tests based on standard IsoMetrics (Using principles of ISO 9241-10) [3] and in the UTAUT - Unified Theory of Acceptance and Use of Technology [4] have allowed to conclud (among others) that:

-immersive musical instruments are the natural environment to move sound sources in the space, and the principles and the chosen type of gestures are a natural and adjusted approach to the interaction with this kind of immersive musical instrument;

-immersive environments allow integrating visual and auditory response as a “multiple unique” concept, giving us the perception of physical, visual and sound immersion;

-immersive musical instruments allow gestural interaction, using several types of movements that facilitate the performance, composition or real-time improvisation with a high degree of control;

-the technology and principles used in the created model of immersive musical instrument, proved to have high acceptance by several groups of people such us students, musicians or even for people that are using other kind of technologies.

References

[1] Chion, Michel, 1983, “Guide des Objects Sonores, Pierre Schaeffer et la recherche musicale”, Éditions Buchet, Chastel, Paris, Nº d’éd. 1221. Nº d’impr. 20569. D.L.: mai 1983.
[2] Cadoz, Claude, 1994, “Le geste canal de communication homme/machine. La communication "instrumentale”, Technique et science informatique, vol. 13, no. 1, pp. 31-61, 1994.
[3] Willumeit H., Gediga G., Hamborg K.-C., 1996, “IsoMetrics: Ein Verfahren zur formativen Evaluation von Software nach ISO 9241/10”, Ergonomie und Informatik, 27, 5-12.
[4] Venkatesh, V., Morris, M., Davis, G., Davis, F., 2003, “User Acceptance of Information Technology: Towards a Unified View”, MIS Quarterly. 27 (3), 425-478.

Processing Iconic Gestures in a Multimodal Virtual Construction Environment

Christian Froehlich, Peter Biermann, Marc E. Latoschik, and Ipke Wachsmuth

Artificial Intelligence Group
Faculty of Technology, University of Bielefeld
D-33594 Bielefeld, Germany
{cfroehli,pbierman,marcl,ipke}@techfak.uni-bielefeld.de

Abstract. In the description of object shapes, humans usually perform iconic gestures that coincide with speech (are coverbal). Marked by a similarity between the gestural sign and the described object, iconic gestures may easily depict content difficult to describe using words alone. Though the expressive potential of iconic gestures in human-computer communication is generally acknowledged, the development of application systems that take non-verbal modalities of natural communication into account has in most cases focused much more on pointing and symbolic gestures. In this paper, we describe how *Imagistic Description Tree* representations can be used in an application system for generating virtual objects with particular spatial features by way of verbal descriptions and iconic gestures.

1 Defining Shape through Iconic Gestures

Based on a comprehensive corpus of speech-gesture shape descriptions acquired from an empirical study [4], the Imagistic Tree (IDT) was proposed as a representation for the semantics of multimodal shape-related expressions, to the end of algorithmically interpretive operational shape descriptions from gesture and speech input modalities. It extends an earlier approach [3], which models the two factors of extent and (partly) profile information in gestures, but which has not included structured spatial organization of gesture and accompanying speech reflecting this factor.

The IDT models object extent, profile, and structure, as the salient semantic elements contained in gesture and speech. The basic level of IDT reprepresentations are object schemata, in which each object is described by a collection of up to three axes which represent the objects extents in one, two, or three spatial dimensions. Using different combinations of axes in an object schema, several basic objects can be represented, such as cubes, cylinders, etc.

Structural aspects of an object are represented in an imagistic description which can recursively embed further imagistic descriptions for object parts, to result in a tree-like structure similar to the hierarchical structure used in the Marr and Nishihara [1] model. The complete tree describing an object including all parts, parts of parts etc. is called Imagistic Description Tree (IDT). A more detailed description of the formal structure can be found in [2].

In our application system – the Virtual Workspace – IDTs are used to represent shape-related information about virtual construction parts. The IDTs are generated online when the system is instructed to create a new object, and are then attached to the object as semantic information, so the object can later be referenced by its shape for further manipulation. Shape-related information is acquired via iconic gestures, which are captured by the use of data gloves and motion trackers. A typical gesture accompanying verbal input processed to create a bent object in virtual reality is shown in Figure 1.

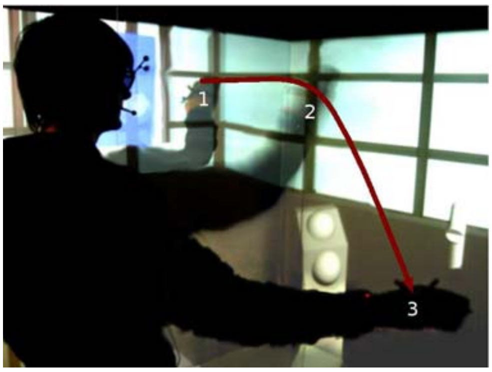

Fig. 1. Defining an object's shape via an iconic gesture, in this case a curved cylinder

Acknowledgment Support by the Deutsche Forschungsgesellschaft (DFG) is gratefully acknowledged.

References

1. H. Marr, D. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. In *Proceedings of the Royal Society*, B 200, pages 269–294, 1978.
2. T. Sowa. Towards the integration of shape-related information in 3-D gestures and speech. In *Proceedings of the Eighth International Conference on Multimodal Interfaces*, New York, 2006. ACM Press.
3. T. Sowa and I. Wachsmuth. Interpretation of Shape-Related Iconic Gestures in Virtual Environments. In I. Wachsmuth and T. Sowa, editors, *Gesture and Sign Language in Human-Computer Interaction*, LNAI 2298, pages 21–33, Berlin, 2002. Springer.
4. T. Sowa and I. Wachsmuth. Coverbal Iconic Gestures for Object Descriptions in Virtual Environments: An Empirical Study. In M. Rector, I. Poggi, and N. Trigo, editors, *Proceedings of the Conference "Gestures. Meaning and Use"*, pages 365–376, Porto, Portugal, 2003. Edições Universidade Fernando Pessoa.

A simple hand posture recognition for efficient single handed pointing and selection

Jean Emmanuel Viallet[1] and Sébastien Carbini[2]

[1] Orange Labs, 22300 Lannion, France
jeanemmanuel.viallet@orange-ftgroup.com
[2] IRISA - ENSSAT, 22300 Lannion, France
Sebastien.Carbini@enssat.fr

Abstract. Pointing and selecting are the two most common gestures performed when interacting with a computer with a mouse. Computer vision techniques allow interaction, far away from a wall size screen, without a handheld device, and the recognition of both pointing and selection allows simulating a mouse which interfaces most computer applications. ISO 9241-9:2000(E) evaluation is performed.

An extended hand is a natural correspondence to pointing. A grasping gesture, closing one's fist is a simple metaphor to associate to a selection event.
A stereo camera located on top of a wall size display allows to continuously track the 3D position of the head and hand of a user in a large interaction space (figure 1) [1]. The head-arm axis enables the user to point any location of a distant screen. But such a view is not appropriate to determine the hand posture (pointing or grasping), even for a human eye. A ceiling camera provides hand images (figure 1) visually easier to classify, with the additional benefit of usually uniform floor background. A standard difficulty encountered with quick single-frame hand posture classification without colored gloves, is which framed image should be considered. Alone skin color filtering or background difference techniques are unable to isolate a hand from a naked forearm (figure 1 far left), and different frames can lead to different classifications. Since the hand 3D position is known, as well as the respective positions and characteristics of both cameras, the 2D position of the hand frame in the ceiling camera image is also known (figure 1 left) allowing hand extraction from the uniform background just in front of the hand (figure 1 middle).
The recognition of the pointing and grasping postures involves a first finite state machine similar to Quek's FingerMouse [2]. Since posture recognition is not error prone, for example during the transition from a posture to another, we introduce a second finite state machine so that two successive similar postures are required to confirm a posture (figure 2 middle). On a one hundred transitions (point to grasp and grasp to point) detection task, over 5000 frames, this constraint leads to an average 2 frame delay and no additional wrongly detected transition. Without the constraint, there is no delay but 24 additional wrong detections. The performance of this single handed input device (figure 2 right *OneHand* curve)

Fig. 1. Far left: 3D body parts tracking camera view. Hand posture classification top camera. Left: view of user pointing. Middle: Background segmented hand, finger-wrist classification (top: user pointing, bottom: user grasping). Right: view of user grasping.

is evaluated with the ISO methodology [3][1], is on average as reliable but 500 ms slower than a previously reported two hand device [1], where selection involves a second hand (figure 2 left, figure 2 right $TwoHands$ curve).
The two camera setup, hand tracking and hand posture classification techniques leads to a sufficiently fast, precise and natural interaction contactless input device suitable for interaction away from a very large display.

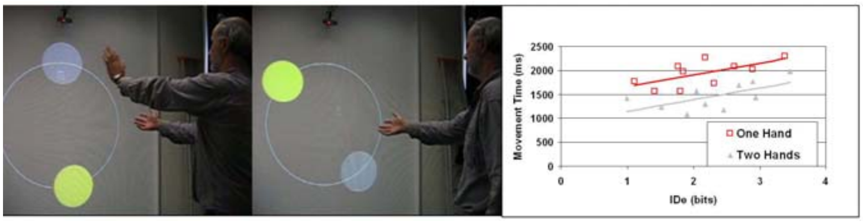

Fig. 2. Left: pointing a Fitts' law based ISO 9241-9 [3] target with a hand and selecting with another ($TwoHands$). Middle: pointing and selecting with $OneHand$. Movement time versus effective index of difficulty for $TwoHands$ and $OneHand$ devices.

Acknowledgements
This work has been supported by EU-funded SIMILAR Network of Excellence.

References

1. Carbini, S., Viallet, J.E.: Evaluation of contactless multimodal pointing devices. The Second IASTED International Conference on Human-Computer Interaction, Chamonix, France, March 14-16, 2006, pp. 226-231.
2. Quek, F., Mysliwiec, T., and Zhao, M.: FingerMouse: A Freehand Pointing Interface. Proceedings of the International Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, June 26-28, 1995, pp. 372-377.
3. ISO 9241-9:2000(E): Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs), Part 9: Requirements for Non-Keyboard Input Devices, International Standards Organization, 2000.