Multimodal User Interface for Augmented Assembly

Sanni Siltanen, Mika Hakkarainen, Otto Korkalo, Tapio Salonen, Juha Sääski, Charles Woodward
VTT Technical Research Centre of Finland
Espoo, Finland
firstname.lastname@vtt.fi

Theofanis Kannetis, Manolis Perakakis, Alexandros Potamianos
Department of Electronics and Computer Engineering
Technical University of Crete
Chania, Greece
{thkannetis, perak, potam}@telecom.tuc.gr

***Abstract*—In this paper, a multimodal system for augmented reality aided assembly work is designed and implemented. The multimodal interface allows for speech and gestural input. The system emulates a simplified assembly task in a factory. A 3D puzzle is used to study how to implement the augmented assembly system to a real setting in a factory. The system is used as a demonstrator and as a test-bed to evaluate different input modalities for augmented assembly setups. Preliminary system evaluation results are presented, the user experience is discussed, and some directions for future work are given.**

Keywords—multimodality;augmented reality, speech control, gesture control, visual feedback,assembly;

Topic area—Multimedia Communication

I. INTRODUCTION

In industrial production, the growing number of product variants, and the need for customized products, shorter life-cycles, smaller lot sizes and accelerated time to market have increased demands on production equipment and concepts. The production companies strive for increasing the performance of production and innovative approaches and technologies are required. One challenge is assembly work that requires skilled manpower to perform work tasks in a specified sequence with careful attention and particular skill. The use of AR (Augmented Reality) has been proposed as a solution to this challenge [5, 6, 7]. AR systems can combine human flexibility, intelligence and skills with the computing and memory capacity of a computer.

One of the main challenges is to the generate concepts for a human worker to operate in complex, short series or in a customized production factory environment. Each individual product may have a slightly different configuration: the order of assembling parts may vary for different products and/or the number of phases in the assembly line may be large. Often the human memory capacity is unable to handle all the required information. The traditional approach is to use assembly drawings (blueprints) and instruction manuals to check content of each work task. The disadvantage is that finding, reading and verifying this assembly information takes time and breaks the actual assembly work. An on site AR system can give the information automatically via a suitable device and the assembly work can be made more fluent and efficient. The challenge is to create a system with a natural user interface and use devices that do not interrupt the actual assembly work, e.g., allow for hand-busy interaction.

The potential of wearable augmented reality has been investigated at the early stages of this research [5,6,7,12], but the wearable AR systems have often been too heavy and big for industrial use. Wearable AR has been used more successfully for fun application and games [11]. However, the rapid development of mobile devices has lead to small devices with enough processing capacity and long lasting batteries to enable light-weight mobile AR systems. Recently PDAs, camera phones [8, 9] and mini PCs [13] have been successfully used in AR applications. At present, mobile augmented reality was listed as one of the ten most potential technologies in the annual MIT Technology Review [10].

Multimodal interfaces allow the user to interact with a computer using more than one input and/or output modes. Multiple modalities offer additional flexibility and make machines readily accessible to a population of naïve or handicapped users. In addition, appropriately designed multimodal interfaces that exploit synergies among modalities can improve efficiency as well as naturalness of interaction [14,15,16]. Most human-computer interfaces employ tactile (keyboard and mouse) input and graphical output. Recently, traditional graphical user interfaces have been augmented or redesigned to include natural language, speech, haptic and gestural input. Speech interfaces are natural and prove especially valuable for mobile application where the devices are too small to support efficient and convenient tactile interaction. This is also true for eyes-busy and/or hands-busy interaction where graphical user interfaces with tactile input are disruptive to the user's task.

In this work, we focus on augmented/virtual reality interfaces and investigate the use of spoken language and gestures as an alternative mode of input for an assembly task [1]. This is a hand-busy, eye-busy interaction and the use of tactile input, e.g., keyboard, to command and control the application is both unnatural and inefficient. Our goal is to investigate if the use of a multimodal spoken dialogue and gestural interface to control an AR application enhances the user experience in terms of efficiency and user satisfaction.

II. PROPOSED METHOD

In this evaluation phase we use a simplified assembly task that simulates a real assembly work. The task is to put 3D parts in a puzzle box (see Fig 1). The user follows instructions and puts piece-by-piece desired parts according to augmented instructions at the right place and in the right order. The parts fit in only if assembled in the right way and order. For system development, this task has various advantages compared to real assembly task. First this set is cheap, portable and adjustments and changes are easy to make. Yet the task is real-enough and the actual devices are used (HDM, camera, etc). With this simplified task we can test different input and output modalities, the robustness of the augmented system and get valuable feedback for designing the actual factory tasks. We use a modified version of ARToolKit [17] with some additional features and improvements for augmentation.

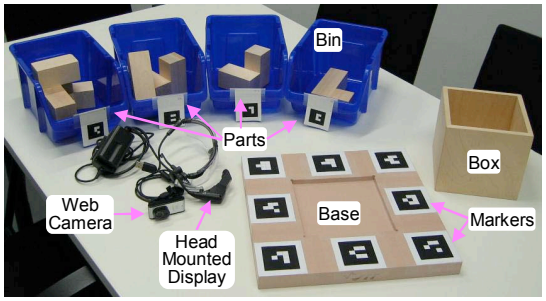

Figure 1: The overview of the demonstration system: the base, parts of assembly in bins, the box containing the parts, web camera and the head mounted display. The computing unit is not shown in this figure.

The assembly line worker often needs to wear safety glasses. We selected a very light weight display that can be attached to safety glasses to ensure minimum amount of parts/devices to be carried by the user. We used MicroOptical SV-3 PC Viewer as display that only weights less than 40 grams. The size of the display is 1cm x 2cm. A Logitech QuickCam for Notebooks Pro was also used. The camera was attached in the middle of the glasses so that the camera view and the user's view were consistent (Fig 2). In the real assembly work the wearable system should be as light weight as possible to enable real work. The interaction with the system should be natural and easy.

Figure 2: The user assembling the 3D puzzle box

For the first round of user evaluation we selected two potential input modalities: speech and gesture control. These were selected due to the hands-busy, eye-busy nature of the task and the limited additional hardware required (a small microphone). The architecture of the system allows adding modalities if required. In addition to augmented instruction, the system provides user feedback in textual and visual format on the display. No audio feedback was integrated in this version.

At startup the user selects the model to assemble (i.e. the correct instruction xml-file). After that the commands that the user can give to the system are to move forward to the next phase, move backward to the previous phase or start color calibration for the gesture recognition. The color calibration for gesture recognition can be started using speech input.

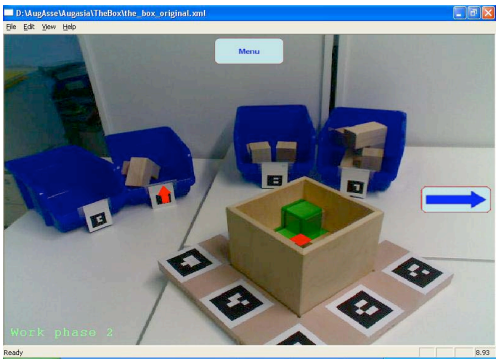

Figure 3: Augmented view: the right arrow indicates that "next phase" control has been detected

Gesture control

The system can be controlled using a head-up display (HUD)-like virtual menu. It consists of icons that can be selected by moving the hand over them. The menu has two states: active and inactive. In the inactive mode, the menu has an activation icon located at the center of the upper edge of the view. The user is able to activate the virtual menu by holding the hand on the activation area. As the menu is activated, two arrows are augmented to the view. The first arrow is located at the left upper corner, and by selecting it the user can move backwards in work phases. Similarly, the second arrow is located at the upper right corner, and it allows the user to jump to the next work phase. After a selection, the virtual menu disappears. If none of the icons is selected, the menu disappears after a small period of time. In practice, people tend to look at their hands while working on assembly task. As the camera is attached in the middle of the safety glasses pointing forwards, the hands appear most of the time in the center of the image. Thus, it is safe to place the menu items at the upper edge, and unintentional selections occur seldom

The hand detection algorithm has to be calibrated before usage. The procedure is carried out by asking the user to hold his hand at the calibration area (similar to menu icons). Then, the system acquires data for n frames. The hand detection is based on histogram back-projection presented in [4]. At the training phase, the hue histogram is constructed from the pixels located at the calibration area. As all the training data is obtained, the histogram is normalized. During usage, we use the value of the histogram as the probability of the pixel belonging to the object of interest. For each frame, we calculate the probability image of the pixels located in the area of menu elements. That is, we give every pixel a probability

value of belonging to the tracked object (hand). As the probability image is constructed, we calculate the percentage of the object area covering the menu item. The value is used to make a decision whether the icon is activated or not. During the user evaluations, the limit was set to 50% in more than 7 out of 10 consecutive frames to apply selection. With the frame rate of 15 fps, this means that the user needs to hold his hand over the item for 0.7 seconds. The application allows us to change all the parameters, so the response time can be adjusted to meet the user preferences. We used skin color in user evaluations. Should the assembly worker use gloves, their color could be used instead.

Figure 4: The gesture regocnition in action.

Speech control

To incorporate speech input in the augmented reality system external automatic speech recognition (ASR) software is required. The speech recognition system operates in a client-server architecture: the controller (ASR client) collects audio data and sends them to the ASR server along with configuration parameters. The server then performs recognition and then passes on the results to the controller. The controller then passes on the (parsed) results to the application. Results and speech information can also be displayed graphical via a GUI interfaces. The overall system architecture is depicted in the next figure.

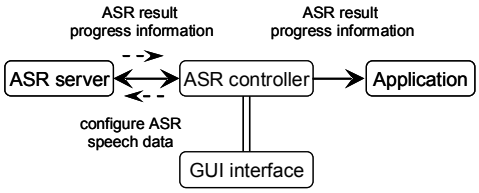

Figure 5: To decouple the application from the many details of handling speech input, the "ASR controller" is used. The GUI interface of the "ASR controller" allows to easily change various ASR parameters and fine tune the recognition process

The Sonic speech recognizer [2, 3] in its client-server mode, is used as "ASR server". In order to identify some simple commands (next, previous, next phase, previous phase etc.) tri-phone acoustic modes trained from males speakers are used and a simple grammar was written. Apart from the speech recognition, ASR server also keeps the log files that we used in the computation of the system word error rate (see evaluation). The "ASR controller", is used as the speech controller who establishes the communication between the system and the speech recognizer. It also allows to easily change various ASR parameters and fine tune the recognition process. Screenshots of the controller's GUI interface are shown next.

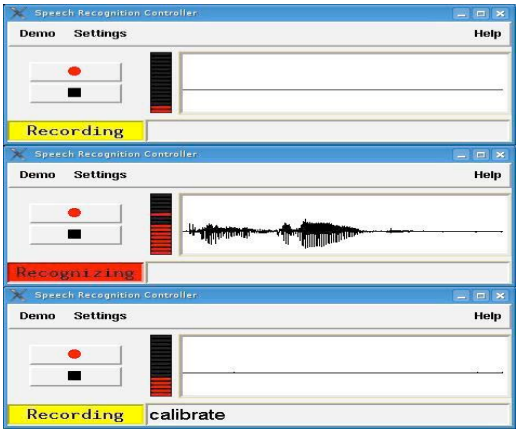

Figure 6: ASR controller GUI interface screenshots a) Recording speech b) Recognition in progress c) Recognition result.

First the user is connected to the ASR server using the controller GUI. In order to start the recognition process, the record button must be pushed. Once recording is started, voice activity detection is employed to determine user speech interaction (Fig 6a). A waveform or a spectrogram (depend on the user choice) is depicted along with a speech signal level meter for user feedback. Upon speech detection, audio data are sent to recognizer and the text "Recognizing" is shown in the left side of controller's GUI status bar (Fig 6b). Once recognition is finished, the recognized text is shown in the right side of the GUI status bar (Fig 6c). ASR controller acts as a proxy to the application by informing it of the recognition result and/or progress information which in turn can be shown by the application for user visual feedback. Finally some common configuration parameters can be easily adjusted using the "Settings" menu at any time (not shown here).

Feedback from the system

The assembly instructions are displayed on the HMD, thus one natural choice for feedback channel was to give visual and textual feedback on the same display. The virtual menu items are shown with red edges when hand is detected in the selection area (Fig 4). An arrow is also shown if the speech command is detected (Fig 3). A work phase count is shown at the lower left corner (Fig 3 and 4).

III. Experimental Results and Discussion

The system was evaluated by five novice users (two female, three male). Their experience of multimodal user interfaces varied from none to some and experience of virtual or augmented reality and previous use of data glasses/see-through displays varied. Two of them had no experience at all, one had tried data glasses before, and two of them were familiar with similar systems. The users tested the system in three experimental setups: first only with gesture or speech control, then only with the other control modality (gesture/speech) and lastly in truly multimodal mode (all modalities were allowed). In the third experiment, the users also calibrated the gesture recognition. The duration of the

task did not depend on the modality used or on the order in which they used the different modalities. We used an acoustic model trained on male speakers (see speech control) that performed poorly with the non-native English speaking females, to investigate if speech recognition performance affected their choice of input modality. Both actual usage mode statistics were collected and the preferred modality was elicited via a questionnaire at the end of the experiments. As expected the recognized output was poor for female users and they preferred to use the gesture control over speech. The opposite was true for male users who had good speech recognition performance and preferred the speech control. In general, in the third phase, users preferred the modality that performed the best in the first two test phases. In the third phase gesture control was used successfully 9 times and speech control 13 times.

Some other users tried to solve the same task with printed instructions (on one page). For this simple task people who used the printed instructions were able to perform quicker than those using the augmented system. However some users were confused by the printed instructions, especially with the orientation and identity of one puzzle piece. This was not an issue for the augmented system where users were able to turn the base box and reveal the structure of the parts.

The users found the location of the virtual menu on the top of the image exhausting for long use, as the user needs to raise a hand to reach the virtual menu. (In fact, the users could also lower their sight to get the hand to appear on the top of the image over the virtual menu, but none of the test users used this feature).

All users found the feedback from the system insufficient. They were often unsure whether the system understood their command or not, and if the system really moved to the next work phase. The display contained a small text on the lower left corner containing the text "Work phase n", but the text was too small for the users to notify the change of the phase count. The system also displayed a right (left) arrow on the upper right (left) edge of the image when user gave the command "next phase" ("previous phase"). The users suggested that the arrow should have blinked or otherwise noticeable reacted to a recognized command. Also audio feedback (a beep) and progress bar were suggested.

The assembly instruction of each part was animated, the part moved from top of the base box downwards to the desired location in correct posture. One of the users suggested a pause feature to the animation, to be able to freeze the animation while comparing the posture of the part in the hand and the augmented part.

IV. Conclusion and Future Work

Comparing printed instructions to augmented instructions the difference was subtle in this simplified example, but all users shared the opinion that it may be useful for more practical assembly tasks (e.g. installing a digital-TV box, putting together furniture, etc.). Also the multimodal input interface was favorably judged by the users.

In an ongoing research project, we will focus on industrial case: assembly of a tractor accessory's power unit. In this case, the assembly worker is guided by virtual objects and visual assembly instructions. In the development of the industrial case, we will take into the consideration the results from this evaluation. We will concentrate on the robustness of input modes and improve on the multimodal output of the system, including the placement of the virtual menu.

Currently it is the user's responsibility to notice if he/she has performed the task correctly by browsing back- and forward in the instructions. In the future, we will use computer vision to recognize whether the user has put the right part in the right place and use the system also for worker training.

Acknowledgment

This work was supported by the EU-IST Muscle NoE.

References

[1] Bolt, R. Put-That-There : Voice and gesture at the graphics interface. Computer Graphics, 14(3): 262-270, 1980.

[2] Pellom, B. SONIC: The University of Colorado Continuous Speech Recognizer. Technical Repor TR-CSLR-2001-01, University of Colorado, March 2001.

[3] Pellom, B., Hacioglu Kadri. Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task. in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong, April, 2003.9.

[4] Swain, M. J. and Ballard, D. H. 1991. Color indexing. *Int. J. Comput. Vision* 7, 1 (Nov. 1991), 11-32.

[5] Azuma R., "A survey of augmented reality", Presence: Teleoperators and Virtual Environments 6, 4 (Aug 1997), pp. 355 385.

[6] Azuma R., Baillot Y., Behringer R., Feiner S., Julier S., MacIntyre B., "Recent advances in augmented reality", IEEE Computer Graphics and Applications 21, 6 (Nov/Dec 2001), pp. 3447.

[7] Bimber O. and Raskar R., "Modern Approaches to Augmented Reality", SIGGRAPH 2005, Course Notes on Spatial Augmented Reality, 86 pp.

[8] Henrysson, A., Billinghurst, N., Ollila, M., "Virtual object manipulation using a mobile phone", Proc. 15th International Conference on Artificial Reality and Telexistence (ICAT 2005), Dec 5th 8th,2005, Christchurch, New Zealand, pp. 164171.

[9] Rohs, M., "Marker-Based Embodied Interaction for Handheld Augmented Reality Games", Proceedings of the 3rd International Workshop on Pervasive Gaming Applications (PerGames) at PERVASIVE 2006, Dublin, Ireland, May 2006

[10] Jonietz, E., "Augmented Reality: Special Issue 10 Emerging Technologies 2007" , MIT Technology Review, March/April 2007.

[11] Haller,M., Billinghurst, M.,Thomas, B., Emerging Technologies of Augmented Reality, 2006, IGI Publishing, Hershey, PA, USA, pp.367-

[12] Thad,S. et al, Augmented Reality through Wearable Computing, Presence, Special Issue on Augmented Reality, 1997

[13] Honkamaa P., Siltanen S., Jäppinen J., Woodward C., Korkalo O., "Interactive outdoor mobile augmentation using markerless tracking and GPS." To appear in Proc. VRIC – Laval Virtual 2007.

[14] A. Potamianos et al, "Design principles and tools for multimodal dialog systems," in *Proc. ESCA Workshop Interact. Dialog. Multi-Modal Syst.*, (Kloster Irsee, Germany), June 1999.

[15] S. Oviatt et al, "Designing the User Interface for Multimodal Speech and Pen-based Gesture Applications: State-of-the-Art Systems and Future Research Directions", 2000.

[16] Sharon Oviatt , "Design Robust Multimodal Systems for Universal Access", Center for Human Computer Communication, Computer Science Department, Oregon Graduate Institute of Science & Technology, Proceedings of the EC/NSF workshop, 2001.

[17] ARToolKit: http://www.hitl.washington.edu/artoolkit/