Natural Communication Technologies Providing Information through Natural Exchanges that Do Not Interfere with Human Activities

Atsushi Sagata, Takashi Sano, Kota Hidaka, Takashi Satou, Shinji Fukatsu, Takafumi Mukouchi, and Hidenobu Nagata

Abstract

NTT Service Evolution Laboratories is conducting research on providing high-presence viewing experiences involving virtual people and objects as if they have been brought to life right in front of the viewers. This article introduces natural communication technologies providing natural and valuable user experiences by creating people, objects, and space that emulate an incredibly realistic presence and seamlessly linking them to the real world through natural interactions without any sense of latency.

Keywords: natural experience, space and object creation, zero latency

1. Towards creating *natural* experiences through the intersection of real and virtual space

We have been focusing on transmitting information of people and objects from remote locations to fully reproduce them in other remote locations to make it feel as if they are right in front of us [1, 2]. Now that highly intelligent applications and services have permeated people's lives and virtual reality (VR) and augmented reality (AR) have become more accessible, we would like to further expand the range of experiences offered to users and offer and have them be more natural. Our aim is not only to faithfully reproduce information of people and objects at a location but also create objects through video projection and audio that give the feeling that they have been brought to life, thus create valuable experiences that transcend reality. We are studying what aspects are important for providing such experiences. While accurately expressing objects is valuable in terms of higher reality, transforming or exaggerating them may result in expressions with more impact. For example, in the "Great Wave off Kanagawa" from the "Thirty-six Views of Mount Fuji," Katsushika Hokusai was able to express the wave as a still image but as if it were moving. Perhaps some people feel the power of this wave more than objects in realist paintings or photographs due to Hokusai's unique expression that gives it movement. Thus, if someone's understandings of a large wave is much more powerful than that of a wave seen in a photograph, Hokusai's painting can be said to have created a real experience for him or her that transcends reality.

1.1 Society 5.0

The advent of the information society has made it possible to digitize various objects in real space and enable complex re-construction and expression in cyberspace. Continuing from the hunting/gathering society (Society 1.0), agricultural society (Society 2.0), industrial society (Society 3.0) and information society (Society 4.0), the Cabinet Office of Japan defines Society 5.0 as "a human-centered society that balances economic advancement with the resolution of social problems by a system that highly integrates cyberspace and physical space" [3]. In other words, in addition to accurately imitating real objects that exist in real space, there is an expectation that reality will be transcended through the fusion of the real and virtual.

1.2 Generative adversarial networks

Generative adversarial networks (GANs) are key for creating video that allows viewers to perceive an object that looks real or expressing objects that transcend the original [4]. Through this technology, computer graphics (CG) images with the quality of real photos is becoming a reality. It is also becoming possible using GANs to create completely different images but with the same characteristics of the reference images. GANs enable processing of tasks, such as sensing the actors on stage, then generating images of animals having the same characteristics of the actors' movements. For example, in the *kabuki* dance performance "Renjishi" (Two Lions), there is a scene where the affection between the parent and child is expressed through the performance of a lion watching over its cub as it runs up the side of a valley. Using GANs, we even can entertain audiences by generating realistic images of the lions as having the same characteristics as the actors. It would be effective to realistically express the kabuki lion as a "lion," but if it were possible to create an experience of Utagawa Hiroshige's woodblock print "shishi no kootoshi" (a lion drops its cub) that seemed real in front of viewer's eyes, it could be considered as an experience that surpassed reality. When presenting an object generated with image-processing technology, to make viewers feel as if life has been breathed into it, it is important for the viewers to feel that the generated object is not just a mere imitation. If it is recognized as a mere imitation, it will be difficult to stir people's emotions. Although it is a major challenge, technology is required to achieve a viewer experience in which generated objects appear to move autonomously. Research on predicting human postures based on the skeleton has begun, and it is somewhat possible to predict future postures based on pre-learned motions and reproducible motions that have some form. We aim to achieve natural exchanges in which the object/person that is sensed moves naturally and autonomously, so that in its interactions with, for example, actual people, there are no distractions due to processing latency, etc.

1.3 Natural communication technologies

To achieve these objectives, we have been researching and developing natural communication technologies. Such technologies consist of the following five elements; (1) space and object creation technology to freely create seemingly real space and objects of the human imagination that transcend reality, (2) zero latency media technology that both reduces physical delays in transmission and processing, thus eliminating sensory delays such as the discomfort people feel due to latency, (3) 2D/3D video display technology that enables both two-dimensional (2D) and three-dimensional (3D) displays to be viewed naturally, (4) information-presentation technology to enable natural interactions between reality and virtual space, and (5) new approaches collectively called "five senses + X transmission technology" for transmitting and presenting not only our five senses but also one's psychological feelings directly and naturally.

2. Space and object creation technology

Space and object creation technology estimates data beyond the obtained sensing data by taking into account the past forms of the same scene or person. We have thus far developed a technology to generate 3D spatial information (CG model) from 2D video in real time using deep learning (**Fig. 1**). This technology was actually used in a kabuki performance called "Cho Kabuki," which is a new type of kabuki performance using information and communication technology. It was held during the Niconico Chokaigi 2019 festival organized by video streaming company Dwango Co., Ltd [5]. Although 3D information is not included in 2D video, this technology creates 3D information based on similar past scenes in real time and *transforms* the characters that appear in Cho Kabuki into other CG characters, enabling a new yet characteristically Cho Kabuki production. In the future, we plan to apply *transformation* to other live-action video by combining 3D spatial information (CG models) and object-extraction technology.

3. Zero latency media technology

Physical delays in transmission and processing is a major issue in achieving natural interaction at remote locations by VR/AR. Although efforts have been made to reduce these physical delays and commercialization has advanced to some extent, it is physically impossible to reduce delay to zero, even at light speed. Thus, to achieve natural interaction, we

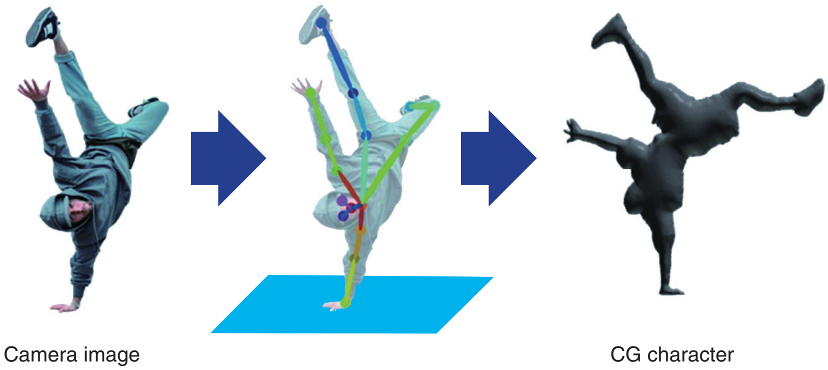

Fig. 1. Real-time 3D information synthesis.

believe that technology is needed that not only eliminates physical delays but also eliminates the discomfort that humans feel from delays as well as eliminating sensory delay. Thus, we are researching zero latency media technology. Specifically, this ongoing research involves clarifying the mechanisms of sensory delay from various information such as peripheral situations and behavior patterns, creating more natural prediction technology that does not cause discomfort due to delay, and clarifying the prediction mechanisms in the brain to eliminate sensory delay due to the world predicted in the brain.

4. 2D/3D video display technology

NTT has developed “HiddenStereo”—a stereo-video-generation technology that uses the characteristics of human vision to enable users to enjoy 3D images when wearing stereo glasses and clear 2D images without stereo glasses [6]. This technology generates left and right images by adding and subtracting a *disparity inducer* to a 2D image to generate left and right parallax. When the disparity-inducer components are combined, they cancel each other out so that only the original 2D image is visible when viewed with the naked eye. Depth information of the 2D image is required to generate the disparity-inducer components. In the case of stereo images, depth information can be obtained using epipolar geometry, etc., but a great deal of operations and ingenuity are required to capture the precise depth information. With normal 2D images taken with a monocular camera, however, the depth information of each pixel cannot be obtained. We are engaged in automatic generation and systemization of “HiddenStereo” that entails depth estimation using deep learning models, extraction of objects through frame and background differences, and instance segmentation using deep learning models for 2D images taken with monocular cameras.

5. Information-presentation technology for natural interactions

To enable natural and realistic information-presentation technologies, we are conducting research and development on 360-degree glasses-free tabletop 3D display technology (360° autostereoscopic 3D) (**Fig. 2**) and sound field synthesis technology. This technology enables viewing of 3D objects with binocular disparity on a screen on a table without using 3D glasses by combining multiple projectors arranged in a circle and a special screen called a *spatially imaged iris plane screen* [7]. A large 120 cm-diameter screen and optical linear blending enable smooth movement of the point of view (smooth motion parallax), even though the number of projectors is 1/4 to 1/10 that of conventional technology. Sound field synthesis is a technology that can reproduce a sound field using a linear loudspeaker array of multiple loudspeakers in a line. This technology can control the distance between a sound source and audience member as well as its direction. In past events and installations, we used this technology to produce virtual sound sources that come close to audience members to reproduce kabuki performances [8] and goalball matches, a team sport played by visually impaired athletes. We have been investigating an extension of this technology to reproduce sound in a limited area. We have succeeded in reproducing a sound field using a multipole-loudspeaker array consisting of multiple small speakers closely arranged in

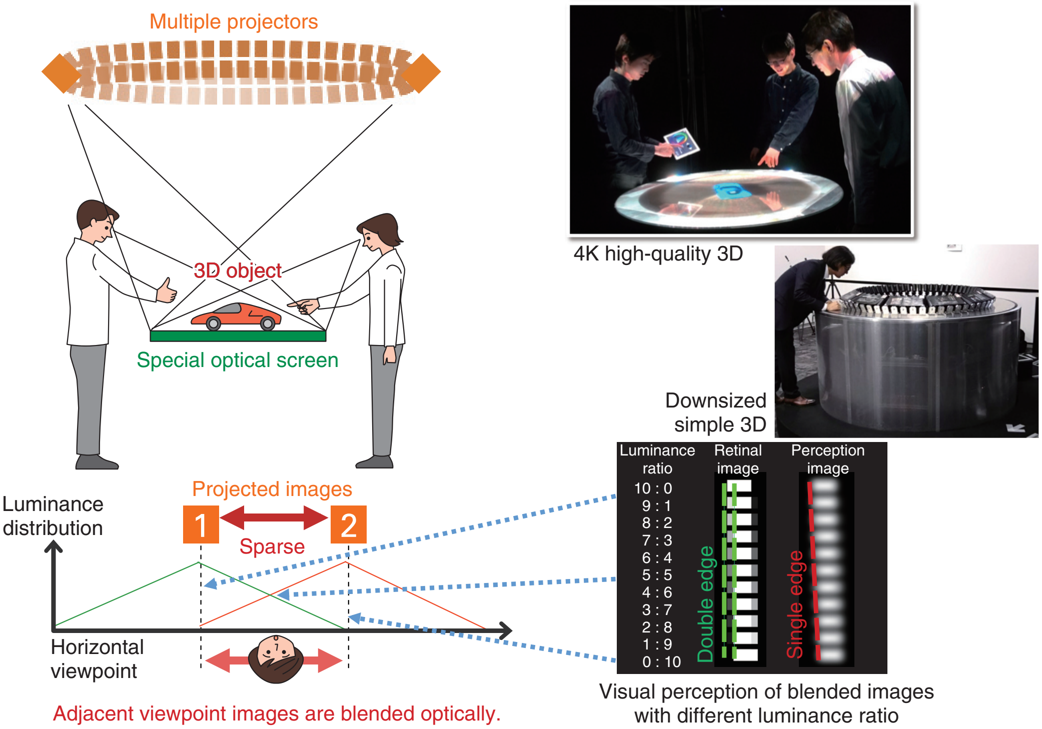

Fig. 2. 360-degree glasses-free tabletop 3D display technology (360˚ autostereoscopic 3D).

a Cartesian grid. This array enables control of the directivity pattern of an arbitrary sound source while being smaller than current speaker arrays [9, 10]. These technologies place importance on achieving natural communications to enable users to enjoy a sense of reality without having to wear special devices such as head mounted displays, 3D glasses, or headphones. In other words, the sophistication of environments surrounding users will make it possible to reproduce natural and realistic presence without burdening them. Such technologies could be applied in the field of entertainment to create more realistic experiences of sporting events or concerts. In business, these technologies should help teleconferencing evolve from screen and audio sharing to space sharing. It may no longer be just a dream to have a remote participant projected by a digital twin next to oneself at a conference and whisper to that person or share written communications. Even in the home, a television (TV) that plays back sound with volume and frequency characteristics optimized only in areas where there are elderly people could become a reality. In other words, rooms and TVs present sounds properly controlled according to each listening area so one would not have to increase the volume to suit the elderly when watching TV with the whole family. It may enable the elderly to enjoy sounds without hearing aids. It would also be possible to control the sound so that it would not leak in the direction of children’s bedrooms so that one could watch TV without the need of using headphones.

6. New approaches to transmit and present the five senses + X directly and naturally

The evolution of sound and video as means of presenting information has been remarkable and has made it possible to experience not only improvements in definition but also 3D effects. However, utilization of the other senses has not progressed. Achieving natural user experience and presentation of various information requires utilization of vision and hearing as well as the other senses. Services that create highly realistic experiences will have to appeal not only to sight and hearing but also touch, smell, and even taste. In terms of information presentation, methods that occupy the important sensory organs for sight and hearing are not always the natural means of conveying information. This means that if information can be transmitted using methods other than

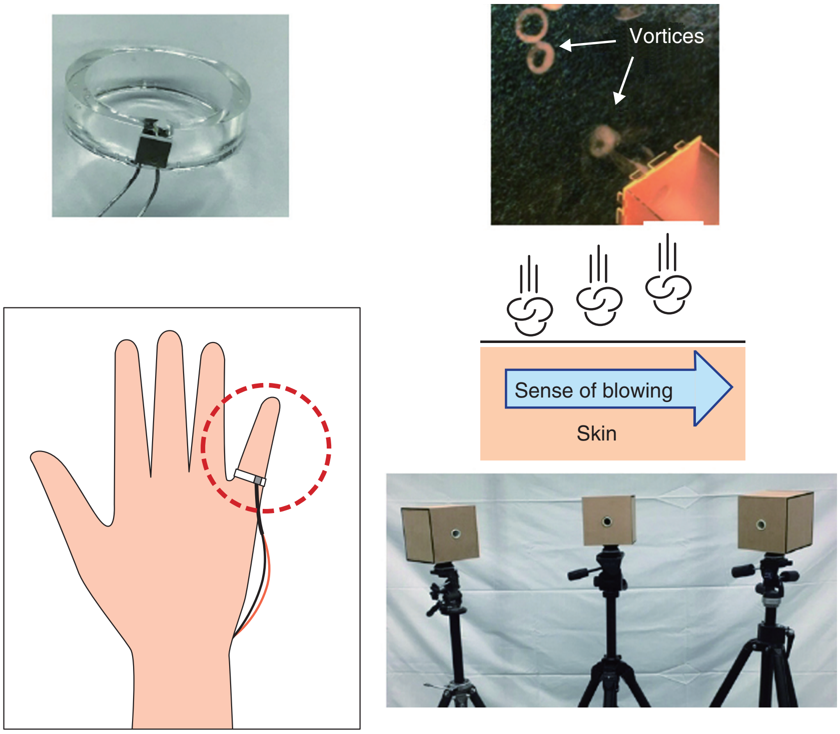

Fig. 3. “ThermalBitDisplay” presents information to the skin.

visual or auditory, or a combination of these, it will be possible to receive information naturally anytime and anywhere. We are currently working on presenting information to the skin (**Fig. 3**). “ThermalBitDisplay” is an information-presentation device that uses temperature without disturbing the senses of sight and hearing and consists of a ring with an embedded thermoelectric element used to obtain information only when pressed against the lips. Unlike push-type notifications, this device enables one to check information only when he/she wants to, and enables one to check information casually because one does not need to use sight or hearing. We are also working on the expression of new reality through stimulation using vortex rings (air movements that can be generated using a so-called air cannon). By controlling the time difference between stimuli, we are also researching technology to create pseudo-sensations as if something is passing close to oneself. By creating sensations other than that of something passing through controls that include stimulation in certain locations, for example presentations combining sounds, we believe it will be possible to further improve the realism of moving sound sources. Going forward, we will continue researching and developing technologies to enable more natural experiences or reception of information by conveying sensations other than through the senses of sight and hearing.

7. Conclusion

This article described the research and development of natural communication technology to achieve free interaction and intersection between real and virtual spaces including information exchange. The virtual space is sometimes described as a *mirrorworld* [11], and we believe it is not just simply an electronic cyber world but also a *parallel world* by *intersecting* with real space. Ideally, we believe that as well as enabling users to experience what they want to be good at, such as singing or dancing, in the parallel world, it is desirable to feed this back to the real world. We will make such intersection between virtual and real world one of our themes and proceed with our research and development of natural communication technology.

References

[1] H. Nagata, H. Miyashita, H. Kakinuma, and M. Yamaguchi, “Real-time

Extraction of Objects with Arbitrary Backgrounds," NTT Technical Review, Vol. 15, No. 12, 2017.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201712fa7.html

[2] J. Nagao, H. Miyashita, T. Sano, K. Hasegawa, and T. Isaka, "Kirari! for Arena: Real-time Remote Reproduction of 4-directional Views with Depth Perception," Journal of the Imaging Society of Japan, Vol. 58, No. 3, pp. 306–315, 2019 (in Japanese).

[3] Website of Cabinet Office of the Government of Japan, "Society 5.0," https://www8.cao.go.jp/cstp/english/society5_0/index.html

[4] I. J. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," Proc. of NIPS2014, pp. 2672–2680, Montreal, Canada, Dec. 2014.

[5] Press release issued by NTT on Mar. 25, 2019 (in Japanese), https://www.ntt.co.jp/news2019/1903/190325b.html

[6] T. Fukiage, T. Kawabe, and S. Nishida, "Hiding of Phase-based Stereo Disparity for Ghost-free Viewing Without Glasses," ACM Transactions on Graphics, Vol. 36, No. 4, pp. 147:1–17, July 2017.

[7] M. Makiguchi and H. Takada, "360-degree Tabletop Glassless 3D Screen System," NTT Technical Review, Vol. 16, No. 12, 2018.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201812fa4.html

[8] K. Tsutsumi and H. Takada, "Powerful Sound Effects at Audience Seats by Wave Field Synthesis," NTT Technical Review, Vol. 15, No. 12, 2017.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201712fa5.html

[9] K. Tsutsumi, K. Imaizumi, A. Nakadaira, and Y. Haneda, "Analytical Method to Convert Circular Harmonic Expansion Coefficients for Sound Field Synthesis by Using Multipole Loudspeaker Array," Proc. of the 27th European Signal Processing Conference (EUSIPCO 2019), A Coruña, Spain, Sept. 2019.

[10] K. Imaizumi, K. Tsutsumi, A. Nakadaira, and Y. Haneda, "Analytical Method of 2.5 D Exterior Sound Field Synthesis by Using Multipole Loudspeaker Array," Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2019, New York, USA, Oct. 2019.

[11] K. Kelly, "AR Will Spark the Next Big Tech Platform—Call It Mirrorworld," Wired, Feb. 2019.
https://www.wired.com/story/mirrorworld-ar-next-big-tech-platform/

Atsushi Sagata

Executive Research Engineer, Natural Communication Project, NTT Service Evolution Laboratories.

He received a B.E. in electronic engineering from the University of Tokyo in 1994 and joined NTT the same year. He has been engaged in R&D of video coding systems and in the development of the digital high-definition TV transmission system at NTT Communications. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE).

Takashi Sano

Senior Research Engineer, Natural Communication Project, NTT Service Evolution Laboratories.

He received a B.E. and M.E. in electronic engineering from Tokyo University of Science in 2002 and 2004. He joined NTT Cyber Space Laboratories in 2004. He has been engaged in R&D on software and hardware design of video coding.

Kota Hidaka

Senior Research Engineer, Supervisor, Natural Communication Project, NTT Service Evolution Laboratories.

He received an M.E. from Kyushu University, Fukuoka, in 1998 and Ph.D. in media and governance from Keio University, Tokyo, in 2009. He joined NTT in 1998. His research interests include speech signal processing, image processing, and immersive telepresence. He was a senior researcher at the Council for Science, Technology and Innovation, Cabinet Office, Government of Japan, from 2015 to 2017.

Takashi Satou

Senior Research Engineer, Supervisor, Natural Communication Project, NTT Service Evolution Laboratories.

He received a Ph.D. in information engineering from the University of Tokyo in 1996. He joined NTT the same year and studied human computer interaction and video handling technologies. He managed R&D of image recognition technologies and innovation process in NTT DOCOMO from 2014 to 2019. He is currently managing R&D of natural communication technologies in NTT Service Evolution Laboratories.

Shinji Fukatsu

Senior Research Engineer, Supervisor, Natural Communication Project, NTT Service Evolution Laboratories.

He received a Ph.D. in engineering from Osaka University in 2002 and joined NTT the same year. He has been engaged in R&D of human interfaces and video streaming services. He has also been engaged in planning and development of video streaming services at NTT Plala and in promoting standardization and ICT international development at the Ministry of Internal Affairs and Communications. He is currently engaged in R&D of natural communication technology.

Takafumi Mukouchi

Senior Research Engineer, Supervisor, Natural Communication Project, NTT Service Evolution Laboratories.

He received an M.E. from Waseda University, Tokyo, in 1995 and joined NTT the same year. He has been engaged in the research of machine learning and development of video communication systems. He is currently engaged in the research of natural communication technology. He is a member of IEICE.

Hidenobu Nagata

Senior Research Engineer, Supervisor, Natural Communication Project, NTT Service Evolution Laboratories.

He received an M.E. in systems and information engineering from the Faculty of Engineering, Hokkaido University in 2001. He joined NTT in 2001 and studied video handling technologies including video searching, video indexing, automatic summarization and their interfaces. From 2008 to 2014, he worked at NTT Electronics and developed professional transcoders and embedded audio IP for consumer devices. He transferred to NTT in 2014 and is currently researching ultra-realistic telecommunication technology.