This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2024.3352240

Understanding the Quality of User Experience in Telepresence Systems from an Information Theory Perspective

Ruiqing Wang, *Student Member, IEEE,* Kebin Liu*, *Senior Member, IEEE,* Ziyue Dang, *Student Member, IEEE,* Xu Wang, *Member, IEEE,* Fan Dang, *Member, IEEE,* Yue Sun, *Student Member, IEEE,* Yuang Tong, *Student Member, IEEE,* Haitian Zhao, *Member, IEEE,* Yunhao Liu, *Fellow, IEEE*

***Abstract*—Efforts to enhance the user experience (UX) of telepresence edge systems in various application scenarios have been significant. However, existing approaches tend to focus on specific aspects, leaving us with a fragmented understanding of UX quality. We address this gap by examining Video Conference Systems (VCSs) for remote collaboration, using an Information Theory Perspective as a lens. We introduce a novel model to quantify the multimodality information users receive while engaged in mobile office environments, enabling an evaluation of existing VCSs. Our approach transforms the assessment of UX quality into the measurement of a set of information *channels*. Based on this insight, we identify new prospects and meaningful guidelines for future multimedia telepresence edge systems and try to induce a new prototype design under cost restriction. To demonstrate the validity of our method, we implement the prototype which seamlessly integrates visual, audio, and olfactory dimension information. Extensive experiments and user studies validate the effectiveness and practicality of our approach.**

***Index Terms*—Telepresence Edge System, Model Analysis, User Experience Quality**

I. Introduction

TELEPRESENCE edge systems have been increasingly common in people's daily lives. More and more conversations, work meetings as well as formal conferences are moving from offline to online. Significant attention has been drawn from both academia and industry to improve the efficiency and user experience (UX) of these systems, especially video conference systems (VCSs) for remote collaboration. As shown in Fig. 1 (a), cloud-based approaches [1] tried to provide high-resolution images and voice streams with low latency. MAJIC [2], Hydra [3], Gaze-2 [4] and MultiView [5] discussed the importance of conveying gaze direction as illustrated in Fig. 1 (b). The fixed viewpoint can result in a poor sense of reality, thus many efforts [6]–[10] have been devoted to capturing and displaying 3D user silhouettes. Fig. 1 (c) demonstrates a stereoscopic solution using a cylindrical display. As shown in Fig. 1 (e), Horizon Workrooms [11] from Meta enabled people to collaborate in virtual reality with their cartoon avatar, and robot surrogates such as the one shown in Fig. 1 (d) was also introduced to appear in remote conference rooms. The What-You-See-Is-What-I-See [12]–[14] VCSs focused on building a shared "screen" for both co-located and remote users, where users' images were embedded into physical and digital contexts for deictic referencing. An example is demonstrated in Fig. 1 (f).

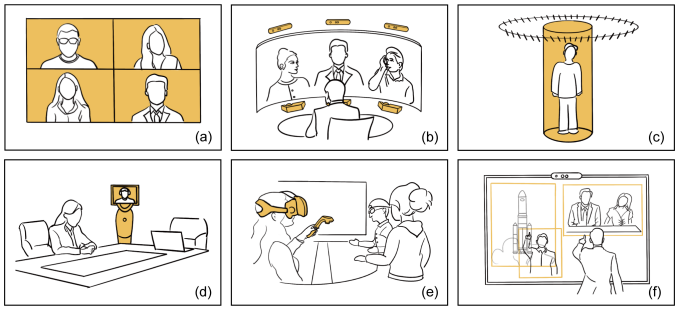

Fig. 1. Visualization of typical telepresence edge systems: (a) Cloud-based approach (b) Mutual gaze aware solution (c) Holographic display for 3D reconstruction (d) Robotic avatar (e) VR collaboration (f) WYSIWIS meeting.

After reviewing all these mainstream techniques, we find that each of them improves the telepresence edge system from a particular perspective and puts emphasis on limiting factors. We, however, still lack a global picture and UX quality measurement method of them. A novel descriptive model is needed for systematic measurement and analysis.

We take the user experience in local meetings, in other words, "being there", as a reference for a "good" mobile officing system. The insights behind this choice are that human beings perceive their surroundings through all of their sensory organs. For example, in an offline meeting, people see each other with their eyes and hear other participants' voices with their ears. Each sensory organ corresponds to a particular modality of information. Therefore, the dimensions of input information to the human brain are limited by the number of sensory organs. Information from all sensory organs forms a complete set of inputs to the human brain, which thus determines the user's experiences. Based on this observation, we propose to revisit VCSs, the most typical telepresence edge system from an Information Theory [15] perspective.

This work is supported in part by the National Key Research and Development Program of China under grant No.2022ZD0114900, and the NSFC under grant No.62202263. *(*Corresponding author: Kebin Liu.)*

Ruiqing Wang, Yue Sun, Yuang Tong, Kebin Liu, Xu Wang, Fan Dang, Yunhao Liu are with Global Innovation Exchange, Tsinghua University, China (e-mail: {wang-rq22, suny21, tya21}@mails.tsinghua.edu.cn, {kebinliu2021,xu_wang, dangfan, yunhao}@tsinghua.edu.cn).

Ziyue Dang is with Computer Science Department, University of California, Los Angeles, America (e-mail: ziyue.dang@cs.ucla.edu).

Haitian Zhao is with Department of Automation, Tsinghua University, China (e-mail: zhaoht2022@mail.tsinghua.edu.cn).

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2024.3352240

It is worth mentioning that our model is different from the traditional Information Theory which calculates the channel capacity of computer networks. We hope to model the information received by all sensory organs of the user during the remote collaboration, which means the information flow is from user to machine to user.

In an online meeting scenario, the separation of physical space blocks people from perceiving information with their sensory organs directly from remote sites. In this case, VCS is responsible for sampling, encoding, transmitting, and finally demonstrating multi-modality information from one meeting room to another. It can be regarded as a set of *channels* for different sensory information. Besides, as we find the trend that digital media can offer richer information and achieve higher efficiency than just "being there," we add new *channels* to our model to consider these features. The concept of "beyond being there" has been proposed in [16], and recent advances in computation speed, AI algorithm functionality, *etc.*, bring us new possibilities [17], [18] to mobile officing applications. With this model, the problem of studying the efficiency and UX of telepresence edge systems can be transformed into measuring the quality of each *channel*.

The contribution of this study are as follows.

- First, we propose a novel Information Theory based model to calculate the multi-modality information received by users and to evaluate the UX quality of telepresence edge systems.
- Second, we analyze 29 typical solutions and discuss insights with heuristic guidance for further remote collaboration improvements.
- Third, we propose a design under the above guidelines that aims to achieve high UX quality at a modest cost.
- Fourth, we implement the prototype and conduct extensive experiments to validate the validity of our methods.

The rest of this paper is organized as follows. Section II describes research related to our work. We present an Information Theory based model for UX quality measurement in Section III. The insights and heuristic guidelines for future design are discussed in Section IV, and we also present a prototype design in this section. We express the prototype system implementation and experimental results in Section V, then conclude our work and future plans in Section VI.

II. Related Work

A telepresence edge system combines high-definition video, audio, and interactive components to create a unique "face-to-face" experience on the web [19]. The most typical telepresence edge system is VCSs for mobile officing.

VCSs that directly capture, transmit and present images at a specific and fixed angle include MAJIC [2], Hydra [3], GAZE-2 [4] and Teleport [20]. They try to convey multiple eye contact and support proper awareness of gaze direction among participants, and each person is represented by a separate camera/projector or camera/monitor pair. These approaches are limited by the number of devices and the fixed viewing perspective, making it difficult to expand to multiple parties.

To better convey non-verbal cues, researchers capture and reconstruct 3D scenes and 360° videos in immersive multimedia systems. Coliseum [21] uses head tracking and IBVH for real-time rendering. Telehuman1 [8] and Blue-C [10] use polarized projection so the user wears shutter glasses to obtain stereoscopic perception. But the above methods are essential "one-to-one" or "one-to-many" modes. VirtualCube [6] uses six RGBD cameras to capture multi-view stereo for more accurate depth estimation and then renders high-quality videos on a surrounding life-size display by using Lumi-Net. Two or three remote users can correctly preserve mutual eye gaze and attention, have side discussions and share "workspace".

Autostereoscopic displays applied in telepresence edge systems can be divided into two categories: flat panel displays (FPDs) and curved surface displays (CSDs). FPDs include retroreflective displays, parallax barrier displays, and lenticular lens displays. Research [5], [22], [23] use the above three technologies, respectively, so that multiple users can view the same stereoscopic scene from different perspectives without wearing any devices, but the overall brightness of the images is low. CSDs offer the opportunity to achieve continuous horizontal motion parallax and holographic scenes. [24], [25] Telehuman2 [9] and Lightbee [26] set up a projector array around a life-size cylindrical light field display. The angle between each projector is 1.3 degrees, which is less than the average distance between adult pupils. However, such designs are costly, difficult to deploy, and currently only support "many-to-one" mode. The sense of ritual also lacks.

2D spatial sharing and fusion follows the principle of "what you see is what I see" (WYSIWIS). From ClearBoard's [14] two separate parties sharing the same drawing board, to HyperMirror's [13] multiple images mirroring and fusion, to MirrorBlender's [12] arbitrarily adjustment of cloud conference interface. The layout, position, scale, and transparency of multi-interfaces can be adjusted by each user.

In addition, there are two main types of telepresence edge robots. Mobile robots [27] generally have a display screen placed in the center part and can be controlled by the remote user while communicating with the local end. Humanoid robots [28]–[30] focus on correctly reproducing "non-verbal cues" presented by the remote user's head. However, an "uncanny valley effect" may easily happen to these robots.

With the rapid development of XR, it has become an emerging way for users to enter and interact freely with others in the virtual world as avatars, such as in Horizon Workrooms [11] and Mesh for Teams. However, long-time use of HMDs very easy to causes fatigue, or even motion sickness, because of the desynchrony of the human brain's motion instructions and sensory feedback. More importantly, it is difficult to transmit users' facial expressions, "non-verbal cues" and the state of perception when they are wearing HMDs, which will greatly reduce the authenticity of the VCSs. Although Holoportation [7] can solve this problem, it requires a large number of specialized and expensive hardware devices.

In this work, we hope to draw a whole picture of the telepresence edge systems for remote collaboration, evaluate their UX quality systematically, and propose a more preferable cross-modality information fusion solution.

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2024.3352240

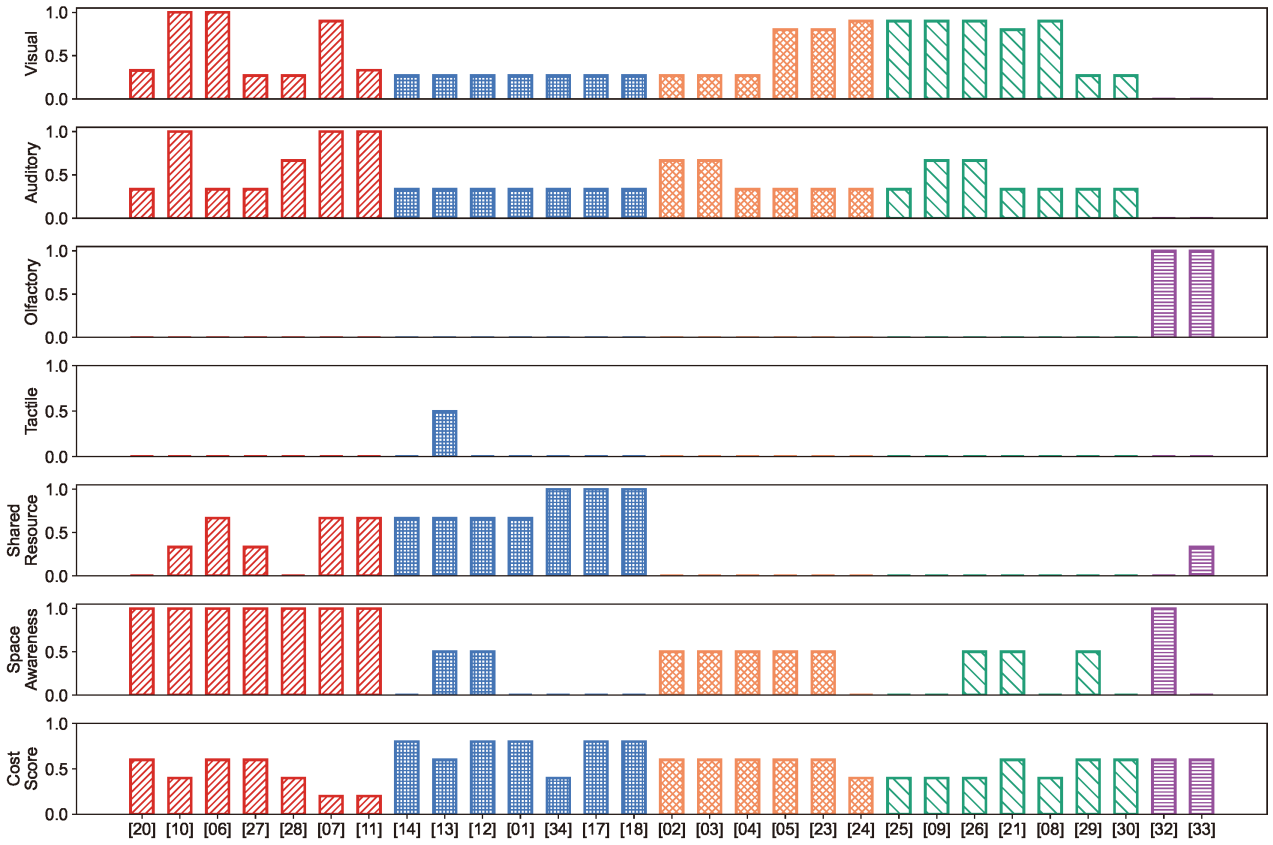

Fig. 2. The evaluation results of 29 typical telepresence edge systems according to our model show the trend and performance of emphasis in this field. Through K-means clustering, we classified similar works and they are Robots and HMDs that focus on spatial information (red); WYSIWIS approaches (blue); gaze-aware focusing (orange); 3D video and motion parallax providing (green); and olfactory aid (purple).

III. An Information Theory Model for UX Quality Measurement

A. Model Definition

In our model, a telepresence edge system is defined as a vector:

$$\vec{T} = (C_v, C_a, C_o, C_t, C_g, C_r, C_s). \quad (1)$$

Each element in $\vec{T}$ refers to a *channel*, corresponding to the information dimension perceived by a particular human sensory organ. In other words, each *channel* can be regarded as a subsystem of each telepresence edge approach that is in charge of conveying a certain modality of information. We consider different information dimensions separately and assign a *channel* to each of them. In detail, C_v denotes the subsystem which distributes visual information to remote participants; C_a is the *channel* for auditory information; C_o denotes the olfactory *channel*; C_t corresponds to the tactile sensation; and, C_g is the gustatory *channel*. We can find that the first five variables in $\vec{T}$ have covered all the information types processed by human sensory organs. Our study also considers richer information beyond traditional local meetings, that is, the shared resources *channel* C_r and the awareness of space *channel* C_s. The shared resources include three kinds of information. Digital materials for demonstration, marks and deictic gestures made by the users, and information generated by AI algorithms. The C_s *channel* indicates the awareness of space, in other words, the capability of making distributed users feel that they are located in the same physical space.

B. Channel Analysis

We evaluate the quality of each *channel* with two metrics: the *capacity* and the *cost*. The concept *capacity* comes from Information Theory, which denotes the maximum amount of information content that can be conveyed from source to destination. We observe that the limitations of many existing approaches, such as lack of non-verbal cues [6], deixis challenges [12], "Mona Lisa effect" [13] by video medium, *et al.*, are mostly caused by information loss. If there are several perfect *channels* that can bring complete sensory information to remote users in an unaware manner, people can hardly determine whether they are joining a local meeting or a video-mediated conference. After building an information content [15] model for each data modality, the second criterion is *cost*, including both the complexity of devices and the extra burden that a system brings to users. Complex and heavy-weight devices will result in high prices and affect user experience in the telepresence edge scenario.

Take the visual dimension as an example to describe the workflow of a *channel*. To facilitate our modeling, the vision subsystem is divided into three major stages: Sampling Stage, Transmitting Stage, and Reconstruction Stage. Information

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2024.3352240

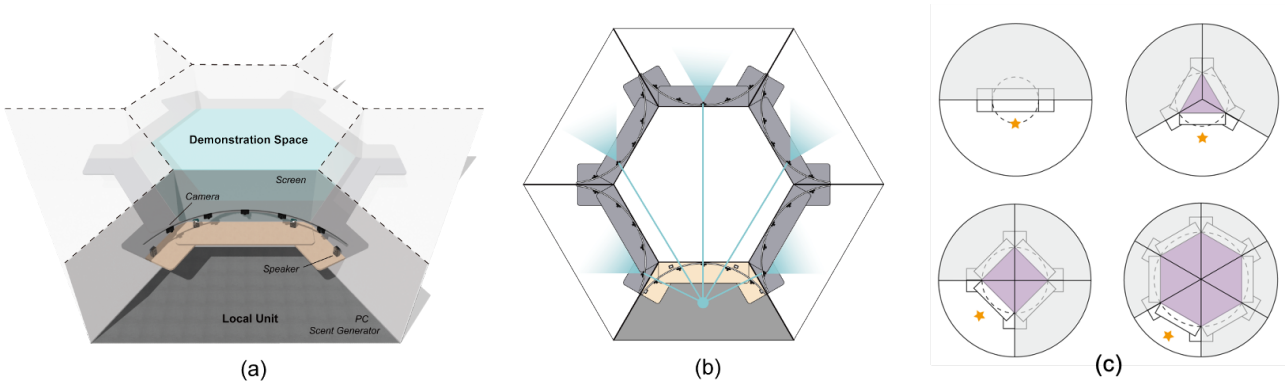

Fig. 3. The architecture of our prototype: (a) Hardware arrangement, including screen, cameras, microphone and speakers, scent generator, table, PC, etc. (b) Video streams selected for a local unit (c) Different layout when different number of parties join the same meeting room.

loss can occur in each stage. However, with the advances in coding algorithms and networking techniques, information loss in the coding and transmitting process can be neglected, so we mainly focus on the sampling and reconstruction stage.

Before evaluating the visual *channel*'s *capacity*, we require an information content model to represent complete visual information in an arbitrary scenario. In this work, we introduce Plenoptic Function [31] as the visual information model which includes all light rays from a particular viewing position.

$$P = P(\theta, \phi, V_x, V_y, V_z). \tag{2}$$

To understand this definition, we can imagine placing an ideal eye at every location (V_x, V_y, V_z) and recording the color value of the rays projecting into the pupil from every possible angle (θ, ϕ). Then the Plenoptic Function indicates a 5D light field that encodes all the image information within a scenario. In fact, by enumerating all possible locations and viewing angles, we can get a set of panoramic images that can be a representation of the 5D light field. An image captured by the camera can be regarded as a projection from the 5D vision space to a 2D plane. Meanwhile, the displaying stage can cause visual information loss as well.

In Information Theory, the information content of a message is determined by the amount of "surprise" or "uncertainties" conveyed by the message. According to Shannon's definition [15] and Plenoptic Function, the information content of a visual scene is represented by the following differential entropy.

$$H_v(X) = -\int_{V_x,V_y,V_z}\int_{\theta,\phi}\int_{r,g,b} p(x)\log p(x)\mathrm{d}x. \tag{3}$$

Then the information delivery rate is measured by the average mutual information between the source and the receiver.

$$I_v(X;Y) = H_v(X) - H_v(X|Y). \tag{4}$$

And the *channel capacity* is defined as the maximum $I(X;Y)$ respect to varying input distributions:

$$C \overset{def}{=} max_{p(x)}\{I_v(X;Y)\}. \tag{5}$$

From the above equations, we can find that the *channel capacity* is determined by the source information content $H(X)$ and residual information content $H(X|Y)$. We have discussed the source information content $H(X)$ before and $H(X|Y)$ can be described as the amount of "uncertainties" a remote user still holds after watching the images provided by the telepresence edge system. In other words, the *channel capacity* indicates the most possible "uncertainties" that can be eliminated. Since "uncertainty" comes from the absence of visual information in the sampled image, we consider all light rays that are out of the camera's field of view when calculating the residual information content.

$$H_v(X|Y) = -\int_{V_x,V_y,V_z}\int_{(\theta,\phi)\notin f}\int_{r,g,b} p(x)\log p(x)\mathrm{d}x. \tag{6}$$

In practice, the above integration is hard to calculate, so we select the portrait region as our *target region* and assume that the information contained in it follows an independent and identical distribution. Therefore, we apply the angle of view of cameras to approximate the coverage information content. We approximate the *channel capacity* of the capturing stage with the following expression.

$$\tilde{C}_c = \frac{N\alpha}{\pi}H(X), \tag{7}$$

where α denotes the angle of view that a camera can cover, N is the number of separate cameras around the target person (here we assume the view angles of these cameras are not overlapped), and $H(X)$ represents the source information content. Now, we have an approximation of the *channel capacity* for the image-capturing stage. Particularly, some multi-view stereoscopic techniques [6], [10] leverage several cameras to reconstruct the integrated 3D visual information of a person, in which case we set $C = H(X)$.

Next, we look at the displaying stage. Similar to that in the capturing stage, the *channel capacity* of the displaying stage C_d is measured by the difference between $H(X')$ and residual uncertainties.

$$\tilde{C}_d = H_v(X') - H_v(X'|Y). \tag{8}$$

$H(X')$ denotes the source information content in displaying stage. We assume $H(X) = H(X')$ to facilitate the following discussion and similar analysis can be conducted while

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2024.3352240

Fig. 4. Implementation and user study setting: (a) Four-parties conference, (b) Six-parties conference, (c) Scent generator, (d) User study environment.

$H(X) \neq H(X')$. Theoretically, perfect display equipment can show all visual information simultaneously, that is, reconstructing the complete light field of a Plenoptic Function. In practice, this can be extremely difficult and expensive and is not necessarily as well. We observe that human cannot sense all light rays in a scene simultaneously. Human eyes can be regarded as two parallel cameras that can only take pictures from one static viewpoint at a time, allowing us to simplify the displaying problem. Thus, we propose to measure the residual by the differences in view angles between the displayed image and user desired one:

$$\tilde{C_d} = \frac{\pi - \beta}{\pi} H_v(X'), \tag{9}$$

$$\tilde{C} = \gamma \tilde{C_d}(\tilde{C_c}), \tag{10}$$

where β denotes the summarized view angle differences and γ is a penalty factor taking some other sources of information loss into account, such as lack of awareness of parallax, *etc.*

We conduct analogous analysis on auditory, olfactory, tactile, and other dimensions.

IV. Model-Based Analysis and Insights

We select 29 typical telepresence edge systems and analyze them based on the proposed model. For *channel* measurement, we first select proper values for parameters in our model. We set $\alpha = \pi/3$ in visual dimension because many multi-view stereo systems can reconstruct high-quality 3D silhouettes with just three commercial cameras. Then, the parameter γ gives a 10% penalty (γ=0.9) to each of the following situations, such as lack of motion parallax or limited display scope. This is because according to the previous indicators, many work results are consistent in the same dimension, but still have some differences or defects when compared with each other. We hope to characterize these defects that are not dominant factors so that they can be distinguished in the final result without affecting the performance of the main factors, thus a small penalty parameter is artificially set for visual dimension. For audio, we treat audio stream, distance, and direction as equally important, so as the shared resource dimension. For tactile information, we estimate the coverage of haptic feedback with respect to the whole human body. For measuring the awareness of space, we set the global coordinates system with a score of 1 and the relative positioning system with a score of 0.5. Finally, we empirically assign cost scores to different approaches according to their *cost*. Based on the above, we then normalize all the results to $[0, 1]$. The normalized *capacity* can be regarded as a score for each dimension.

Note that the scores of different dimensions naturally form a feature vector, we conduct a K-means clustering of these methods using these feature vectors. As shown in Fig. 2, we find correlations among methods within each group. The first cluster attaches significant importance to awareness of space, which usually applies robots or HMDs in their systems, resulting in extra burden to users and relatively high cost. The second cluster focuses on building a shared 2D space where images and other types of information are blended. In this way, rich information about shared resources can be conveyed. These methods generally leverage low-cost COTS devices. The third group applies 2D images as well but is aware of gaze direction through some relative positioning schemes. The fourth cluster pay more attention to capturing and reconstructing rich visual information such as 3D video. Motion parallax is also supported by some of them. The final category of efforts mainly considers olfactory dimension [32], [33].

Based on the above observation, we conclude three main shortcomings of current VCSs with opportunities for future remote collaboration.

- The visual dimension is the main area where the existing work is focused. Many SOTA approaches require expensive and complex devices while supporting only a limited number of users, or resulting in extra burdens to users such as wearing shutter glasses or HMDs that bring uncomfortable feelings and inconvenience. Based on this insight, we should attach more importance to the validity of *cost* and information delivery.
- The information transmission of the auditory dimension mainly focuses on verbal content and seldom pays attention to the aspect of the direction and distance of sound, declining the authenticity of the conversation and the awareness of the shared space.
- Besides, other dimensions are being seriously neglected in current mobile officing. We believe that information from different modalities is complementary to constructing a good conference environment and improving the quality of UX. In addition, cross-modality information deduction and fusion could bring potential opportunities.

To go further, providing a sense of ritual, coexistence, and free from constraint is of significant importance to remote collaboration. Also, there has been a significant trend to augment online applications with rich information from AI analyzers [34] such as auto-summary [18], talk tracking [17]

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2024.3352240

TABLE I
SENTIMENT CLASSIFICATION PERFORMANCE WITH DIFFERENT KERNELS

Metrics	Kernels			
	Linear	**Polynomial**	**Sigmoid**	**Gaussian**
Accuracy	**83.43%**	79.01%	79.01%	82.87%
Balanced Accuracy	**77.15%**	74.71%	69.35%	76.77%
F1-Score (macro)	**78.32%**	76.85%	70.89%	78.27%
95% Confidence Interval of Accuracy	**[78.01%, 88.85%]**	[73.08%, 84.94%]	[73.08%, 84.94%]	[77.38%, 88.36%]

and mood recognition. We believe this would be a good direction to raise the meeting efficiency and thus help people achieve a "beyond being there" solution.

V. IMPLEMENTATION AND EXPERIMENT

A. System Implementation

We first present some *cost* restrictions:

- We only use COTS components in our prototype.
- We do not bring extra burdens to users or change their habits during the meeting process.
- The proposed system is not dedicated to "one-to-one" meetings but supports flexible multi-party conferences.

For the visual part, as illustrated in Fig. 3 (a-b), the shared conference space is a fusion of real meeting rooms and virtual ones corresponding to remote parties. A 3200mm x 3200mm x 2400mm CAVE-like structure formed the interface. The angle between the side screen and the front screen can be adjusted within 90 to 180 degrees through articulation. As the number of attending parties changes, the structure changes to achieve seamless integration as shown in Fig. 3 (c). The virtual scene is pre-built on the Unreal Engine 5 platform, and the number of parties supported is 2-6 in consideration of the optimal meeting efficiency and user experience. To ensure the visual acquisition of the correct perspective and avoid the "Mona Lisa effect", we set up a ring of cameras at the height of 1250mm, which is flush with the observer's general sight height when sitting. Webcams are placed on tripods for real-time shooting with a resolution of 1080p/30fps. The number and angle of the webcams are determined by the number of attending parties, always one less than the number of parties. Different video streams are sent to corresponding remote participants according to their relative positions in the global space. Mutual eye contact and other nonverbal information are preserved. Although the cameras are in the user's field of view, the later experiment proved that it does not affect the user's acquisition and transmission of information.

For the acquisition and playback of surround sound, we collect the sounds from each party and implement audio panning based on the Pydub package. Given the number of participants and their directions, audio streams can be panned respectively to achieve a spatial sense by using two speakers. In addition, the sound can be zoned during group discussions to automatically block the sound of other groups.

We also place a scent generator containing two different scents, mint and sandalwood. Sandalwood can relieve anxiety and settle people's central nervous system; mint can refresh the mind and eliminate sleepiness. After collecting the visual and auditory streams, we analyze multi-modality information to define the whole conference's atmosphere and automatically release the scent [35]. The essential oil burets are connected to the Arduino UNO board through the atomization pieces, so that the category of scent release can be controlled. The detailed method will be described in Section 5.2, and the final prototype is shown in Fig. 4(a)-(c).

B. Atmosphere Analysis and Experimental Results

Our method takes video streams as input and features are extracted using a neural network structure named MI-MAMO [36]. The obtained features are represented with the Arousal-Valence space. We apply an SVM classifier to recognize the sentiment of users and thus determine the atmosphere of the whole conference.

To construct the train and test dataset, we collect and segment video clips of real-life video conferences from the Internet, and the final dataset includes 454 video segments in total. The dataset is randomly split into a train set that contains 60 % data and a test set containing the rest 40 %. The evaluation results are shown in Table I, according to which we find that the Linear kernel achieves the best performance than other kernels. In fact, Linear kernels behave similarly to Gaussian kernels, and there might be no significant difference in the overall implementation of the system, so we ultimately chose the former with simpler structures and faster computations.

C. User Study

The user study aims to qualitatively model the telepresence edge systems' performance and explore users' direct perceptions and feedback when receiving multidimensional information. We hope to gain a more in-depth understanding of the UX quality of our model-induced prototype and to prove the validity of our IT model.

Participants. We recruited 20 participants (8 women, 12 men, age range 21-52, M=28.15, SD=9.38) from our university.

Procedure. We first asked the participants to sign the consent and instruction form. Then, they were asked to hold a five-minute online meeting with five other people using Zoom. The UX quality in this scenario will serve as a baseline for comparison. The next step is to get participants into the local CAVE to hold a conference with five remote parties shown in Fig. 4(d). After that, the scene switched to a four-party meeting. Surround sound and scent was added gradually so that the participants could feel the seamless integration of virtual, auditory , and olfactory senses in the conference room.

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2024.3352240

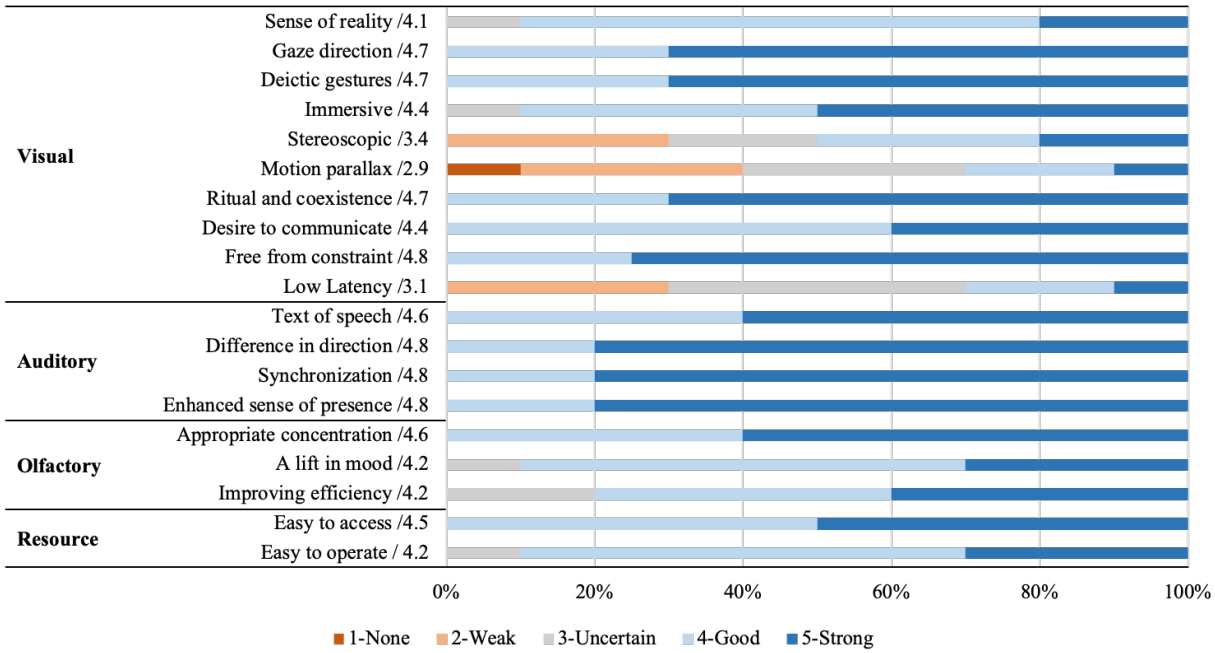

Fig. 5. The results of the user experience study, using more precise metrics of user experience from visual, auditory, olfactory, and resource dimensions.

In this scenario, they are also asked to make specified eye contact, gestures, and conversations just the same as in Zoom. In the end, participants filled out our UX questionnaire.

Questionnaire, analysis, and results. Our questionnaire is based on a 5-level Likert Scale and measures UX quality from four dimensions including 19 criteria. These metrics are all obtained through our intensive analysis of the existing VCSs and theories. The experimental results shown in Figure 5 meet our expectations. The model-induced prototype achieved more than 4 points in each sensory aspect and participants rated the overall experience at an average of 4.5 out of 5. Participants scored particularly high in free from constraint (4.8), gaze and gesture direction awareness, sense of ritual and coexistence (4.7), immersive, desire to communicate (4.4) as well as the perception of different directions of sound based on audio and video information (4.8). There is truly a lift in mood and concentration caused by odor release (4.6), and digital resources are easy to access (4.5).

We must admit that there are still shortcomings in some aspects, such as providing stereoscopic and low latency images, and achieving horizontal and vertical motion parallax. We will continue to improve them in the future.

VI. Conclusion and Future Works

In this work, we propose an Information Theory based model to characterize and evaluate the UX quality of the existing telepresence edge approaches from a novel perspective. The information flow changes from traditional machine-machine to user-machine-user. We also discuss insights and heuristic guidelines derived from this descriptive model for future remote collaboration improvement. Then, based on the above modal-based analysis, we present a prototype design under certain cost restrictions. Finally, we implement the prototype and the validity of our method is validated by extensive experiments and user studies. Although the model-based analysis is inevitably biased, our study provides meaningful references for future telepresence edge system design. For our future work, we will first improve our model by making more accurate and detailed estimates. Second, we would like to improve our design, particularly paying attention to cross-modality information fusion and AI-based analysis.

References

[1] "Tencent meeting," https://meeting.tencent.com, (Accessed on Sep. 2022).

[2] K.-I. Okada, F. Maeda, Y. Ichikawaa, and Y. Matsushita, "Multiparty videoconferencing at virtual social distance: Majic design," in *Proceedings of the 5th ACM Conference on Computer Supported Cooperative Work*, 1994, p. 385–393.

[3] A. Sellen, B. Buxton, and J. Arnott, "Using spatial cues to improve videoconferencing," in *Proceedings of the 10th SIGCHI Conference on Human Factors in Computing Systems*, 1992, p. 651–652.

[4] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung, "Gaze-2: Conveying eye contact in group video conferencing using eye-controlled camera direction," in *Proceedings of the 20th SIGCHI Conference on Human Factors in Computing Systems*, 2003, p. 521–528.

[5] D. Nguyen and J. Canny, "Multiview: Spatially faithful group video conferencing," in *Proceedings of the 23rd SIGCHI Conference on Human Factors in Computing Systems*, 2005, p. 799–808.

[6] Y. Zhang, J. Yang, Z. Liu, R. Wang, G. Chen, X. Tong, and B. Guo, "Virtualcube: An immersive 3d video communication system," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 5, pp. 2146–2156, 2022.

[7] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, V. Tankovich, C. Loop, Q. Cai, P. A. Chou, S. Mennicken, J. Valentin, V. Pradeep, S. Wang, S. B. Kang, P. Kohli, Y. Lutchyn, C. Keskin, and S. Izadi, "Holoportation: Virtual 3d teleportation in real-time," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, 2016, p. 741–754.

[8] K. Kim, J. Bolton, A. Girouard, J. Cooperstock, and R. Vertegaal, "Telehuman: Effects of 3d perspective on gaze and pose estimation with a life-size cylindrical telepresence pod," in *Proceedings of the 30th*

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2024.3352240

SIGCHI Conference on Human Factors in Computing Systems, 2012, p. 2531–2540.

[9] D. Gotsch, X. Zhang, T. Merritt, and R. Vertegaal, "Telehuman2: A cylindrical light field teleconferencing system for life-size 3d human telepresence," in *Proceedings of the 36th CHI Conference on Human Factors in Computing Systems*, 2018, p. 1–10.

[10] M. Gross, S. Würmlin, M. Naef, E. Lamboray, C. Spagno, A. Kunz, E. Koller-Meier, T. Svoboda, L. Van Gool, S. Lang, K. Strehlke, A. V. Moere, and O. Staadt, "Blue-c: A spatially immersive display and 3d video portal for telepresence," *ACM Trans. Graph.*, vol. 22, no. 3, p. 819–827, Jul 2003.

[11] "Workrooms — vr for business meetings," https://www.oculus.com/workrooms/, (Accessed on Sep. 2022).

[12] J. E. Grønbæk, B. Saatçi, C. F. Griggio, and C. N. Klokmose, "Mirrorblender: Supporting hybrid meetings with a malleable video-conferencing system," in *Proceedings of the 39th CHI Conference on Human Factors in Computing Systems*, 2021.

[13] O. Morikawa and T. Maesako, "Hypermirror: Toward pleasant-to-use video mediated communication system," in *Proceedings of the 7th ACM Conference on Computer Supported Cooperative Work*, 1998, p. 149–158.

[14] H. Ishii and M. Kobayashi, "Clearboard: A seamless medium for shared drawing and conversation with eye contact," in *Proceedings of the 10th SIGCHI Conference on Human Factors in Computing Systems*, 1992, p. 525–532.

[15] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[16] J. Hollan and S. Stornetta, "Beyond being there," in *Proceedings of the 10th SIGCHI Conference on Human Factors in Computing Systems*, 1992, p. 119–125.

[17] S. Chandrasegaran, C. Bryan, H. Shidara, T.-Y. Chuang, and K.-L. Ma, "Talktraces: Real-time capture and visualization of verbal content in meetings," in *Proceedings of the 37th CHI Conference on Human Factors in Computing Systems*, 2019, p. 1–14.

[18] S. Samrose, D. McDuff, R. Sim, J. Suh, K. Rowan, J. Hernandez, S. Rintel, K. Moynihan, and M. Czerwinski, "Meetingcoach: An intelligent dashboard for supporting effective & inclusive meetings," in *Proceedings of the 39th CHI Conference on Human Factors in Computing Systems*, 2021.

[19] Q. Zhao, "A survey on virtual reality," *Science in China Series F: Information Sciences*, vol. 52, no. 3, pp. 348–400, 2009.

[20] S. J. Gibbs, C. Arapis, and Christian, "Teleport – towards immersive copresence," *Multimedia Systems*, vol. 7, no. 3, pp. 214–221, 1999.

[21] H. H. Baker, N. Bhatti, D. Tanguay, I. Sobel, D. Gelb, M. E. Goss, W. B. Culbertson, and T. Malzbender, "Understanding performance in coliseum, an immersive videoconferencing system," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 1, no. 2, p. 190–210, May 2005.

[22] D. J. Sandin, T. Margolis, J. Ge, J. Girado, T. Peterka, and T. A. DeFanti, "The varriertm autostereoscopic virtual reality display," *ACM Trans. Graph.*, vol. 24, no. 3, p. 894–903, Jul 2005.

[23] P. Lincoln, A. Nashel, A. Ilie, H. Towles, G. Welch, and H. Fuchs, "Multi-view lenticular display for group teleconferencing," in *Proceedings of the 2nd International Conference on Immersive Telecommunications*, 2009, p. 1–8.

[24] Y. Pan and A. Steed, "A gaze-preserving situated multiview telepresence system," in *Proceedings of the 32nd SIGCHI Conference on Human Factors in Computing Systems*, 2014, p. 2173–2176.

[25] A. Jones, M. Lang, G. Fyffe, X. Yu, J. Busch, I. McDowall, M. Bolas, and P. Debevec, "Achieving eye contact in a one-to-many 3d video teleconferencing system," *ACM Trans. Graph.*, vol. 28, no. 3, Jul 2009.

[26] X. Zhang, S. Braley, C. Rubens, T. Merritt, and R. Vertegaal, "Lightbee: A self-levitating light field display for hologrammatic telepresence," in *Proceedings of the 37th CHI Conference on Human Factors in Computing Systems*, 2019, p. 1–10.

[27] "Test-driving willow garage's telepresence robot - cnet," https://www.cnet.com/culture/test-driving-willow-garages-telepresence-robot/, (Accessed on Sep. 2022).

[28] D. Sakamoto, T. Kanda, T. Ono, H. Ishiguro, and N. Hagita, "Android as a telecommunication medium with a human-like presence," in *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction*, 2007, p. 193–200.

[29] P. Lincoln, G. Welch, A. Nashel, A. Ilie, A. State, and H. Fuchs, "Animatronic shader lamps avatars," in *Proceedings of the 8th IEEE International Symposium on Mixed and Augmented Reality*, 2009, p. 27–33.

[30] M. Otsuki, K. Maruyama, H. Kuzuoka, and Y. SUZUKI, "Effects of enhanced gaze presentation on gaze leading in remote collaborative physical tasks," in *Proceedings of the 36th CHI Conference on Human Factors in Computing Systems*, 2018, p. 1–11.

[31] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, 1995, p. 39–46.

[32] J. Brooks, S. Nagels, and P. Lopes, "Trigeminal-based temperature illusions," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–12. [Online]. Available: https://doi.org/10.1145/3313831.3376806

[33] D. Dmitrenko, E. Maggioni, G. Brianza, B. E. Holthausen, B. N. Walker, and M. Obrist, "Caroma therapy: Pleasant scents promote safer driving, better mood, and improved well-being in angry drivers," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–13. [Online]. Available: https://doi.org/10.1145/3313831.3376176

[34] Y. Xiong and F. Quek, "Meeting room configuration and multiple camera calibration in meeting analysis," in *Proceedings of the 7th International Conference on Multimodal Interfaces*, 2005, p. 37–44.

[35] B. Sun, Q. Ma, S. Zhang, K. Liu, and Y. Liu, "iself: Towards cold-start emotion labeling using transfer learning with smartphones," *ACM Transactions on Sensor Networks (TOSN)*, vol. 13, no. 4, pp. 1–22, 2017.

[36] D. Deng, Z. Chen, Y. Zhou, and B. Shi, "Mimamo net: Integrating micro- and macro-motion for video emotion recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 2621–2628.

Ruiqing Wang is currently a dual master's degree student in Data Science and Information Technology at Global Innovation Exchange, Tsinghua University, Beijing, and University of Washington, Seattle. Her research interests include Human-Computer Interaction and AIoT.

Kebin Liu is a Research Associate Professor in Global Innovation Exchange, Tsinghua University, Beijing, China. He received his M.S. and Ph.D. degrees from Shanghai Jiaotong University, China. His research interests include Internet of Things, Pervasive Computing, and Network Diagnosis.

Ziyue Dang received his B.E. degree in computer science and technology from Tsinghua University, Beijing, in 2023. He is currently a graduate student in Computer Science Department, University of California, Los Angeles. His research interests include mobile computing, pervasive computing, and Internet of Things.

Xu Wang received his B.E. and Ph.D. degrees in software engineering from Tsinghua University, Beijing, in 2015 and 2020, respectively. He is a research assistant professor at the Global Innovation Exchange, Tsinghua University, Beijing. His research interests include the Industrial Internet, Edge Computing, and Internet of Things.

Fan Dang received his B.E. and Ph.D. degrees in software engineering from Tsinghua University, Beijing, in 2013 and 2018, respectively. He is a research assistant professor at the Global Innovation Exchange, Tsinghua University, Beijing. His research interests include the industrial Internet, edge computing, and mobile security.

Yue Sun received his bachelor's degree in electronic and computer engineering in 2021. He is currently a dual master's student pursuing a degree in data science and information technology at Global Innovation Exchange, Tsinghua University. His research interest is 3D computer vision.

Yuang Tong received his B.E. degree in Department of Automation from Tsinghua University, Beijing, in 2021. He is now a dual master's student at Global Innovation Exchange, Tsinghua University, Beijing, and University of Washington, Seattle.

Haitian Zhao received his Ph.D. degree from the School of Architecture, Tsinghua University, Beijing, in 2022. He is currently a Post-Doctoral Researcher in Department of Automation, Tsinghua University. His research interests include smart building and AIoT.

Yunhao Liu is Chair Professor at Tsinghua University. Yunhao received his B.S. degree in Department of Automation from Tsinghua University, an M.S. and a Ph.D. degree in Computer Science and Engineering from Michigan State University. He is a fellow of ACM and IEEE.

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.