Multimed Tools Appl (2009) 44:361–388
DOI 10.1007/s11042-009-0281-0

Towards Environment-to-Environment (E2E) multimedia communication systems

Vivek K. Singh · Hamed Pirsiavash · Ish Rishabh · Ramesh Jain

Published online: 5 May 2009

Abstract We present an approach to connect multiple remote environments over web 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. This object position and activity information is used by a scalable event-based multimodal information system called EventServer to share the appropriate experiential information with other environments as well as to present incoming multimedia information on right displays and speakers. 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.

Keywords Environment-to-Environment communication · Connecting environments · Experiential interaction · Telepresence

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 video-conferencing have

V. K. Singh (✉) · H. Pirsiavash · I. Rishabh · R. Jain
University of California, Irvine, Irvine, CA, USA
e-mail: singhv@uci.edu

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 [24, 35, 41], interaction, peripheral vision [20], spatio-semantic integrity and even information flow [21]. 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 [25]), or exploring the other extreme of moving all the communication to the virtual world (e.g. SecondLife [34]). 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' [43] 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. Such a *web-based* architecture has multiple far-reaching consequences in terms of the use of web-based technologies like 'ontologies' for handling experiential data, the adoption and scalability of the approach as well as supporting 'serendipitous interoperability' across environments. 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. The implementation of two different applications also allows us to identify the components which should be handled at a generic level and which are highly application dependent.

Our contributions in this paper are two-fold:

1. We propose E2E as a new paradigm for experiential web-based 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 [34] 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 [25], 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 [6] and Multi-perspective-imaging [16] 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.

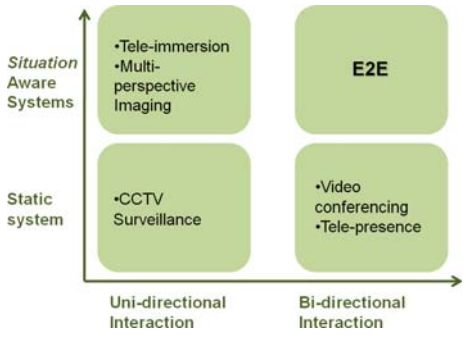

Fig. 1 Comparison of E2E with related works

The 'Office of the future' project [28], 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 [4, 8, 40]. 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 [8]. 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 [8], and people generally resort to using exaggerated movements to communicate over video [10]. 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 [22], ISI's Multimedia Conference Control (mmcc) [33], 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.

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 [18], gloves [6] or unnatural hardware devices like surrogate [13] to support interaction.

Ambient intelligence, ubiquitous computing [43], Smart/Aware Home research areas (e.g. [7, 17]) on the other hand have made many advancements in understanding user behaviors within an environment for applications like tele-medicine, 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. Pervasive Communication approaches like iRos [27], Gaia [30] also look at connecting multiple devices within physical spaces for communication, but none of them has a web based approach to create a web of physical spaces and devices which can interoperate freely to support bidirectional experiential communication and knowledge capture. Semantic web community [3, 36] on the other hand is highly interested in capturing and utilizing all the textual information present on the Internet. However, they do not consider the approach of enhancing and supporting experiential communication as a way of capturing experiential knowledge for future use.

Some interesting works within the Multimedia community have been proposed for tele-immersive dance and music performance [37, 45] and [32]. Works like [45] and [37], 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 [32], 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. [12] describe how event based organization can support communication across time and space. Similarly, Boll et al. [5] 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.

An earlier version of this paper appeared as [39]. While that version described our preliminary ideas, the current version reflects the developments in our thought process regarding various components. Specifically it gives proper impetus to our ideas on a web-based architecture for the communication, a more comprehensive description for the components of Environment Modeling and Situation Modeling and provides detailed implementation description for the Office Collaboration scenario.

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 the users should be allowed to interact via their natural physical spaces rather than fabricated cyber spaces. Similarly, the users 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. [16]), 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 [25] 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 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 link age 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

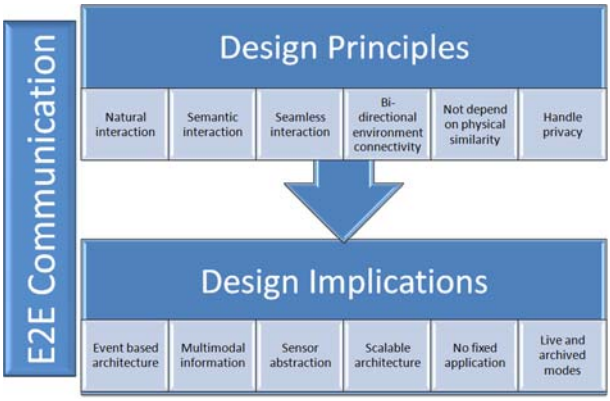

Fig. 2 A summary of design principles and applications for E2E

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 Architecture for E2E communication paradigm

Based on the design principles described in the preceding 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 in any environment, we adopt an event-based architecture to support it. Similarly, the growth and scalability of E2E paradigm leverages on the web-based architecture for communication across nodes.

4.1 Environment node architecture

Figure 3 shows an overview of the architecture for each E2E node i.e. an environment which supports the E2E communication paradigm. The **'Data acquisition and analysis' (DAA)** component gathers the relevant information from various sensors and

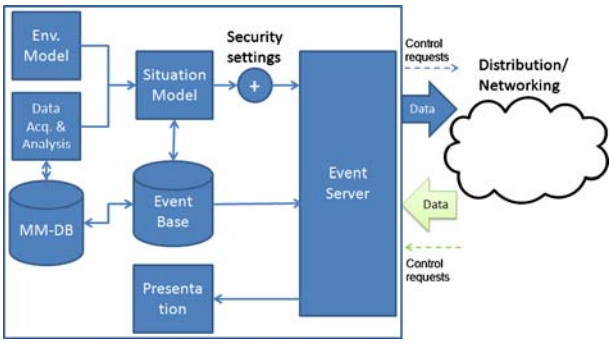

Fig. 3 A high-level architecture diagram for E2E

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 a 'person presence' and 'person talking', the environment model is useful in deciding which location these sub-events originate from and hence whether they refer to the same person. The combination of information coming DAA and EM can be used to define events at the elemental level which is context or application independent. The semantic understanding of the event requires additional contextual information to be added by the specific **Situation Model (SM)**. The SM represents all the 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. The explicit modeling of environment and situation modeling are important to decouple the business-logic from being hard-coded to specific sensors/devices or specific entities.

The generated event are filtered based on the security/privacy settings 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.

4.2 Web based communication architecture

The individual environment nodes in E2E are connected together to form of web of physical spaces, which are virtually joined together in a 'Joint Situation Model' as shown in Fig. 4. 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

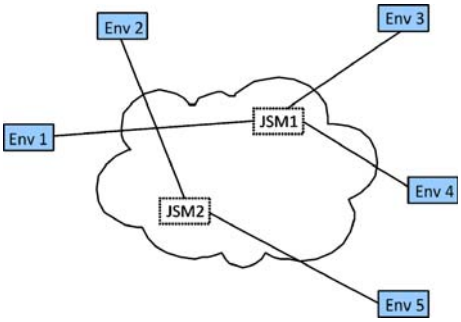

Fig. 4 Multiple E2E clients at different locations can connect through the Internet and create correct collaborative 'situations' at each location using the Joint Situation Model

or may not hold much relevance, their combined effect in the JSM will be of critical importance.

The use of a web based architecture brings multiple advantages to the system. Firstly, a web based architecture allows for Internet scale growth of E2E communication systems wherein each user environment with any set of devices can easily join this web and start communicating with others. We conceptualize the EventServer as a next generation Web Server which can make available any web-services (i.e media streams/resources) as it may feel appropriate to others. Hence the EventServer starts to share experiential information coming from sensors and devices directly on the web just like the current-day Web Servers which share textual information.

Next, a web based architecture allows us to harness the power of emerging web-based technologies which will act as integral building blocks for our sentient communication systems. The Semantic Web [3] based technologies like OWL and RDF based ontologies. Tools like DBpedia [2] which organize the entire wikipedia knowledge into relational tuples, can be used by the system to gain semantic knowledge about a variety of events, entities and relationships between them. Such contextual knowledge is very useful for creating the appropriate Situation Model for the various environments. Similarly an organization of different types of sensors and ontology of different physical locations allows us to generically approach Environment Modeling (e.g. [38]). Also, the fact, that such resources are freely available on the Web means, that:(1) We can start using them without re-creating them, (2) Each environment can use the same resource, without any need for maintaining separate copies or rework, (3) Each ontology or resource update gets automatically transmitted to each node and lastly (4) Each node uses similar abstractions to allow interoperability across different nodes of E2E and even other architectures which adopt similar approach.

While envisioning such planetary scale multimodal communication systems, we also realize the challenges and the opportunities which lie in efficient data organization. While real-time data organization for such disparate data sources is a major challenge, it is also an opportunity to use the right architecture to start collecting all the experiential data being generated in the world from now on. This is where we start realizing an *EventWeb* and move ahead of what is propounded by the semantic web community. While the Semantic Web community is actively engaged in making available (and usable) all the information which is present in an explicit text media form anywhere on the Internet; we also want to capture the multimodal experiential data which may not amenable to any explicit text based rendering. Thus, while semantic web is useful to move from data to information, we want to move from

information to insights and experiences.We build upon the sources or contextual knowledge as made available by various ontologies and use *events* and *entities* as the common abstractions across different environments to bridge the information which may get fragmented on the current web due to its application context, physical location or user accounts. For example, earlier instant messenger services used to save the user contacts and their details locally on a user PC, but soon they moved to a web-based approach wherein a user can access his/her contacts and their profile pictures etc. wherever he/she logs in. We want to extend such an approach such that experiential information is made available at different physical locations (i.e environments), across different devices (e.g. PDAs, phones, Computers, Televisions), to multiple users and across applications (say surveillance, life-blogging and telepresence) by a common web based architecture.

Such an architecture also influences the way we approach our data transmission. While the current web is largely *pull*-based with only a few *pseudo-push* approximations (e.g. RSS or 'Facebook' event feeds). We envision a *publish-subscribe* based architecture to make available all experiential communication information on the Web. The publish-subscribe organization can be based on social networks or be explicit user input. Thus instead of 'sending' the event media (like photographs, video etc.) to any contacts, we envision their automatic selection, ordering and sharing across the web with different people with distinct access rights. Thus users can connect to each other in both synchronous and asynchronous modes to multimodal data and learn whats happening with others in their daily environments.

Lastly, the use of a common web based architecture creates opportunities for serendipitous inter-operation across devices and resources which are physically separate. For example, sophisticated earthquake measuring equipment in a California laboratory could be connected to field remote sensors in Hawaii to provide shared visualization to users in Germany and Singapore. Such shared visualization can lead to shared inputs and emergent semantics which would not be possible without a multimodal planetary level architecture. In certain sense we borrow the concepts of a room-based operating system from the pervasive communications community (e.g. iRos, Interactive Room Operating System [27]), where all devices in a room like PDA, printer, PC, Whiteboard etc. are controlled by a common meta-operating system and extend the ideas to creating a planetary level interconnected version where all devices can inter-operate to make the most appropriate resources available to right person at the right location and the right time.

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. We give specific details for the Environment Modeling and Situation Modeling components to highlight our

Table 1 The coverage for each device

Device	Coverage
Cam_1	3, 4
Cam_2	3, 4
Cam_3	1, 2, 3, 4
Mic_1	1, 2, 3, 4
$Display_1$	1, 2, 3, 4
Spk_1	1, 2, 3, 4

ideology that the systems should be designed at a generic level with specific context added through late binding to specific applications.

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, gesture recognition and audio volume detector modules as shown in Tables 1 and 2.

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

Table 2 The list of devices covering the space

Space label	Devices
1	- Cam_3
	- Mic_1
	- $Display_1$
	- Spk_1
2	- Cam_3
	- Mic_1
	- $Display_1$
	- Spk_1
3	- Cam_1, Cam_2
	- Mic_1
	- $Display_1$
	- Spk_1
	- Table
4	- Cam_2, Cam_1
	- Mic_1
	- $Display_1$
	- Spk_1
	- Table

mere evidences of the event happening. A preliminary description of this approach was presented in [1].

5.2 Environment model

The sensory and presentation devices, along with information about the ambient physical space define the environment of the system. Environment Model, thus, is an abstraction of the sensory and presentation space. We have identified the important constituents of EM:

1. *Devices*: Each device has a descriptor called Information Source Registry (ISR) that completely defines the device. Each ISR contains the following fields:
   - *Device ID*: must be unique for each device.
   - *Device type*: Sensory or presentation device.
   - *Signal type*: Visual, audio, motion, RFID, etc.
   - *Movability*: Fixed or mobile.
   - *Sharablility*: Some sensory devices may be used only for monitoring the environment, and not for communication. This information is reflected in the ISR.
   - *Coverage*: It defines the *part of physical environment which the device covers*. This is an important characteristic of a device and will be discussed in detail soon.
   - *Accessibility information*: IP address, make/model, driver information.
2. *Physical environment*: This describes the physical space in which the Event Server (ES) resides along with the sensory and presentation devices. We can further subdivide this information into two separate fields:
   - *3-D floorplan*: 3-D floorplan or maps are increasingly becoming popular and they provide important contextual information about the environment. In our implementation, however, we have not incorporated the floor plans.
   - *3-D object map*: In a room, there may be several objects like desk, bed, whiteboards, which can be considered fixed for all practical purposes. Hence their location and coverage in space can be indicated on the floor plan. This information can be very important for detecting activities (like using the desk/whiteboard). This has been depicted in Tables 1, 2 and 3.
3. *Inverted coverage*: This is an inverted index of the physical space and consists of information about the devices and objects present in that part of the space. Thus, one can readily obtain the list of devices and objects placed in a given space. Note that this can be computed from the coverage information of all devices and objects.

Each device and object covers a certain part of the physical space it is kept in. By using this information, one can capture the implicit relationships between devices as well as the objects.

Table 3 Relationship of objects with the surrounding space is shown

Object	Space covered	Adjacent space
Table	3, 4	1, 2

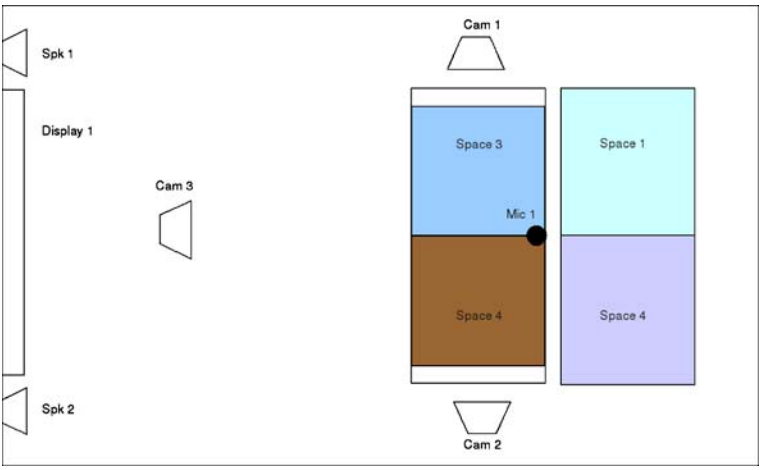

Fig. 5 Space indices shown for the Environment 2 which was examination room and studio respectively for the two implementations

As already mentioned, each device and object has a description of coverage of the physical space. Space in the physical environment can be divided into various parts and a label or index be assigned to each such part. ISR of each device should maintain a list of space indices which the device covers. Various calibration techniques can be used to determine the coverage indices of the devices. However, it should be noted that it is the coverage, and not the calibration parameters that are important. In fact, many a times, even calibration may not be necessary. During initial setup, an administrator can manually index physical space and define the coverage of the device. Our implementation illustrates this point. The system also maintains an inverted index of volume space indices. Thus, each volume index stores a list of devices and objects that cover that volume. This is needed for efficient implementation, just as an inverted file index is needed for efficient text retrieval.

Tables 1 and 2 show coverage information for Environment 2 which was mapped as 'studio' and 'examination room' for the office collaboration and telemedicine applications respectively. Physical spaces were labeled with different labels as shown in Fig. 5.

EM serves to decouple the environment from the application and hence provides flexibility in defining the *rules* or *business logic* of application. Using this, we can define rules like *if there is an activity in region x, then select the device that best covers region y*. Without EM, one would need to define rules as *if there is activity in device p, then select device q*. As we can see, the latter includes the devices explicitly and hence is inflexible. If, later on, device q is moved to cover a different region and device r replaces it, this would require to change the rule itself. Note that changing the rule amounts to changing the application code.

However, when EM is used, even if device q is replaced by another device r, only the relevant entries in EM configuration file (Tables 1 and 2) need to change.

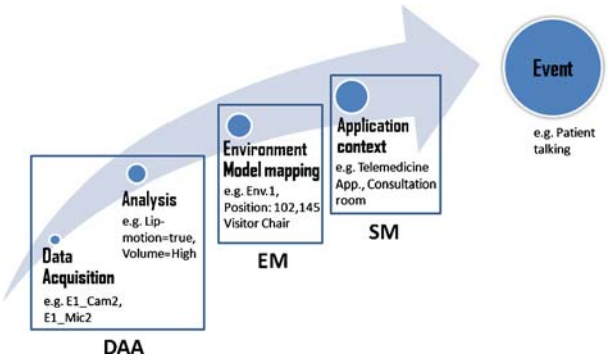

Fig. 6 Event detection process in E2E

5.3 Situation model

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

The process of the combination of information from DAA, EM and SM to detect the events has been shown in Fig. 6. As shown, the information captured from the sensors in the data acquisition step is analyzed to detect *elemental* level events [31] i.e. events which are independent of application context and hence are at a level which is common across different applications. An example of such an event is E2.enters(L2), where E2 and L2 are labels for a particular Entity and Location. The translation of E2 and L2 to their semantic equivalents would require application based context (e.g. E2.role= Architect, L2.role= studio), and hence the event can be used by the system as 'the architect entered the studio' event. A summary of the various events detected in the current implementation and their respective information components coming from DAA, EM mapping and event detection at both elemental and application level has been shown in Table 7. Note that while elemental level event may be common across applications (e.g. Event 3 and Event 12 in Table 7), their Application level mapping is different based on the situational context as shown in Table 5.

An important aspect of Situation Modeling hence is to identify the *types* of contextual information which are critical for translating elemental level events into application level events. We identify the relevant context based on 6 important notions of *What, Who, Where, When, How and Why* and we argue that (refer Table 4) *Entities, Location, Time-duration and Relationships* are the most critical

Table 4 Components of a situation model

	Entity	Location	Time	Relationships
What	X	X	X	
Who	X			
Where		X		
When			X	
How				X
Why				X

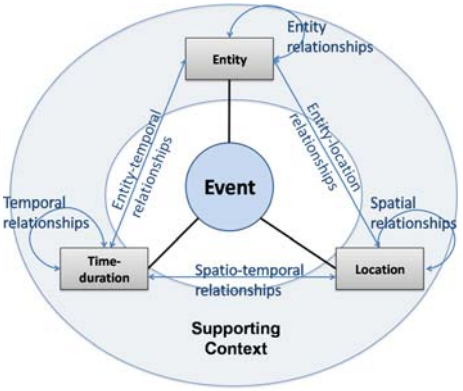

Fig. 7 Events and surrounding context

types of contextual information to support such queries (refer Fig. 7). We assign each instantiation of Entity, Location and time-duration a *label*, *type*, a *structural* level identifier and a semantic level *role*. Examples of such descriptions are shown in Table 5. Note that the system level identifier is something which is assigned automatically by the system each time it comes across a new entity, location or event. The *type* of a contextual element is defined based on a list of supported classes.

The structural level identifier is a real world identifier which is unambiguous with respect to a system defined reference frame and usable across multiple applications. The 'type' and structural identifier information is added by the Environment Model itself upon detection and is application independent. The semantic role on the other hand is the specific translation in the particular application context. Hence we need world knowledge, data sources, common sense, rules or human input to specifically provide this information. In the current office collaboration example, for the 'Person' type of information, we used names as the structural identifiers and used an employee database to translate names to the semantic roles of 'Manager', 'Architect' etc. Similarly we used a (dummy) departmental layout table to obtain the semantic location roles of 'Studio' etc. The semantic labels for the previous events were obtained from an EventBase as further described soon. Also note that we

Table 5 Contextual information elements used in office collaboration application

Entity Id	Type	Structural label	Semantic role
E1	Person	Bob	Manager
E2	Person	Alice	Architect
E3	Person	Charles	Specialist
E4	Chair	949123	Manager's chair
E5	Chair	949124	Visitor's chair
E6	Model	949126	Aircraft model
E7	Board	949127	Architect's projection board
E8	Board	949127	Specialist's projection board
L1	Location	DBH.2.059	Manager's office
L2	Location	DBH.2.060	Studio
L3	Location	CalIT2.3.011	Specialist's room

Table 6 Relationship between different contextual information elements used in office collaboration application

Relationship	Type(elements)	Type(logical)	Attribute 1	Attribute 2
Is boss of	Entity–entity	Semantic	Manager	Architect
Gives feedback	Entity–entity	Semantic	Specialist	Architect
Is nearby	Location–location	Semantic	Manager's room	Studio
Precedes	Temporal–temporal	Structural	Architect's entry	Architect talking
Belongs to	Location–entity	Semantic	Manager's room	Manager
Sits upon	Entity–entity	Semantic	Manager	Manager's chair

decided to use names as the choice for unambiguous reference at structural for persons, while using RFID Asset tags for the chairs, model and boards. For location we followed the convention of using buildingname.floor.roomNumber as prevalent in our university. Hence the Structural references were unambiguous within the chosen reference frame and independent of applications considered though not universally unique.

Relationships across the different information types can also be at two levels-semantic and structural (refer Table 6). The structural level relationships are those which make sense even when the entities (or location, time-durations) are represented in their structural form. Examples of such relationships include 'same', 'different', 'inside', 'outside', 'before', 'after' etc. Semantic relationships on the other hand describe relationship between semantic labels. Examples of such attributes include 'is-boss-of', 'gives-feedback', 'is-nearby', 'belongs-to' etc. While the structural relationships are defined at system design time, the list of valid/relevant semantic relationships to be supported for each application need to be entered or selected from an ontology by the system administrator at configuration time. Lastly, the relationships can be between attributes of the same type (e.g. 'spatial', 'temporal') or across different modalities (e.g. 'spatio-temporal', 'entity-temporal', 'spatio-entity-temporal'). Relationships across modalities open doors for higher level understanding (and exploitation of) relationships like 'my fiance's birthday' or 'My Boss's office' etc.

Besides allowing similar elemental events to be mapped differently across applications by using contextual information, the role of explicit Situation Modeling is also important in E2E to decouple the system's business logic from the specific contextual information. While the rules in application business logic can be defined at application level e.g. 'Do X, when *Architect* is talking', they need not be hard-coded using specific entities like E2 or person Alice. Hence, if a new person (say David), joins the company and takes up the architect role, the rules can work just as it is.

5.4 Event server

The ES received the event-related streams and physical & semantic location data from the EM and SM and 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.

Table 7 Various events detected in the current E2E implementation

No.	Data acquisition	Data analysis	EM mapping	Elem. event	App. event
Telemedicine					
1.	E1_Cam1	Face, Blob	E4(chair)	Person detected (E1)	Nurse present
2.	E1_Cam1 , E1_Cam2	Face, Blob	E4(chair), E5(chair)	2 persons present (E1, E2)	Nurse, Patient present
3.	E1_Cam1 , E1_Mic1	Volume, Lip-tracking	E4(chair)	Talking near E4	Nurse talking
4.	E1_Cam2 , E1_Mic2	Volume, Lip-tracking	E5(chair)	Talking near E5	Patient talking
5.	E1_Cam1 , E1_Cam2	Face, Blob	L1(room)	No person in L1	Exit from consultation room
6.	E2_Cam1 , E2_Cam2 , E2_Cam3	Face, Blob	L2(room)	Person present in L2	Entry into exam room
7.	E2_Cam3	Blob	E6(table)	Person movement near E6	Nurse movement
8.	E3_Cam1 , E3_Cam2	Face, Blob	L3(room)	Person position in L3	Doctor's position
9.	E3_Cam2	Gesture recog.	E6 (board)	Interaction with E6	Interaction with X-ray
Office collaboration application					
10.	E1_Cam1	Face, Blob	E4(chair)	Person detected (E1)	Manager present
11.	E1_Cam2	Face, Blob	E5(chair)	Person detected (E2)	Architect's entry
12.	E1_Cam1 , E1_Mic1	Volume, Lip-tracking	E4(chair)	Person talking near E4	Manager talking
13.	E1_Cam2 , E1_Mic2	Volume, Lip-tracking	E5(chair)	Person talking near E5	Architect talking
14.	E1_Cam2	Face, Blob	E5(chair)	Person exit	Architect's exit
15.	E2_Cam1 , E2_Cam2 , E2_Cam3	Face, Blob	L2(room)	Person entry into L2	Entry into studio
16.	E2_Cam1 , E2_Cam2 , E2_Cam3	Face, Blob	L2(room)	Person position in L2	Architect's position
17.	E1_Cam1 , E1_Cam2	Face, Blob	L1(room)	Person position in L1	Manager's position
18.	E3_Cam1 , E3_Cam2	Face, Blob	L3(room)	Person position in L3	Specialist's position
19.	E2_Cam3	Gesture recog.	E7(board)	Gestures on E7	Architect interacting
20.	E1_Cam2	Gesture recog.	E7(board)	Gestures on E8	Specialist interacting

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). Figure 4 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

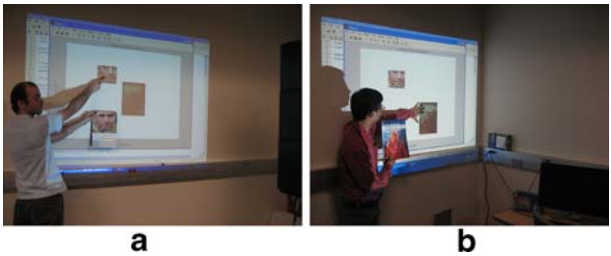

Fig. 8 JSM for shared visualization of media artifacts from both environments in Office Collaboration application (**a**, **b**)

a JSM for the connected environments. Figure 8, shows an example of bi-directional interaction between two environments, wherein they are sharing their resources with others and at the same time accessing others. While the resources were physically present in two separate individual environments, their virtual combination existed as one common JSM on the web with appropriate access information. For more details on JSM creation please refer [26].

The triggers for undertaking various information sharing, removal, editing etc. in the shared visualization space were naturalistic hand-gestures. While the mapping of such action triggers to different system behavior was manually configured at this time, in future we aim to make it generic. 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.

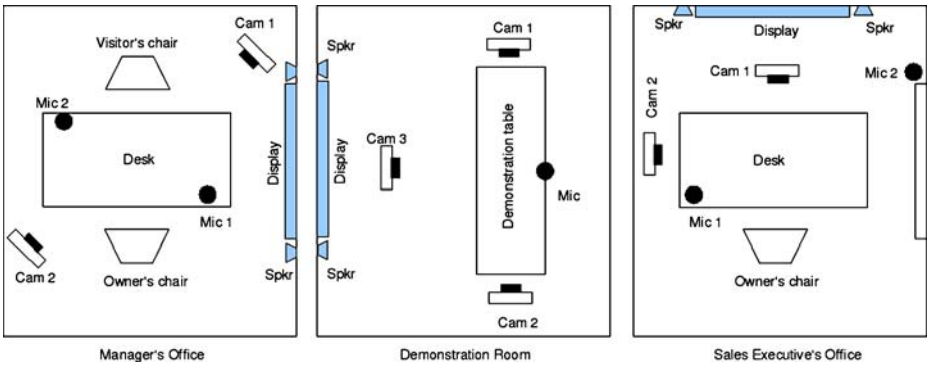

Fig. 9 Layout of the environments 1, 2 and 3 which map to Consultation room, Patient examination room and Doctor's room resp. (telemedicine application) and as Bob's room, Alice's studio and Charles' office (office collaboration application).

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

5.7.1 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. 9.

Fig. 10 Connection between 'Consultation room' and 'Doctorś room' environments.

Fig. 11 Connection between 'Examination room' and 'Doctorś room' environments.

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. 10a). 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. 10b.

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. 11 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.

5.7.2 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 Bling747 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 by looking at it from different angles. They also use the projection board to share different relevant design documents and ideas before finally converging on one suitable design.

Fig. 12 Connection between the three environment's in office collaboration application

Just like the telemedicine application, the initial discussion appears like a virtual 3 way communication at both the offices. When Alice reaches her studio, automatic triggers are used to immediately connect her environment to the two other environments. Similarly, all the users are presented with the most appropriate video feed at each time instant as Alice is interacting with the model. The undertaken environment modeling and Situation modeling have already been described. Some snapshots of the connected environments are shown in Fig. 12. A summary of the events detected in the current implementations is shown in Table 7.

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. 9.

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 implementations relate closely to the promulgated design principles.

The interaction between the manager, architect and the specialist (viz. 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. Specialized operations upon a shared visualization were also handled using hand-gestures without

any additional equipment. Similarly, the users 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. Similarly in office collaboration application, the manager, architect and the specialist started of with an across the table feel of interaction. This was changed to display of appropriate information like specialist feedback etc. on the relevant front screens.

The system maintained *seamless interaction* even when the architect and manager moved between the manager's room environment and the studio. The new environment was connected automatically. In fact, the display screen in studio split or joint automatically to provide big or small images of other users depending on number of other users currently connecting to the architect. The architect and the specialist also had sufficient freedom to move within their environment's room and the patient/nurse could continuously maintain contact.

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. Similarly media artifacts useful for the discussion on the aircraft model were added by both the architect and the specialist into a common virtual space.

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. Similarly, in the shared visualization, the architect and the specialist maintained their control rights over the artifacts originating from their environment, even though they were stored on the web.

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.

Table 8 Research challenges and road-map summary

Area	Challenges	Approaches to be developed based on/extended from:
DAA	– Handling of multimodal data	– 'Out of the box' handling of heterogeneous data [11]
	– Assimilation	– *Time-line* as opposed to sensor (e.g. video) based assimilation [1]
	– Sensor set selection	– *Sensors selected based on the type of events to detected*
	– Selective data sampling	– Experiential sampling [14]
EM	– Automatic registration of important points	– Based on CAD models, architectural blueprints and helped by sensor registration
	– Sensor registration	– Sensor specs, web info., and extension of works like [29]
	– Behavioral constraints on *actionability*	– Developing formal language which extends a conceptual spatio-temporal model [9] with additional constructs.
SM	– Representing *situations* as an evolving series of events generated by objects	– Extension of models like SNOOP and Amit.
	– Minimize decision making at *compile-time*	– *Indices* concept from real-time data management community
ES	– Scalable indexing for multimodal data coming from different environments	– Scalable indexing like [32]
	– Multimodal information fusion into higher level multi-dimensional index structure	– Extension of multidimensional indices like HYDRA [32]
	– Event schema language	– Based on ontology of languages such as RDF schema and OWL from the semantic web domains [44]
	– Privacy/ Security issues	– Automated approach which is *dynamic, flexible*, can be *feed-backed on*, and allows *user control* [23].
Presentation & interaction	– Finding relevant yet feasible positions for display	– Automatic detection of suitable surfaces for data presentation [19]
	– Easy tools for user to specify desired view-point	– Natural interfaces like VRML based in [15]
	– Usability of camera switching	– Eye perception studies like [42]
Networking & distribution	– Best effort internet issues like latency, congestion and availability	– Exploiting correlation between different sensors to reconstruct missing information and predicting best sensors to select in near future.
	– Scale to large number of participating environments	– P2P concepts like swarms to serve multiple environments simultaneously
	– Novel means to reduce burden on network bandwidth	– Exploiting 'social-network' features to characterize/channel environment data.

6 Research challenges and road-map

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 8. 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.

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.

Acknowledgements We would like to thank Professor Aditi Majumder for multiple feedbacks and guidance in creating the joint situation model.

References

1. Atrey PK , Kankanhalli MS, Jain R (2005) Timeline-based information assimilation in multimedia surveillance and monitoring systems. In: VSSN ’05: proc ACM international workshop on video surveillance & sensor networks, pp 103–112
2. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2008) Dbpedia: a nucleus for a web of open data. In: ISWC ’07 + ASWC ’07: in proceedings of 6th international semantic web conference, 2nd Asian semantic web conference, November, pp 722–735
3. Berners-Lee T, Hendler JA, Lassila O (2001) The semantic web. Sci Am 284(5):34–43
4. Bly S, Harrison S, Irwin S (1993) Media spaces: bringing people together in a video, audio, and computing environment. Commun ACM 36(1):28–46
5. Boll S, Westermann U (2003) Mediaether: an event space for context-aware multimedia experiences. In: ETP ’03: proc ACM SIGMM workshop on experiential telepresence, pp 21–30
6. Chastine JW, Nagel K, Zhu Y, Yearsovich L (2007) Understanding the design space of referencing in collaborative augmented reality environments. In: GI ’07: proceedings of graphics interface 2007, pp 207–214

7. de Silva GC, Yamasaki T, Aizawa K (2005) Evaluation of video summarization for a large number of cameras in ubiquitous home. In: MULTIMEDIA '05: proc ACM international conference on multimedia, pp 820–828
8. Gaver W, Moran T, MacLean A, Lovstrand L, Dourish P, Carter K, Buxton W (1992) Realizing a video environment: Europarc's rave system. In: Proceedings of CHI'92, pp 27–35
9. Gregersen H (2006) The formal semantics of the timeer model. In: APCCM '06: proc Asia-Pacific conference on conceptual modelling, pp 35–44
10. Heath C, Luff P (1992) Disembodied conduct: communication through video in a multi-media office environment. In: Proceedings of CHI'92, pp 651–652
11. Jain R (2003) Out-of-the-box data engineering events in heterogeneous data environments. In: Proceedings. 19th International conference on data engineering, 5–8 March 2003, pp 8–21
12. Jain R, Kim P, Li Z (2003) Experiential meeting system. In: ETP '03: proc ACM SIGMM workshop on experiential telepresence, pp 1–12
13. Jouppi NP, Iyer S, Thomas S, Slayden A (2004) Bireality: mutually-immersive telepresence. In: MULTIMEDIA '04: proc ACM international conference on multimedia, pp 860–867
14. Kankanhalli M, Wang J, Jain R (2006) Experiential sampling on multiple data streams. IEEE Trans Multimedia 8(5):947–955
15. Katkere A, Moezzi S, Kuramura DY, Kelly P, Jain R (1997) Towards video-based immersive environments. Multimedia Syst 5(2):69–85
16. Kelly PH, Katkere A, Kuramura DY, Moezzi S, Chatterjee S (1995) An architecture for multiple perspective interactive video. In: MULTIMEDIA '95: proc ACM international conference on multimedia, pp 201–212
17. Kidd CD, Orr R, Abowd GD, Atkeson CG, Essa IA, MacIntyre B, Mynatt ED, Starner T, Newstetter W (1999) 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, pp 191–198
18. Liu W, Cheok AD, Mei-Ling CL, Theng Y-L (2007) Mixed reality classroom: learning from entertainment. In: DIMEA '07: proc international conference on digital interactive media in entertainment and arts, pp 65–72
19. Majumder A, Stevens R (2005) Perceptual photometric seamlessness in projection-based tiled displays. ACM Trans Graph 24(1):118–139
20. Mark G, DeFlorio P (2001) An experiment using life-size hdtv. In: Proc IEEE workshop on advanced collaborative environments (WACE)
21. Mark G, Abrams S, Nassif N (2003) Group-to-group distance collaboration: examining the 'space between'. In: Proc European conference of computer-supported cooperative work, pp 14–18
22. McCanne S, Jacobson V (1995) Vic: a flexible framework for packet video. In: MULTIMEDIA '95: proc third ACM international conference on multimedia, pp 511–522
23. Moncrieff S, Venkatesh S, West G (2007) Privacy and the access of information in a smart house environment. In: MULTIMEDIA '07: proc international conference on multimedia, pp 671–680
24. Nguyen D, Canny J (2007) Multiview: improving trust in group video conferencing through spatial faithfulness. In: Proceedings of CHI'07, pp 1465–1474
25. Packard H (2007) Hp halo overview
26. Pirsiavash H, Singh V, Majumder A, Jain R (2009) Shared Visualization Spaces for Environment to Environment Communication, Workshop on Media Arts, Science, and Technology (MAST), Santa Barbara, CA
27. Ponnekanti SR, Johanson B, Kiciman E, Fox A (2003) Portability, extensibility and robustness in iros. In: PERCOM '03: proceedings of the first ieee international conference on pervasive computing and communications, p 11
28. Raskar R, Welch G, Cutts M, Lake A, Stesin L, Fuchs H (1998) 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, pp 179–188
29. Raskar R, Brown MS, Yang R, Chen W-C, Welch G, Towles H, Seales B, Fuchs H (1999) Multi-projector displays using camera-based registration. In: VISUALIZATION '99: proc 10th IEEE visualization 1999 conference (VIS '99)
30. Roman M, Campbell RH (2000) Gaia: enabling active spaces. In: EW 9: proceedings of the 9th workshop on ACM SIGOPS European workshop, pp 229–234
31. Saini M, Singh V, Jain R, Kankanhalli M (2008) Multimodal observation systems. In: ACM multimedia conference

32. Sawchuk AA, Chew E, Zimmermann R, Papadopoulos C, Kyriakakis C (2003) From remote media immersion to distributed immersive performance. In: ETP '03: proc ACM SIGMM workshop on experiential telepresence, pp 110–120
33. Schooler E (1991) A distributed architecture for multimedia conference control. Technical report, University of Southern California
34. SecondLife (2009) SecondLife homepage. http://secondlife.com/
35. Sellen BBA, Arnott J (1992) Using spatial cues to improve videoconferencing. In: Proceedings of CHI'92, pp 651–652
36. Shadbolt N, Berners-Lee T, Hall W (2006) The semantic web revisited. IEEE Intell Syst 21(3):96–101
37. Sheppard R, Wu W, Yang Z, Nahrstedt K, Wymore L, Kurillo G, Bajcsy R, Mezur K (2007) New digital options in geographically distributed dance collaborations with teeve: tele-immersive environments for everybody. In: MULTIMEDIA '07: proc international conference on multimedia, pp 1085–1086
38. Sheth A, Henson C, Sahoo SS (2008) Semantic sensor web. IEEE Internet Computing 12(4): 78–83
39. Singh VK, Pirsiavash H, Rishabh I, Jain R (2008) Towards environment-to-environment (e2e) multimedia communication systems. In: 1st ACM international workshop on semantic ambient media experiences, pp 31–40
40. Stults R (1986) Media space. Technical report, Xerox PARC
41. Vertegaal R, der Veer GV, Vons H (2000) Effects of gaze on multiparty mediated communication. In: Proc graphics interface, pp 95–102
42. Vertegaal R, Weevers I, Sohn C, Cheung C (2003) 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, pp 521–528
43. Weiser M (1999) The computer for the 21st century. SIGMOBILE Mob. Comput Commun Rev 3(3):3–11
44. Westermann GU, Jain R (2006) Events in multimedia electronic chronicles (e-chronicles). Int J Semantic Web Inf Syst 2(2):1–27
45. Yang Z, Wu W, Nahrstedt K, Kurillo G, Bajcsy R (2007) Viewcast: view dissemination and management for multi-party 3d tele-immersive environments. In: MULTIMEDIA '07: proc international conference on multimedia, pp 882–891

Vivek K. Singh is a Phd Student in the department of Computer Science, (ICS) at University of California, Irvine. He earlier worked as a Research Assistant working in the area of "Coopetitive Visual Surveillance" at the National University of Singapore. From 2002–2006, Vivek was a Lecturer in Multimedia Technology Department at Institute of Technical Education (Macpherson), Singapore. Vivek obtained his B.Eng (Comp Eng.) and M.Computing (part-time) degrees from the National University of Singapore in 2002 and 2005 respectively. His research interests lie in active media sensing and application domains of video surveillance and experiential telepresence.

Hamed Pirsiavash received his B.Sc. degree in electrical engineering from Iran University of Science and Technology, Tehran, Iran in 2003 and his M.Sc. degree in electrical engineering from Sharif University of Technology, Tehran, Iran in 2006. He is currently working towards his Ph.D. degree in the Department of Computer Science (ICS) at the University of California Irvine under Prof Ramesh Jain. His research interests are event detection in multimedia information systems, computer vision, and machine learning.

Ish Rishabh is pursuing his doctoral degree at Bren School of ICS, University of California, Irvine. He completed his undergraduate studies in Electronics Engineering in 2002 from Institute of Technology, Banaras Hindu University, India. Before joining UCI, he worked for four years as a research scientist at Centre for Artificial Intelligence and Robotics (CAIR), Bangalore. His research interests include multimedia communication, computer vision and image processing. He is a member of Association for Computing Machinery (ACM).

Ramesh Jain joined University of California, Irvine as the first Bren Professor in Bren School of Information and Computer Sciences in 2005. Ramesh has been an active researcher in multimedia information systems, image databases, machine vision, and intelligent systems. While professor of computer science and engineering at the University of Michigan, Ann Arbor and the University of California, San Diego, he founded and directed artificial intelligence and visual computing labs. He was also the founding Editor-in-Chief of IEEE MultiMedia magazine and Machine Vision and Applications journal and serves on the editorial boards of several magazines in multimedia, business and image and vision processing. He has co-authored more than 250 research papers in well-respected journals and conference proceedings. Among his co-authored and co-edited books include Machine Vision, a textbook used at several universities. Ramesh has been elected Fellow of ACM, IEEE, IAPR, AAAI, and SPIE.