Towards Environment-to-Environment (E2E) Multimedia Communication Systems

Vivek Singh,
singhv@uci.edu

Hamed Pirsiavash,
hpirsiav@uci.edu

Ish Rishabh,
irishabh@uci.edu

Ramesh Jain,
jain@ics.uci.edu

Department of Information and Computer Science,
University of California, Irvine.

ABSTRACT

We present an approach to connect multiple remote environments for natural interaction among people and objects. Focus of current communication and telepresence systems severely restrict user affordances in terms of movement, interaction, peripheral vision, spatio-semantic integrity and even information flow. These systems allow information transfer rather than experiential interaction. We propose Environment-to-Environment (E2E) as a new paradigm for communication which allows users to interact in natural manner using text, audio, and video by connecting environments. Each Environment is instrumented using as many different types of sensors as may be required to detect presence and activity of objects and this object position and activity information is used to direct multimedia information to be sent to other Environments as well as present incoming multimedia information on right displays and speakers. The mediation for the appropriate data capture and presentation is done by a scalable event-based multimodal information system. This paper describes the design principles for E2E communication, discusses system architecture, and gives our experience in implementing prototypes of such systems in telemedicine and office collaboration applications. We also discuss the research challenges and a road-map for creating more sophisticated E2E applications in near future.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation (e.g., HCI)**]: Multimedia Information Systems; H.4.3 [**Information Systems Applications**]: Communications Applications

General Terms

Design, Human factors

SAME'08, October 31, 2008, Vancouver, British Columbia, Canada.

Keywords

E2E, Environment to Environment communication, multimedia communication systems

1. INTRODUCTION

With the influx of technology, human communication has moved from person-to-person communication to device-to-device communication. Devices like phones for telecommunication or even cameras and display devices for videoconferencing have been valuable in transmitting the information across physical spaces. In doing so however, these devices have restricted the *affordances* available to the users in terms of physical movement [23, 32, 1], interaction, peripheral vision [20], spatio-semantic integrity and even information flow [19]. For example, users in a video conference need to conscientiously stay within field-of-view, focus and zoom range of the camera. This restricts their physical movement and makes the interaction *unnatural*. Similarly, the fact that all information is presented on just one screen, depletes it of its context and the spatial/semantic coherence. Thus simple instructions like 'look *there*' are lost in translation across environments as there are no easy ways to perceive such spatial/semantic notions.

Recently, there have been some efforts at enhancing the feeling of co-presence across physical space, either by using specially fabricated meeting rooms which look like mirror images of each other (e.g. HP:HALO[24]), or exploring the other extreme of moving all the communication to the virtual world (e.g. SecondLife [29]). However, both of these options remove us from the grounded reality of natural environments in which we would ideally like to interact.

Hence, we propose E2E as the new form of communication which allows users to connect their natural physical environments for communications. In E2E, multiple heterogeneous sensors, devices and technology are used. However, their abundance and the underlying design architecture push them into a supporting role in the background to maintain the focus on natural *human-human-interaction*. Thus the users need not worry about staying within proximity, field of view, audible distance and so on of a sensor or an output device (e.g. screen, speaker etc.) but rather just interact in their natural settings and let the system find the most appropriate input and output devices to support communication. Thus in a way we create a realization of the Weiser's

vision of 'most profound technologies are those that disappear' [34] and extend it to connect multiple environments across space.

To realize E2E communication many heterogeneous sensors analyze data to detect and monitor objects and activities. The system analyzes this sensor information to detect events in the physical environment, and assimilates, stores, and indexes them in a dynamic real-time *EventBase*. The sensor information and EventBase for each environment are shared by an *Event Server* over the Internet to create a *Joint Situation Model* which represents a combined environment. Thus, a person in one environment can interact with objects and observe activities from other environments by interacting with the appropriate ES in a natural setting.

We also discuss our experiences with realizing E2E communication, via one telemedicine and one office collaboration scenario. The telemedicine application connects a doctor's clinic (or home) environment with that of a far-flung health center where a nurse and a patient are present. The nurse and the patient can move between the consultation room and the medical examination room and still be seamlessly connected with the doctor as if she is present with them. Similarly, doctor's clinic environment seamlessly adapts to the different environments where the patient and nurse are present and can continue interacting with them in a naturalistic setting. The office collaboration scenario also provides similar affordances, though in a different context. The implementations help us clearly visualize how E2E communication will be fundamentally different from other forms of communications and also appreciate the practical challenges.

Our contributions in this paper are two-fold:

1. We propose E2E as a new paradigm of communication, formulate its design principles and thence propose an architecture to support it.

2. We describe experiences with implementing such systems, discuss the research challenges posed and then suggest a road-map towards solving them.

The organization of the rest of the paper is as follows. In section 2, we discusses the related work. Section 3, discusses the design principles for E2E communication, which leads to the proposed architecture in section 4. We describe our implementation experiences in section 5. Research challenges expected for E2E systems and a road map towards solving them is given in section 6 before concluding in section 7.

2. RELATED WORK

In this work we are interested in connecting physical natural spaces, hence we do not consider virtual spaces like SecondLife [29] etc. in related work.

On the surface, E2E systems might look comparable to video-conferencing systems or tele-immersive works. However, E2E fundamentally differs from both of them. Video-conferencing/telepresence systems like HP's Halo [24], Microsoft's Roundtable, and Cisco's Telepresence support bi-directional interactivity but are totally oblivious to the *situations* (i.e. semantics of the multimodal content) they connect. Hence they result in systems which are rigid in terms of required set-up (e.g. specially crafted meeting rooms), applications supported, and the bandwidth required. On the other hand, tele-immersive [5] and Multi-perspective-imaging [15] works often understand user objectives to support enhanced user interaction, but they do so only uni-directionally. E2E communication systems support enhanced user affordances bi-directionally based on a semantic understanding of the environments connected. This has been illustrated in Fig. 1.

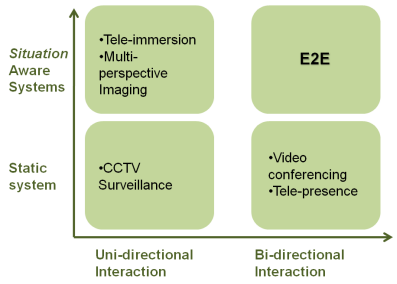

Figure 1: Comparison of E2E with related works

The 'Office of the future' project[26], made multiple advancements in creating bidirectional tele-immersive environments. They employed 'seas' of cameras and projectors to create panoramic image display, tiled display systems etc. for 3D immersive environments. Their focus however was on the 3D visualization aspects while we focus on understanding the *situations* of the environments being connected to employ the best sensors. Further, in E2E we have a wider scope and also consider issues like event understanding, data management, networking and so on which were not considered in their project.

Since 1980s researchers have experimented with connecting remote environments in the form of media spaces [3, 7, 31]. Media spaces in general use a combination of audio, video, and networking to create a 'virtual window' across a distance and into another room. However, the combination of technologies typically used in media spaces restricts naturalistic behavior [7]. A video image of a remote scene has a restricted field of view limiting peripheral vision. The absence of stereo information hinders the users' depth perception of the remote environment. Sound is presented through limited channels, which constrains users' ability to localize speech or sounds in the remote environment. The fixed positions of limited cameras constrain interactive movement. Robotic or pan-tilt cameras offer more options for remote views but still are limited by their reactive speed and direction of focus. Thus, to date, interaction in media spaces is discontinuous as opposed to smooth and seamless [7], and people generally resort to using exaggerated movements to communicate over video [9]. We intend to change each of these with E2E systems.

Multimedia networking community has also made some interesting contributions for remote collaboration. Berkeley's *vic/vat* tools[21], ISI's Multimedia Conference Control (mmcc) [28], the Xerox PARC Network Video tool, (nv) and the INRIA Video-conferencing System (ivs) have all provided interesting ideas. However, these works were based on support of IP multicast and ideally required a connection to IP Multicast Backbone (MBone). Unfortunately, IP Multicast never materialized and today's internet is still best effort. We counter this issue by allowing for graceful degradation of system depending on available resources. Further, we have a broader vision for supporting experiential interaction which go beyond networking aspects.

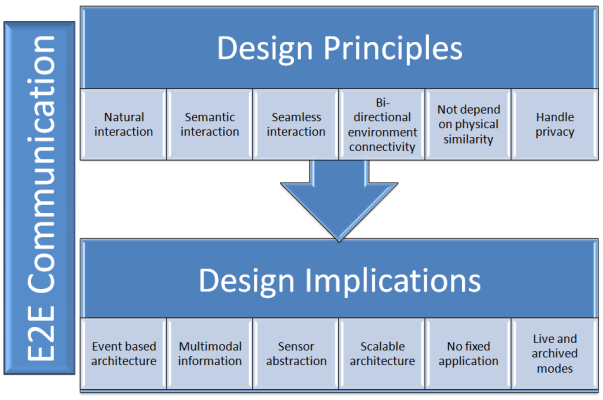

Figure 2: A summary of design principles and applications for E2E

Areas like wearable computing, augmented reality etc. have provided tools to enrich user's experiences. However, we want the communication to be natural, hence do not want to use specialized goggles[17], gloves[5] or unnatural hardware devices like surrogate[12] to support interaction.

Ambient intelligence, ubiquitous computing [34], Smart and Aware Home research areas (e.g. [6], [16]) on the other hand have made many advancements in understanding user behaviors within an environment for applications like telemedicine, monitoring and assisted living. While the use of context is well studied in these areas, they have not focused on bidirectional semantic interaction across environments.

Some interesting works within the Multimedia community have been proposed for tele-immersive dance and music performance[36, 30] and [27]. Works like [36] and [30], however focus more on *extracting* the user data out of their environments to create a combined dance performance rather than connecting the various components of the environment to support more general purpose interaction. In HYDRA[27], the focus was more on studying the networking/delay issues in transferring such large performance data rather than the two way interaction.

There has been a growing interest in Event based architectures for combining information across space for telepresence. Jain et al. [11] describe how event based organization can support communication across time and space. Similarly, Boll et al. [4] describe an architecture for event organization. We in fact adopt an event based architecture to support the many levels of dynamics required by the E2E systems. However, the focus now is on synchronous two-way communication across environments.

3. DESIGN PRINCIPLES

In this section, we list down the design principles for Environment-to-Environment communication systems.

1. *Natural Interaction*: The system should allow the users to interact in their *natural environments*, in a *natural way*. Thus users should be allowed to interact in their natural physical spaces rather than fabricated cyber spaces. Similarly, they need not wear any special gadgets or employ un-natural techniques to communicate.

2. *Semantic Interaction*: The interaction should be facilitated at the human intelligence level. Thus the system should label all events happening in the environment at the human understandable level. Similarly, it should present all incoming information in a way which makes most sense to human users.

3. *Seamless Interaction*: The system should allow for seamless interaction as the user moves between physical spaces. Thus, not only should the correct sensors and devices get actuated as the user moves within one environment, but also when she moves from one environment to another. This is analogous (though many times more sophisticated) to a mobile phone user maintaining her call as she moves from one location to another.

4. *Bi-directional environment connectivity* should be allowed by the system. Thus *both* the participating environments should have elements of the other environment mapped onto appropriate positions in their environments. This is different from the typical approach in immersive reality and Multi-perspective-imaging works (e.g. [15]), where focus is on uni-directional immersion of *one* remote environment onto the other.

5. *Interaction should not depend on physical similarities.* Thus, unlike many current tele-presence [24] systems which rely heavily upon physical similarities (e.g. crafted 'meeting' rooms) to support feeling of co-presence, E2E systems would focus on semantic coherence to present information. This is analogous to the idea that in real world people visiting each other's places do not expect replicas of same environment, but rather just a general consistency of treating visitors e.g. being offered a chair to sit on.

6. *Privacy rights* of users should be maintained, and easy tools should be provided to configure such settings.

These design principles, lead us to a set of supporting design implications.

1. The system should support an *event-based architecture*. This is important to ensure that the dynamic aspects of the interaction get adequately captured (in addition to just 'static' aspects as typically covered by 'object' based architectures). Handling dynamic events is central to the whole theme of E2E communication as this allows the system to actively reconfigure itself to react to the events happening in the user environments. An event-based architecture is required to allow the system to dynamically choose appropriate sensors and presentation devices based on the user actions and movements within the environment.

2. In order to support the freedom to express in a naturalistic setting, the system must support *multi-modal information* capture and presentation modes. Thus the system should be able to handle any type of sensors and devices as required by the application scenario

3. *Abstracted interaction*: The interaction should not be tied up to any sensor or even a group of sensors. In fact, dynamic reconfiguration of sensors and devices to aid experiential interaction can be possible only if the users do not directly control the sensors/devices but rather employ an intelligent information system

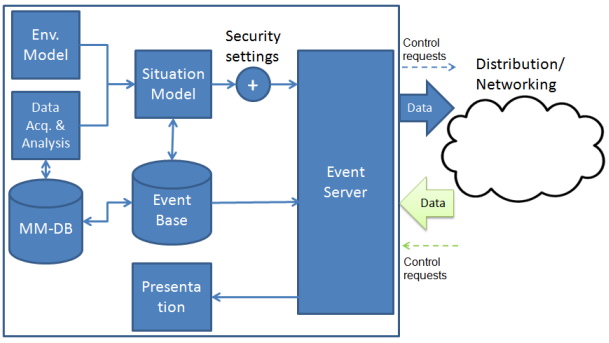

Figure 3: A high-level architecture diagram for E2E

to handle it. For example, the task of finding the best user feed in Env. 1 and presenting it at the best location in Env. 2 can not be handled by a static linkage between sensors and devices. There is a need for an intelligent mediator to explicitly handle such translations. Similarly, such an information system allows dynamic creation of macroscopic views of situation to support semantic interaction even when any of the micro views might not be able to capture it.

4. *Scalable architecture*: The system should work in a scalable manner with no centralized bottlenecks. The system should scale up and down gracefully as the sensor variety or the available bandwidth is varied. Thus, the system should automatically determine its capabilities and then request appropriate feeds from other environments. For example, it may provide a single low-bitrate stream to a user connecting his PDA 'environment' while providing multiple high definition streams to a user connecting a more sophisticated environment.

5. *No fixed application.* The system should not limit itself to any particular application or scenario. Rather it should allow the event markups and behaviors to be configured which allow it to handle multiple applications like official collaboration, tele-medicine, family get-togethers, interactive sports and so on.

6. It should work in *live as well as recorded modes.* The system should continuously archive and index all generated multi-modal data. This can be immediately useful as a tool for periodic review even while the communication is progressing. In a longer term, archival allows for review, summarization, re-interpretation and record-keeping where relevant. Further, the data recorded at a current instance could also become useful contextual information to aid future communications.

The design principles and design implication for E2E communications are summarized in Fig. 2.

4. SYSTEM ARCHITECTURE

Based on the design principles described in the preceding section section, we developed an appropriate system architecture for E2E systems. As the true power of E2E systems lie in their flexibility and the ability to react to the various *events* happening inside it, we adopt an event-based architecture to support it.

Fig. 3 shows a high-level architecture diagram of our E2E approach. The **'Data acquisition and analysis' (DAA)** component gathers the relevant information from various sensors and undertakes the necessary processing on it. It is the information ingestion component of the system. The translation of sensor inputs into meaningful event triggers for E2E communication however does not happen in DAA.

It first needs additional input in terms of physical model of the sensors and the environment. This is handled via the **Environment Model (EM)** which creates linkages between the various sensors and their physical location. Thus if a camera and a microphone detect the sub-events of 'person present' and 'person talking', the environment model is useful in deciding if these sub-events refer to the same person. Further, the actual semantic understanding of the event requires additional contextual information to be added by the specific **Situation Model (SM)**. The SM represents all domain-dependent information which is required to support application functionality. Thus the information coming from multiple sensors and their physical locations will be combined with application specific contextual information to create event triggers by the Situation Model. It captures the current *situation* of the environment by recording all the events happening in the environment at each time instant.

The generated event are filtered based on the security/ privacy settings before before being put up on the Internet by the **Event Server(ES)**. The ES will be responsible for routing out the most appropriate data streams as well as for routing the incoming data streams to be presented at *most appropriate locations* in conjunction with the **presentation module**. ES is also responsible for arbitrating and controlling incoming 'control requests' for available environment resources as well as for making such requests to other environments.

All the generated multimodal information is archived in a **multimedia database (MMDB)**, while the semantic level labels for all the events generated are stored in an **EventBase**. The EventBase does not store any media by itself but maintains links to relevant data in the MMDB.

The events act as *triggers* to initiate communication sessions across environments as well as to activate selection of appropriate sensors and presentation devices across environments. The *'control requests'* for accessing resources in other environments are also understood from event triggers rather than manually requested.

The actual **distribution** of the data is undertaken via peer-to-peer links over Internet between the various ESs. We abstract, construct, and utilize each physical environment as a peer (i.e server and a client). Each sensor and rendering device is seen as a web-service and is used by other EventServers. The sharing of Event Servers over the Internet allows the users to collaborate across environments in their natural settings via virtualized *'Joint Situation Models'*. The JSMs allow users opportunities to interact, collaborate and create new media and events which exist in totality only in the Joint Space. Thus while their components in the respective environments may or may not hold much relevance, their combined effect in the JSM will be of critical importance.

5. IMPLEMENTATION EXPERIENCE

In this section we describe our early implementation experiences with E2E communication. The purpose of this

implementation is to ground the theoretical ideas proposed into a real working system. To ensure that we do not move away from the architecture and start implementing for any single application domain, we considered two different application scenarios. While the specific configuration/settings for the two implementations (telemedicine and office collaboration) were different, they were based on the same enveloping architecture. In this section we describe how the various components of E2E have been currently implemented.

5.1 Data Acquisition and Analysis

For the first implementation, we have focused on audio-visual sensors in different environments and chosen enough number of sensors to provide us reasonable coverage, so that users need not keep their placement etc. in mind while undertaking their interactions. For *data analysis*, we have used face-detector, blob-detector, lip-tracking, hand-tracking and audio volume detector modules as shown in table 1.

For the detection of events, we have adopted a time-line segmentation approach as opposed to media segmentation approach. This approach signifies that we do not consider events as happening in any particular media stream (e.g. video) but rather in a real world time-line. The various media streams are mere evidences of such an event taking place rather than the primary entities themselves. For example 'person talking' is an event which happens in the real world on a real time-line. The audio and video streams which capture the person's presence and audio energy data are mere evidences of the event happening. A preliminary description of this approach was presented in [2].

5.2 Environment Model

EM captured the physical properties of the environment and served two important functions. First, it contained a collection of information regarding the various sensors being employed (i.e. cameras and microphones), their coverage, and signal acquisition characteristics, together with the location and geometry of the room. This allowed us to map the sensor signals obtained via DAA component onto their physical locations. This was useful in correlating the information coming from multiple data sources as knowing their relative physical locations allows better interpretation and combination their readings. Further, the knowledge of number and variety sensors/devices allowed the system to request for appropriate type of information from other environments. This was useful to allow the system to gracefully scale up and down as the device sophistication level changes.

The second purpose of EM was to correctly identify the location and geometry of the objects of interest (e.g. desk) in the environment. Knowing the semantic labels for various objects of interest allowed the system to capture and present sensor data from/to the correct semantic destination even though the corresponding objects had different positions in different environments. Examples of important labels created in our implementation are 'owner chair', 'examination table', and 'X-ray projection board'.

In our current implementation we manually entered the various parameters discussed in a configuration file, but this step will be automated in near future.

5.3 Situation Model

Situation Model represents all domain-dependent information required to support application functionality. As we considered two different applications in our implementation viz. telemedicine and office collaboration, the role of Situation Model was critical in translating the various inputs coming from DAA and EM into domain specific events.

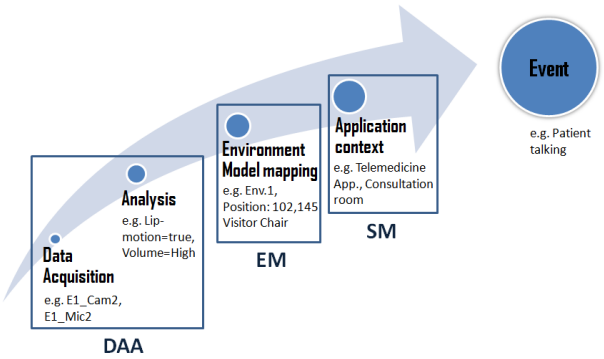

Figure 4: Event detection process in E2E

The process of the combination of information from DAA, EM and SM to detect the events has been shown in Fig. 4. As shown, the information captured from the sensors in the Data acquisition step is analyzed, EM mapping is undertaken on top of this data and finally the application context is added before the final detection of event is undertaken. A summary of the various events detected in the current implementation and their respective information components coming from DAA, EM mapping and Application context have been shown in table 1. Note that, some events had very similar DAA and EM mappings but different application context added from SM changed the events detected.

5.4 Event Server

The ES received the event-related streams and physical & semantic location data from the EM and SM and then determined the most appropriate data to be sent. Similarly, it used the physical layout from EM and the presentation module parameters to decide on the most appropriate locations to present the incoming information from other Environments.

In both telemedicine and office collaboration scenarios, it was desirable to store and archive all data for later analysis such as to study 'patient case history' or to re-look at the meeting from a different perspective. All the sensory information was stored in the multimedia database (MMDB). An index of all the events with explicit linkages to related sources was stored in the *EventBase*. EventBase provided the central facility to organize, and search the multimodal data handled by the system.

The critical end-product of the use of E2E architecture was 'Joint Situation Model' (JSM). Fig. 5 describes how multiple environments can create collaborative *'situations'* using the JSM. The JSM maintains the communication session across multiple environments. The individual SMs (as shared through ES) are combined to represent a JSM for the connected environments. As shown in Fig. 5, each of the three environments experiences bi-directional interaction with other environments by sharing its resources with others and at the same time accessing theirs. The requests for resources were based on event *triggers* like 'entry into studio' rather than any manual input.

Table 1: Various events detected in the current E2E implementation

(a) Telemedicine

S. No.	Data acquisition	Data Analysis	EM Mapping	Appl. Context	Event detected
1.	E1_Cam1	Face, Blob	Owner chair	Telemedicine	Nurse present
2.	E1_Cam1 , E1_Cam2	Face, Blob	Owner Chair, Visitor chair	Telemedicine	Nurse, Patient present
3.	E1_Cam1 , E1_Mic1	Volume, Lip-tracking	Owner chair	Telemedicine	Nurse talking
4.	E1_Cam2 , E1_Mic2	Volume, Lip-tracking	Visitor chair	Telemedicine	Patient talking
5.	E1_Cam1 , E1_Cam2	Face, Blob	Entire Env._1	Telemedicine	Exit from consultation room
6.	E2_Cam1 , E2_Cam2 , E2_Cam3	Face, Blob	Entire Env._2	Telemedicine	Entry into Exam room
7.	E2_Cam3	Blob	Examination table	Telemedicine	Nurse movement
8.	E3_Cam1 , E3_Cam2	Face, Blob	Entire Env._3	Telemedicine	Doctor's position
9.	E3_Cam2	Hand tracking	Projection Board	Telemedicine	Interaction with X-ray

(b) Office Collaboration application

S. No.	Data acquisition	Data Analysis	EM Mapping	Appl. Context	Event detected
10.	E1_Cam1	Face, Blob	Owner chair	Office Coll.	Bob present
11.	E1_Cam2	Face, Blob	Visitor chair	Office Coll.	Alice's entry
12.	E1_Cam1 , E1_Mic1	Volume, Lip-tracking	Owner chair	Office Coll.	Bob talking
13.	E1_Cam2 , E1_Mic2	Volume, Lip-tracking	Visitor chair	Office Coll.	Alice talking
14.	E1_Cam2	Face, Blob	Visitor chair	Office Coll.	Alice's exit
15.	E2_Cam1 , E2_Cam2 , E2_Cam3	Face, Blob	Entire Env._2	Office Coll.	Entry into Studio
16.	E2_Cam1 , E2_Cam2 , E2_Cam3	Face, Blob	Entire Env._2	Office Coll.	Alice's position
17.	E1_Cam1 , E1_Cam2	Face, Blob	Entire Env._1	Office Coll.	Bob's position
18.	E3_Cam1 , E3_Cam2	Face, Blob	Entire Env._3	Office Coll.	Charles' position
19.	E2_Cam3	Hand tracking	Environment2_Whiteboard	Office Coll.	Alice writing
20.	E1_Cam2	Hand tracking	Env._1 whiteboard	Office Coll.	Bob writing
21.	E3_Cam2	Hand tracking	Env._3 whiteboard	Office Coll.	Charles writing

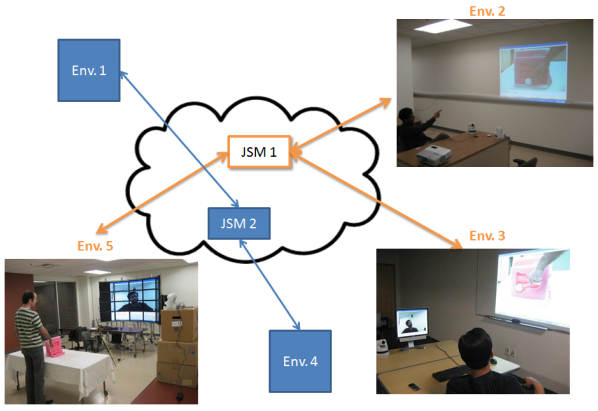

Figure 5: Multiple E2E clients at different locations can connect through the Internet and create correct collaborative 'situations' at each location using the Joint Situation Model. Environments 1, 3, and 5 show three different environments connected in an office collaboration scenario.

While the events to be detected, their triggers and corresponding actions were manually configured in the current implementation we intend to undertake a generic event definition approach in near future. Similarly, while joint-environment events were detected across pre-assigned environments , we intend to adopt an domain based ontology approach for correlating events across multiple environments for JSM in near future.

The user's privacy/access control settings were also handled by the Event Server. It was used to share only certain type of data and devices in the JSM while restricting others. For example, in our telemedicine application, the doctor had more access rights than the nurse.

5.5 Presentation and Interactions

Depending on the activities in the remote environment and local environment, E2E systems presented different information to the users. One important requirement was to find best *feasible* device and presentation position. For example, the projectors need a planar surface to project or we may have only a limited number of display devices. Thus these factors were considered by the Presentation module before presenting the actual information.

The system had a default information presentation mode, but users were also offered a semantic selection of information. For example, in the telemedicine scenario, default video mode was to present doctor with images of the patient body area currently being examined by the nurse. However, the doctor could choose from other labels like 'nurse view', 'patient head', 'patient leg' etc. Further, in office collaboration application, streams from different sensors (one capturing face and the other capturing whiteboard) were presented in different locations in the remote site so the user could see any of the streams just by turning the head and need not choose explicitly what he/she wants to see.

5.6 Distribution/Networking

The distribution of the data was handled via a peer-to-peer like architecture running over the Internet. We did not

want to route the multimodal information via any central servers which may become a bottleneck soon, as the number of connected environments increases. We adopted a Skype like hybrid P2P model for the connection of environments and the distribution of information. The environments registered themselves with a central name server to indicate their status. However once the communication started between two environments all the data was transferred in a P2P manner with no central bottleneck.

5.7 Application Scenarios

Telemedicine application

The considered scenario was that of a remote health center being connected to a specialist doctor's clinic. In the scenario we consider 3 different environments, two of which are the consultation and the medical examination room at the remote health center and the third is the doctor's office. We assume that each of the 3 environments has adequate sensors and output devices. The layout of the three environments is shown in Fig. 6.

Nurse can connect her consultation room environment to the doctor's room by simply clicking a button. Doctor's audio and video feeds are made available to the patient and the nurse in such a way that they feel like having an 'across the table' 3 way communication (Fig. 7a). Similarly doctor also experiences an 'across the table' communication. Doctor asks patient the relevant health questions and the nurse meanwhile enters all the important health statistics into an electronic health report on her computer monitor, which gets transmitted and displayed automatically at the doctors own monitor as shown in Fig. 7b.

The doctor asks nurse and patient to move to examination room for closer checkup. However, the patient and nurse's movement does not disrupt their communication with the doctor as the examination room *automatically* gets connected to the doctor's environment.

The nurse provides the archived X-ray image of the patient which is transmitted and displayed in the doctors's environment. The doctor can annotate the X-ray projection as required and uses this to discuss the findings with the nurse and to direct her to undertake more detailed medical examinations. In effect, the X-ray acts as a *handle* for the doctor to describe to the nurse the locations and measurements to be undertaken and the nurse reacts accordingly. Depending on the nurse's actions and the patient body parts being observed, different camera feeds are *dynamically* selected and presented to the doctor. For example Fig. 8 shows the doctor labeling X-ray and asking the nurse to check the 'metatarsus' region of the leg, and nurse's actions lead to the appropriate camera selection whose feed is shown in the monitor display in doctor's environment.A video demonstration of environment connections as described above is available at http://www.ics.uci.edu/singhv/vids.

Office collaboration application

The scenario for the office collaboration also employs 3 different environments and works as follows. Alice H. is a dynamic design architect working on the next model of Bling787 aircraft. To discuss her ideas she goes to meet the sales manager Bob J. in his office. They both discuss the necessary requirements in the office and decide to involve their collaborator Charles S. from Singapore to get his inputs. After the preliminary discussion, Alice H. realizes that she indeed has model in her lab which might suit the requirements very well. She goes back to her lab and connects her environment to that of Bob's and Charles' respective offices. All three of them go through the details of the current model and discuss the various positives and few necessary changes which would make the model perfectly suitable for the project.

Just like the telemedicine application, the initial discussion appears like a virtual 3 way communication at both the offices. The sophisticated ideas on requirements are discussed via the shared whiteboard. When Alice reaches her lab, she immediately connects back to the two environments. The other users are presented with the most appropriate video feed at each time instant as Alice is interacting with the model. A brief overview of the three connected environments can be seen in Fig. 5. Further details on this implementation are omitted here due to space constraints.

5.8 The practical experience

In this initial implementation we focused on audio-visual content and used a total of 7 cameras, 4 microphones, 4 speakers and 5 display devices (1 multi-tiled display, 2 projectors and 2 PC monitors) spread across the three environments . One PC in each environment acted as an Environment Server and undertook the necessary processing. The detailed layouts of the environments and the locations of the input and output devices are shown in Fig. 6 .

The implementation was undertaken across 2 buildings (Donald Bren Hall and CalIT2) within our university. X-ray image was used as an example of relevant contextual (non-sensory) data which may become useful in undertaken application(s). All the input and output devices were IP based (IP Cameras, IP microphones, IP speakers were realized using Axis Communication 214PTZ duplex-audio support cameras). Epson 2315 IP-based projector and other Internet connected PC monitors were used to handle the display requirements. The use of IP based sensors/devices eased the implementation for the ES and also allowed the system to be scalable. Network delay was minimal across two buildings in same university campus, but may become increasingly relevant as the scale of E2E environments grows.

5.9 Discussion

The undertaken implementation relates closely to the promulgated design principles.

The interaction between the nurse, patient and the doctor was totally *natural* in the sense there were no specialized equipment, gloves, goggles etc. which needed to be worn by them. Similarly, they interacted in their physical 3D natural spaces rather than any concocted environments.

The interaction was also *semantic* as data was presented in the manner most appropriate. Audio-visual feeds of the patient, nurse and the doctor were presented in places appropriate to give an 'across the table' co-presence feeling. Similarly the patient health report was presented onto the doctor's PC while the X-ray was projected onto a projection board in the doctor's room.

The system maintained *seamless interaction* even when the patient and nurse transferred between the consultation room environment and the medical examination room. The new environment was connected automatically. Similarly, the doctor had sufficient freedom to move within his room and the patient/nurse could continuously maintain contact.

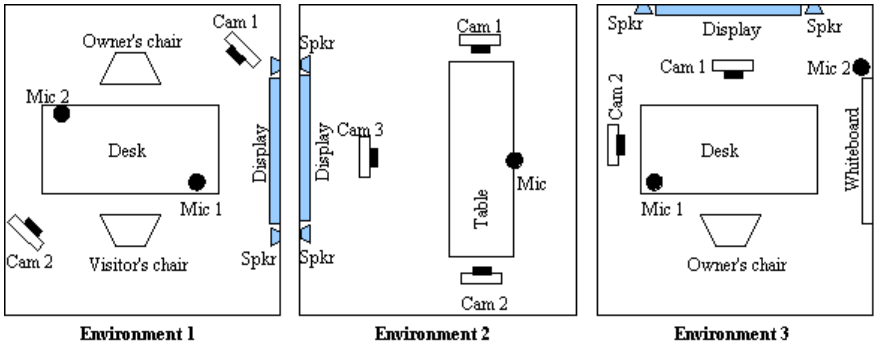

Figure 6: Layout of the environments 1, 2 and 3 which map to Consultation room, Patient examination room and Doctor's room resp. (telemedicine application) and as Bob's room, Alice's studio and Charles' office (office collaboration application).

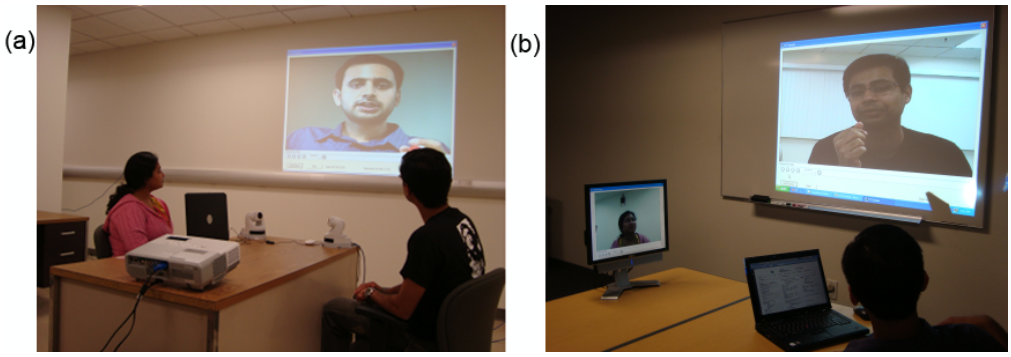

Figure 7: Connection between 'Consultation room' and 'Doctorś room' environments.

The system also clearly allowed *bi-directional connectivity*. This also allowed creation of JSM (Joint Situation Model) which not only connected the two environments but also created opportunity for creation of new type of media which can only be created in such joint spaces. The virtual projection of the archived X-ray originating from nurse's environment was physically annotated by the doctor in his environment. Such combination of archived-virtual and live-physical entities across environments to create new events and entities which do not belong to any one environment but rather the JSM is indeed an interesting artifact.

The interaction undertaken was not dependent on physical similarity. The doctor's room environment dynamically reconfigured itself to connect to the new environment as required. It changed from showing two live video feeds to one X-ray image and one (most) appropriate video feed.

The *privacy* aspect was handled by allowing users to configure their sharing setting in the event server. Hence, while the doctor was able to see the contents from the nurse's computer the reverse was not true in our implemented system.

The design implications were also adhered to as the architecture was event-based and multimodal. Abstracted interaction via the event server allowed the various input/outputs to change dynamically. It was also a scalable architecture working on the Internet and the system was able to scale up and down with device sophistication. For example, we used PC monitors, projectors and multi-tiled display walls as different video output devices in our current implementation. The system was able to request appropriate streams and support the various sophistication levels as required. Multiple applications (telemedicine and office collaboration) were implemented and tested and the system supported data storage for revisits.

Thus, all the design principles promulgated in section 2 were adhered to and demonstrated (albeit in a preliminary form) via the implementation.

6. RESEARCH CHALLENGES AND ROADMAP

While we have described successful initial implementation experience with E2E systems, there are multiple research challenges which need to be handled effectively for creation of more sophisticated E2E systems. A summary of the relevant challenges expected in different areas of E2E have been summarized in table 2. It also lists the possible approach to solve the relevant problem or mentions the preliminary work in that direction undertaken (both by our group and others in the research community) in that direction.

An important point to note is that though challenges in some aspects of the components outlined have been handled before, we need to look at them afresh with an E2E perspective. Also, putting the pieces together presents some novel challenges for individual areas as well as for developing interconnections among the components, cross-optimizing components, meeting the real-time requirement for true interactivity, and developing a new paradigm for using these systems. Most importantly, it brings a new perspective. This holistic perspective results in the Gestalt: a unified concept, a configuration that is greater than the sum of its parts.

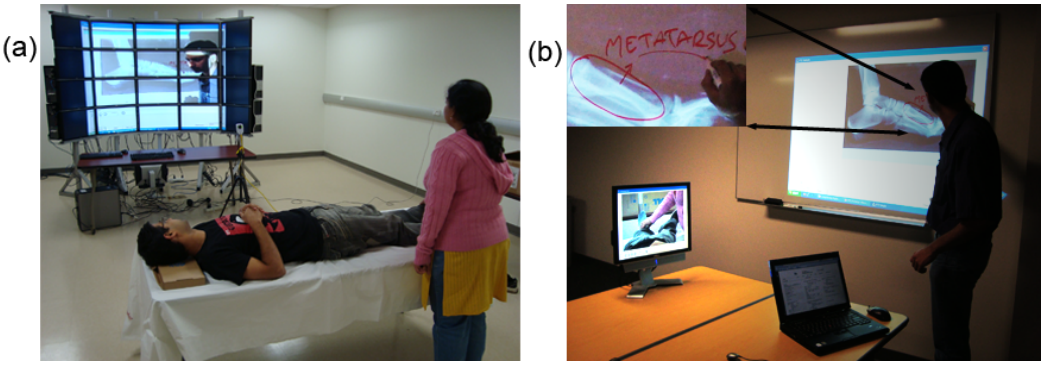

Figure 8: Connection between 'Examination room' and 'Doctorś room' environments.

Table 2: Research challenges and road-map summary

Area	Challenges	Approaches to be developed based on/extended from:
DAA	- Handling of multimodal data - Assimilation - Sensor set selection - Selective data sampling	- 'Out of the box' handling of heterogeneous data[10] - *Time-line* as opposed to sensor (e.g. video) based assimilation[2] - *Sensors selected based on the type of events to detected* - Experiential Sampling[13]
EM	- Automatic registration of important points - Sensor registration - Behavioral constraints on *actionability*	- Based on CAD models, architectural blueprints and helped by sensor registration - Sensor specs, web info., and extension of works like [25] - Developing formal language which extends a conceptual spatio-temporal model [8] with additional constructs.
SM	- Representing *situations* as an evolving series of events generated by objects - Minimize decision making at *compile-time*	- Extension of models like SNOOP and Amit. - *Indices* concept from real-time data management community
ES	- Scalable indexing for multimodal data coming from different environments - Multimodal information fusion into higher level multi-dimensional index structure - Event schema language - Privacy/ Security issues	- Scalable indexing like [27] - Extension of multidimensional indices like HYDRA [27] - Based on ontology of languages such as RDF schema and OWL from the semantic web domains [35] - Automated approach which is *dynamic, flexible*, can be *feed-backed on*, and allows *user control* [22].
Presentation& Interaction	- Finding relevant yet feasible positions for display - Easy tools for user to specify desired view-point - Usability of camera switching	- Automatic detection of suitable surfaces for data presentation[18] - Natural interfaces like VRML based in [14] - Eye perception studies like[33]
Networking& Distribution	- Best effort Internet issues like latency, congestion and availability - Scale to large number of participating environments - Novel means to reduce burden on network bandwidth	- Exploiting correlation between different sensors to reconstruct missing information and predicting best sensors to select in near future. - P2P concepts like swarms to serve multiple environments simultaneously - Exploiting 'social-network' features to characterize/channel environment data.

7. CONCLUSIONS

In this paper, we have described a new form of communication which supports natural human interaction by connecting environments to environments (E2E) rather than specific devices. We formulated the critical design principles for such communication as being natural, semantic, seamless, bi-directional, privacy-aware and independent of physical similarities. We proposed an abstracted, event-based, multimodal and scalable architecture to support such communications. The key ideas were demonstrated via an implementation which supported telemedicine and an office collaboration applications. The specific research challenges anticipated in creation of more sophisticated E2E systems were listed and a road map was suggested.

8. REFERENCES

[1] B. B. A. Sellen and J. Arnott. Using spatial cues to improve videoconferencing. In *Proceedings of CHI'92*, pages 651–652, 1992.

[2] P. K. Atrey, M. S. Kankanhalli, and R. Jain. Timeline-based information assimilation in multimedia surveillance and monitoring systems. In *VSSN '05: Proc. ACM international workshop on Video surveillance & sensor networks*, pages 103–112, 2005.

[3] S. Bly, S. Harrison, and S. Irwin. Media spaces: bringing people together in a video, audio, and computing environment. *Communications of the ACM*, 36(1):28–46, 1993.

[4] S. Boll and U. Westermann. Mediaether: an event space for context-aware multimedia experiences. In *ETP '03: Proc. ACM SIGMM workshop on Experiential telepresence*, pages 21–30, 2003.

[5] J. W. Chastine, K. Nagel, Y. Zhu, and L. Yearsovich. Understanding the design space of referencing in collaborative augmented reality environments. In *GI '07: Proceedings of Graphics Interface 2007*, pages 207–214, 2007.

[6] G. C. de Silva, T. Yamasaki, and K. Aizawa. Evaluation of video summarization for a large number

of cameras in ubiquitous home. In *MULTIMEDIA '05: Proc. ACM international conference on Multimedia*, pages 820–828, 2005.

[7] W. Gaver, T. Moran, A. MacLean, L. Lovstrand, P. Dourish, K. Carter, and W. Buxton. Realizing a video environment: Europarc's rave system. In *Proceedings of CHI'92*, pages 27–35, 1992.

[8] H. Gregersen. The formal semantics of the timeer model. In *APCCM '06: Proc. Asia-Pacific conference on Conceptual modelling*, pages 35–44, 2006.

[9] C. Heath and P. Luff. Disembodied conduct: Communication through video in a multi-media office environment. In *Proceedings of CHI'92*, pages 651–652, 1992.

[10] R. Jain. Out-of-the-box data engineering events in heterogeneous data environments. *Data Engineering, 2003. Proceedings. 19th International Conference on*, pages 8–21, 5-8 March 2003.

[11] R. Jain, P. Kim, and Z. Li. Experiential meeting system. In *ETP '03: Proc. ACM SIGMM workshop on Experiential telepresence*, pages 1–12, 2003.

[12] N. P. Jouppi, S. Iyer, S. Thomas, and A. Slayden. Bireality: mutually-immersive telepresence. In *MULTIMEDIA '04: Proc. ACM international conference on Multimedia*, pages 860–867, 2004.

[13] M. Kankanhalli, J. Wang, and R. Jain. Experiential sampling on multiple data streams. *Multimedia, IEEE Transactions on*, 8(5):947–955, Oct. 2006.

[14] A. Katkere, S. Moezzi, D. Y. Kuramura, P. Kelly, and R. Jain. Towards video-based immersive environments. *Multimedia Syst.*, 5(2):69–85, 1997.

[15] P. H. Kelly, A. Katkere, D. Y. Kuramura, S. Moezzi, and S. Chatterjee. An architecture for multiple perspective interactive video. In *MULTIMEDIA '95: Proc. ACM international conference on Multimedia*, pages 201–212, 1995.

[16] C. D. Kidd, R. Orr, G. D. Abowd, C. G. Atkeson, I. A. Essa, B. MacIntyre, E. D. Mynatt, T. Starner, and W. Newstetter. The aware home: A living laboratory for ubiquitous computing research. In *CoBuild '99: Proceedings of the Second International Workshop on Cooperative Buildings, Integrating Information, Organization, and Architecture*, pages 191–198, 1999.

[17] W. Liu, A. D. Cheok, C. L. Mei-Ling, and Y.-L. Theng. Mixed reality classroom: learning from entertainment. In *DIMEA '07: Proc. international conference on Digital interactive media in entertainment and arts*, pages 65–72, 2007.

[18] A. Majumder and R. Stevens. Perceptual photometric seamlessness in projection-based tiled displays. *ACM Trans. Graph.*, 24(1):118–139, 2005.

[19] G. Mark, S. Abrams, and N. Nassif. Group-to-group distance collaboration: Examining the 'space between'. In *Proc. European Conference of Computer-supported Cooperative Work*, pages 14–18, 2003.

[20] G. Mark and P. DeFlorio. An experiment using life-size hdtv. In *Proc. IEEE Workshop on Advanced Collaborative Environments (WACE)*, 2001.

[21] S. McCanne and V. Jacobson. vic: a flexible framework for packet video. In *MULTIMEDIA '95: Proc. third ACM international conference on Multimedia*, pages 511–522, 1995.

[22] S. Moncrieff, S. Venkatesh, and G. West. Privacy and the access of information in a smart house environment. In *MULTIMEDIA '07: Proc. international conference on Multimedia*, pages 671–680, 2007.

[23] D. Nguyen and J. Canny. Multiview: Improving trust in group video conferencing through spatial faithfulness. In *Proceedings of CHI'07*, pages 1465–1474, 2007.

[24] H. Packard. Hp halo overview, 2007.

[25] R. Raskar, M. S. Brown, R. Yang, W.-C. Chen, G. Welch, H. Towles, B. Seales, and H. Fuchs. Multi-projector displays using camera-based registration. In *VISUALIZATION '99: Proc. 10th IEEE Visualization 1999 Conference (VIS '99)*, 1999.

[26] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs. The office of the future: a unified approach to image-based modeling and spatially immersive displays. In *SIGGRAPH '98: Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 179–188, 1998.

[27] A. A. Sawchuk, E. Chew, R. Zimmermann, C. Papadopoulos, and C. Kyriakakis. From remote media immersion to distributed immersive performance. In *ETP '03: Proc. ACM SIGMM workshop on Experiential telepresence*, pages 110–120, 2003.

[28] E. Schooler. A distributed architecture for multimedia conference control. Technical report, University of Southern California, 1991.

[29] SecondLife. http://secondlife.com/.

[30] R. Sheppard, W. Wu, Z. Yang, K. Nahrstedt, L. Wymore, G. Kurillo, R. Bajcsy, and K. Mezur. New digital options in geographically distributed dance collaborations with teeve: tele-immersive environments for everybody. In *MULTIMEDIA '07: Proc. international conference on Multimedia*, pages 1085–1086, 2007.

[31] R. Stults. Media space. Technical report, Xerox PARC, 1986.

[32] R. Vertegaal, G. V. der Veer, and H. Vons. Effects of gaze on multiparty mediated communication. In *Proc. Graphics Interface*, pages 95–102, 2000.

[33] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung. Gaze-2: conveying eye contact in group video conferencing using eye-controlled camera direction. In *CHI '03: Proc. SIGCHI conference on Human factors in computing systems*, pages 521–528, 2003.

[34] M. Weiser. The computer for the 21st century. *SIGMOBILE Mob. Comput. Commun. Rev.*, 3(3):3–11, 1999.

[35] G. U. Westermann and R. Jain. Events in multimedia electronic chronicles (e-chronicles). *International Journal of Semantic Web and Information Systems (IJSWIS*, 2(2):1–27, Oct. 2006.

[36] Z. Yang, W. Wu, K. Nahrstedt, G. Kurillo, and R. Bajcsy. Viewcast: view dissemination and management for multi-party 3d tele-immersive environments. In *MULTIMEDIA '07: Proc. international conference on Multimedia*, pages 882–891, 2007.