A Real-Time Framework for Natural Multimodal Interaction with Large Screen Displays

N. Krahnstoever[1], S. Kettebekov, M. Yeasin, R. Sharma

Pennsylvania State Univ., Dept. of Comp. Science and Eng.
220 Pond Lab, University Park, PA 16802, USA
Phone: (814) 865-9505, Fax: (814) 865-3176

Advanced Interface Technologies, Inc.
403 South Allen St.
Suite #104, State College, PA 16801, USA

{krahnsto,kettebek,yeasin,rsharma}@cse.psu.edu

Abstract

This paper presents a framework for designing a natural multimodal human computer interaction (HCI) system. The core of the proposed framework is a principled method for combining information derived from audio and visual cues. To achieve natural interaction, both audio and visual modalities are fused along with feedback through a large screen display. Careful design along with due considerations of possible aspects of a systems interaction cycle and integration has resulted in a successful system. The performance of the proposed framework has been validated through the development of several prototype systems as well as commercial applications for the retail and entertainment industry. To assess the impact of these multimodal systems (MMS), informal studies have been conducted. It was found that the system performed according to its specifications in 95% of the cases and that users showed ad-hoc proficiency, indicating natural acceptance of such systems.

Keywords: Continuous Gesture Recognition, Multimodal HCI, Speech-Gesture Co-Analysis, Visual Tracking, Real-Time System.

1. Introduction

Although humans communicate with each other through a range of different modalities, human-machine interfaces rarely reflect this. A successful HCI system should mimic the kind of effortless and expressive natural interaction that humans are accustomed to in everyday communication. It is well known that speech and gesture compliment each other and when used together, create an interface more powerful than either modality alone. Hence, coverbal gesticulation has the best prospects of achieving effortless and expressive HCI. Integration of speech and gesture has tangible advantages in the context of HCI, especially when coping with the complexities of spatial representations.

The requirements of the natural interactive system therefore include the ability to understand multiple modalities such as speech and gesture, where information is somehow distributed across the modalities. Up to date there have been several designs of multimodal systems. However, Bolt's "put that [point] there [point]" [1] paradigm still prevails. While there were some advances on including speech recognition into limited domains, most of the gesture recognition work is limited to understanding artificially imposed signs and gestures, e.g. [2,3] and often involve cyber gloves or electronic pens. The resulting MMS are far from satisfying the "naturalness of interaction" criterion. For instance, studies on pen-voice interface for information query and manipulation of electronic maps indicate, that linguistic patterns significantly deviated from canonical English. In contrast, we believe that by fusing remotely sensed natural gestures with speech along with adequate and timely feedback through a large screen display, users are able to achieve a much more natural level of interaction.

Psycholinguistic studies have shown that natural gestures do not string together in syntactic bindings [4]. Rather, they fall under the category of deictic gestures as opposed to symbolic (predefined) gestures as in sign language [5]. Hence, our system is trained to recognize unconstrained natural *deictic* gestures that a user is exhibiting when interacting with a large display. Our system differs from related systems [6] in that the gesture recognition is based on learned statistical gesture models and a trained speech gesture co-occurrence analysis.

The framework presented in this paper has evolved from a number of previous multimodal speech gesture interfaces. The Campus Map [7] was a direct result of the weather narration keyword/gesture co-occurrence analysis [8]. The subsequent Crisis Management system (XISM) [9] was first system to be running on a single processing platform. Rather than keyword spotting, continuous speech recognition was employed. Further improvements addressed the problems of user initialization and error recovery and finally removed the need for any user devices by using ceiling mounted microphone domes. Other systems related to this work are MIT's Smart projects [10] and Microsoft's Easy Living system [11].

2. Design of the Proposed Framework

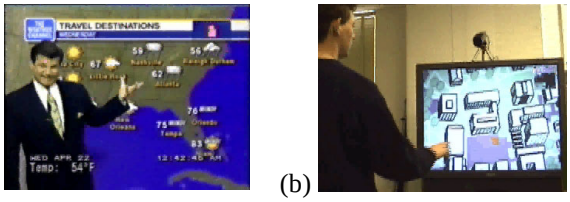
The main feature of the proposed framework is the use of speech and gesture to create a natural interface. The system is designed to accommodate the use natural gestures and speech commands of an experienced as well as an inexperienced user to increase the usability of the system in domains where user training is not feasible. Another important aspect is the use of a large screen display to provide appropriate feedback to the user. Large screen displays are a natural choice for many applications, especially interaction with spatial and geocentric data, immersive virtual reality environments and collaborative sys-

[1] Corresponding author.

tems that allow interaction with multiple users simultaneously.

2.1. Considerations for Multimodal HCI

The problems with designing a MMS that functions outside the laboratory range from conceptual to system design issues. A slow progress in addressing conceptual problems in HCI has caused an inherent lack of adequate user interaction data or the so-called chicken and egg problem [12]. To design a natural MMS, e.g., using statistical techniques, one needs valid multimodal data. However, it is impossible to collect the data unless a system exists. Several solutions to this dilemma are possible, including Wizard-of-Oz style experiments where a human behind the scene plays the role of a hypothetical MMS. However, this method does not guarantee a timely and accurate system response, which is desirable for eliciting adequate user interaction. Instead, Bootstrap and Evolve strategies were used. Comparative analysis of the weather channel narration broadcast is closely related to the desired type of gestural interaction [12]. It led to the development and statistical training of appropriate gesture recognition models at the bootstrapping stage [8]. These studies have revealed that in natural HCI with a large display, corresponding gestures and keywords exhibit some temporal pattern of their alignment that helps in disambiguating the meaning of utterances [12].



(a) (b)

Figure 1: Weather map analysis (a) and the Campus Map (b).

From a system design perspective it has to be acknowledged, that resource constraints in general severely limit the choice of sensing technology. If integration into a single processing unit is desired, many compromises have to be made. In addition, smooth and automatic interaction initialization, robust real-time visual processing, error recovery and graceful interaction termination are very important for ensuring user satisfaction. Unexpected system behavior almost always leads to a sharp drop in perceived system quality and ultimately acceptance failure. This requires a holistic approach to multimodal HCI system design, which however comes at the cost of system and integration complexity. This paper presents a design of a multimodal framework that attempts to address all of the above issues.

3. System Components

To capture speech and gesture commands the *i*Map framework utilizes a directional microphone and a single active camera. The tasks that the system needs to perform vary over time and are given an empirically evolved state transition model of an interaction session between a user and the system (Figure 2).

3.1. Interaction Session

An interaction session consists of three main phases. During the **initialization phase** the interaction dialogue between a new user and the system is established. It is followed by **interaction phase,** where the actual communication between the user and the system takes place. Finally, the **termination phase** is entered when the user (or the system) decides to conclude the dialogue.

Initialization Phase: In the absence of any users within the sensor range, the system is in the **wait state**. User detection is achieved through face detection [13]. Not only does the detection of a face indicate the presence of a person in front of the system, but it also assures that the person is looking at the system, which is strong evidence for a users desire to initiate an interaction dialogue. Each detection leads to subsequent head tracking. If at least one person has been detected, the system enters the **attraction state** in which it tries to establish a dialogue with one of the currently tracked people. In addition, the system continues to detect and track new arrivals. If at any point all people leave the sensor area, the system falls back into the waiting state.

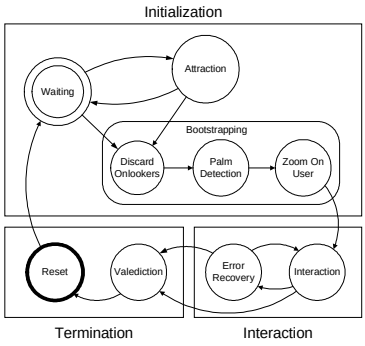


Figure 2: State transition model of an interaction session between a user and the system.

The placement of the camera and the microphone sensor require a user to interact with the system from a certain location to ensure optimal performance. Therefore, the system encourages detected people to step closer and guides them towards the proper location for interaction. Once a person has stepped into the proper spot and found to be facing the system, the system enters the final **bootstrapping state** of the initialization phase. The system immediately discards all processing of onlookers since all the available resources have to be invested into the dialogue with the main user. It furthermore performs palm detection to obtain an initial location of the user's active hand(s) and initializes the hand-tracking algorithm. Finally, it adjusts its active camera to adjust to the exact location, height and size of the user to allow optimal sensor utilization after which the interaction phase is initiated.

Interaction Phase: During the interaction phase, the actual dialogue between the system and the user commences. The system utilizes speech recognition, motion analysis and gesture recognition as its main interaction modalities. The vision-based modalities mainly rely on robust continuous head and hand tracking based on motion and color cues. From the hand trajectory data, a gesture recognition module continuously extracts free hand gestures using stochastic gesture models (cf. [18]). Rec-

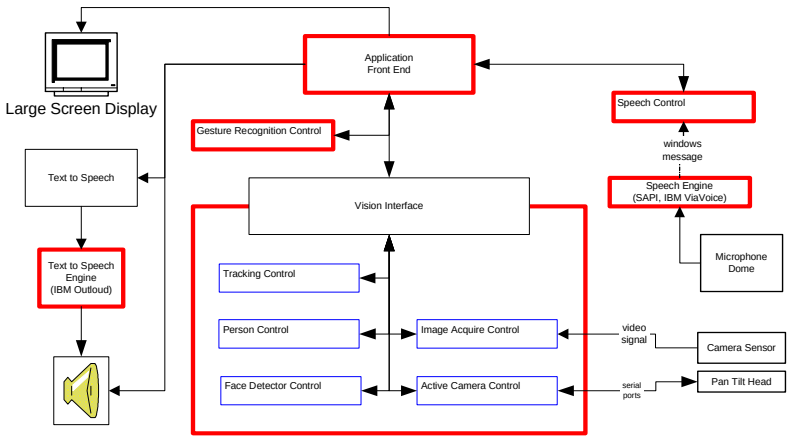


Figure 4: Overview of the iMap system architecture. Each of the bold framed boxes constitutes a separate thread of execution.

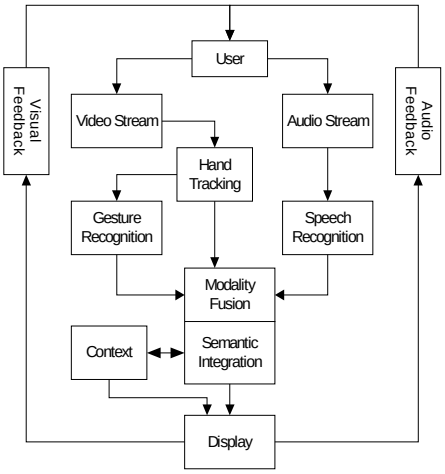


Figure 3: Logical flow of the system.

ognized gestures are combined with speech recognition results by a speech-gesture modality fusion module (Figure 3). The semantic integration of the final user commands depends heavily on the application and the time varying context of the system, which constrains the set of possible user actions for increased response reliability.

Termination Phase: An HCI system may gracefully terminate an interaction process, for example because the user has actively indicated that the session is concluded. A much more difficult problem is the sudden termination by a user, for example because the user chooses to walk away. It is necessary to constantly run a diagnostic error recovery module that decides whether or not the user is still present. It is important that this decision is made with an as low as possible false alarm rate, because the wrongful termination of an interaction session leads to user confusion and hence dissatisfaction.

Upon termination, the system informs the user or potential onlookers, that the session is terminated ("valediction"). Then it resets its internal states, positions the active camera in wide-angle initialization mode and switches back to the initialization phase.

3.2. Visual Components

To ensure the smooth progression of an interaction sessions as outlined above, a large number of vision (face detection, palm detection, head and hand tracking) and speech (command recognition, audio feedback) related components have to cooperate together under tight resource constraints on a single processing platform. The link between system responsiveness and user satisfaction mandates a strict adherence to the maximum possible processing rate (30 frames/sec or possibly 60 fields/sec) with respect to motion tracking and the associated visual feedback. Since all systems are integrated onto a single standard PC the allowable complexity of motion tracking methods is limited, especially, because the system latency has to be minimized to avoid a "sluggish" interface experience.

Face Detection: One of the most important and powerful components in the system is the face detector for robust user detection and continuous head track status verification. The implementation [13] is based on neural networks and favors a very low false positive ROC of <0.5% .

Palm Detection: With the proper camera placement and a suitable skin color model extracted from the face region, strong priors can be placed on the potential appearance and location of a user's active hand in the view of the camera. The automatic palm detection rests on the assumption that the object to be detected is a small skin colored blob-like region below and slightly off center with respect to the users head. In addition, the palm detector favors but does not rely on the occurrence of motion at the location of the hand and integrates evidence over a sequence 60 frames (cf. [14] for details).

Head and Hand Tracking: The algorithms for head and face tracking are based on similar but slightly different approaches. Both trackers are based on rectangular tracking windows whose location is continuously adapted using Kalman filters to follow the users head and hand. While the head tracker relies solely on skin color image cues, the hand tracker is a continuous version of the Palm Detector [14] and optimized to track skin colored moving objects. Prior knowledge about the human body is utilized in avoiding and resolving conflicts and interference between the head and hand tracks. The tracking methods used are based on simple imaging cues but extremely efficient and require less than 15% processing time of a single CPU.

Continuous Gesture Recognition: The main visual interaction modality is continuous gesture recognition. Unlike with previous gesture recognition systems [1], the user does not have to adhere to specific predefined gestures. It has been trained to recognize natural gestures, i.e., gestures that a person has a natural tendency to perform when interacting with large screen displays. This approach increases the naturalness of our system tremendously. However, the gesture recognition component is no longer able to solely carry the complete intent of the user. Rather, the semantics of a command or request becomes distributed across the speech and gesture modalities such that gesture recognition and speech recognition have to be tightly coupled to extract reliable command and request information.

Based on our experience with examining weather narration broadcast [8] (Figure 6) we temporally modeled deictic gestures based on a set of fundamental gesture primitives that pose a minimal and complete basis for the large-display interaction tasks considered by our applications.

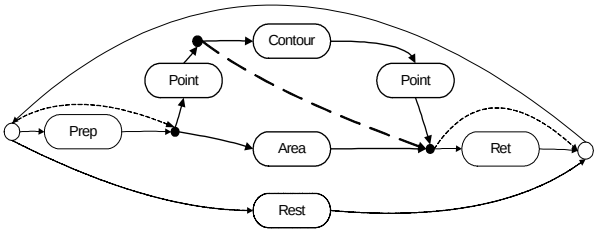


Figure 6: Statistical deictic gesture model.

More specifically, the system has been trained to learn pointing gestures (selection of a single item, reference to a single location), area gestures (selection of a number of items or an item extensive in size, reference to an area) and contour gestures (a compound point-contour-point gesture used to semantically connect references and selections).

The statistical gesture model and continuous recognition is based on continuous observation density Hidden Markov Models [16] and token passing [17] and is based on [18]. After bootstrapping, refinement of the HMM and the recognition network were performed using a customized application that was designed to "pull" desired gestures from a user. The system extracted gesture and speech data and automatically segmented the thus obtained gesture training data. After a training of HMMs on the isolated gesture data, a final embedded training of the compound network was performed.

3.3. Audio Components

Speech Recognition: Speech recognition has improved tremendously in recent years and the robust incorporation of this technology in multimodal interfaces is becoming feasible. The presented system has been operating with both speaker dependent and speaker independent recognition engines (cf. [14] for details). While speaker dependent systems were found to be superior in performance, speaker independence is essential for domains where potential users are unknown and speech training is infeasible (e.g., for commercial systems operating in public).

The set of all possible utterances is defined in a context free grammar with embedded annotations. This allows constraining the necessary vocabulary that has to be understood by the system while retaining flexibility in how speech commands can be formulated. The speech recognition module of the system only reports time-stamped annotations to the application front end, which is responsible for the modality fusion and context maintenance.

Audio Feedback: All applications that have been developed on top of the proposed framework provide audio feedback to the user. Audio feedback can be as simple as sound effects that confirm the successful capture of a user's commands (e.g., a selection noise when a button was selected) or in the form of pre-recorded speech from a narrator or text to speech synthesis, narrated by an animated character (Figure 10, top-right). The choice of appropriate feedback depends on the application front-end. While sound effects are sufficient for an interactive game, a speaking and animated avatar is much more appropriate in for example a shopping assistant application.

3.4. Modality Fusion

In order to correctly interpret a user's intent from his or her utterances and gestural motions, the two modalities have to be fused appropriately (Figure 3). Due to the statistical method employed for continuous recognition, both the speech recognition and gesture recognition systems emit their recognition results with time delays of typically 1 sec. Verbal utterances such as "show me **this** region in more detail" taken from a typical geocentric application (see below) have to be associated with co-occurring gestures such as "<Preparation>-<Area Gesture Stroke>-<Retraction>". The understanding of the temporal alignment of speech and gesture is crucial in performing this association. While in pen based systems [3], gesture have been shown to occur before the associated deictic word ("this"), our investigations from HCI and Weather Narration [8] showed that for large screen display systems, the deictic word occurred during or after the gesture in 97% of the cases. Hence modality fusion can reliably be triggered by the occurrence of verbal commands.

The speech recognition system emits streams of time stamped annotation embedded in the speech grammar; for the above case one would obtain

…[ZOOM, t_0, t_1] [LOCATION, t_1, t_2] [REGION, t_2, t_3]…

The annotation "LOCATION" occurring around the time $t_s=(t_1+t_2)/2$ corresponds to the occurrence of the deictic keyword "this". Similarly, the gesture recognition might report

…[PREP, s_0, s_1] [AREA, s_1, s_2] [RETRACTION, s_2, s_3]…

indicating that an area gesture was recognized in the time interval $[s_1, s_2]$.

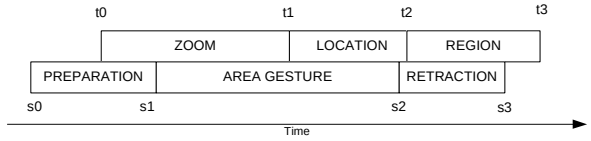


Figure 7: Speech gesture modality fusion.

Using the time stamp of the deictic keyword, a windowed search in the gesture recognition result history is performed. Each past gesture stroke is checked for co-occurrence with appropriate annotations. Given for example time stamps $[s_1, s_2]$ for a gesture stroke, association with a keyword that occurred at time t_s is assumed if

t_{se} *c* $[s_1-dt_b, s_2+dt_e]$. Where dt_b and dt_a are constants learned from training data. This approach allows the occurrence of the keyword a short time before the gesture and a longer time delay after the gesture.

Upon a successful association, the physical content of the area gesture, namely hand trajectory data for the time interval $[s_1, s_2]$ is used to obtain the actual gesture conveyed components of the compound speech gesture command. In the case of for example an area gesture, a circle is fitted to the thus obtained gesture data in order to determine which region of the screen actually to show in more detail.

3.5. System Integration

The presented framework requires only moderate computational resources. All presented systems run comfortably on Dual Pentium III 500 Mhz or correspondingly faster single processing platforms with less resources required if the system runs with not all system modules enabled. For a detailed description of the system components see [14].

The systems main tasks were separated into a set of separate execution threads as shown in Figure 4. Most resources are consumed by the vision components of the system and especially the face detection algorithm. Since many of the components run on different time scales (especially the Speech Recognition, Face Detector and Active Camera Control), the architecture was designed to take advantage of multi-threaded parallel execution. Communication between components is performed using message passing and straightforward thread synchronization.

4. Case Studies

A number of successful HCI applications were build based on the presented framework. While some systems are used for scientific purposes and user studies [19] other systems have led to actual commercial products operating unattended in public spaces .

Figure 8: XISM, a multimodal crisis management system.

4.1. Geocentric Systems

Many information representation and access systems are geocentric in nature. Large screen display systems are especially suited for interaction with spatially or geographical data. The first system that was developed at Penn State based on the *i*Map framework was the Campus Map [7], which helped visitors find their way around the University Park Campus (Figure 1). The system could be queried using speech and gesture commands (e.g., "What is the name of this department?", "How do I get from here to the library?").

The Multimodal Crisis Management System (XISM) is a dynamic system in which the user takes the role of an emergency center operator using speech and gesture commands to dispatch emergency vehicles to rapidly occurring crisis centers in a virtually generated city (Figure 8). In contrast to static systems, where the progression of interaction is determined by the user the operator has to react to rapidly occurring events under time pressure. This system has been and is currently being used for conducting cognitive load studies in which different aspects of multimodal interaction can be measured accurately and compared to traditional and alternative interaction methods under variable but controlled conditions.

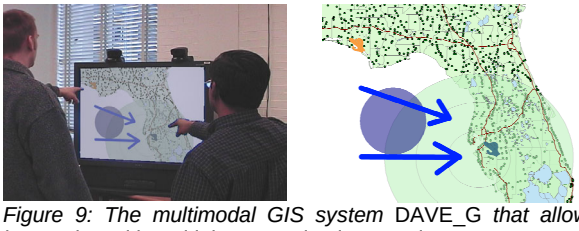


Figure 9: The multimodal GIS system DAVE_G *that allows interaction with multiple users simultaneously.*

In an ongoing project, the XISM system is extended to operate with multiple users simultaneous interfacing to a geographical information system (GIS). The system supports collaborative task planning and decision making [15].

4.2. Retail And Entertainment Systems

Finally, a number of commercial applications have been developed and deployed in public environments (Figure 10) showing that advanced multimodal HCI technology has reached sufficient maturity and robustness for extended and unattended public use in retail and entertainment environments.

Figure 10: Commercial embodiments of the presented multimodal HCI framework. Left: Physical structure. Top right: GiftFinder, a retail system with a virtual avatar that helps user navigation. Bottom right: SuperBounz, interactive multimodal entertainment.

The system proved its usability in several fairs and expositions with over 10000 interaction sessions with unique

users. Informal customer observations and surveys showed that 80% of users had successful interaction experiences. In addition, observations revealed that the system behaved according to its specifications in 95% of the cases.

Furthermore, returning users showed a dramatic increase in interaction proficiency, indicating that once the initial novelty barrier has been overcome, the acceptance of such systems is high with little or no difficulties in understanding the "mechanics" of multimodal interaction. Formal user studies are currently in progress.

5. Discussion and Conclusion

This paper describes issues related to the development of a robust real-time framework that exploits natural gestures and spoken command as input and a large screen display for visual feedback. The framework has been validated by implementing a number of prototype systems, which transfer real-world interactions to novel metaphors thus bridging the gap between digital environments and user interactions. It was found that a careful design, integration and due considerations of possible aspects of a systems interaction cycle can yield a successful system.

Room for improvement still exists for the speech recognition module that can perform unreliably in very noisy public environments even if using advanced sound acquisition devices such as microphone arrays or -domes. The introduction of multiple users introduces additional challenges, especially when users are spatially close to each other. The system must resolve ambiguity in identifying and attaching motion and spoken command to the right user. Model based head tracking for extracting lip motion, and gaze tracking to localize attention are currently investigated to improve disambiguation. Also, model-based articulated tracking is being developed to extract reliable information from visual data [14]. Finally, a prosody based speech-gesture co-analysis is under investigation to improve on continuous gesture recognition.

Acknowledgements

The authors wish to thank E. Schapira, E. Polat, H. Raju and M. Leas who have contributed to the development of parts of this work and I. Rauschert and S. Olenoski for providing the photos for Figures 9 and 10 respectively.

The financial support of this work in part by the National Science Foundation CAREER Grant IIS-97-33644 and NSF IIS-0081935 is gratefully acknowledged.

References

[1] R. Bolt, " Put-that-there: Voice and gesture at the graphic interface," *In SIGGRAPH-Computer Graphics*, 1980.

[2] K. Nguyen, "Methods and apparatus for real-time gesture recognition," US Patent.

[3] S. L. Oviatt, "Multimodal interfaces for dynamic interactive maps," in *Proc. of the Conf. on Human Factors in Comp. Systems*, ACM Press, 1996, pp. 95-102.

[4] J. Cassell, "What you need to know about natural gesture," in *Intl. Conf. on Automatic Face and Gesture Recognition: Keynote Address*, 1996.

[5] T. E. Starner and A. Pentland, "Visual Recognition of American Sign Language Using Hidden Markov Models," in *Intl. Workshop on Automatic Face- and Gesture-Recognition*, 1995.

[6] M. Lucente, "Visualization Space: A Testbed for Deviceless Multimodal User Interface," *Computer Graphics*, vol. 31, 1997.

[7] R. Sharma, I. Poddar, E. Ozyildiz, S. Kettebekov, H. Kim, and T. Huang, "Toward Intrepretation of Natural Speech/Gesture: Spatial Planning on a Virtual Map," in *Proc. of the ARL FedLab Symposium*, 1999.

[8] I. Poddar, Y. Sethi, E. Ozyildiz, and R. Sharma, "Toward Natural Gesture/Speech HCI: A Case Study of Weather Narration," in *Proc. 2nd Workshop on Perceptual User Interface*, 1998, pp. 1-6.

[9] S. Kettebekov, N. Krahnstoever, M. Leas, E. Polat, H. Raju, E. Schapira, and R. Sharma, "i2Map: Crisis Management using a Multimodal Interface," in *Proc. of the 4th ARL FedLab Symposium*, 2000.

[10] MIT Media Laboratory, http://whitechapel.media.mit.edu/vismod/demos/demos.html.

[11] B. Brumitt, B. Meyers, J. Krumm, A. Kern, and S. Shafer, "EasyLiving: Technologies for intelligent environments," 2nd Intl. Symp. on Handheld and Ubiquitous Computing (HUC), Bristol, UK, 2000.

[12] S. Kettebekov and R. Sharma, "Toward Natural Gesture/Speech Control of a Large Display", in Engineering for Human-Computer Interaction, Lecture Notes in Computer Science, Springer Verlag, 2001.

[13] M. Yeasin and Y. Kuniyoshi, "Detecting and tracking human face using a space-varying sensor and an active head,", CVPR, 2000.

[14] N. Krahnstoever, S. Kettebekov, M. Yeasin, and R. Sharma, "iMap: A Real-Time Framework for Natural Multimodal Interaction with Large Screen Displays," Dept. of Comp. Science and Eng. Technical Report CSE-02-010, Pennsylvania State University, May 2002.

[15] I. Rauschert, P. Agrawal, S. Fuhrmann, I. Brewer, R. Sharma, G. Cai, and A. MacEachren, "Designing a Human-Centered, Multimodal GIS Interface to Support Emergency Management," in 10th ACM International Symposium on Advances in Geographic Information Systems, McLean, VA, 2002.

[16] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, pp. 257-286, 1989.

[17] S. J. Young, N. H. Rusell, and J. H. S. Thornton, "Token Passing: a Conceptual model for Connected Speech Recognition," Cambridge University Engineering Dept, CUED/F-INFENG/TR38, 1989.

[18] I. Poddar, "Continuous Recognition of Natural Hand Gestures for Human Computer Interaction," MS Thesis, Pennsylvania State University, 1999.

[19] E. Schapira and R. Sharma, "Experimental evaluation of vision and speech-based multimodal interfaces," in *Workshop on Perceptive User Interfaces, ACM Digital Library*, 2001.