Viet Anh Nguyen, Jiangbo Lu, Shengkui Zhao,
Douglas L. Jones, and Minh N. Do

Teleimmersive Audio-Visual Communication Using Commodity Hardware

Natural human communication involves complex visual and audio behavior, and often context and joint interaction with the surrounding environment, to create a rich and satisfying experience. However, widely used virtual meeting systems such as WebEx and Skype still provide rather limited functionalities and hardly maintain the experience of an in-person meeting. In particular, traditional systems lack a sense of colocation and interaction as in a face-to-face meeting due to the separate displays of remote participants and poor integration with the shared collaborative contents. As a result, teleimmersive (TI) systems that aim to provide natural user experiences and interaction have attracted increasing research interest [1]. High-end telepresence products such as Cisco TelePresence or HP's Halo were expressly designed to create the perception of meeting in the same physical space. But to achieve such an experience, these systems require a proprietary installation and high setup costs. Recently, some three-dimensional (3-D) TI systems have been developed to enhance remote collaboration by merging remote participants into the same 3-D virtual space [2]–[4]. However, these systems still fall short of simulating a face-to-face collaboration with the presence of shared contents. Also, the required bulky and expensive hardware with nontrivial calibration and setup hinders their wide adoption. With the wide availability of low-cost, commodity computing devices with embedded video cameras, microphones, and ubiquitous Internet access adequate for real-time media, the dream of high-quality TI communication should finally be within our reach.

Achieving the dream of natural interactive communication at a distance with low-cost personal computing devices presents several new challenges that high-cost, dedicated systems can avoid. Since (unlike the Halo system with its dedicated, carefully purpose-designed physical environment, acquisition and display hardware, and dedicated network communication link) such a system may be used in any distracting environment and background, so a commodity telepresence system must detect and segment the user(s) from any type of background scene. It must also successfully and reliably do so with any common camera device in any variety of illumination conditions without prior calibration. Since the communication must utilize the Internet, it must be robust to network bandwidth variations, latency, and temporary dropouts, and the data must be efficiently compressed into a low bandwidth. We also desire to the ability to enable multiparty participation and interaction with virtual electronic objects such as presentations or imagery. A natural and high-quality audio experience is critical to achieve a sense of truly "being there," as well as to enable users to parse a complex scene to know who is speaking and when, so the audio should maintain a correct 3-D spatial rendering and remove background noise and interference. Finally, all of these elements must be seamlessly integrated in a system of low enough computational complexity to operate in real time on commodity hardware.

> THE DREAM OF HIGH-QUALITY TI COMMUNICATION SHOULD FINALLY BE WITHIN OUR REACH.

In this column, we present a real-time immersive telepresence system for entertainments and meetings (ITEM) based on a low-cost, flexible setup (e.g., a Webcam and/or a depth camera and a desktop/laptop connected to the Internet). The system puts remote participants into the same virtual space and seamlessly integrates them with any shared contents for a more natural person-to-person interaction. With the addition of a depth camera and a low-cost microphone array, ITEM supports spatialized 3-D audio, active speaker detection, and gesture-based controls to reproduce nonverbal signals and interactions with the shared contents in an intuitive and effective manner. The key points of ITEM are highlighted in Table 1 compared to existing TI solutions.

The remainder of this column describes a complete design and implementation of such a system by addressing the challenges and key components in the whole pipeline of media processing, communication, and display.

SYSTEM OVERVIEW

Our focus is on the system aspects in building such a lightweight practical TI system that maximizes the end-user experience, optimizes the system and network resources, and enables a variety of TI application scenarios. We consider major practical requirements in our design to build a system that supports

Digital Object Identifier 10.1109/MSP.2014.2340232
Date of publication: 15 October 2014

1053-5888/14©2014IEEE

[TABLE 1] A COMPARISON BETWEEN AN ITEM AND THE EXISTING VIDEO TELECONFERENCING SOLUTIONS.

SOLUTIONS	SETUP COST	HARDWARE	NETWORK	AUDIO/VIDEO QUALITY	QUALITY OF EXPERIENCE
HIGH-END TELEPRESENCE (CISCO, HP's HALO)	EXTREMELY HIGH	DEDICATED SETUP, PROPRIETY HARDWARE	DEDICATED BANDWIDTH	LIFE-SIZE VIDEO QUALITY, STUDIO ROOM QUALITY	IMMERSIVE ILLUSION, PERIPHERAL AWARENESS
NTII [2], TEEVE [3], BEING THERE [4]	HIGH	BULKY, EXPENSIVE 3-D CAMERA SETUP, BLUE SCREEN	INTERNET2 NETWORK	RELIABLE VIDEO CUTOUT, LOW 3-D VIDEO QUALITY, LOW FRAME RATE, STANDARD AUDIO	3-D IMMERSIVE RENDERING, BODY INTERACTION AND COLLABORATION
2-D TI SYSTEMS (VIRTUAL MEETING [5], COLISEUM [6], CUTE CHAT [7])	LOW	STANDARD PCs—AUDIO, VIDEO PERIPHERALS, MULTIPLE CAMERAS (IN COLISEUM)	INTERNET	UNRELIABLE VIDEO CUTOUT, LOW VIDEO RESOLUTION, LOW FRAME RATE, STANDARD AUDIO	IMMERSIVE DISCUSSION, WITHOUT SUPPORTING NONVERBAL SIGNALS/CUES AND COLLABORATION
STANDARD VIDEO CONFERENCING (SKYPE)	LOW	STANDARD PCs—AUDIO, VIDEO PERIPHERALS	INTERNET	MEDIUM-TO-HIGH VIDEO QUALITY, STANDARD AUDIO	NONIMMERSIVE, WITHOUT NONVERBAL COLLABORATION
ITEM	LOW	STANDARD PCs—AUDIO, VIDEO PERIPHERALS, DEPTH CAM (OPTIONAL), MICROPHONE ARRAY	INTERNET	ROBUST, RELIABLE CUTOUT, SUPPORT HD RESOLUTION HIGH FRAME RATE, SPATIAL AUDIO, SPEAKER DETECTION	IMMERSIVE, NATURAL CONVERSATION WITH NONVERBAL COLLABORATION

multimodality (audio/video, shared contents), scalability for a large number of participants and concurrent meetings, flexibility in a system setup [two-dimensional (2-D) color Webcam and/or 3-D depth camera], and desirable functionality to best suit different application contexts. As an end result, we present several interesting applications and user experiences created by ITEM. Figure 1(a) gives an overview of the ITEM system, where only the pair of a sender and a receiver is shown for simplicity. At the sender site, a commodity Webcam (or a depth camera) captures a live video stream, which is processed with our video object cutout technique to segment out object A in real time. The foreground object stream is then encoded using chroma key-based object coding prior to transmission over the Internet, reaching the receiver site under the management of an enhanced video delivery scheme. After decoding the video object, a clean segmentation map recovery method is applied to reconstruct a clean foreground segmentation map, which would otherwise contain boundary artifacts caused by compressing object video with background chroma keying. Finally, the user has a range of options on how to compose the final video to be displayed. He or she can choose to merge his/her own object video into the final frames, while the background (e.g., slides, photos) can be selected either from the local store, or streamed dynamically from the Internet as shared resources. In a group teleconferencing scenario, a low-cost compact microphone array is used to provide 3-D audio capture and reproduction as well as active speaker detection and tracking. The new communication experiences and compelling functionalities created by ITEM can be seen in a video on YouTube (http://youtu.be/cuRvXcLUIR4). For all technical details and quantitative evaluation of our technologies and the comparison with other existing approaches, we refer readers to a technical report [8].

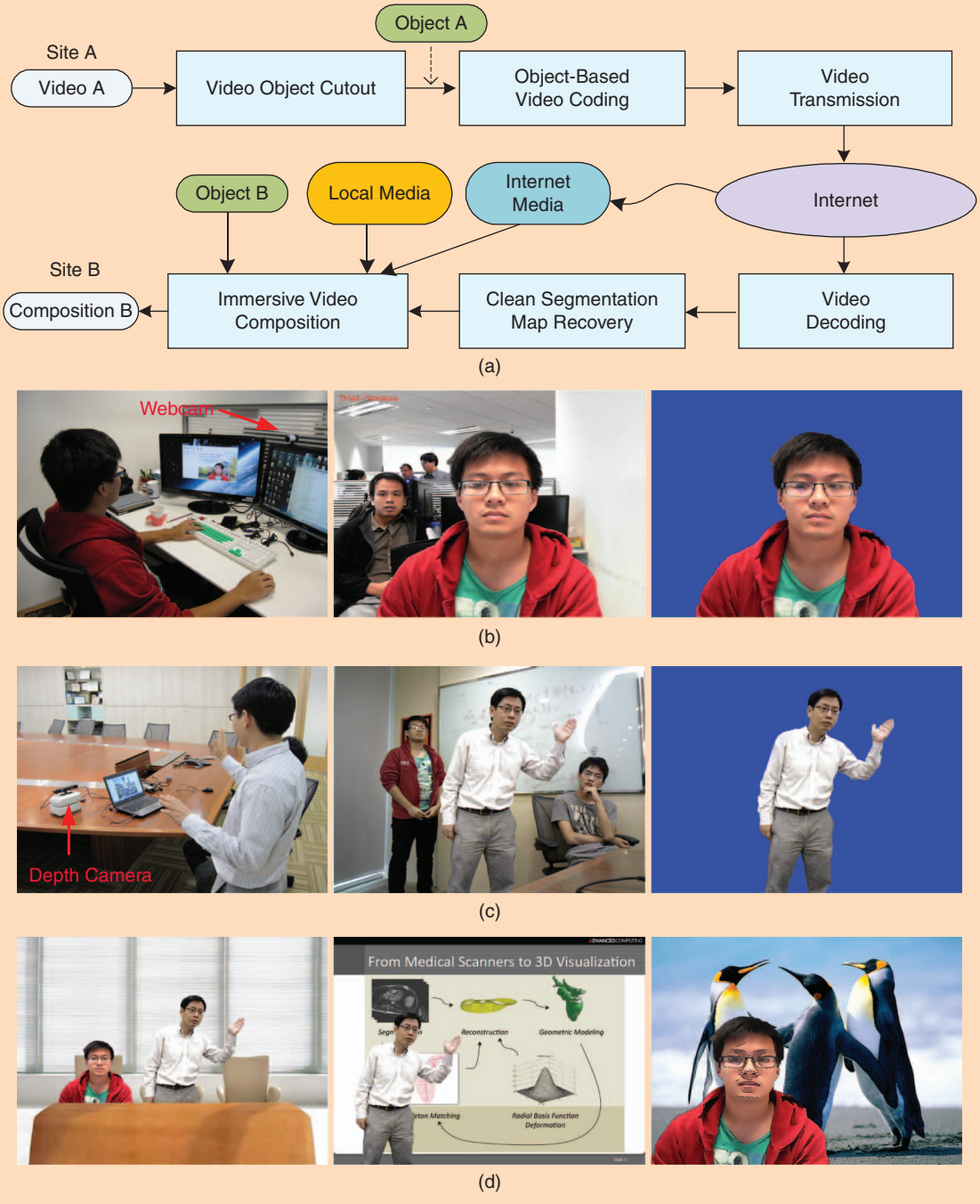
SEPARATE CODING AND DELIVERY OF MEANINGFUL FOREGROUND OBJECTS DECOMPOSED FROM CAPTURED SIGNALS IS ALSO CRUCIAL IN TI SYSTEMS.

VIDEO OBJECT CUTOUT

Video object cutout is essential to enable the immersive experience in the ITEM system. Assuming that the background is known and the Webcam is static, we have developed a practical solution for segmenting a foreground layer in real time from a live video captured by a single Webcam. Though this assumption appears somewhat constrained, a good solution can be widely deployed and it enables the aforementioned exciting systems with no additional cost. In fact, segmenting the foreground layer accurately from a complex scene where various changes can happen in the background is still rather challenging. Compared with state-of-the-art segmentation techniques, our technology has the following advantages: 1) reliable segmentation with high accuracy under challenging conditions, 2) real-time speed [18–25 frames per second (FPS) for video graphics array (VGA) resolution, 14–18 FPS for high-definition (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. Based on a unified optimization framework, our technology probabilistically fuses different cues together with spatial and temporal priors for accurate foreground segmentation. In particular, the proposed technology consists of two major steps, i.e., layer estimation and labeling refinement. 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 1(b) and (c) shows foreground segmentation results using different setups (e.g., with/without using a depth camera) under challenging test conditions.

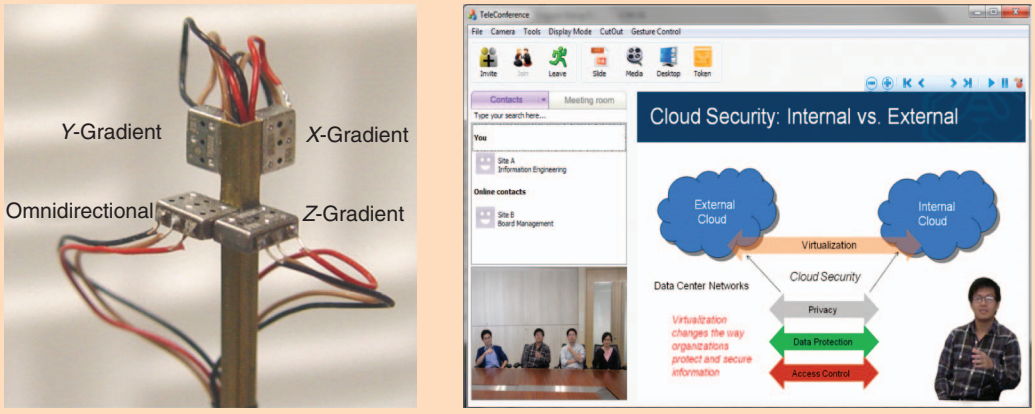
OBJECT-BASED CODING

Separate coding and delivery of meaningful foreground objects decomposed from captured signals is also crucial in TI systems.



[FIG1] An overview of the ITEM system (a) with the key real-time video object cutout technology using (b) a normal Webcam or (c) a depth (plus RGB) camera. From left to right: system setup, input video frame, and cutout result. Some video composition effects are shown in (d).

This not only facilitates object-based processing such as immersive rendering in a shared environment but also eases the network traffic by discarding irrelevant background information. In ITEM, we use an efficient object-based video coding using a chroma-key-based scheme with a high coding efficiency H.264 codec, where a chroma-key color is used as the background [9]. A new, fast-mode decision method speeds up the encoding process by considering the characteristics of real-life conferencing videos (e.g., containing sudden, complex motion such as hand gestures, facial expressions) to eliminate unnecessary coding modes, which reduces both the bandwidth and the computational requirements by factors of three to four [7],



[FIG2] Speaker detection based on visual cue and localization using our (a) low-cost compact microphone array to enhance the communication experience by immersively rendering the active speaker with (b) the shared background.

[9]. This is important to reduce the complexity of highly computational H.264 encoder, leaving more central processing unit (CPU) resources for other tasks. 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.

3-D SOUND CAPTURE, RENDER, LOCALIZATION

Multiparty TI systems are greatly enhanced by spatial sound and the detection of speakers, especially with multiple participants at a single site. Unlike the existing systems requiring a large spatially separated microphone array, we built a very compact microphone array [10], where four collocated miniature microphones approximate an acoustic vector sensor (AVS) [Figure 2(a)]. This AVS array consists of three orthogonally mounted acoustic particle velocity gradient microphones X, Y, and Z and one omnidirectional acoustic pressure microphone O, which is referred to as the XYZO array. Gradient microphones provide additional spatial acoustic information in the amplitude as well as the time difference compared to standard microphone arrays. As a result, the XYZO array offers better performance in a much smaller size. In this system, we, for the first time, deploy and evaluate the XYZO array for the 3-D capture and sound source localization (SSL) Our 3-D audio capture and reproduction utilizes beamforming techniques to reconstruct each beam through the corresponding head-related transfer function (HRTF) to emulate human sound localization based on the filtering effects of the human ear. Meanwhile, 3-D SSL is based on a frequency-domain 3-D spatial search to estimate direction of arrival (DOA). Interested readers can refer to [10] and [11] for more details.

In group teleconferencing, our system not only supports 3-D audio perception but also active speaker detection. With the addition of a depth sensor, the visual content of the active speaker can be accurately tracked and segmented by fusing both audio and visual cues. This will enable a compelling and effective communication experience by immersively rendering the active speaker with the shared contents [see Figure 2 (b)] [12].

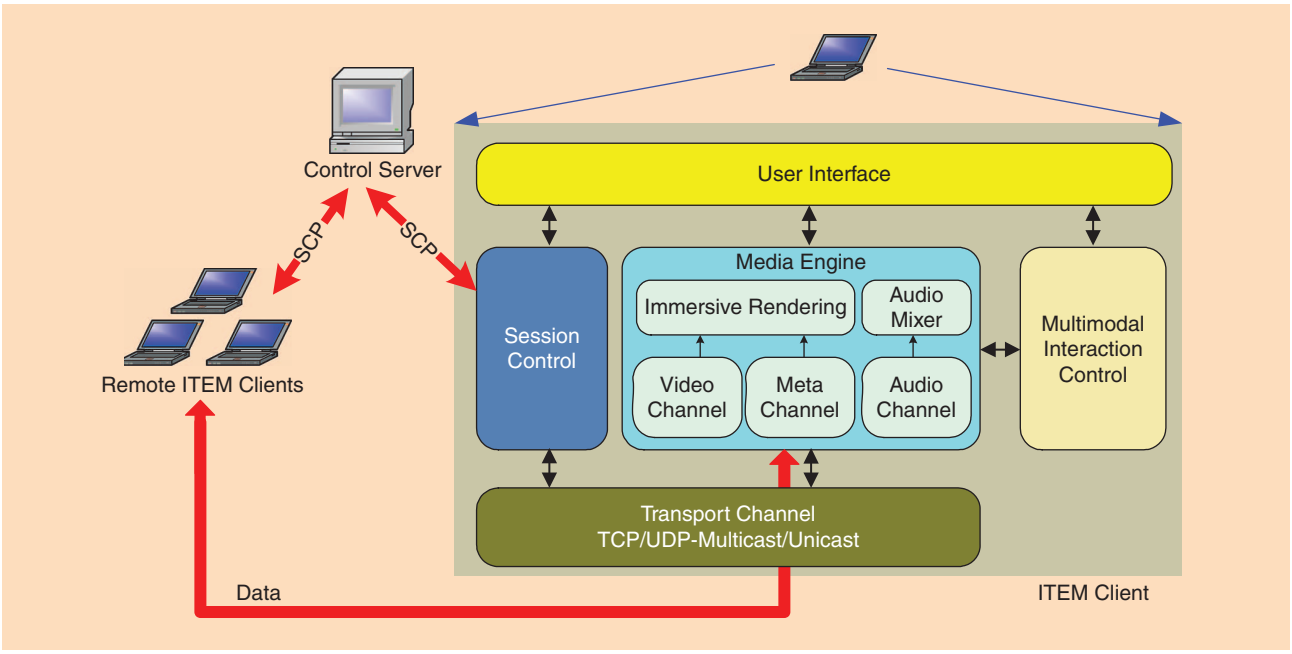
MULTIPARTY NETWORKING STRUCTURE

Media data between two clients in ITEM are exchanged in a peer-to-peer (P2P) manner. To provide the scalability, our system design supports a mechanism to flexibly specify the transmission structure for media data during a session initialization. Currently, we support two architectures for data transmission among multiple users: 1) decentralized ad hoc structure for a small group meeting, and 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 and inexpensive compared with a centralized solution with a dedicated multipoint control unit (MCU), while it avoids the computing and bandwidth bottleneck with an increased number of concurrent sessions. Compared with a full-mesh connection, the uplink bandwidth at each node is significantly reduced and independent of 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 the future.

SYSTEM DESIGN AND IMPLEMENTATION

A modular design approach is used to improve reusability, extensibility, and reconfigurability in various application contexts. Figure 3 depicts the simplified data flows and major components of an ITEM client.

The session control manages the initialization and control of a communication



[FIG3] The simplified diagram of the key components of an ITEM client and the main data flow in the ITEM system architecture.

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).

The role of the media engine is to process both the local media prior to transmission and the incoming media from remote users for immersive composition. 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. Object-based video processing (e.g., video cutout and object coding) are processed within video channel while audio channel accommodates 3-D sound processing and SSL for spatialized audio and speaker detection. 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. The segmented user objects from different sources are merged with shared contents from the metachannel in an immersive and interactive manner [Figure 1(d)] through the immersive rendering module. While refreshing the composed frame upon receiving new data from any source is desired for low-latency rendering, such a rendering strategy will overload the CPU usage due to a high rendering frame rate incurred. Thus, a master clock is used to update and render the composed frame at some frame rate (e.g., 30 FPS) without introducing any noticeable delay.

For a more natural interaction with the shared contents, the use of a keyboard and a mouse to interact with the system should be avoided whenever appropriate. With the addition of a depth camera, we employ hand gestures to interact with the system and provide users a comfortable experience through the multimodal interaction control module. Currently, we support several hand gestures to control the shared contents (e.g., paging through the slides or changing the virtual room background).

The 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 real-time transport protocol (RTP) in conjunction with real-time transport control protocol (RTCP) built on top of user datagram protocol (UDP) packetization.

[TABLE 2] AN ANALYSIS OF LATENCY (MS).

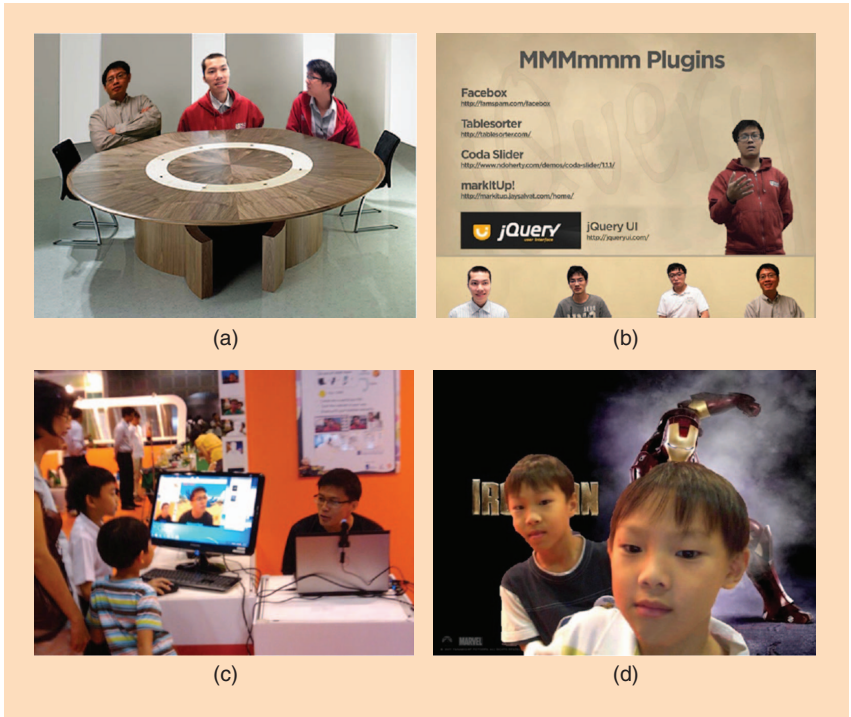
VIDEO OBJECT CUTOUT	38–54
VIDEO OBJECT ENCODING	24–38
NETWORK (JITTER, RELAY, ETC.)	28–43
RENDERING AND DISPLAY	12–30
END-TO-END LATENCY	102–165

SYSTEM PERFORMANCE

We deployed ITEM to conduct multiparty conferencing over the Internet using the decentralized ad hoc structure for the overall system performance evaluation. With the compressed video bit rate of 500 kbits/s, ITEM can easily support up to six participants within a session and run comfortably at a real-time speed. The typical end-to-end latency 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 a 10–15% increase in CPU workload [for ten connections over the local area network (LAN) that is consumed by

the decoding and rendering processes]. The results show that our system has better scalability compared with [4], where almost 100% of the CPU is utilized for only about three participants, leading to a significantly lower frame rate with more participants.

In the case of group conferencing with multiple participants at a site, we observed that the 3-D SSL module could detect the angle of a speaker with an accuracy of 6° in real time. Together with the user-tracking information obtained through a depth sensor, this always led to accurate active speaker identification and his/her correct video content extraction. It is worth mentioning that the speaker's video switched too often whenever there was a noise sound source within a short period, which was distracting. To eliminate this undesired effect, we used a threshold to verify the 3-D SSL new detection output. The system only switches to the new speaker video when a consistent detection output is presented within a certain period.



[FIG4] **(a)–(d) The multiparty immersive video communication for an effective business online meeting and fun video chat is shown.**

USER EXPERIENCE

To validate the system from a user perspective, we customized and deployed ITEM in a variety of application scenarios such as business meetings, group teleconferencing, and video chats to conduct extensive user studies.

For business meeting scenarios, we designed several rendering modes to naturally put participants in the same designed virtual meeting space or shared contents, allowing participants the freedom to navigate around the virtual background (Figure 4). The system supports a variety of collaborative contents from slides, documents, media-rich contents, and even desktop windows through the metachannel. We deployed our system for internal trials and collected some useful initial feedback. Users liked the virtual meeting room design, which gave them a strong sense of presence in the same physical space without any distracting, private backgrounds. Although users were not fully satisfied with the current layout in the content-sharing mode when there were many remote participants, this mode was often preferred due to the need of sharing collaborative content during a meeting and its effectiveness for conveying the gesture signals to the shared contents. It was observed that when there was a single participant at a location, the user preferred a simple setup of a Webcam without using a depth camera for gesture-based control.

In the case of group teleconferencing, we employed spatial audio and active speaker detection. Users liked the immersive visual composition of the active speaker into the shared presentation slides. They felt the quality of the meeting was more effective due to how simple it was to keep track of both the slide contents and the speaker video at the same time. Meanwhile, in discussions without any shared contents, users felt that immersive audio could assist them in quickly identifying who was talking and perceiving the direct conversation flow.

In addition to teleconferencing solutions for the business domain, ITEM was also deployed in the consumer space as a lightweight TI video chat application to bring fun, exciting additions to the video chat experience. The system allowed friends and distant 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 users from the background, the application let users apply video effects such as blurring the background or stylizing the user video (see right-most image in Figure 1(d)]. We have demonstrated our application at various technical festivals and conducted user experience surveys (Figure 4). Users liked this new feature of instantly sharing something as fun and exciting as the background while conducting video chats at the same time. They were impressed by the high quality of the real-time segmentation of the foreground from live videos, and felt that his/her Webcam had been transformed into one that was intelligent. Users also enjoyed the immersive video chat features, where they felt more tightly connected with remote friends.

ACKNOWLEDGMENTS

This study is supported by the research grant for the Human Sixth Sense Programme at

(continued on page 136)

crucial for the improvement of existing applications and the search for novel procedures for diagnosing and treating effectively life-threatening diseases. Therefore, further research in this field is encouraged.

ACKNOWLEDGMENT

We acknowledge financial support from the Research Council of Norway given to the MELODY Project (Phase II) under contract number 225885.

AUTHORS

Raúl Chávez-Santiago (raul.chavez-santiago@rr-research.no) is a researcher at the Intervention Centre, Oslo University Hospital, Norway.

Ilangko Balasingham (ilangko.balasingham@medisin.uio.no) is the head of the Biomedical Sensor Network Research Group at the Intervention Centre, Oslo University Hospital, Norway, and professor in the Department of Electronics and Telecommunications, Norwegian University of Science and Technology, Trondheim.

REFERENCES

[1] C. N. Paulson, J. T. Chang, C. E. Romero, J. Watson, J. F. Pearce, and N. Levin, "Ultra-wideband radar methods and techniques of medical sensing and imaging," in *Proc. SPIE Int. Symp. Optics East*, Boston, MA, 2005, pp. 96–107.

[2] E. Zastrow, S. K. Davis, and S. C. Hagness, "Safety assessment of breast cancer detection via ultrawideband microwave radar operating in pulsed-radiation mode," *Microw. Opt. Technol. Lett.*, vol. 49, no. 1, pp. 221–225, 2007.

[3] A. Daisuke, K. Katsu, R. Chávez-Santiago, Q. Wang, D. Plettemeier, J. Wang, and I. Balasingham, "Experimental evaluation of implant UWB-IR transmission with living animal for body area networks," *IEEE Trans. Microw. Theory Techn.*, vol. 62, no. 1, pp. 183–192, 2014.

[4] R. Chávez-Santiago, and I. Balasingham, "The ultra wideband capsule endoscope," in *Proc. IEEE Int. Conf. Ultra-Wideband (ICUWB)*, Sydney, Australia, 2013, pp. 72–78.

[5] M. S. Chae, Z. Yang, M. R. Yuce, L. Hoang, and W. Liu, "A 128-channel 6 mW wireless neural recording IC with spike feature extraction and UWB transmitter," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 17, no. 4, pp. 312–321, 2009.

[6] B. Moussakhani, "On localization and tracking for wireless capsule endoscopy," Ph.D. dissertation, Dept. Electronics and Telecommunications, Norwegian Univ. Sci. and Technol., Trondheim, Norway, 2013.

[7] S. N. Pavlov, and S. V. Samkov, "Algorithm of signal processing in ultra-wideband radar designed for remote measuring parameters of patient's cardiac activity," in *Proc. 2nd Int. Workshop Ultra Wideband Ultra Short Impulse Signals*, Sevastopol, Ukraine, 2004, pp. 205–207.

[8] L. E. Solberg, S.-E. Hamran, T. Berger, and I. Balasingham, "Minimum variance signal selection for aorta radius estimation using radar," *EURASIP J. Adv. Signal Process.*, vol. 2010, no. 49, pp. 1–13, 2010.

[9] S. A. Alshehri, S. Khatun, A. B. Jantan, R. S. A. R. Abdullah, R. Mahmood, and Z. Awang, "Experimental breast tumor detection using NN-based UWB imaging," *Prog. Electromagn. Res. (PIER)*, vol. 111, pp. 447–465, 2011.

[10] S. Brovoll, T. Berger, Y. Paichard, Ø. Aardal, T. S. Lande, and S.-E. Hamran, "Time-lapse imaging of human heartbeats using UWB radar," in *Proc. IEEE Biomedical Circuits Systems Conf. (BioCAS)*, Rotterdam, The Netherlands, 2013, pp. 142–145.

[11] M. Converse, E. J. Bond, B. D. Van Veen, and S. C. Hagness, "A computational study of ultra-wideband versus narrowband microwave hyperthermia for breast cancer treatment," *IEEE Trans. Microw. Theory Techn.*, vol. 54, no. 5, pp. 2169–2180, 2006.

[12] H. D. Trefná, J. Vrba, and M. Persson, "Time-reversal focusing in microwave hyperthermia for deep-seated tumors," *Phys. Med. Biol.*, vol. 55, no. 8, pp. 2167–2185, 2010.

SP

applications CORNER (continued from page 123)

the Advanced Digital Sciences Center (ADSC) from Singapore's Agency for Science, Technology and Research (A*STAR). We thank Tien Dung Vu, a software engineer at ADSC, for his assistance in implementing our ITEM prototype system.

AUTHORS

Viet Anh Nguyen (vanguyen@adsc.com.sg) is a researcher at the Advanced Digital Sciences Center, Illinois at Singapore.

Jiangbo Lu (jiangbo.lu@adsc.com.sg) is a researcher at the Advanced Digital Sciences Center, Illinois at Singapore.

Shengkui Zhao (shengkui.zhao@adsc.com.sg) is a researcher at the Advanced Digital Sciences Center, Illinois at Singapore.

Douglas L. Jones (dl-jones@illinois.edu) is a professor at the University of Illinois at Urbana-Champaign.

Minh N. Do (minhdo@illinois.edu) is a professor at the University of Illinois at Urbana-Champaign.

REFERENCES

[1] J. G. Apostolopoulos, P. A. Chou, B. Culbertson, T. Kalker, M. D. Trott, and S. Wee, "The road to immersive communication," *Proc. IEEE*, vol. 100, no. 4, pp. 974–990, 2012.

[2] D. E. Ott and K. Mayer-Patel, "Coordinated multistreaming for 3D tele immersion," in *Proc. ACM Multimedia*, 2004, pp. 596–603.

[3] Z. Yang, K. Nahrstedt, C. Yi, B. Yu, J. Liang, S.-H. Jung, and R. Bajscy, "TEEVE: The next generation architecture for tele-immersive environments," in *Proc. 7th IEEE Int. Symp. Multimedia*, 2005, pp. 112–119.

[4] C. Kuster, N. Ranieri, Agustina, H. Zimmer, J. C. Bazin, C. Sun, T. Popa, and M. Gross, "Towards next generation 3D teleconferencing systems," in *Proc. 3DTV-Conf.: True Vision-Capture, Transmission, Display 3D Video (3DTV-CON)*, 2012.

[5] C. W. Lin, Y. J. Chang, C. M. Wang, Y. C. Chen, and M.-T. Sun, "A standard-compliant virtual meeting system with active video object tracking," *EURASIP J. Adv. Signal Process.*, vol. 2002, no. 6, pp. 622–634, 2002.

[6] H. Baker, N. Bhatti, D. Tanguay, I. Sobel, D. Gelb, M. Goss, B. Culbertson, and T. Malzbender, "Understanding performance in coliseum, an immersive videoconferencing system," *ACM Trans. Multimedia Computing Commun. Appl. (TOMCCAP)*, vol. 1, no. 2, pp. 190–210, 2005.

[7] J. Lu, V. A. Nguyen, Z. Niu, B. Singh, Z. Luo, and M. N. Do, "CuteChat: A lightweight tele-immersive video chat system," in *Proc. ACM Multimedia*, 2011, pp. 1309–1312.

[8] V. A. Nguyen, J. Lu, S. Zhao, T. D. Vu, H. Yang, D. L. Jones, and M. N. Do, "ITEM: Immersive telepresence for entertainment and meetings—A practical approach," Tech Rep., Advanced Digital Sciences Center, Aug. 2014, arXiv:1408.0605.

[9] V. A. Nguyen, J. Lu, and M. N. Do, "Efficient video compression methods for a lightweight tele-immersive video chat system," in *Proc. IEEE Int. Symp. Circuits Systems (ISCAS)*, 2012, pp. 149–152.

[10] S. Zhao, A. Ahmed, Y. Liang, K. Rupnow, D. Chen, and D. L. Jones, "A real-time 3D sound localization system with miniature microphone array for virtual reality," in *Proc. IEEE Industrial Electronics Applications (ICIEA)*, 2012, pp. 1853–1857.

[11] S. Zhao, R. Rogowski, R. Johnson, and D. L. Jones, "3D binaural audio capture and reproduction using a miniature microphone array," in *Proc. 15th Int. Conf. Digital Audio Effects (DAFx)*, 2012, pp. 151–154.

[12] V. A. Nguyen, S. Zhao, T. D. Vu, D. L. Jones, and M. N. Do, "Spatialized audio multiparty teleconferencing with commodity miniature microphone array," in *Proc. ACM Multimedia*, 2013, pp. 553–556.

SP