ITEM: Immersive Telepresence for Entertainment and Meetings with Commodity Setup *

Viet Anh Nguyen[1], Tien Dung Vu[1], Hongsheng Yang[1], Jiangbo Lu[1], Minh N. Do[2]
[1]Advanced Digital Sciences Center, Singapore
[2]University of Illinois at Urbana-Champaign, IL, USA
{vanguyen,johan.vu,hs.yang,jiangbo.lu}@adsc.com.sg, minhdo@illinois.edu

ABSTRACT

This paper presents an Immersive Telepresence system for Entertainment and Meetings (ITEM). The system aims to provide a radically new video communication experience by seamlessly merging participants into the same virtual space to allow a natural interaction among them and shared collaborative contents. With the goal to make a scalable, flexible system for various business solutions as well as easily accessible by massive consumers, we address the challenges in the whole pipeline of media processing, communication, and displaying in our design and realization of such a system. Extensive experiments show the developed system runs reliably and comfortably in real time with a minimal setup requirement (e.g., a webcam, a laptop/desktop connected to the public Internet) for tele-immersive video communication. With such a really minimal deployment requirement, we present a variety of interesting applications and user experiences created by ITEM.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Video; I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms

Design, Experimentation, Performance

Keywords

Video conferencing, tele-immersive system, video object cutout

1. INTRODUCTION

Face-to-face meeting has been necessary for effective communication, but time, financial cost, and also environmental concerns are becoming less tolerable. With the advent of network and multimedia technologies, virtual meeting has become increasingly popular to enable more frequent and less costly person-to-person communication. However, most existing virtual meeting systems in both business and consumer spaces hardly maintain the experience of an in-person meeting due to the poor integration of remote participants with the shared collaborative contents and virtual environments, leading to a limited nonverbal interaction among them.

To address these issues, high-end telepresence products such as Cisco TelePresence [1] are designed to create the perception of meeting in the same physical space, which demand a proprietary installation and high-cost setup. Recently, some 3D tele-immersive (TI) systems have been developed to enhance the remote collaboration by presenting remote participants into the same 3D virtual space [6, 8, 9]. However, these systems still fall short of stimulating a face-to-face collaboration with the presence of shared contents. Also, requiring bulky and expensive hardware with nontrivial calibration and setup hinders their wide deployment in real life.

Motivated by these challenges and more, we present in this paper an Immersive Telepresence system for Entertainment and Meetings (ITEM) based on a low-cost, flexible setup (e.g., a webcam and/or a depth camera, a desktop/laptop connected to the public Internet). The system allows putting two or more remote participants into the same virtual space and seamlessly integrating them with any shared collaborative contents for a more natural person-to-person interaction. With an addition of a depth camera, ITEM employs gesture-based controls to reproduce nonverbal signals and interactions with the shared contents in an intuitive and effective manner.

This work shall describe a complete design and realization of such a system by addressing the challenges in the whole pipeline of media processing, communication, and displaying. We consider major practical requirements in our design to build a system that supports *multimodality* (audio/video, shared contents), *scalability* for a large number of participants and concurrent meetings, *flexibility* in a system setup (2D color webcam and/or 3D depth camera), and desirable *functionality* to best suit different application contexts. Taking a systematic and integrated approach, we seamlessly combine various key components from the high-quality video object cutout, efficient media delivery, and immersive composition to improve system performance and enhance immersive experience. As an end result, we present several interesting applications and user experiences created by ITEM.

1.1 Related Work

Though 3D TI systems [6, 8, 9] provide the remote collaboration with immersive experience, they are usually very bulky with nontrivial setup requirements, bandwidth-demanding, and computationally intensive, which seriously limit their wide applications in daily practice. Alternatively, prior research has also been conducted to develop 2D TI systems (e.g., [2, 3]) based on commodity hardware, but the key foreground segmentation techniques adopted

*This study is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR).

MM'12, October 29–November 2, 2012, Nara, Japan.

are often very rudimentary [3] or computationally intensive [2] and provide a quality-compromised segmentation performance. Recently, Lu *et al.* [4] have developed a more advanced segmentation technique to realize a practical 2D TI system. Without supporting a flexible setup of incorporating a depth camera when available, it sometimes does not handle challenging situations well by using only a single webcam (e.g., background/foreground with very similar colors or difficult hand gestures). Furthermore, their system does not support multimodality and allows only limited interactions with shared contents, while the current design lacks capabilities to support many concurrent meeting sessions and flexibility in functionality extension in different application contexts.

2. ITEM SYSTEM

ITEM is designed not only to support basic requirements of a typical conferencing solution (e.g., high video/audio quality, ease of use), but also to provide compelling features for immersive experience. A modular design approach is used in the realization of this system to improve reusability, extensibility, and reconfigurability in various application contexts. Figure 1 depicts the simplified data flows and major components of an ITEM client.

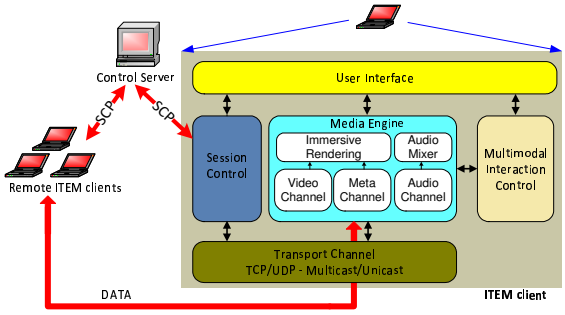

Figure 1: System overview.

2.1 Session control

The session control module manages the initialization and control of a communication session including both resource information (e.g., media capabilities, transmission paths) and process management (e.g., initiation, termination) by communicating with a control server through session control protocol (SCP). Without handling media data, the light-weight control server easily supports many concurrent sessions. We design a simplified set of SCP to facilitate a client interaction with a session such as *creation*, *termination*, *joining*, and *leaving*. Our design provides the reconfigurability for various application scenarios by creating and storing in a control server the media transmission structure, channel properties (e.g., unicast or multicast), and participants' media capabilities for each session. An appropriate connection is established for a client that opts to join a session either as a *passive participant* by only receiving media data or as an *active participant* by also sharing her media data. The control server is also used for new user registration and login authentication through a membership database.

2.2 Media engine

The role of media engine is to process both the local media prior to transmission and the incoming media from remote users for interactive playback. The engine provides seamless audio/video communication among users through a video/audio channel, while shared contents (e.g., documents, media-rich information) are processed through a meta channel. Regardless of channel types, video-audio contents are transmitted using RTP/RTCP protocol in order to support QoS (e.g., packet loss, jitter, bandwidth adaptation) and

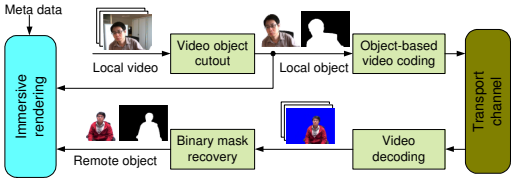

Figure 2: Block diagram and data flows of the video channel.

Figure 3: ITEM's real-time video object cutout technology using a normal webcam (top row) or a depth (plus RGB) camera (middle row). From left to right: system setup, input video frame, and cutout result. Bottom row shows some video composition effects.

audio/video synchronization. Basically, the meta channel inherits many similar modules from the audio/video channel to process shared audio/video contents, while other data types can also be easily handled through a secure protocol like TCP. To meet the low-latency requirement and to improve the system performance, multithreading techniques are employed to independently handle different channels and incoming data from each remote user.

Audio channel: Audio stream is processed in this channel to provide audio capability that is not supported in [4]. We use G.711 codec for audio compression, while a wide range of audio codecs is supported (e.g., G.729, Speex, etc.) for decoding the incoming audio streams from shared contents. A simple audio mixer is also provided to merge audio streams from multiple sources and synchronize with video streams using information from RTCP packets.

Video channel: This is the most critical module in ITEM that enables a variety of functional features for desired immersive experience by processing the user video on an object basis. Figure 2 shows the simplified processing flow of the video channel. Our ***video object cutout*** technology is first applied to segment the user object in real time from a live video captured by a single webcam or a depth camera such as Microsoft Kinect. Compared with the state-of-the-art segmentation techniques, our technology has shown a few important advantages that make ITEM competitive and practical: 1) reliable segmentation with high accuracy under challenging conditions, 2) real-time speed (18-25 FPS for VGA-resolution, 14-18 FPS for HD-resolution) on a commodity hardware such as a laptop/desktop, and 3) ease of use with little or no user intervention in the initialization phase. The technology has been registered as a trade secret recently and a detailed description is out of the scope of this paper. Basically, based on a unified optimization framework, our technology probabilistically fuses different cues together with spatial and temporal priors for accurate foreground layer segmentation. When a depth camera is available, the current framework can also be automatically configured to utilize the important depth information for more reliable inference, while leveraging several other key components also shared by the

webcam-based object cutout flow. Figure 3 shows some foreground segmentation results using different setups (e.g., with/without using a depth camera) under challenging test conditions.

For efficient delivery, ***object-based video coding*** is proposed to encode the foreground object stream using a chroma-key-based scheme with H.264 codec, where a chroma-key color is used as the background. We have developed a new, fast mode decision (MD) method to speed up the encoding process by considering the characteristics of real-life conferencing videos (e.g., containing sudden, complex motion such as hand gestures, face expression) to effectively eliminate unnecessary coding modes [5]. For the incoming object videos, a nonlinear neighborhood filter is used in the ***binary mask recovery*** to attain a clean segmentation map by removing speckle labelling noise due to video coding quantization artifacts.

Immersive rendering: This module is responsible for merging user objects from different sources with shared contents from the meta channel in an immersive and interactive manner (see Figure 3). For low-latency renderization, it is desired to refresh the composed frame upon receiving new data from any sources. With multiple asynchronous video streams, such a renderization strategy may overload the CPU usage due to a high rendering frame rate incurred. Thus, we use a master clock to update and render the composed frame at some frame rate (e.g., 30 FPS) without introducing any noticeable delay. For the ease of object manipulation, an object index map is used to indicate the object location in the composed frame for each media source.

2.3 Multimodal interaction control

For a more natural interaction with the shared contents, the cumbersome of using a keyboard and a mouse to interact with the system should be avoided whenever appropriate. With an addition of a depth camera, we employ hand gestures to interact with the system and provide users a comfortable lean back experience. Currently, we support several hand gestures to control the shared contents (e.g., paging through the slides). For more reliable tracking of hand motions, we have developed a fast hand gesture detection approach, which performs consistently better than the baseline OpenNI SDK, also in a more responsive manner.

2.4 Transport channel

Transport channel communicates with the session control and media engine modules to create an appropriate connection for data transmission in various channels based on the transmission architectures and data types. The module assigns and manages the list of destinations (e.g., a remote user address or a multicast address, if available). Real-time audio/video data is transmitted using UDP protocol in RTP/RTCP packetization. Meanwhile, SCP data and other types of shared data in the meta channel such as text are transmitted using TCP protocol for reliable transmission.

2.5 Networking and multiparty structure

To provide the scalability, our system design supports a mechanism to flexibly specify the transmission structure for media data during a session initialization using SCP. Currently, we support two architectures for data transmission among multiple users: 1) Decentralized ad-hoc structure for a small group meeting, 2) Multicast-based structure for one-to-many connections. In the decentralized ad-hoc structure, we use a node as a designated translator, which establishes P2P connections to other nodes. Each node in a session will only transmit data to the designated translator, which in turn relays the data back to all nodes. The design is simple, inexpensive compared with a centralized solution with a dedicated MCU, while it avoids the computing and bandwidth bottleneck with an increased number of concurrent sessions. Compared with a full-mesh connection in [2, 4] the uplink bandwidth at each node is significantly reduced and independent on the number of users, except for the translator node. Meanwhile, the multicast-based structure is used to support a large number of passive users (e.g., in e-learning), where overlay multicast techniques are employed, if necessary. The current design makes it easy to enhance and extend the networking capabilities in future.

Table 1: Execution speed (FPS)

webcam only		w/ depth camera	
Laptop	Desktop	Laptop	Desktop
17.1	18.4	25.4	25.9

(a) Video object cutout speed

Original video	Object cutout	
	w/o fast MD	w/ fast MD
16.5	41.2	50.7

(b) Object-based video coding speed

Table 2: Analysis of latency (ms)

Video object cutout	38-54
Video object encoding/decoding	24-38
Network (jitter, relay, etc.)	28-43
Rendering and display	12-30
End-to-end latency	102-165

2.6 System performance

We evaluated the performance of the key components and the whole system using a commodity setup (a laptop/desktop with four-core CPU running at 2.7 GHz), a single webcam, and a Kinect depth camera. Table 1 reports the execution speeds of video object cutout and coding. Without fully optimizing the code, our methods can perform comfortably at a real time speed. Compared with [4], we achieved about 41% and 23% faster execution speeds with our depth-based video object cutout and fast mode decision in object coding, respectively. To evaluate the overall system performance, we conducted multiparty conferencing over the public Internet using the decentralized ad-hoc structure. With the compressed video bitrate of 500 kbits/s, ITEM can easily support up to six participants within a session. The typical end-to-end lattency is reported in Table 2. We also measured the total CPU usage of about 35%-40% (about 15% for video object cutout, 10% for video coding/decoding and rendering, 10% for other tasks). With an increased number of participants in a session, we observed about 10%-15% increase in CPU workload (for ten connections over the LAN network) that is consumed by the decoding and rendering processes. The results show that our system has better scability performance compared with [2], where almost 100% CPU is utilized for only about three participants, leading to a significant low frame rate with an increased number of participants.

3. APPLICATIONS AND CASE STUDY

This section presents the utilization of ITEM to enhance the communication experience and effectiveness in various business and consumer solutions through an appropriate configuration and customization to meet the practical requirements.

3.1 Business group meeting

Stimulating in-person meeting characteristics (e.g., interaction among participants and collaborative contents with a sense of belonging to the same place) with scalability support is the critical element for effective tele-meeting. We discuss here some choices when configuring and customizing ITEM to realize such required

Figure 4: Multiparty immersive video communication for an effective business online meeting.

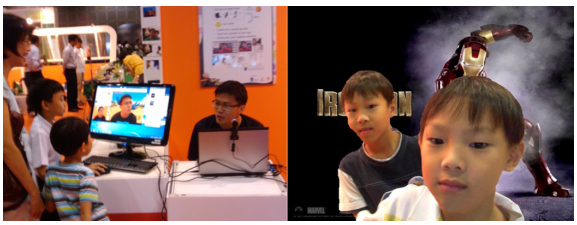

Figure 5: Demo at TechFest for children where the kids were excited and enjoyed themselves with our video chat application.

characteristics. For networking structure, we use the decentralized ad-hoc structure to: 1) easily support concurrent meeting scalability, and 2) reduce total bandwidth within a session. To support TI functionalities, we currently customize several rendering modes to naturally put participants in the same designed virtual meeting space or shared contents, allowing participants a freedom to navigate around (Figure 4). ITEM supports a variety of collaborative contents from slides, documents, media-rich contents, and even desktop windows through the meta channel. We have deployed our system for the internal trials and collected some initial useful feedbacks. Users like the virtual meeting room design that gives them a strong sense of presence in the same physical space without any distracting, private backgrounds. Although users are not fully satisfied with the current layout in the content sharing mode when having many remote participants, this mode is often preferred due to the need of sharing collaborative contents during a meeting and its effectiveness for conveying the gesture signals to the shared contents. It is observed that when there is a single participant at a location, the user prefers a simple setup of a webcam without the need of using a depth camera for the gesture-based control.

3.2 Distance learning and education

Fusion of users into the shared contents and virtual spaces in real time offers a variety of uses in distance learning and education. By merging a lecturer into the course materials, the gesture signals of a lecturer and her interactions with the slides are easily perceived as in a physical classroom, which is an important factor for an effective e-learning course. Inviting another lecturer from a different location to join and discuss during a live lecture is made possible by simply adding her as an active ITEM client. In a long-distance training course (e.g., yoga or rehabilitation), a proper rendering mode is designed to effectively put participants in a virtual class room, allowing a learner to easily see and follow the trainer's movement that is hardly feasible or effective with the existing commodity video communication systems. When natural interaction with the course contents or full-body tracking is required, a depth camera is recommended for the active ITEM participants. To support a large number of students (passive ITEM participants) in an e-course, multicast-based networking structure is employed. For the case of yoga or rehabilitation training or a small virtual class room requiring interactive collaborations, the decentralized ad-hoc structure is employed instead.

3.3 Entertainment

We have deployed ITEM as a light-weight tele-immersive video chat application to bring fun, exciting additions to the video chat experience. The system allows friends and long-distance family members to experience a sense of togetherness by creating a virtual space to let them see, interact, and do something fun together. Being able to separate an user from the background, the application lets users change the background, swap in another, and apply cool, fun video effects such as blurring the background or stylizing the user video (Figure 3 - bottom-right image). With a similar, simple setup as any video chat application in the consumer space, our application creates endless exciting uses with just one click from sharing vacation photos, covering up a messy room, pretending at some places, or hiding someone else who is watching and listening. We have also added a recording feature to let users easily create a fun, immersive video clip and share with friends and relatives on social networks. Creating a photo, video clip with all family members becomes easier than ever regardless of their distant locations.

We have demonstrated our application at various Technical Festivals (TechFest) (Figure 5) and conducted user experience surveys. The results show users like this new feature of instantly sharing something fun, exciting as the background while conducting video chats at the same time. They are impressed by the real-time segmentation of foreground at high quality from live videos, feeling that his/her webcam has been transformed into an intelligent one magically. Users also really enjoy the immersive video chat features, in which they feel more tightly connected with remote friends and selected background contents. Being simple and fun, the application attracts much attention from users (especially children) and gets them involved longer in a long-distance video chat. We find this observation is consistent with a recent HCI study [7].

4. CONCLUSION

In this paper, we have presented the complete design and realization of an immersive telepresence system for entertainment and meetings based on truly commodity capturing, computing, and networking setups. We addressed the challenges in the key components of the proposed ITEM system, but also exploited the synergy between different modules to maximize the overall system performance. Our future plan is to further improve the current system and related key technology to ultimately make the TI system practical and accessible to massive users. For this purpose, we also plan to conduct a more extensive evaluation of user experience.

5. REFERENCES

[1] Cisco TelePresence. http://cisco.com/telepresence.

[2] H. Baker *et al.* Understanding performance in Coliseum, an immersive videoconferencing system. *ACM TOMCCAP*, 2005.

[3] C. W. Lin *et al.* A standard-compliant virtual meeting system with active video object tracking. *EURASIP Journal on ASP*, 2002.

[4] J. Lu *et al.* CuteChat: A lightweight tele-immersive video chat system. In *Proc. of ACM Multimedia*, 2011.

[5] V. Nguyen *et al.* Efficient video compression methods for a lightweight tele-immersive video chat system. In *ISCAS*, 2012.

[6] D. E. Ott and K. Mayer-Patel. Coordinated multi-streaming for 3d tele-immersion. In *Proc. of ACM Multimedia*, 2004.

[7] H. Raffle *et al.* Hello, is Grandma there? Let's read! StoryVisit: Family video chat and connected e-books. In *Proc. CHI*, 2011.

[8] W. Wu *et al.* MobileTI: A portable tele-immersive system. In *Proc. of ACM Multimedia*, 2009.

[9] Z. Yang *et al.* TEEVE: The next generation architecture for tele-immersive environments. In *Proc. of 7th ISM*, 2005.