Human Machine Interaction using Head Pose Estimation

Simon-Alexander Zerawa, Stefan Kohlhauser, Charlotte Roesener, Andreas Perner

Institute of Computer Technology, Vienna University of Technology, Austria

[zerawa, kohlhauser, roesener, perner]@ict.tuwien.ac.at

***Abstract*— In course of the educational research project XINU (eXcellent Interface for Non-haptic Use) high school students, university students, teachers and researchers developed a novel way to interact with and control elevators. Using standard web cams and state-of-the-art head pose estimation, distinctive and predefined head gestures are interpreted as explicit commands for a mechanical prototyped elevator model. For usability tests of the HMI (human machine interface), a cursor-based application was created to allow a mouse-like selection of possible commands by only moving the head. In this article, the system approach and the project setting is described and the current implementation is presented.**

***Keywords*— head pose estimation, user interface, elevator control, gesture, webcam, cursor;**

I. Introduction

To initiate partnerships between senior high schools and universities (and other research institutions), the Austrian Federal Ministry for Science and Research (BMWF) introduced a research program called "Sparkling Science" in 2007 and following years [1]. The basic idea is that young people shall not only get direct contact to up-to-date research but also be actively involved in research projects. Groups of both domains, research and education, participate in socially founded and cutting-edge technological developments partly situated in the learning environment of higher education. Based on this idea, the 2-year-project XINU (eXcellent Interface for Non-haptic Use) was started in October 2009. The project is conducted within a partnership between the Institute of Computer Technology at the Vienna University of Technology and the School Centre Ungargasse where physically challenged (physically, visually or acoustically impaired and others) and physically fit students are provided with economical or technical education [2][3].

The amount of features offered by today's applications steadily grows but the flexibility of their user interfaces usually does not keep pace. As our society more and more depends on the services of our half or fully automated living environment the usability of systems in our daily life is challenged. In this context, the objective of XINU is to develop a novel method of contactless operation – in this project – of an elevator by using head gestures only. As there is no need to touch/push any buttons or levers for controlling a system, this concept can be beneficial in further scenarios where physical interaction is inconvenient, dangerous or requires unnecessary physical effort the operator lacks. The research presented, offers a new and flexible way to control such systems, focussing on an elevator application and using head gestures as additional mode of user interaction. Introduced in [4] the approach used can not only be applied to elevators but several other applications of the building automation domain as well. Although the flexible model of the system allows alternative types of interfaces as discussed in [5], the authors will focus on visual head pose estimation for purpose of this article.

This article will look at the project from both sides, the educational side highlighting the project approach and the cooperation, and the research side looking at the concepts, implementations and the current types of interaction supported. The authors will conclude with discussing current results and future work.

II. Project Context and State of the Art

A major question in modern society is how to relate and to embrace the next generation which is now educated in schools of today, so that they might form a cultural identity with a technological advanced society that requires a continuous advancement and potential of future scientists. Science Education – a special form of science communication – is an emerging area of research, originating increasing activities e.g. case studies and investigations, that shall allow crossing borders where the students' life-world culture move into the world of science filling gaps between the students worldview and the worldview embraced by the scientific community [6]. Science education shall help with informal settings, shall assist to create a "public awareness and understanding of science" in a positive way. There are three aspects, which shall be emphasized: The understanding of science content, understanding the methods of enquiry and the understanding of science as a social enterprise [7].

The novel idea of the "Sparkling Science" program [1] shall enforce the integration of cutting edge research into educational science that allows the early contact of young people with interesting research projects helping them to be more interested in and better understand their (science based) world around them engaging in the discourse of and about cutting edge science. It shall reduce reservations and other barriers between these domains. This cooperation supports a creative atmosphere, where high school students and university researchers can learn from each other. But scientists shall not only talk about their science and fulfil an educational part in this setting: The fresh minds of young people give researchers the opportunity to get new input and angle of vision. The beauty of this project in focus is that students have

the power to test and suggest optimization of the prototyping system. Their participation can help their bodily challenged peers and might prove to be beneficial to the usability aspects of this project.

Growing computational power and optimized algorithms allow creating systems that are able to track movements of humans with high accuracy. The developments of such systems find the way into various application domains. Building automation offers a wide area to install head pose estimation systems for user interaction. Although concepts exist to alleviate the usage of building automation systems, a focus on the assistance of handicapped people is still missing. Project XINU uses head pose estimation to offer new possibilities in exactly this area of research.

Head pose estimation systems are used to help human system interaction for robotic applications [8], [9]. By use of head pose estimation the control-system of the robot is enabled to estimate human gestures or movements in order to react to the human. The systems used to detect the movement of the head can be categorized in approaches which process a whole image and approaches which focus on distinct points [10], [11]. In order to allow a real time processing of the camera-input the second mentioned approach has advantages for control applications like in project XINU. Other approaches like the usage of markers which are placed on the users head are also used in research projects but are not applicable for the project described since not every user can be equipped with identifier points on their heads.

For the usage in real-world environments where light conditions and backgrounds cannot be precisely defined, the use of thermal images proved to be an applicable solution. Another advantage of this approach is the low computational overhead, like outlined in [12].

For systems which cannot be combined with thermal cameras but must work under restricted observable areas the work of [13] offers a promising solution. By the usage of a multi-camera system, partly observable spots can still be used for image processing algorithms.

In case of project XINU the movements of the head are important since based on the movements control commands are interpreted and sent to the elevator. As described in [14] movements of the head can be described with six degree of freedom. This allows the definition of an individual command set for each user in order to respond to certain needs of a specific user.

III. Project Development and Approach

One of the project's major goals is the creation of a prototype system. Therefore, two intervening approaches were established: first, an application based approach targeting the implementation of a model application with a following evaluation phase was conducted. Second, a broader approach focuses on creating a multipurpose platform in order to allow the integration of different user interfaces with multiple applications.

A. Prototype setup

In order to give access to this system for a wide range of different users using all kind of their own equipment, standard off-the-shelf components like notebooks and webcams were chosen to create the actual hardware for the user interface. At the Institute of Computer Technology a physical 1:20 model of an elevator was constructed providing a fieldbus based sensory network and control system using industrial sensors and actuators.

As an incentive, students of the School Centre Ungargasse were directly involved in the early state of solution design of the user interface and the following evaluation steps. To ensure communication and interdisciplinary working atmosphere, several workshops and regular meetings were held in which students and researcher could interact, exchange and evaluate ideas and challenges in a coequal way.

As several methods were evaluated, a commercial head pose estimation product was selected for development (see section IV) and the implementation of an early prototype providing basic functionality.

B. Flexible and distributed system layout

Beside efforts to build and test a prototype of the system, a broader application of the underlying concept was performed. Key components were identified to allow decoupling of the systems components (see IV.C). In a first stage this can be achieved by using client-server network or fieldbus architecture (see [4]). This allows logical separation of the actual user interaction and the specific application. Due to a modular design with standardized interfaces, the potential systems can provide various types of interaction (e.g. buttons, head pose estimation, speech recognition) on one hand and support all forms of applications (e.g. multiple elevators, air conditioning, heater) on the other. Within this precise approach, the detailed concept provides three layers (see [5]):

- A Human Computer Interaction (HCI) layer implementing the actual interface for the user,
- an application layer providing the services of the application,
- and a communication layer ensuring a proper communication between HCI and application layer.

All layers communicate their data via defined interfaces and protocols. Therefore, a distribution on different hardware systems and even mobile devices is possible. To allow the support of multiple user interfaces and different applications at the same time, a registration and administration server enables the storage and management of different user configuration sets and application-specific software. The server is able to provide an initial communication infrastructure and handles the process of establishing a communication between a specific user (with a specific method for user interaction) and the services of a specific application.

IV. System Design and Implementation

The project's objective is the implementation of a system, which provides physically challenged people with additional means to control an elevator without the need of physically strength and mobility required for pressing a button on control panels. There are various ways to achieve this objective. The focus of this project is the human computer interaction via head movement recognition. However, the design and implementation of a face tracking software itself is not in the scope of this project. Based on an extensive requirements analysis for application and user specific demands, a suitable product (API) had to be found which can provide the following criteria:

- **Robust:** The product has to handle various types of faces coping with variations in skin colour, facial deformation, beards, glasses, hairdo etc. Furthermore, occlusions or bad lighting conditions, which are common in real elevators, have to be handled well.
- **Real time:** The product has to be able to convert the movement of a head into coordinates which can then be used to control a graphic user interface (GUI). This conversion has to be done without any delays, in order to provide a smooth user experience. To be able to acquire and detect head movements a frame rate close to or above 30 frames per second is desired.
- **Simple:** Since price is one of the main issues when considering barrier-free services, low cost is an important point to pierce any resistance. This means the system should run on a standard personal computer and use a webcam coping with simple products and existing infrastructure.

Several commercial and non-commercial products have been considered. The one finally selected was the program faceAPI from Seeing Machines. It provides fast and reliable face tracking, makes no difference between varying head shapes, size, and facial features, runs with standard Intel processors supporting standard webcams and offers a well-documented application programming interface (API). It is implemented in C++ and the provided library can easily be integrated and used with other projects. The program provides tracking values of a head along the X-, Y- and Z-axis as well as rotation along these axes and an additional confidence value, which gives an estimation of how accurate the system rates these values. A commercial and a non-commercial version of the faceAPI exist. A licence was acquired, however, the functionality of the non-commercial version proved to be sufficient.

A. Gesture implementation

In the first step, the system was implemented in C++ as a local application on a single computer. It included the face tracking module, the GUI and a simple elevator simulator. It is still used as a test bed to control the mechanical model elevator instead of the software simulator during integration tests. A more detailed description can be found in [5]. This first implementation used head gestures for the human computer interaction. Nodding the head upwards or downwards selects the desired floor and nodding the head to the left commits the command to the system for execution. Using solely three different head gestures, this approach depended strongly on the knowledge of the gestures and the ability to perform them. Gestures/command sets can also be customized and trained on individual users for a mobile setup, where every user has a control unit. However, this command setup is hardly feasible in an environment where the camera is fixed and directly embedded in the elevator.

B. Cursor-based implementation

The second implementation developed considers several new approaches for improvement of user interaction. Predefined commands should be avoided and a new form of feedback was intended, namely the controlling of a cursor on a GUI. The first enhancement was to not recognize complete head gestures as specific commands but instead to interpret the coordinates provided by the faceAPI as cursor position. In this approach, the movement of the head was converted into input for the cursor which was moved over a GUI for the elevator. In this scenario the user had always a direct visual feedback and the process was much more intuitive as with the implementation described before. Therefore there was no need to agree on and learn predefined head movements as with the gesture based solution above. However, this entailed the necessity of an additional display for feedback needed for the GUI designed for the elevator.

In the GUI rectangular buttons represent the different floors for selection and other control commands of the elevator control. Functions can be selected by moving the circular cursor over the button and holding it there for about two seconds. This process is visualized by small sectors increasingly surrounding the floor number while hovering over a button (see Fig. 1). The size and transparency of the cursor represent the distance from the camera to the head and the confidence of the head pose estimation process. Once a floor button is selected the rectangular area gets highlighted until the elevator has served the floor.

An additional objective of the project is to place a mobile solution at the physically impaired student's disposal. Therefore, it should be possible to use a mobile device like a laptop, notepad, tablet PC or a smart phone for command input. The logical modules of the system were separated in order to be able to run them on distributed devices.

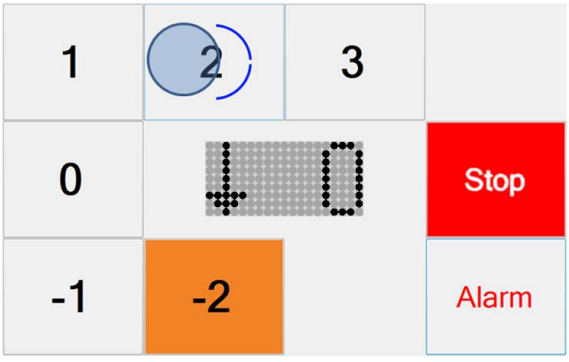

Fig. 1 GUI of the cursor based implementation

The first logical module represents the head pose estimation itself and is running on a PC with a webcam and directly uses the faceAPI framework. This module converts the face movements from the video input from the camera into coordinates, parses them into an XML data structure and provides them over the network.

The second module is the GUI server. It receives the XML-coordinates from the first module and transfers them into cursor movements, considering display size and smoothing algorithms. These cursor movements are then used to control the GUI, as mentioned above. The GUI itself can either be displayed on a screen within an elevator or it can likewise be displayed on the same device (e.g. PC) where the first module is running. By decoupling the modules, either solution is possible.

The third functional module is either the elevator simulator or the mechanical elevator prototype. When a button is selected on the GUI, the command is sent to the elevator which executes the command and sends an updated status back to the GUI. The modules and their interactions are displayed in Fig. 2. Based on this conceptual changes as mentioned above extensive refactoring of the existing code would have been necessary. This led to the idea of a complete redesign and implementation of the GUI server and the simulator as the software part of the model elevator. This step provided several additional possibilities.

The new implementation was done in C# in general. The only module not ported to C# was the actual faceAPI interface itself. For this module only a socket was added to send the coordinates in an XML data structure via the network. Any device can be used as an input for the GUI, independently of its implementation as long as it is able to send coordinates in an XML structure over a TCP (Transmission Control Protocol) socket.

Nonetheless, the GUI module, the simulator and the interface for the model elevator were ported to C#. The first advantage was that it is easier, less error prone and thus faster to write a program in C#. As a corollary it is easier to write reusable code for planed future additions.

Another major benefit of using C# was that it provided the possibility to integrate the high school students better into the actual development process: Most of them were already familiar with the C# programming language, whereas only a few had reasonable knowledge about C++. This lowers the barrier for the students to get in contact not only with the software itself but also with its source code.

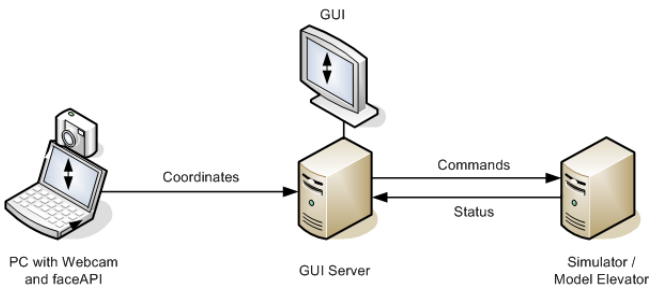

Fig. 2 The XINU modules

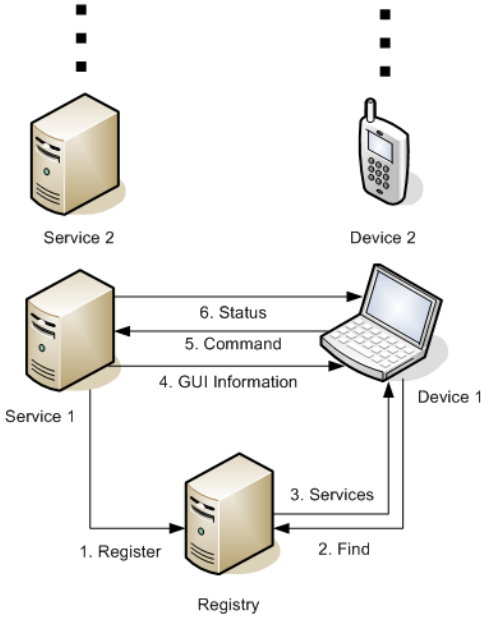

Fig. 3 Dynamic system design

This is important if the students encounter errors and do active bug tracking during evaluation. If the students have a reasonable knowledge about the programming language they can possibly provide in-depth bug reports with a precise error allocation within the source code or even might be able to fix the error themselves entailing a better insight and vast better integration of the high school students in the software engineering process.

In the current implementation the interpretation of the module output as cursor coordinates allows the controlling of any GUI. In the previous implementation only a fixed set of gestures for commands could be defined.

C. Generic service and interface implementation

This section provides an outlook for this project. As already mentioned, one of the major goals of the project is prototyping for an interaction framework. The prototype shall be able to control a wide range of services from a wide range of devices.

How the second part can be achieved was outlined already in the previous section. The goal was the usage of any mobile device, like laptops or smart phones, as an input device for the interaction with the control system. With a modular design and explicit interfaces this can be achieved. E.g. several platforms like Apple's iPhone or Google's Android have head tracking features in development. However, for this approach a novel way how to control a wide range of services has to be reconsidered. An important factor is the fact that several services have specific and diverse interaction possibilities and therefore require a special interface design.

This approach is similar to Service Oriented Architectures (SOA): Several independent services are offered in a network. They send an XML message with their IP-Address (Internet Protocol) and the service name to a registry server via broadcast. When a device joins the network, it can send a

broadcast to find the registry server and get a list of available services. The options are inserted into a GUI framework which basically provides only a GUI stub. This stub is modified by the information received from the registry. Once a service is selected, the device can request for a specific service of the GUI from the service server. Again, the GUI stub is modified according to the information provided by the service adding buttons, labels, etc. as needed. By pressing a button a command is submitted to the service where the according action is performed and an updated status can be returned to the device.

The setup and interactions of the approach is shown in Fig. 3. This basic architecture will be implemented by senior researchers and students at the university. As soon as the interfaces for the services are defined, extra functionalities for the system can then easily be implemented. An example would be a web interface so the services can be used by any standard browser.

V. Discussion And Results

In October 2010 a first model setup was presented allowing the operation of the mechanical model elevator using head movements only. Following this stage, students could work directly on implementation tasks during their internships at the institute in 2010. While the crucial implementation tasks were mainly accomplished by researchers and students at the university, the high school students had the chance to use the software and carry out surveys while conducting tests with other students in school addressing different classes and physically challenged students. In several meetings, the students presented their findings to the project team. As the first prototype was set up at the institute and a prove-of-concept was operational, the mechanical elevator model was separated from the head pose estimation control itself. For evaluation and testing purposes preconfigured laptops were given to the students. Containing an elevator software simulation and several versions of webcam control software, evaluations and tests could be performed directly at the school. These were accompanied with surveys to get direct feedback from the participants, mainly peers of the same school who volunteered for testing.

The results were presented and discussed with researchers in regular meetings. It showed that once the users were familiar with the predefined head gestures the time needed to make a selection (a few seconds) was perceived appropriate by most of the participants. Additionally we found that the choice of gestures used in the selection process is very crucial and need to be configurable to cope with the needs and preferences of the individual user.

To provide better visual feedback and a more intuitive control possibility a mouse-like interface could be introduced to the hardware prototype in May 2011. This setup was presented to undergraduate students as part of a lecture at the university. Additional testing with students showed that this allowed reducing the time to make a selection between three and four seconds (including the time needed to hover over the selection).

The first mobile setup was introduced in June 2011 consisting of a Wi-Fi enabled laptop with integrated webcam. Within the range of the wireless-infrastructure, a user could select between two different applications available (the hardware model elevator and an elevator simulation). The selection of the application itself and the related control functions was performed entirely by head movements.

Throughout the project the students of the participating school were generally very eager and motivated to find a new solution to help their physically challenged colleagues with help of this project. It proved to be very beneficial conducting regular meetings both at the university and the school in order to discuss the project status and distribute tasks. The meetings helped the high school students to see their own work within a larger perspective and increase accountability for their tasks. The authors and project members also noticed a very application based mindset focussing at very specific solutions, unfortunately sometimes unrelated to the project goals itself (e.g. voice recognition), but nonetheless emerging interesting ideas. Here it was important to point out methodology and the focus of the project but still welcome and support input from all sides.

With assistance of motivated teachers, it was even possible to integrate parts of the project work in the regular classes at the school and a positive effect on motivation and productivity of the students within those periods was noticeable.

During the project's period 4-week internships at the university could be provided to motivated students during school holidays, in 2010. The interns were given specific programming and testing tasks related to the project. Concluding their assignments they presented their work to the scientific staff.

VI. Future Work

Based on good results and positive feedback from the first group of internships it was to initiate a second period of internships staring in July 2011. Following the main system implementation, smaller project tasks for the interns will be provided. To make sure the task are appropriate, the project members intend to offer topics related to HTML (Hypertext Markup Language) or ASP.NET (Active Server Pages for .NET) and other technologies that are already familiar to the high school students.

This adds to the philosophy of the XINU project that the general scientific and engineering work is accomplished the university. However, smaller tasks with a shorter time frame and well defined goals can be handed over to the high school students. These tasks extend their existing knowledge about techniques used in praxis and leave enough space for them to learn practical skills and deepen their knowledge. Some very time consuming tasks, e.g. extensive testing, give reason to be outsourced to students, who did not participate directly in the implementation and are less prone to following logical traps during evaluation. Additionally, researchers from the university enjoy the contact and fresh ideas of students, who help them to get a new view of their own research.

With regards to research content, the evaluation and development of a mobile solution will be continued. It is intended to provide a web interface to allow operation of the elevator model from any device capable to process internet webpages.

VII. Conclusion

The authors presented an educational project using state-of-the-art head pose estimation to control an elevator. Together, high school students, university students, teachers and researchers developed a prototype application consisting of a mechanical elevator and a PC with webcam-based control software. Two possible implementations were presented. First, defined head gestures can be used to trigger elevator control commands without pressing buttons. Second, a mouse-like selection of the floor is possible by only moving the head to position a cursor over a corresponding button on a screen.

Acknowledgment

This study is supported by the research program "Sparkling Science" by the Austrian Federal Ministry for Science and Research (BMWF).

Special thanks to Seeing Machines Inc. for their great support and interest in our research work.

References

[1] Internet reference: Sparkling Science. http://www.sparklingscience.at, May 2011

[2] Internet reference: Institute of Computer Technology, Vienna University of Technology. http://www.ict.tuwien.ac.at, May 2011

[3] Internet reference: School Centre Ungargasse. http://www.szu.at, May 2011

[4] C. Roesener, A. Perner, S. Zerawa, and S. Hutter, "Interface for non-haptic control in automation," in *Industrial Informatics (INDIN), 2010 8th IEEE International Conference on*, 2010, pp. 961 –966.

[5] S.-A. Zerawa, A. Perner, and C. Roesener, "Non haptic interaction system," in *20th International Symposium on Industrial Electronics, (In Press)*, 2011.

[6] G. S. Aikenhead, "Students' ease in crossing cultural borders into school science," *Science Education*, vol. 85, pp. 180–188, 2001.

[7] T. W. Burns, D. J. O'Connor, and S. M. Stocklmayer, "Science Communication: A Contemporary Definition," *Public Understanding of Science*, vol. 12, no. 2, pp. 183–202, 2003.

[8] M. Baklouti, M. Bruin, V. Guitteny, and E. Monacelli, "A human-machine interface for assistive exoskeleton based on face analysis," in *Biomedical Robotics and Biomechatronics, 2008. BioRob 2008. 2nd IEEE RAS EMBS International Conference on*, 2008, pp. 913 –918.

[9] M. Baklouti, S. Couvet, and E. Monacelli, "Intelligent camera interface (ICI): A challenging HMI for disabled people," in *Advances in Computer-Human Interaction, 2008 First International Conference on*, 2008, pp. 21 –25.

[10] V. Graveleau, N. Mirekov, and G. Nikishkov, "A head-controlled user interface," in *Proceedings of the 8th Int. Conf. on Humans and Computers, HC 2005*. Aizu-Wakamatsu, 2005, pp. 306–311.

[11] M. N. Mamatha and S. Ramachandran, "Automatic eyewinks interpretation system using face orientation recognition for human-machine interface," in *IJCSNS International Journal of Computer Science and Network Security*, vol. 9, no. 5, 2009, pp. 155–163.

[12] X. Yu, W. K. Chua, L. Dong, K. E. Hoe, and L. Li, "Head pose estimation in thermal images for human and robot interaction," in *Industrial Mechatronics and Automation (ICIMA), 2010 2nd International Conference on*, vol. 2, May 2010, pp. 698 –701.

[13] M. Voit and R. Stiefelhagen, "A system for probabilistic joint 3D head tracking and pose estimation in low-resolution, multi-view environments," in *Proceedings of the 7th International Conference on Computer Vision Systems: Computer Vision Systems*, ser. ICVS '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 415–424.

[14] Y. Matsumoto, N. Sasao, T. Suenaga, and T. Ogasawara, "3D model-based 6-DOF head tracking by a single camera for human-robot interaction," in *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, May 2009, pp. 3194 –3199.