# Multimodal Human-Computer Interaction for Crisis Management Systems

N. Krahnstoever[*,1,2], E. Schapira[2], S. Kettebekov[1], R. Sharma[1,2]

[1] *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*

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

*{krahnsto,schapira,kettebek,rsharma}@cse.psu.edu*

## Abstract

*This paper presents a multimodal crisis management system (XISM). It employs processing of natural gesture and speech commands elicited by a user to efficiently manage complex dynamic emergency scenarios on a large display. The developed prototype system demonstrates the means of incorporating unconstrained free-hand gestures and speech in a real-time interactive interface.*

*This paper provides insights into the design aspects of the XISM system. In particular, it addresses the issues of extraction and fusion of gesture and speech modalities to allow more natural interactive behavior. Performance characteristics of the current prototype and considerations for future work are discussed. A series of studies indicated positive response with respect to ease of interacting with the current system.*

## 1. Introduction

Management of crisis and emergency response tasks requires an operator to process vast amounts of information in a time critical manner. Crisis situations need to be assessed and effective response plans need to be generated in a timely fashion. In a real setting these kinds of scenarios are usually associated with resource allocation problems. To maximize the effectiveness of this process and minimize the chance of human error, computers should play an important role in automating the steps in the involved assessment and decision making process. However, means of the traditional human-computer interaction (HCI) (i.e., mouse and keyboard interfaces) present a bottleneck at the interface level by encumbering the information exchange. Meanwhile, humans communicate with each other through a range of different modalities with no effort. Hence, the ultimate goal of novel interfaces is to mimic expressiveness and easiness of everyday human-to-human communication in HCI.

It is well known that speech and gesture compliment each other and when used together, create an interface more powerful than either of the modalities alone. Integration of speech and gesture has tangible advantages in the context of HCI, especially when coping with complexities of the visual environment [1]. Therefore, coverbal gesticulation has the best prospects of achieving effortless and expressive HCI.

Thus, a computer system should have the ability to understand multiple modalities, i.e., speech and gesture when information is somehow distributed across the modalities. Up to date there have been several designs of multimodal systems. However, the rigidity of Bolt's "put that [point] there [point]" [2] paradigm still prevails in the majority of the designs. While there were some advances of including speech recognition into limited domains, most of the gesture recognition work is limited to understanding artificially imposed signs and gestures, e.g. [3, 4 ]. In addition, it often involves cyber gloves or electronic pens. The resulting systems 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 [5].

In contrast, we present a multimodal HCI system that allows a human operator to naturally interact with a large screen display through the simultaneous use of speech and gesture. The high-resolution display provides real-time visualization of current data and events allowing an operator to maintain an active view of the current state of affairs and to establish relationships between events. The operator's movements are visually tracked via an active camera situated on top of the display. Since previous studies [6] [10] have shown that deictic gestures are more suitable for a large display manipulation as opposed to symbolic gestures as in sign language, e.g. in [7], our system is trained to recognize unconstrained *deictic* gesture primitives.

The XISM system presented in this paper has evolved from a previous multimodal speech gesture interface system (Campus Map [8]), which was developed based on a weather narration keyword/gesture co-occurrence analysis [6]. Our system differs from related multimodal systems [9] in that the gesture recognition is based on learned statistical gesture and speech-gesture co-occurrence models.

## 2. Crisis Management System

The multimodal crisis management system presented in this paper is a research prototype system designed to study the suitability of advanced multimodal interfaces for typical

[*] Corresponding author.

crisis management tasks. The user takes the role of an emergency center operator and is able to use speech and gesture commands to dispatch emergency vehicles to rapidly occurring crisis in a virtually generated city (*Figure 1*). The operator is standing at a distance of about 5 feet from the display in the view of a camera located on top of the unit. Speech commands are captured with a microphone dome suspended from the ceiling.

*Figure 1: XISM, a multimodal crisis management system.*

The operator has a birds-eye view of the city but has the ability to zoom in on locations to get a more detailed view. The goal of the operator is to acknowledge and attend to incoming emergencies, indicated by animated emergency symbols and accompanying audible alarms (*Figure 2*).
Alarms are acknowledged by pointing at them and giving an appropriate verbal command ("Acknowledge this emergency." or simply "Acknowledge this!"). The acknowledgement indicates to the system that the operator is aware of the crisis and that it will be attended to shortly. The visual as well as audio signals emitted by the alarm are reduced to a lower level. The speed at which each emergency is attended to, ultimately determines the final performance of the operator. Hence, the emergencies have to be resolved as quickly as possible by sending an appropriate emergency vehicle to the crisis location. For that, the operator has to decide which type of unit to dispatch and from which station to dispatch it. Emergency stations (hospitals, police and fire stations) are spread throughout the city and have limited dispatch capacities. Units are dispatched through speech gesture commands such as "Dispatch an ambulance from this station to that location." accompanied with an appropriate deictic contour gesture.
The XISM system allows a variation of the following crisis characteristics:

- city size and urbanization density
- amount of available emergency stations
- amount of dispatch units per station
- alarm occurrence rate
- deployment time penalty

This allows the creation of different scenarios ranging from sparsely populated country regions with few emergency stations to large, densely populated cities with many rapidly occurring emergencies. If displayed information is too dense to perform accurate dispatch actions, an operator needs to be able to get more detailed views of city locations. The required speech gesture commands to complete this task are of the form "Show this region in more detail." or "Zoom here." with an accompanying area or pointing gesture. While the pointing gesture enlarges the view to the maximum for a given location, the area gesture allows controlling the level of detail by naturally conveying which area to enlarge.

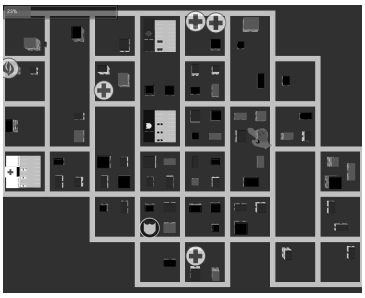

*Figure 2: City with six active emergencies.*

## 3. System Components

To capture speech and gesture commands the XISM system utilizes a directional microphone and a single active camera. 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.

From a system design perspective smooth and automatic interaction initialization, robust real-time visual processing and error recovery are very important for the success of advanced interface approaches for crisis management applications, because unexpected system behavior is unacceptable for mission critical systems. The iMap framework that the XISM system was build upon takes a holistic approach to multimodal HCI system design, attempting to address all of the above issues.

### 3.1. Initialization

In the initialization phase, user detection is achieved by detecting a face. The detection leads to subsequent head tracking initialization. Once the operator has stepped into the proper spot and found to be facing the system, the system enters the bootstrapping state of the initialization phase. The system immediately performs palm detection to obtain an initial location of the user's active hand 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.

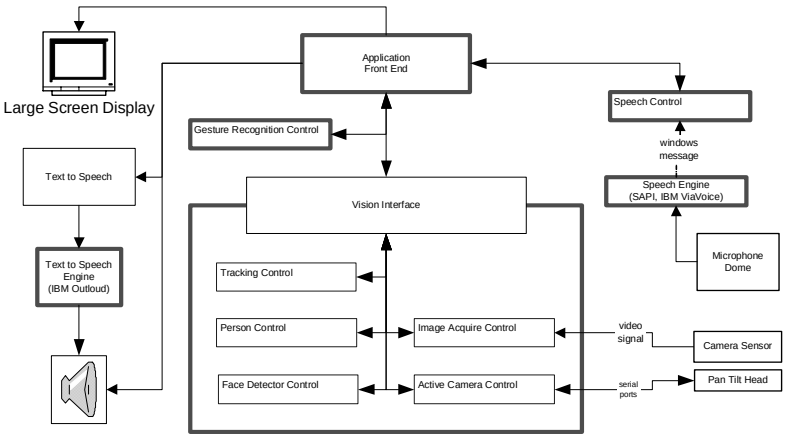


*Figure 4: Overview of the iMap system architecture. Each of the bold framed boxes constitutes a separate thread of execution.*

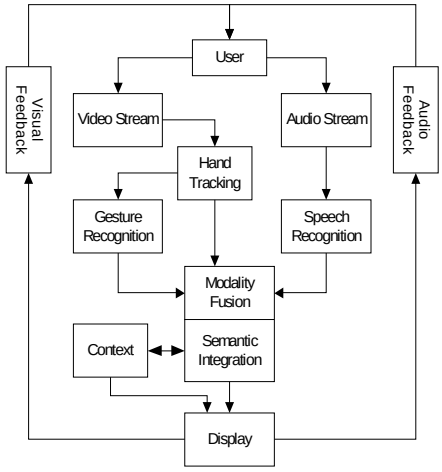


*Figure 3: Logical flow of the system.*

After the initialization, the actual dialogue between the system and the user commences. 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. Recognized gestures are combined with speech recognition results by a speech-gesture modality fusion module (*Figure 3*). The semantic integration maintains a time varying context in order to constrain the set of possible user actions for increased response reliability.

### 3.2. Vision Components

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 components in the system is the face detector for robust user detection and continuous head track status verification. The implementation 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. Palm detection is based on the Viterbi algorithm and is performed as follows: For each frame I[.], two probability density functions are calculated with support ranging over the expected image locations x of the hand, one describing the probability of pixels belonging to the hand

$$p(x \mid color) \propto G_c \otimes N(I[x]; \mu_c, \Sigma_c), \qquad (1)$$

the other describing the a probability of seeing the palm based on motion evidence obtained through frame differencing

$$p(x \mid motion) \propto G_m \otimes N(\Delta I[x]; \sigma_m). \qquad (2)$$

In both cases, N(.) are Gaussian distributions with $\mu_c \in \mathbb{R}^3, \Sigma_c \in \mathbb{R}^{3\times3}$ (the skin color model) obtained from the face detector and variance $\sigma_m$ determined empirically for the motion cue.

The motion and color cues tend to show responses in an image at different locations. While the color cue leads to the strongest response in the middle of the palm, motion energy is observed mostly at the edges. To allow a combination of both cues, both responses are spatially smoothed with Gaussian masks $G_c, G_m$ of width c and m, respectively. The final probability density of observing a hand at a location x is finally set to

$$p(\text{palm} \mid x) \propto w_c p(x \mid \text{color}) + w_m p(x \mid \text{motion}) + \quad (3)$$

$$(1 - w_c - w_m)\, p(x \mid \text{motion})\, p(x \mid \text{color}).$$

From this distribution, a number of palm location hypothesis are generated for each frame given by the local maxima within each block of a regular 8x8 tiling. Each hypothesis is associated with a corresponding normalized probability p(palm | x). The (time varying) location of the palm is then given by the optimal (most probable) hypothesis-connecting path through the set of frames under consideration. The probability of the path depends on the probability of each hypothesis plus an additional cost associated with the spatial shift in location from one frame to the next. The optimal path can be found efficiently using dynamic programming using the Viterbi algorithm.

**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 and geared towards skin colored moving objects. Prior knowledge about the human body is utilized for 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 [2], 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.

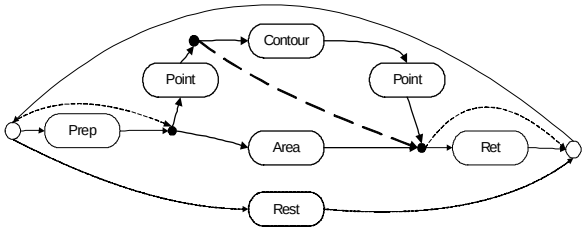


*Figure 5: Statistical deictic gesture model.*

In order to be able to model natural gestures using statistical techniques, one needs valid multimodal data. However, it is impossible to collect the data unless a system exists that allows the respective user interaction. For this work, a bootstrap-and-evolve strategy was used. A comparative analysis revealed that weather channel narration broadcast is closely related to the desired type of gestural interaction [10]. It led to the development and statistical training of appropriate gesture recognition models at the bootstrapping stage [6]. Based on our experience, 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 5).

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 (HMM).

### 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 utilizes a speaker dependent voice recognition engine for reliable speech acquisition after a short speaker enrollment procedure. The set of all possible utterances is defined in a context free grammar with embedded annotations. The grammar constrains 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.

**Audio Feedback:** Audio feedbacks in the form of sound effects and/or speech are important components for multimodal interfaces. The current XISM system utilizes audio effects of varying volume as a means of notifying an operator of occurring emergencies on the one hand and in order to create a task appropriate noise environment (e.g., sirens) that an actual operator would be subjected to on the other.

### 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 typically with time delays of 1 sec. Verbal utterances such as "show me **this** region in more detail" 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 [4], gesture have been shown to occur before the associated deictic word ("this"), our investigations from HCI [10] and Weather Narration [6] showed that for large screen display systems, the deictic word occurred during or after the gesture in 93% 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 $[s_1, s_2]$.

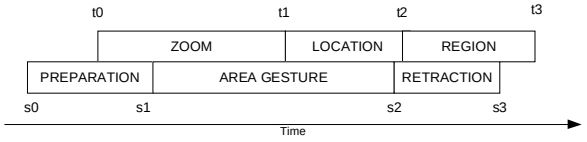


*Figure 6: Speech gesture modality fusion.*

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 $dt_e$ 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.

## 4. Discussion and Conclusion

XISM 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 [8].

The XISM 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.

Informal user studies with the crisis management and related multimodal applications [8] have shown 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. In an ongoing project, the XISM system is extended to operate with multiple users simultaneous interfacing to a geographical information system (GIS), which will allow to extent the applicability of the system to the area of natural disaster scenarios.

While the current system was developed with crisis and emergency management in mind, the framework easily translates to a number of related application areas. Flight traffic control, command post systems and entertainment application all share characteristics with the presented system.

## Acknowledgements

The authors wish to thank E. Polat, H. Raju and M. Leas who have contributed to the development of parts of this work.

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] A. D. Angeli, W. Gerbino, G. Cassano, and D. Petrelli, "Visual display: pointing and natural language: The power of multimodal interaction," presented at Advanced Visual Interfaces, L'Aquila, Italy, 1998.

[2] R. A. Bolt, "Put-That-There: voice and gesture at the graphics interface," in *ACM Computer Graphics*, vol. 14: ACM Press, New York, 1980, pp. 262-270.

[3] K. H. Nguyen, "Methods and apparatus for real-time gesture recognition," in *United States Patent*. US: Electric Planet, Palo Alto, CA, 2001.

[4] S. L. Oviatt, "Multimodal interfaces for dynamic interactive maps," in *Proceedings of the Conference on Human Factors in Computing Systems (CHI'96)*: ACM Press, New York, 1996, pp. 95-102.

[5] S. Oviatt, A. D. Angeli, and K. Kuhn, "Integration and synchronization of input modes during multimodal human-computer interaction," presented at Conference on Human Factors in Computing Systems (CHI'97), 1997.

[6] I. Poddar, Y. Sethi, E. Ozyildiz, and R. Sharma, "Toward Natural Gesture/Speech HCI: A Case Study of Weather Narration," in *Proc. Second Workshop on Perceptual User Interface (PUI'98)*, 1998, pp. 1-6.

[7] T. E. Starner and A. Pentland, "Visual Recognition of American Sign Language Using Hidden Markov Models," in *International Workshop on Automatic Face- and Gesture-Recognition IWAFGR95*, 1995.

[8] N. Krahnstoever, S. Kettebekov, M. Yeasin, and R. Sharma, "A Real-Time Framework for Natural Multimodal Interaction with Large Screen Displays," Dept. of Computer Science and Engineering, 220 Pond Lab, University Park, PA, Technical Report CSE-02-010, May 2002.

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

[10] S. Kettebekov and R. Sharma, *Toward Natural Gesture/Speech Control of a Large Display*: Engineering for Human-Computer Interaction (EHCI'01),Lecture Notes in Computer Science, Springer Verlag, 2001.