THE INTELLIMEDIA WORKBENCH - A GENERIC ENVIRONMENT FOR MULTIMODAL SYSTEMS

Tom Brøndsted (tb@cpk.auc.dk), Lars Bo Larsen (lbl@cpk.auc.dk), Michael Manthey (manthey@cs.auc.dk), Paul Mc Kevitt (pmck@cpk.auc.dk), Thomas Moeslund (tbm@cpk.auc.dk), Kristian G. Olesen (kgo@vision.auc.dk) (authors in alphabetical order)

Institute of Electronic Systems, Aalborg University, Denmark.

ABSTRACT

The present paper presents a generic environment for intelligent multi media applications, denoted "The Intellimedia WorkBench". The aim of the workbench is to facilitate development and research within the field of multi modal user interaction. Physically it is a table with various devices mounted above and around. These include: A camera and a laser projector mounted above the workbench, a microphone array mounted on the walls of the room, a speech recogniser and a speech synthesiser. The camera is attached to a vision system capable of locating various objects placed on the workbench.

The paper presents two applications utilising the workbench. One is a campus information system, allowing the user to ask for directions within a part of the university campus. The second application is a pool trainer, intended to provide guidance to novice players. Introduction

1. INTRODUCTION

Background.

Driven by the current move towards multi modal interaction an activity was initiated at Aalborg University to integrate the expertise present in a number of previously separate research groups [11]. Among these are speech and natural language processing, spoken dialogue systems, vision based gesture recognition, decision support and machine learning systems.

This activity has resulted in the establishment of an "Intellimedia WorkBench". The workbench is a physical as well as a software platform enabling research and education within the area of multi modal user interfaces. The workbench makes a set of tools available which can be used in a variety of applications. The devices are a mixture of commercially available products (e.g. the speech recogniser and synthesiser), custom made products (e.g. the laser system) and modules developed by the project team (the machine learning system, microphone array, the gesture recogniser and the natural language parser).

Purpose.

The workbench serves a number of purposes. Primarily it is a platform for researching aspects of multi modal user interaction. For example to do experiments of integration of speech and gestures across various applications. The workbench facilitates this, so experiments can easily be set up as most of the modules will be present already. Also, new applications can be developed within a limited time scope, utilising some or all of the available soft- and hardware modules. This is important, e.g. when the workbench is used for student projects, thus allowing the students quickly to develop and test their own applications, or integrate a new module into the architecture.

In the following sections the workbench is described.

2. BASIC ARCHITECTURE AND EXAMPLE APPLICATIONS

In the following, two examples of applications are described. A campus information system and an automatic pool training system. The hardware devices and some of the software modules are used in both cases, but there are significant differences in the way the overall system management has been implemented

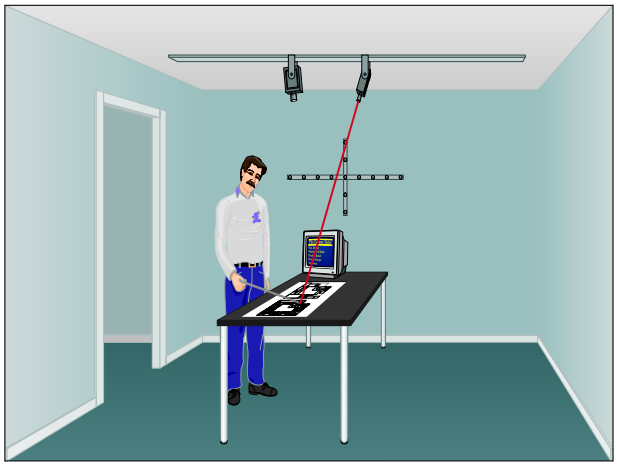

Figure 1: Physical layout of the workbench shown with the campus information application. The table is substituted with a pool table, but otherwise only minor changes are made in the pool training example

Figure 1 shows what the workbench actually looks like. It is placed in a specially designed laboratory. The room has a movable wall, allowing for optimal conditions across different experimental setups. A camera and a laser are mounted in the ceiling. A microphone array is placed on the wall. Additional cameras can be mounted to monitor and record the behaviour of test subjects from an adjacent laboratory, which acts as a control room.

2.1 The Campus Information System

In this case, the application is a multi modal campus information system [3], [4]. A model (blueprint) of a building layout is placed on the workbench table (see figure 1). The system allows the user to ask questions about the locations of persons and offices, labs, etc. Typical inquiries are about routes from one location to another, where a given person's office is located, etc. Input is simultaneous speech and/or gestures (pointing to the plan). Output is synchronised speech synthesis and pointing (using a laser beam to point and draw routes on the map). The central module is a blackboard, which stores information about the system's current state, history, etc. All modules communicate through the exchange of semantic frames with other modules or the blackboard. The process synchronisation and intercommunication is based on the DACS IPC platform, developed by the SFB360 project at Bielefeld university [9]. DACS allows the modules to be distributed across a number of servers.

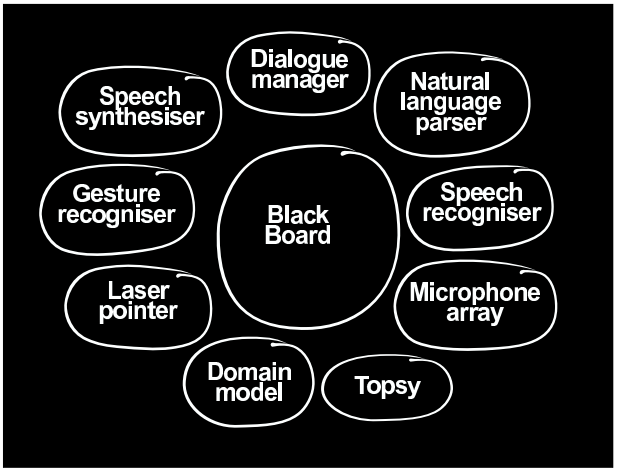

Figure 2: Software architecture of the workbench

Figure 2 shows the blackboard as the central element with a number of modules placed around it. Presently modules for speech recognition, parsing, speech synthesis, 2D visual gesture recognition and a laser pointing device are directly integrated into the application.

Furthermore, a sound source locator (microphone array [14]) and a machine learning system (called Topsy) [15] are included in the workbench.

Frame semantics

A frame semantics has been developed for integrated perception in the spirit of Minsky [17] consisting of (1) input, (2) output, and (3) integration frames for representing the meaning or semantics of intended user input and system output. Frames are produced by all modules in the system and are placed on the blackboard where they can be read by all modules. The format of the frames is a predicate-argument structure and we have produced a BNF definition of that format.

Frames represent some crucial elements such as module, input/output, intention, location, and time-stamp. Module is simply the name of the module producing the frame (e.g. parser). Inputs are the input recognised whether spoken (e.g. “Show me Hanne's office”) or gestures (e.g. pointing coordinates) and outputs the intended output whether spoken (e.g. “This is Hanne's office.”) or gestures (e.g. pointing coordinates). Time-stamps can include the times a given event commenced and completed. The frame semantics also includes two keys for language/vision integration: reference and spatial relations.

2.2 The Automatic Pool Trainer

The aim of this application is to provide guidance for novice pool players [7]. A pool table is placed directly under the laser and camera. The system locates the position of the table edges, the balls and the cue using the camera mounted above the table. When the user points the cue towards the cue ball for a specified time, the pool trainer calculates the trajectories of the balls given the direction of the cue.

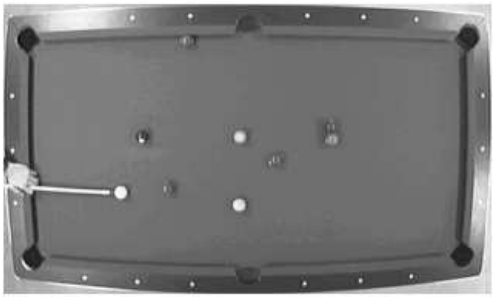

Figure 3: The pool table as seen from the camera (from [7]).

The trajectories are directly drawn on the baize using the laser. Thus, the player is given feedback in a very direct and natural way, simply by having the result of his shot shown to him directly on the surface of the table. If he changes the direction of the cue slightly, new trajectories will be calculated and drawn. The user can issue spoken commands to the system, and receives feedback by the laser and speech. A system with similar aims has been developed at the MIT Media Lab [13]. However, in that case the user is required to use a head mounted video unit (camera and display). The computed trajectories are superimposed on the image and shown on the display unit, thus creating an immersive environment. We believe that the setup described here, although much simpler, provides a more natural and intuitive interface.

Fully developed, this application will provide a number of lessons in a manner similar to e.g. chess exercises [7]. The user is instructed to place balls at specific positions (shown with the laser). When all has been set up, the user is instructed to e.g. pot a specific ball. Of course, the system can also be utilised in a multi-player situation.

One obvious limitation of the pool trainer is that trajectories can only be computed for shots not applying spin.

Although most modules are similar to those of the campus information system, the overall control is distributed among a number of communicating processes, as opposed to the blackboard architecture. Each process is modelled as an autonomous state machine, communicating through the Java RMI (Remote Method Invocation) package [8].

3. SYSTEM MODULES

The workbench presently includes the modules described below.

Speech recogniser.

Speech recognition is handled by the grapHvite [18] real time continuous speech recogniser. It is based on Hidden Markov Models of triphones for acoustic decoding of English and Danish (in the present case). The recognition process focuses on recognition of speech concepts and ignores non content words or phrases. In the campus information application speech concepts are routes, names and commands which are modelled as phrases.In the pool trainer spoken input is simple commands, spotted as keywords. A finite state network describing the phrases is created in accordance with the respective domain model and the grammar for the natural language parser described below.

Speech Synthesiser.

The speech synthesiser used within the platform is the Infovox [12], which in the present version is capable of synthesising Danish and English languages. It is a rule based formant synthesiser, and can simultaneously cope with multiple languages, e.g. pronounce a Danish name within an English utterance.

Natural Language Parser.

The workbench includes a number of general Natural Language Processing (NLP) modules the core of which is a natural language parser and a grammar converter deriving grammar networks to be used by continuous speech recognisers from grammars used for parsing. Various parsing and recognition grammar formats are supported. The modules are partly rooted in The Danish Spoken Language Dialogue Project 1991-94 [1] and are utilised also within the EU-funded project REWARD [5].

The sub language of the campus information system demonstrator is implemented in a compound feature based (so-called unification) grammar format that is currently the most powerful formalism supported by the NLP modules. The grapHvite standard lattice (SDL) format grammar constraining the speech recogniser (see above) is generated from the unification grammar. The natural language parser extracts semantics from the one-best output written by the speech recogniser to the blackboard. The parser carries out a syntactic constituent analysis of input and subsequently maps values into semantic frames of the type described in section 2.1. The rules used for syntactic parsing, are based on a subset of the EUROTRA formalism (lexical rules and structure building rules) [2]. Semantic rules define certain syntactic sub-trees and which frames to create if the sub-trees are found in the syntactic parse trees. For each syntactic parse tree, the parser generates only one predicate and all created semantic frames are arguments or sub-arguments of this predicate. If syntactic parsing cannot complete, the parser can return the found frame fragments to the blackboard.

The natural language processing are described in greater detail in [6]. The parser is not utilised in the Pool Trainer, due to the simple mapping from keywords to commands.

Gesture Recogniser.

A design principle of imposing as few physical constraints as possible on the user (e.g. data gloves or touch screens) lead to the inclusion of a vision based gesture recogniser. It tracks a pointer (or, in the case of the pool trainer; the cue and the balls) via a camera mounted in the ceiling. Using one camera, the gesture recogniser is able to track 2D pointing gestures in real time.

In the campus information application there are two gestures; pointing and not-pointing. In future versions system other kinds of gestures like marking an area, indicating a direction, etc. will be included.

The camera continuously captures images which are digitised by a frame-grabber. From each digitised image the background is subtracted leaving only the motion (and some noise) within this image. This motion is analysed in order to find the direction of the pointing device and its edge. By temporal segmenting of these two parameters, a clear indication of the position, that the user is pointing to at a given time, is found. The error of the tracker is less than one pixel (through an interpolation process) for the pointer.

In the case of the pool trainer, different versions of the Hough Transform [10] are used to locate the balls and the cue [7].

Laser Pointer.

A laser system is mounted next to the camera, acting as a "system pointer". It is used for showing positions and draw routes on the map/pool table. The laser beam is controlled in real-time (30kHz). It can scan frames containing up to 600 points with a refresh rate of 50 Hz thus drawing very steady images on the workbench surface. It is controlled by a standard Pentium host computer. The tracker and the laser pointer are carefully calibrated in order to work together. An automatic calibration procedure has been set up, involving both the camera and laser.

Sound source locator.

A microphone array [14] is used to locate a sound source, e.g a person speaking. (This module is not hooked-up at present). Depending upon the placement of a maximum of 12 microphones it calculates the position in 2 or 3 dimensions. It is based on measurement of the delays with which a sound wave arrives at the different microphones. From this information the location of the sound source can be identified. Another application of the array is to use it to focus at a specific location, thus enhancing any acoustic activity at that location.

Domain Model.

The campus information system domain model holds information on the institute's buildings and the people that works there. The purpose of the model is to be able to answer queries about who lives where etc. The domain model associates information about coordinates, rooms, persons etc.

The model is organised in a hierarchical structure: areas, buildings and rooms. Rooms are described by an identifier for the room (room number) and the type of the room (office, corridor etc.). For offices there is also a description of tenants of the room by a number of attributes (first and second name, affiliation etc.).

The model include functions that return information about a room or a person. Possible inputs are coordinates or room number for rooms and name for persons, but in principle any attribute can be used as key and any other attribute can be returned. Further a path planner is provided, calculating the shortest route between two locations.

The domain model of the pool trainer holds information on the physical dimensions and colours of the table, balls, etc. It contains an algorithm for calculating the resulting trajectories when two balls collide. Basic pool playing rules, e.g that the user is only allowed to hit the white (cue) ball are also included.

Dialogue Manager.

The dialogue manager is only explicitly present in the campus information application, as decisions are distributed among the individual modules in the pool trainer architecture.

In the campus information system, the dialogue manager makes decisions about which actions to take and accordingly sends commands to the output modules via the blackboard. In the present version the functionality of the dialogue manager is mainly to react to the information coming in from the speech/ NLP and gesture modules by sending synchronised commands to the laser pointer and the speech synthesiser modules. Phenomena such as clarification sub dialogues are not included at present.

Topsy.

The basis of the Phase Web paradigm [15], and its incarnation in the form of a program called Topsy, is to represent knowledge and behaviour in the form of hierarchical relationships between the mutual exclusion and co-occurrence of events. (In AI parlance, Topsy is a distributed, associative, continuous-action, partial-order planner that learns from experience.) Relative to multimedia, integrating independent data from multiple media begins with noticing that what ties such otherwise independent inputs together is the fact that they occur simultaneously (more or less). This is also Topsy's basic operating principle, but this is further combined with the notion of mutual exclusion, and thence to hierarchies of such relationships[16].

4. CONCLUSIONS AND FUTURE DIRECTIONS

As mentioned in the introduction, two major goals are behind the establishment of the workbench. One is to facilitate research on especially the integration of visual and linguistic (spoken) information, and the other is to make a platform available for post graduate student projects. A M.Sc post graduate programme in intelligent multimedia has recently been set up [IMM 1997] and the workbench plays an important role by enabling students to rapidly build advanced user interfaces including multiple modalities. The pool trainer is the result of such a project, and was essentially designed and implemented over a period of 3 months in the spring semester of 1998. Developments will continue on the workbench.

Future plans include the inclusion of a large screen projector, more advanced speech recognition and high-quality synthetic speech in Danish. A full implementation of the blackboard architecture will also be included.

5. ACKNOWLEDGEMENTS

The authors wishes to thank Jacob Buck, Søren B. Christiansen, Adam Cohen, Sajid Muhammed, Sergio Ortega and Susanna Thorvaldsdottir from the Intelligent MultiMedia programme at Aalborg University for their contribution on the pool trainer. Niels Jungclaus, SFB360 at Bielefeld for help with the DACS system. This work was partly funded by Center for PersonKommunikation and partly by the Technical faculty at Aalborg University.

6. REFERENCES

[1] Baekgaard, A. et al. "The Danish spoken language dialogue project - a general overview". ESCA WS on Spoken Dialogue Systems: theory and applications, Vigsø, Denmark 1995, pp. 89-92.

[2] Beck, A. "Description of the EUROTRA Framework" In: Studies in Machine Translation and Natural Language Processing, vol. 2 1991, ed. C. Copeland et al.

[3] Brøndsted, T. et al. "A platform for developing Intelligent MultiMedia applications" Technical Report R-98-1004, May 1998, Aalborg University. (http://www.kom.auc.dk/~tb/articles/tim2spe.ps)

[4] Brøndsted, T. et al. "The Intellimedia WorkBench - an environment for building multimodal systems". In proc of CMC'98 workshop, January 1998, Tilburg, the Netherlands.

[5] Brøndsted T., B. Bai, J. Ø. .Olsen: "The REWARD Service Creation Environment. An Overview". These Proceedings.

[6] Brøndsted, T.: "The Natural Language Parsing Modules in REWARD and IntelliMedia 2000+". S. Kirchmeier-Andersen, H.E. Thomsen (eds.): Proceedings from the Danish Society for Computational Linguistics (DALF), Copenhagen Business School, Dep. of Computational Linguistics, 1998. In press

[7] Buck, J. et al. "Intelligent Multimedia Based Pool Trainer", IMM, Aalborg University, May 1998.

[8] Erckel, B. "Thinking in Java". Prentice-Hall 1998.

[9] Fink, G.A. et al: "A Distributed System for Integration of Speech and Image Understanding" In Rogelio Soto (ed.): Proceedings of the International Symposium on Artificial Intelligence, Cancun, Mexico 1996, pp. 117-126.

[10] Gonzalez, R.C. and R.E. Woods. "Digital Image Processing" Addison-Wesley Publishing company 1993

[11] http://www.kom.auc.dk/CPK/Speech/MMUI/

[12] "INFOVOX Text-to-speech converter. User's manual" Telia Promoter Infovox 1994.

[13] Jebara, T. et al. "Stochastics: Augmenting the Billiards Experience with Probabilistic Vision and Wearable Computers" in Proc of the Intl. Symposium on Wearable Computers, Cambridge MA 1997.

[14] Leth-Espensen P. and Lindberg B. "Application of microphone arrays for remote voice pick-up - RVP project, final report" Center for PersonKommunikation, Aalborg University 1995

[15] http://www.cs.auc.dk/topsy/

[16] Manthey, M. "The Phase Web Paradigm". Int'l J. of General Systems, special issue on General Physical Systems Theories, K. Bowden Ed. In press.

[17] Minsky, M. 1975 "A framework for representing knowledge" The Psychology of Computer Vision, P.H. Winston (Ed.), 211-217 New York: McGraw-Hill.

[18] Power et al. "The grapHvite Book" for grapHvite Version 1.0 Entropic Cambridge Research Laboratory Ltd, 1997.