Development of Remote Support Service by Augmented Reality Videophone

Atsushi Fukayama, Shunsuke Takamiya, Junichi Nakagawa, Shinyo Muto, and Naoki Uchida
NTT Service Evolution Laboratories
Nippon Telegraph and Telephone Corporation
1-1, Hikari-no-oka, Yokosuka-shi, Kanagawa, Japan
{fukayama.atsushi, takamiya.shunsuke, nakagawa.junichi, muto.shinyo, uchida.naoki}@lab.ntt.co.jp

***Abstract*—This paper presents the development of a remote support service based on an existing videophone system enhanced with AR (Augmented Reality) technology. The service enables a support person to point to a real object in a remote site by overlaying a virtual object, which is named as "air stamp," onto video image. Besides, the air stamp stays on the object it was attached to at first based on image-based AR technology, even when a camera of mobile device in the remote site is moved. These features make the communication between the support person and the onsite worker much more efficient. Video latency and effective frame rate of the videophone image were compared between the original videophone system and AR-enabled system, which indicated the addition of the AR function did not affect the performance significantly.**

Keywords-videophone; augmented reality; network value-added service; remote support; telepointing.

I. INTRODUCTION

The recent spread of fixed and mobile broadband networks and dissemination of smart telecommunication devices facilitates the people's easy access to videophone services. These conditions could launch a rapid rise of videophone service use, finally after communication service providers' long struggle for it. But, it seems to be a little further to come.

One possible way to encourage the videophone use is offering new use cases by extending basic videophone functionality. Ordinary videophone services just transfer live audio and video media back and forth as it is captured. But, recent improvement of media processing technologies [1][2] allows real-time and real-world media processing which is needed to enhance plain videophone service.

For instance, an acoustic speech recognition technology estimates "who speaks when and what" based on audio media in real time in natural multi-party meeting configuration [1]. A face recognition technology can continuously detects registered person's face in live video image under various types of situation change such as face posture and illumination [2].

By combining with the wide variety of media processing technologies, a single videophone service can match wide variety of use cases. But, each use case is not as general as ones covered by traditional basic videophone services. This means that the expansion of existing videophone service must be realized in low cost.

Against these issues, we propose a service architecture to enhance existing videophone service with Augmented Reality (AR) technology [3] that detects real objects in videophone image and overlays related virtual objects on the real objects. This architecture facilitates the low-cost videophone service expansion by enabling reuse of their existing asset such as Multi-point Connecting Unit (MCU) and client software on terminal devices [4].

As an application of our AR videophone service architecture, we developed a remote support service by which an instructor can place a virtual object, which we call as "air stamp," onto a real object being shot in video image. The air stamp can be used to convey what the instructor talked about to a remote onsite worker. Based on image recognition technology, the air stamp stays on the object it firstly attached to even when a camera of mobile device that captures the remote site scene is moved, which permits free movement of the onsite worker who is holding the device.

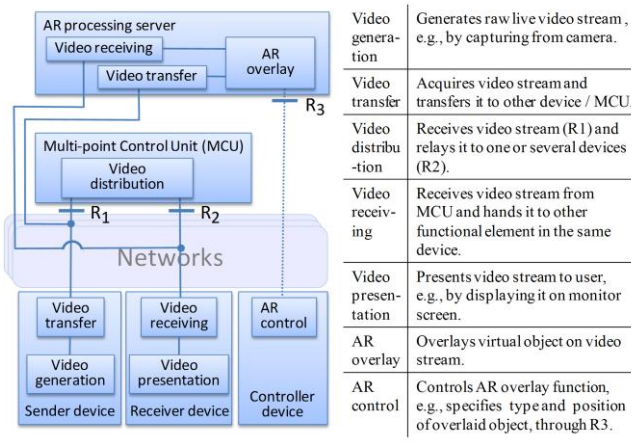
In this report, after introducing related works in Section II, design of the remote support service based on collected requirements is discussed in Section III. Performance evaluation result is shown in Section IV to confirm that AR function added on base videophone system does not degrade system performance.

II. RELATED WORK

An example of the efforts the telecommunication industry has been making to realize media-enhanced network value-added services was recently seen in the Rich Communications Services (RCS) program enacted by the GSM Association (GSMA). In the program, they are exploring enriched communication services other than mere voice communication and trying to standardize technical realization [5]. Proposed examples of network value-added services include enriched content sharing, in which pictures shared between RCS-enabled devices are converted in the IMS network, and enriched chat with text conversion services such as language translation. The Telecommunication Technology Committee (TTC) is discussing technical realization with respect to network value-added media services. It has proposed an architecture in which an Application Server (AS) for network value-added services intervenes the media flow between a device and other AS dedicated for existing services [6].

Intelligent media processing services combined with telecommunication services are about to appear from the telecom industry. For example, NTT DoCoMo announced and demonstrated an automatic live translation service available during mobile telephone conversation, which

 ISBN: 978-1-61208-239-4



Video genera-tion	Generates raw live video stream, e.g., by capturing from camera.
Video transfer	Acquires video stream and transfers it to other device / MCU.
Video distribu-tion	Receives video stream (R1) and relays it to one or several devices (R2).
Video receiv-ing	Receives video stream from MCU and hands it to other functional element in the same device.
Video presen-tation	Presents video stream to user, e.g., by displaying it on monitor screen.
AR overlay	Overlays virtual object on video stream.
AR control	Controls AR overlay function, e.g., specifies type and position of overlaid object, through R3.

Figure 1. Basic architecture of AR videophone service.

performs speech recognition, translation, and synthesis on a network cloud [7]. This prototype service enhances the audio medium of voice telephone in a network. Our study proposes an enhancement of visual media in network cloud.

As one of many visual media processing technologies, AR has been the subject of a huge number of references in the literature, including some that applied AR to teleconferencing [8][9] and remote collaboration scenarios [10]. Historically, many AR-based conferencing and collaboration studies have employed specialized devices such as head mounted displays (HMDs). Against the background of recent trends in commoditized Internet video-chat and smart devices, an attempt was made to combine a client-based AR system with a video-chat system [9]. Defining our study from this viewpoint, we applied a server-based AR technology to existing managed videophone service.

In the Internet service domain, we can find some communication services that incorporate media processing technologies in the middle of communication channels. Google Talk enhancement with language translation functionality seems to match with our concept [11]. The translation functionality in the text chat system is implemented as translation "bot," which is apparently same as human chat participants but actually an autonomous agent automatically responding to others. The translation bot listens to other participant's words, translates the words, and utters the translated words. Addition of this translation functionality does not affect the base text-chat system at all. It can lower deployment cost for the service provider and entry barrier for users.

III. SYSTEM DESIGN

A. Basic Architecture

Figure 1 depicts the architecture of AR videophone service we proposed [4]. Based on this architecture, the AR processing server connects with other system components in the same way as normal terminal device. This can minimize system modification for inserting the AR functionality in the middle of video image transfer channel.

The architecture comprises three types of terminal devices, a Multi−point Control Unit (MCU), and an AR processing server. Although devices are categorized into sender, receiver, and controller in terms of device function, one implemented device can belong to more than one functional category. For instance, ordinary videophone device has both functions of sender and receiver.

B. Service Concept

Figure 2 shows core functionality of our remote support service based on the AR videophone architecture [4]. The target service of this system is remote support, where instructors in a support center give instructions to support their customers or onsite workers. At the remote site, a video image is taken using a mobile device, e.g., smart-phone and tablet, and shared with the instructor. The instructor can put virtual objects, i.e., air stamps, onto actual objects seen in the video so that the user in the remote site can understand what the instructor is talking about. The stamps must be stuck to the actual objects because the remote user often moves the mobile camera.

In the previous report [4], we already proposed the concept of this remote support service and presented its prototype. Development of a practical system following the prototype is introduced in this report.

C. Requirements

By using the prototype visualizing our service concept, we interviewed and derived requirements to the remote support service from people working in related domains such as Information Technology (IT) maintenance and manufacturing as well as customer support.

1) Number of devices: Many instructors should be involved in some cases such as failure analysis of complicated system. Therefore, several instructor devices should be able to join one session. Some installation works of networked systems require cooperation between onsite workers in different sites. To support this type of work, two workers should be able to join one same session and show situation in their working sites to instructors in turn.

2) Media quality: In most remote support use cases, high frame rate is not required because the objects to work on are static. Video image resolution that is enough for instructors in remote support sites to recognize printed

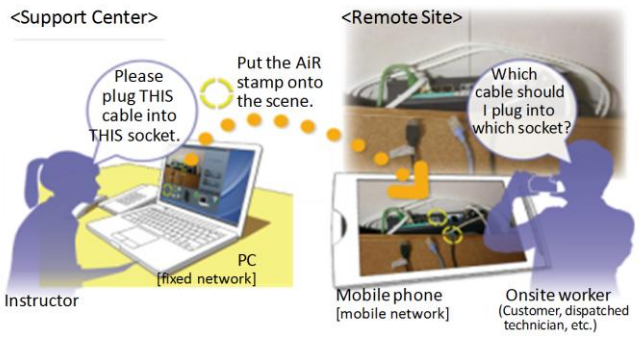


Figure 2. Remote support service based on AR videophone.

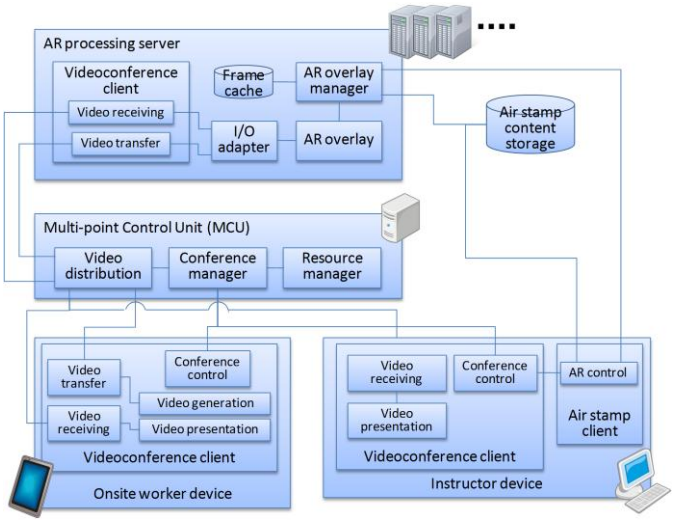


Figure 3. Developed remote support system diagram.

letters on working object is required. As it is assumed that onsite worker uses mobile device (Figure 2), it would be sufficient if the letters can be read when they are captured in close-up shot. Clear voice is required for efficient remote support communication. Though the air stamps can facilitate smooth communication, it would play a complementary role with verbal instruction.

3) Onsite network environment: Wide range of network access for the onsite mobile device is required. For working sites in outside and third-party's premise, 3G / LTE mobile access is required. In one's own premise, WiFi is sufficient. Wired access is required when use of wireless communication is forbidden for security reason or to avoid possible harm of radio wave on electronic equipment.

4) Additional functions: Many additional functions supplementary with the core air-stamp functionality were found. Recording of audio / video stream, document sharing, and text messaging were common requirements.

D. Implementation

Figure 3 shows system diagram of developed remote support system.

We deployed a commercial videoconferencing system for MCU and video generation / transfer / receiving / presentation functions shown in Figure 1. Analyzing the requirements about number of devices and additional functions, we found that many high-end or middle-class videoconferencing systems already meet those requirements.

Figure 4. User interface on instructor device.

Considering the media quality requirements and onsite network environment, we configured video resolution to VGA (640x480 pixels), which a preliminary test indicated was the bottom line to allow users to read letters in the video image. It was expected that VGA would spend too much bandwidth for applying to 3G mobile access with standard frame rate such as 15 or 30 frames per second (fps). In that case, we can adjust the frame rate to lower values.

The videoconference client on the onsite-worker and instructor devices is the standard client software of the deployed videoconferencing system. Figure 4 shows user interface on the instructor computer, which consists of the standard videoconference client on the right side and an air stamp client on the left side.

When a user sets up a normal videoconference, the user requests the conference manager function through the conference control function in the client software. The conference manager function controls the video distribution function to make a call to devices of conference participants including the requesting party.

In the case the user requests an AR-enabled videoconference, the conference manager inquires available AR processing server to the resource manager function and obtain its address. Then, a conference including the participants and the available AR server is established.

After a videoconference is set up, raw video stream is transferred to the AR processing server through the MCU. Raw video frames decoded by the client software in the server are input to the AR overlay function through the Input / Output (I/O) adapter function which is needed when I/O interface of deployed AR technology and videoconference client does not match.

The AR overlay function searches specified reference images in the input video image and overlay an air stamp on each discovered image. Because this processing is quite general among image-based AR technologies, various AR technologies can be used for the AR overlay function.

The AR overlay manager function receives information about instructor's operation to put an air stamp on a sub-region in the video image from the AR control function. The reference image is extracted from raw input video image based on the sub-region information, which is passed to the AR overlay function together with the air stamp image.

Video transfer delay prevents the AR overlay function from making the reference image from the right frame of the right timing. The isolation between the base videoconferencing system and AR functionality added on the system makes it difficult to use timestamp information to compensate the delay, though it brings the reusability of the base system. To solve this problem, the AR overlay function caches raw input frames during the user's air-stamp operation and the AR control sends the frame image when the air stamp is stamped, which enables the AR overlay function to retrieve the right input frame from the cache.

 ISBN: 978-1-61208-239-4

IV. EVALUATION

To evaluate possible performance degradation caused by insertion of the AR processing server in the middle of video source and destination, we compared video delay and effective frame rate between AR-enabled conference and standard video-only conference. Evaluation was performed in local network environment consisting of WiFi for onsite worker device's access and wired gigabit Ethernet for other part. AR processing server, MCU, and instructor device were built on a desktop computer with Intel Core-i processor and Windows 7. A mobile tablet with 1.0GHz ARM Cortex-A8 processor and Android 2.3 was used for onsite worker device.

The results of video delay comparison are shown in Figure 5. The video delay measurement was almost in the same level between the video-only and AR-enabled conference. Number of air stamps affixed on the video image less than six did not increase the delay. Android-based mobile tablet device showed longer delay than Windows-based desktop computer because of less codec performance. Longer delay in the video-only conference caused by difference in the tablet's codec performance between front and rear camera employed in the video-only and AR-enabled conference.

Effective video frame rate was also compared. The base videoconferencing system showed 10 fps for VGA resolution, which was restricted by capability of the Android-based tablet. In an AR-enabled conference, effective frame rate was 8-10 fps.

These results denied any significant effect on video transfer performance in this configuration which the AR processing server intervention was expected to cause.

Current major determining factor of the delay and effective frame rate was performance of video codec especially on the mobile tablet device. In near future, if performance of mobile device improves faster than performance of the AR processing server, which is quite likely to happen, the result could change.

V. CONCLUSION AND FUTURE WORK

We developed a remote support service system based on the architecture we have proposed for videophone enhancement with Augmented Reality (AR) technology. We discussed requirements to develop a practical system and presented the system design determined based on the requirements. The system did not give significant modification on base videophone system, which suggests that the architecture can facilitate system reusability in enhancing existing videophone system. Evaluation indicated that the addition of AR functionality on top of existing videophone system did not cause performance degradation. The developed system will be tested in actual remote support scenarios to evaluate its effect on work performance and accuracy.

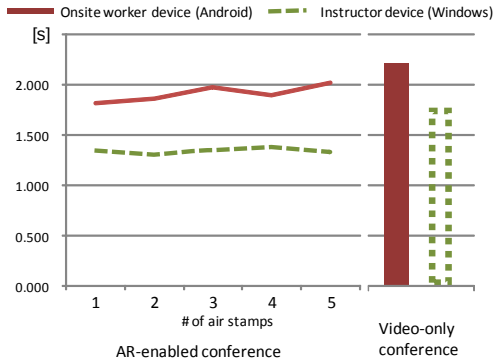


Figure 5. Round-trip video delay.

REFERENCES

[1] S. Araki, T. Hori, M. Fujimoto, S. Watanabe, T. Yoshioka, T. Nakatani, and A. Nakamura, "Online meeting recognizer with multichannel speaker diarization," *Conference Record of 44th Asilomar Conf. on Signals, Systems and Computers (ASILOMAR 2010)*, pp.1697-1701, 2010.

[2] H. Imaoka, Y. Morishita, and A. Hayasaka, "Real-time face recognition demonstration," *Proc. 2011 IEEE International Conf. on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, pp.344, 2011

[3] G. Simon, A.W. Fitzgibbon, and A. Zisserman, "Markerless tracking using planar structures in the scene," *Proc. IEEE and ACM International Symposium on Augmented Reality, 2000. (ISAR 2000)*, pp.120-128, 2000

[4] A. Fukayama, S. Takamiya, J. Nakagawa, N. Arakawa, N. Kanamaru, and N. Uchida, "Architecture and prototype of augmented reality videophone service," *Proc. 15th International Conference on Intelligence in Next Generation Networks (ICIN2011)*, pp. 80-85, 2011.

[5] "Rich Communications Suite Release 4 - Service Definition -," Version 1.0, GSM Association, 2011.

[6] "RCES Phase 2 Stage 2 / 3 Specification Network Value Added Services," TS-1014, Version 1.1, Telecommunication Technology Committee, 2010.

[7] "Cloud-based Translator Phone," NTT docomo, http://www.nttdocomo.com/features/mobility36/, 2012. [retrieved: July, 2012]

[8] M. Billinghurst, A. Cheok, S. Prince, and H. Kato, "Real world teleconferencing," *IEEE J. Computer Graphics and Applications*, Vol. 22, Issue 6, pp. 11-13, 2002.

[9] I. Barakonyi, T. Fahmy, and D. Schmalstieg, "Remote collaboration using augmented reality videoconferencing," *Proc. Graphics Interface 2004, GI '04*, pp. 89-96, 2004.

[10] S. Bottecchia, J. Cieutat, and J. Jessel, "T.A.C: augmented reality system for collaborative tele-assistance in the field of maintenance through internet," *Proc. 1st Augmented Human International Conference, AH'10*, 2010.

[11] "Adding bots to your chat list," http://support.google.com/talk/bin/answer.py?hl=en&answer=172257-12-19-n41.html, 2007. [retrieved: July, 2012]

 ISBN: 978-1-61208-239-4