A Head-Mounted Multimodal Augmented Reality System for Learning and Recalling Faces

Daniel Sonntag
Christian Schulz
German Research Center for AI
Saarbrücken, Germany
sonntag@dfki.de,
chschulz@dfki.de

Markus Weber
Takumi Toyama
German Research Center for AI
Kaiserslautern, Germany
Markus.Weber@dfki.de,
Takumi.Toyama@dfki.de

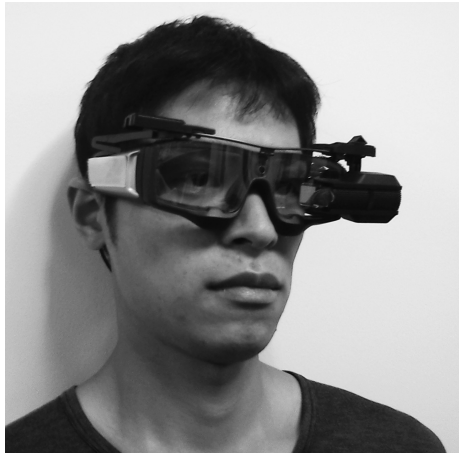

Figure 1: User with head-mounted HMD / eye-tracker combination.

CHI'13, April 27 – May 2, 2013, Paris, France.
ACM 978-1-XXXX-XXXX-X/XX/XX.

Abstract
We present a new augmented reality (AR) system for knowledge-intensive location-based expert work. The multimodal interaction system combines multiple on-body input and output devices: a speech-based dialogue system, a head-mounted augmented reality display (HMD), and a head-mounted eye-tracker. The interaction devices have been selected to augment and improve the expert work in a specific medical application context which shows its potential. In the sensitive domain of examining patients in a cancer screening program we try to combine several active user input devices in the most convenient way for both the patient and the doctor. The resulting multimodal AR application has the potential to yield higher performance outcomes and provides a direct data acquisition control mechanism. It effectively leverages the doctor's capabilities of recalling the specific patient context by a virtual, context-based patient-specific "external brain" for the doctor which can remember patient faces and adapts the virtual augmentation according to the specific patient observation and finding context. In addition, patient data can be displayed on the HMD—triggered by voice or object/patient recognition.

ACM Classification Keywords
H.5.2 [User Interfaces]: Input Devices and Strategies, Natural Language, Graphical HCIs, Prototyping

Author Keywords
Augmented Reality, Medical Healthcare, Realtime Interaction

General Terms
Experimentation, Human Factors, Performance

Introduction
In ubiquitous computing, it can be said that most profound technologies are those that disappear by weaving themselves into the fabric of everyday (professional) life. It would be even better if we could carry and wear those technologies on our bodies which would make us rather independent of the location in which they are used. For documentation purposes in the "myths of the paperless office,[10]" digital pens, for example, have been invented and used to implement paper-based interactions in digital environments [11, 12]. The problem is that you cannot easily replace a screen-based (laptop) computer, because the display is often missing for a proper interaction, and time-based transient interaction modes such as speech dialogue cannot replace them properly. Recent display systems are available as head-mounted displays (HMDs) which provide new ubiquitous possibilities for interaction and real-time systems often referred to as cyber-physical systems [7]. In this paper, we propose a new multimodal interaction system that can also learn important, task-relevant (visual) information. The multimodal interaction system combines multiple on-body input and output devices: a speech-based dialogue system, a head-mounted augmented reality display, and a head-mounted eye-tracker. The interaction devices have been selected to augment and improve the expert work for a particular application domain, the physical examination of patients during cancer screening. The scenario is shown in figure 2.

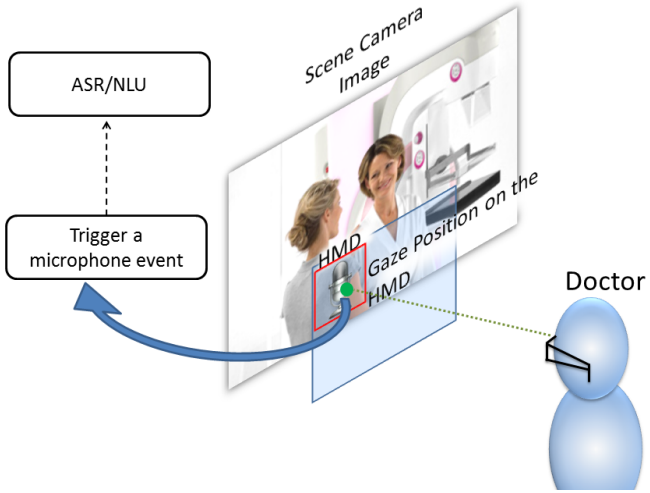

Figure 2: Patient Examination Scenario where the doctor wears a head-mounted eye-tracker and an HMD.

Our main contribution is that we bring two major, normally separated, goals of intelligent interaction together, thereby combining active and passive user input devices in the most convenient way for both the patient and the doctor. First, we want to leverage the doctor's capabilities to recall patients and inform himself about a specific patient record. For this purpose, we implemented the first online, head-mounted face learning system which uses a mobile eye-tracker. In the most elaborate interface mode, we allow for a "real-time interactive" face detection. Second, in the interactive experience with the doctor and the patient, the system should improve the performance of the human-computer interaction and the usability in the patient context. For this purpose, we use the gaze position on the HMD in combination with an automatic speech recognizer (ASR) as part of the multimodal interaction structure. As a result we

implemented the first multimodal interaction system which combines a mobile eye-tracker with a head-mounted display, and this in combination with speech-based interaction which can now be evaluated for its task-based usability. Towards this goal, we provide a preliminary evaluation of the face learning and detection mode to identify avenues for further improvements of the technical machine learning approach. The design of the learning and detection system could lead to new visions of other application domains interacting with mobile HMD technology and provide more dedicated scenarios where the object/face detection task can be facilitated.

Related Work

Motivated by previous findings showing the relevance of eye-gaze in multimodal conversational interfaces [9], we extended the passive input idea to active user input in the medical augmented reality realm. This also extends the work of using the gaze information to resolve the ambiguities of users speech [17]. In the medical domain, several experimental settings with HMDs (neither using eye-trackers nor speech input) have been investigated. Consider the following application example: with the standard working position at the ultrasound machine with a normal display, a doctor can look at the display to see the results of a scan. But he or she has to turn the head towards the patient when repositioning the probe. Thus in [3], the usage of a HMD in ultrasound scanning task has been investigated as there is no need to turn the head any more. Image guided surgery has been introduced successfully in many modern operating rooms. In order to improve the navigation on a standard computer monitor, [15] enhanced the direct sight of the physician by an HMD overlay of the virtual data onto the doctor's view in the context of the patient. In [5], HMDs have been used in various forms to assist surgeons and other medical personnel to support and improve the visualisation of the workplace related procedures.

Complete Scenario and Multimodal Dialogue System Architecture

Over the last several years, the market for speech technology has seen significant developments [8] and powerful commercial off-the-shelf solutions for speech recognition (ASR) or speech synthesis (TTS). For industrial application tasks such medicine discourse and dialogue infrastructures are available [13].

The run-time environment we use in this special I/O scenario is based on a middleware platform which connects system components and follows a hub-and-spoke architecture. The architecture comprises a number of functional components that deal with tasks like modality-specific interpretation, context-based interpretation, interaction and task management, target control, presentation management, and modality-specific generation. All functional components are generally application-independent and are configured by respective models. We use a distributed dialogue system architecture, where every major component can be run on a different platform, increasing the scalability of the overall system (figure 3). Thereby, the dialogue system also acts as middleware between the clients and the backend services. This architectural characteristic is very suitable to monitor internal and external messages of the three major parts: the multimodal interface, the dialogue system, and the event bus. The event bus basically provides a data streaming functionality for the automatic speech recogniser (ASR/NLU) and the text-to-speech synthesis (TTS).

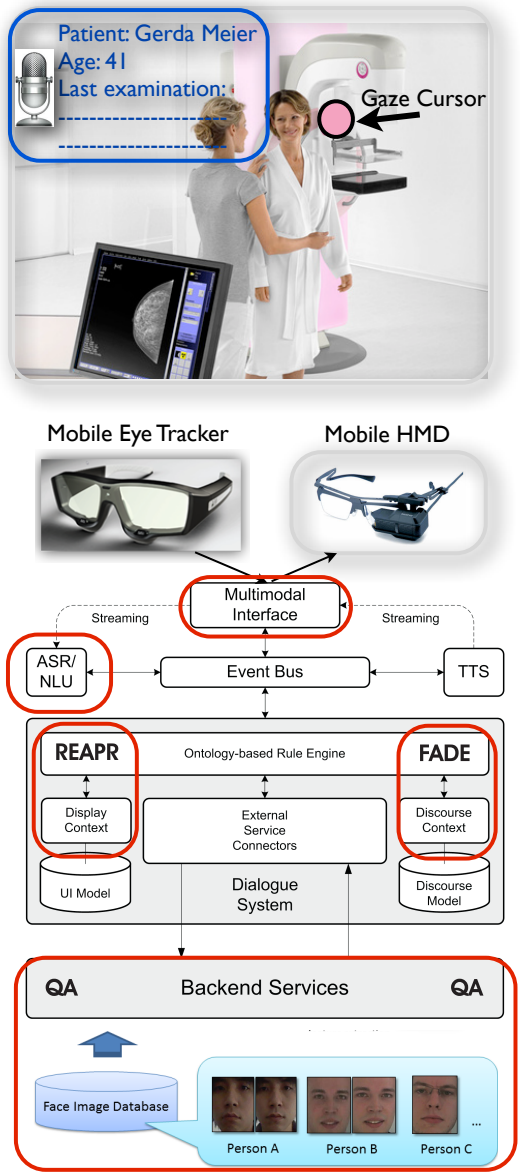

Figure 3: HCI Architecture and Online Learning Framework. The red boxes indicate the active dialogue modules during the HMD interaction. The resulting augmentation in the see-through HMD contains patient name, age and last examination information (background scene photo courtesy of Siemens AG). The gaze cursor is not visible in the HMD.

One of the functional components of the dialogue system is the fusion and discourse engine (FADE), while another is the reaction and presentation component (REAPR). Both are important for the mobile eye-tracker and mobile HMD scenario.

FADE is needed to compute the discourse-context information. It is based on a fusion approach which waits for appropriate multimodal input to be fused together for a interpretation that can be stored in the question and answer (QA) database.

Dialogue processing in REAPR means that the dialogue information state is implemented. Information state theory of dialogue modelling basically consists of a description of informal components (e.g., obligations, beliefs, desires, intentions), their formal representations (data structures such as ontology instances as in our case), dialogue moves (often conceived of as speech acts), update rules (trigger moves and updates of state information), and update strategies (in which order to apply the applicable update rules) [6]. While obligations and dialogue moves do no have a role to play in the current version, the integration with context-based presentation of patient information (or the recognition itself) is straightforward. Essentially, the display context defines what the user sees in the HMD. Upon recognition of an object (or person) by the eye-tracker and recognisers, REAPR triggers the context-dependent display in the mobile HMD. Other context factors, such as patient and examination context,

can be smoothly integrated into the context model. In addition, REAPR is also responsible for initiating a potential TTS synthesis. For example "that's patient x, age 37". In our scenario, we focus on the multimodal dialogue interactions which directly relevant to active learning part of the HMD and eye-tracker scenario:

1. The user gazes at the microphone button and starts the ASR (we use dwell time for selection, 500ms; the Midas touch problem is circumvented by the size and positioning of the microphone).
2. The user says: "learn a new person," which issues a respective command in the multimodal interface and the eye-tracker connection.
3. Upon face recognition, FADE gets informed about a *new* face and remembers the database instance which is stored in the service backend.
4. The user looks again on the microphone and starts the ASR.
5. The user says: "This is a new patient, Peter Meier, " which the FADE module fuses into a face image database command now containing the face classification features and the name of the newly created patient database instance.

Eye-Tracker and HMD Display

Over several decades, researchers investigated a lot in the area of eye tracking and gaze-based interfaces. As a result of recent progress of this research area, a light-weight and compact mobile eye-tracker is available today; it enables us to use gaze as an interface in various scenarios [14, 2].

In our multimodal dialogue system, we use the SMI Eye Tracking Glasses (ETG)[1] in order to recognise which person the doctor is looking at (in the examination room) and to obtain the gaze position in the HMD. ETG is a binocular eye tracker, which captures the images of both eyes and computes the gaze position in a scene image (which, in turn, is captured by the scene camera located in the center of the glasses.) In order to obtain accurate gaze positions, the user is required to do a system calibration before using it. The calibration is done by looking at one (or three) point(s) indicated by the system. Brother recently released a product of new head mounted display, whose feature is the transparency of the display. The user can see the environment through the display when nothing is displayed. We combined this HMD with the ETG.

Display Calibration
One of the most primitive and intuitive ways for calibrating HMDs is to use a gaze cursor. By using user gaze as an input source for commands, the user can control the system intuitively. In order to display a gaze cursor on the HMD, we need to calibrate the HMD so that the system can compute gaze position on the HMD.

Examples of scene image and the user view are shown in figure 4. From the user perspective, the HMD can be seen as shown in the top of the image (light blue rectangle) but for the scene camera, where the HMD is located, it is unknown. We propose two different methods for display calibration, gaze based calibration and scene image based calibration as shown in figure 4. We implemented both of them for a comparison. From the SMI eye tracker, we then receive a scene image (1280x960) and a gaze position (coordinate of the scene image, e.g., x:220,y:341).

[1] http://eyetracking-glasses.com/

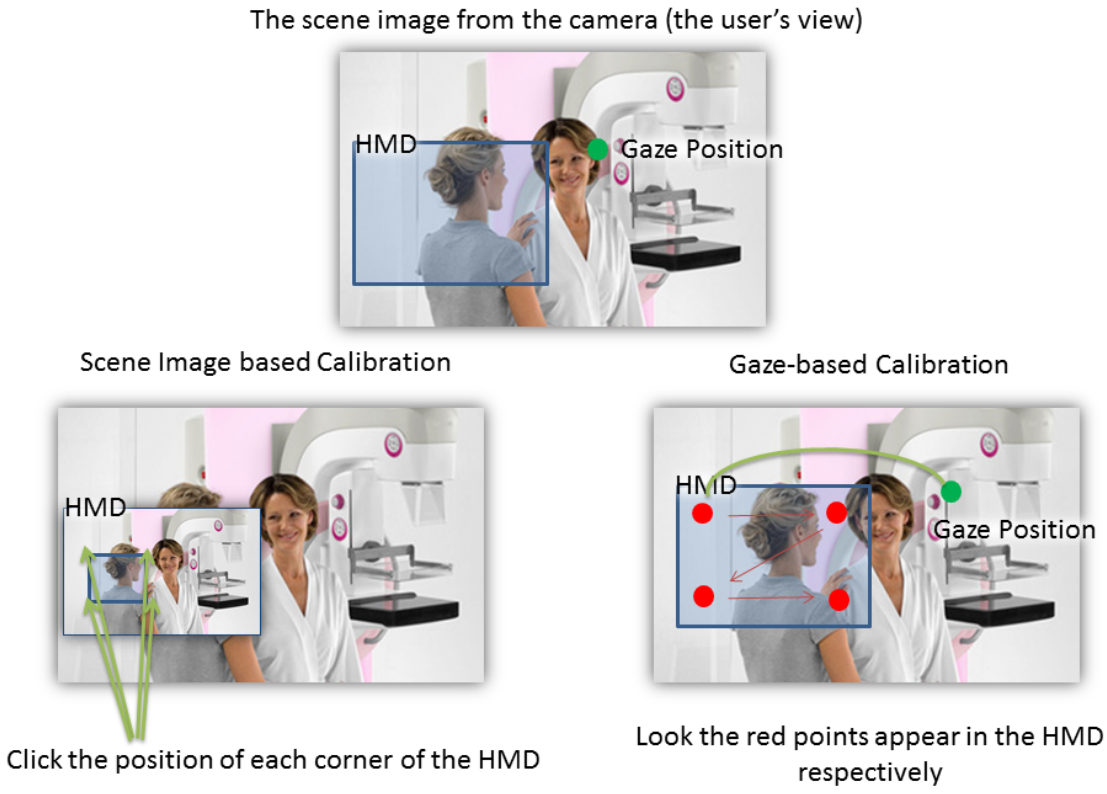

Figure 4: Display Calibration

Gaze-based Calibration The gaze-based calibration can be seen on the the lower right image of figure 4. A red point appears in the HMD when the calibration process starts. The user has to look the point, which means he has to look at it form a few seconds. Then, the system computes the mean position of the gaze samples received from the eye tracker. By repeating this process four times with different point locations, we get four corresponding points between eye tracker and the HMD. These corresponding points supply us with the calibration parameters in the mathematical formulas of perspective transformation. The drawback of this method is that the performance dependents fully on the accuracy of the eye tracker. If you get noisy gaze data, you will most likely get wrong transformation parameters.

Scene-based Calibration In scene-based calibration, the whole scene image captured by the camera is shown in the HMD (figure 4, lower left). The user has to click on the position of each corner of the HMD *in the HMD view*. In this way, the calibration system knows where the HMD is located in the scene image. (We use a mouse for the calibration interaction.) Once you get the location of the HMD in the scene image, you can easily transport the gaze position from the eye-tracker to the HMD. This method seems to be more promising because you dont have to rely on the performance of the eye tracker. Furthermore, you may also overlay an image in the augmented reality vision.

We used the second method in the application system and we think this a better option from the performance point-of-view. However, the process of calibration might be complicated for users in a working environment. Therefore we still need to evaluate the different methods from a usability and ease-of-use perspective.

After calibration, we get the gaze position in the HMD, so we can trigger the microphone event when the user gazes at the microphone button (which appears in the left top of the HMD).

Face Recognition

In this system, we combine a face recognition framework with the eye-tracking system in order to recognise the patient being examined, i.e., to provide the doctor with the image-content-based "external brain."

The face recognition procedure follows the following procedure:

1. The scene image and the gaze position is obtained from the eye tracker. Then:
2. Find the face closest to the gaze position. The detection of the faces in the video image(s) is done via Haar-like features and an AdaBoost learning cascade (see [16] for further details: we used the OpenCV2.3 implementation, in combination with a raster raster scan method.)
3. Compute the Local Binary Pattern (LBP) [1] features from the face images.
4. Execute a nearest neighbor search and find the nearest face from the database (to improve the speed of search, we may later use an approximate nearest neighbour search like [4] once distinctive feature are identified on a larger population).
5. Get a result and send it from the multimodal input interface to the REAPR display context which triggers the presentation of real-time results in the HMD.

It is to be noted that in video streams, the system sometimes returns false results (for example, if we have 85% accuracy, 15 frames out of 100 frames are expected to be false results). In order to remove this kind of noisy results and to obtain more accurate results, we also applied an attention detection approach [14]. This simple heuristic counts the number of frames that have the same recognition results and if the number reaches the threshold, the system assumes that the user is really looking at an object (or person). [14] report that this approach effectively detects the object that drew the user attention in combination with a 3D object recognition framework. In our tests, we observe this also holds for the face recognition framework in a mobile, non-stationary, environment of the head-mounted camera.

Preliminary Evaluation of the Learning System

In our view, a preliminary evaluation has to answer the following question: "Can we produce an active learning environment with the head-mounted user interface?"

We conducted a preliminary test for (online) face recognition to answer this question partly. Thereby, we focussed on the general applicability of the employed technical methods for a mobile, head-mounted active learning environment. In this preliminary test, we used 5 images from 8 different persons for training. Test images were taken in the same condition as in the training (in the same room and with the same lighting conditions similar to the examination room in figure 2). We achieved 100% precision and 68% recall. This results indicate that if the training is done in the same condition as the use-case, this online face recognition method performs well.

The recall rate is low when the patient's face is too dark so that the face detection cannot be successfully performed. However, in our examination scenario, we can provide for suitable conditions for face recognition (good lighting to reveal many individual face textures), thus this problem could be solved in a real examination environment.

Methods and Materials

The study was conducted in our research lab. We used two standard laptop computers (ThinkPads X201) to connect to the eye tracker and the HMD, to classify the incoming data streams, and to synchronise the behaviour with the speech-based dialogue system. The multimodal dialogue system (except for the multimodal interface) runs on a third ThinkPads X201 laptop computer. Standard Java network packages take care of the data exchange between the head-mounted I/O devices and the dialogue system.

Conclusion

We combined multiple on-body input and output devices, namely a speech-based dialogue system, a head-mounted augmented reality display, and a head-mounted eye-tracker, and implemented a complete scenario in a specific medical application context which shows its potential. The choice of the HMD device and the eye-tracker is very important because of the calibration need. One has to consider the precision of the mapping of the HMD result to the screen as well as the handling of the eye-tracker gaze position. The resulting interaction should, however, be very natural to the user, like looking on thinks and imagining their relevance to the working context. A little gaze and speech input "does the rest" to make daily routine a bit more effective and yield higher performance outcomes on similar knowledge intensive tasks.

Currently we use a simple nearest neighbour search method but this can be extended to an approximate nearest neighbour method such as [4], in order to expand the size of the test database to become productive. By using this kind of approximate method, it is assumed to be able to cover more than 100 person faces which would, in combination with the clinical records that can be retrieved from the QA database, become an "external brain" to cognitively loaded working domains like cancer screening.

Acknowledgements This research has been supported in part by the THESEUS Program in the Radspeech Project, which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016.

References

[1] Ahonen, T., Hadid, A., and Pietikäinen, M. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell. 28*, 12 (2006), 2037–2041.

[2] Bonino, D., Castellina, E., Corno, F., Gale, A., Garbo, A., Purdy, K., and Shi, F. A blueprint for integrated eye-controlled environments. *Universal Access in the Information Society 8*, 4 (2009), 311–321.

[3] Havukumpu, J., Vähäkangas, P., Grönroos, E., and Häkkinen, J. Midwives experiences of using hmd in ultrasound scan. In *NordiCHI*, A. I. Mørch, K. Morgan, T. Bratteteig, G. Ghosh, and D. Svanaes, Eds., ACM (2006), 369–372.

[4] Indyk, P., and Motwani, R. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing* (Dallas, Texas, USA, 1998), 604–613.

[5] Keller, K., State, A., and Fuchs, H. Head mounted displays for medical use. *J. Display Technol. 4*, 4 (Dec 2008), 468–472.

[6] Larsson, S., and Traum, D. R. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Nat. Lang. Eng. 6*, 3-4 (2000), 323–340.

[7] Lee, E. A. Cyber physical systems: Design challenges. Tech. Rep. UCB/EECS-2008-8, EECS Department, University of California, Berkeley, Jan 2008.

[8] Pieraccini, R., and Huerta, J. Where do we go from here? research and commercial spoken dialog systems. In *Proceedings of the 6th SIGDdial Worlshop on Discourse and Dialogue* (September 2005), 1–10.

[9] Prasov, Z., and Chai, J. Y. What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, IUI '08, ACM (New York, NY, USA, 2008), 20–29.

[10] Sellen, A. J., and Harper, R. H. *The Myth of the Paperless Office*. MIT Press, Cambridge, MA, USA, 2003.

[11] Signer, B., and Norrie, M. C. PaperPoint: a paper-based presentation and interactive paper prototyping tool. In *TEI '07: Proceedings of the 1st international conference on Tangible and embedded interaction*, ACM (New York, NY, USA, 2007), 57–64.

[12] Sonntag, D., Liwicki, M., and Weber, M. Interactive paper for radiology findings. In *Proceedings of the 16th international conference on Intelligent user interfaces*, IUI '11, ACM (New York, NY, USA, 2011), 459–460.

[13] Sonntag, D., Reithinger, N., Herzog, G., and Becker, T. *Proceedings of IWSDS2010—Spoken Dialogue*

Systems for Ambient Environment. Springer, LNAI, 2010, ch. A Discourse and Dialogue Infrastructure for Industrial Dissemination, 132–143.

[14] Toyama, T., Kieninger, T., Shafait, F., and Dengel, A. Gaze guided object recognition using a head-mounted eye tracker. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, ACM (New York, NY, USA, 2012), 91–98.

[15] Traub, J., and Sielhorst, T. Advanced display and visualization concepts for image guided surgery. *Display Technology, ...* (2008).

[16] Viola, P. A., and Jones, M. J. Rapid object detection using a boosted cascade of simple features. In *CVPR (1)* (2001), 511–518.

[17] Zhang, Q., Imamiya, A., Go, K., and Mao, X. Overriding errors in a speech and gaze multimodal architecture. In *Proceedings of the 9th international conference on Intelligent user interfaces*, IUI '04, ACM (New York, NY, USA, 2004), 346–348.