A Framework of Delay Perception in 3D Tele-Immersion

Leave Authors Anonymous
Institute
City, Country
example@email.com

Leave Authors Anonymous
Institute
City, Country
example@email.com

Leave Authors Anonymous
Institute
City, Country
example@email.com

ABSTRACT

3D Tele-Immersion (3DTI, e.g., Holoportation [65]) allows distributed users to communicate and interact with each other in the same virtual space. It grows rapidly in recent years. However, no work has studied network delay perception in 3DTI. Network delay is an important factor that affects user experience. In this paper, we explore users' perception of network delay in 3DTI. We propose a conceptual framework that classifies 3DTI tasks into three levels by their network delay requirement: *synchronous tasks*, *turn-based audiovisual tasks* and *turn-based visual-only tasks*. They require a network delay within about 50 ms, 250 ms, and 300 ms respectively. There are two tendencies in 3D delay perception. First, more applications fall into the first level *synchronous tasks* in 3DTI. We should pay attention to support their very low network delay. Second, the users are less sensitive and more tolerable to the network delay in 3D conversation. Based on these findings, the framework gives suggestions for network design. Finally, we describe a controlled study as illustrating examples and validation.

CCS CONCEPTS

• **Human-centered computing** → *HCI theory, concepts and models*;

KEYWORDS

Delay perception; 3D tele-immersion

ACM Reference Format:
Leave Authors Anonymous, Leave Authors Anonymous, and Leave Authors Anonymous. 2019. A Framework of Delay Perception in 3D Tele-Immersion. In *Proceedings of ACM SigCHI conference (CHI'19).* ACM, New York, NY, USA, 13 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

The past centuries have witnessed the growth of communication technology. The invention of the telephone has saved a great deal of time and money by displacing physically face-to-face meeting. In recent two decades, 2D audiovisual communications are getting popular, such as teleconference [1, 56], telecollaboration [2, 17], robotic telepresence [36, 58, 59], and so on.

3DTI emerged in the past decade [46, 53, 54, 67]. It allows distributed users to communicate and interact with each other in the same virtual space. Both the improvement of pipeline and hardware make 3DTI hopeful to be practical in the near future. Microsoft's Holoportation [65] is a typical 3DTI pipeline with high-quality and real-time performance. For computing, GPUs are getting more powerful. For rendering, immersive displays such as Head-Mounted Displays (MHDs) are becoming popular.

Network delay is one of the critical problems in communication technology. On the one hand, network delay is a crucial factor that affects user experience [7, 72, 73, 82]. For example, the delay should be within 150 ms to provide a good user experience for an audio-mediated application [17, 69]. On the other hand, new communications generally require higher bandwidth. It challenges the requirement of low network delay [42].

Numerous works explored delay perception in telephone and 2D communications. However, no work has been done to study users' perception of network delay in an advanced 3DTI system, i.e., with reconstruction and rendering in full 3D. It is necessary to rebuild the framework of delay perception in 3D because there is a big difference between 2D and 3D. First, an advanced 3DTI system support co-present, that is, the two users feel like exactly at the same virtual space. Second, 3DTI offers much more visual information. These features may lead to unknown changes of delay perception in 3D.

In this paper, we explore users' perception of network delay in a full 3D tele-immersion system. We systematically reviewed previous works on delay perception in 2D. Previous works inspire us so that we had several hypotheses about the problem. To validate the hypotheses, we first built a 3DTI system with an ideal network delay of 50 ms, and then conducted a controlled study. Finally, we proposed a conceptual framework of network delay perception in 3DTI.

The framework suggests that users perceive network delay by cues. A user perceives the network delay if he feel that his partner response to the cues abnormally, because the perceived response time is mixed with the network delay.

CHI'19, April 2019, ElPaso, Texas USA
2019. ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

In previous works, there are mainly three types of cues: synchronous actions, conversation and visual feedback. Among them synchronous actions are the strongest cues to reveal the network delay, while the visual feedback is the weakest one. Thus, we classifies 3DTI tasks into three levels: *synchronous tasks*, *turn-based audiovisual tasks* and *turn-based visual-only tasks*. We suggest that they require network delays within about 50 ms, 250 ms and 300 ms respectively.

The framework infers significant changes of network delay perception in 3D. First, much more applications fall into the first level *synchronous tasks*. We should pay attention to support their low network delay. Second, the users are less sensitive and more tolerable to the network delay in a conversation. Based on these findings, we give suggestions on network design to improve the user experience.

Our contribution is threefold: first, the framework infers significant changes of network delay perception in 3D. 3DTI practitioners can assess their applications through our framework to improve the network design. Second, we give suggestions on network engineering for each situation. These suggestions can help saving network resource and improving the user experience. Third, our project is open-source [?]. We give the necessary explanation in the system overview to make sure that the readers can easily build up a similar system.

In the remainder of the paper, we first present our framework (section 2). We next give an overview of our 3DTI system (section 3). We then describe the controlled study (section 4). We supplement related works on 3DTI systems and existing studies on 3DTI delay perception (section 5). The paper concludes by discussing our limitation and the future work (section 6).

2 A CONCEPTUAL FRAMEWORK OF DELAY PERCEPTION IN 3DTI

The framework classify 3DTI tasks by their requirement of network delay. There are three levels:

- **(L1) *synchronous tasks***: The tasks with synchronous voices or actions, e.g., shaking hands, dancing together and musical collaboration.
- **(L2) *turn-based audiovisual tasks***: The tasks with turn talking and actions (the audio channel is available), e.g., teleconference and remote interview.
- **(L3) *turn-based visual-only tasks***: The tasks with turn actions (the audio channel is not available), e.g., playing chess and remote surgery simulation.

In this section, we first answer the question how users perceive network delay in communications. Next, we summarize the new features in 3D. Then, we predict two tendencies of delay perception in 3D: (1) more tasks fall into the most delay-sensitive level *synchronous tasks*; (2) users are less sensitive and more tolerable to network delay in a conversation. Last, we gives suggestions on network engineering for each situation.

How users perceive network delay?

Users perceive network delay by cues. The cues can be divided into two categories: synchronous cues and turn-based cues. A synchronous cue is that the two users speak or gesture at the very same time. It is the strongest cue to reveal the network delay. The user notices the delay once he feel that the partner acts slower than himself. Thus, we define the most delay-sensitive level of tasks as (L1) *synchronous tasks*, which involves synchronous cues in the interaction.

A turn-based cue is that the users take turn speaking or gesturing. The turn gap is defined as the time between a user finish a turn and his partner response. It is uncertain but somehow predictable for both the users. In a networked communication, the network delay prolongs the perceived turn gap. The perceived turn gap is the sum of actual turn gap and the round-trip delay. When the network delay is low, the user can not judge if the prolonged turn gap is caused by the network delay or by the slow response of his partner. When the network delay becomes higher, the user feels that the turn gap is abnormal, hence he notice the network delay. Because of the uncertain turn gap, turn-based cues are the weaker cues for delay perception compared to synchronous cues.

[图] maybe there should be a figure to illustrate that: perceived turn gap = actual turn gap + round-trip delay. The shorter and easier to predict the actual turn gap is, the easier the network delay can be perceived.

Turn-based cues contain conversation (with audio channel) and visual feedback (without audio channel). Previous theories suggests that the turn gap of a conversation is generally easier to predict compared to the visual feedback. The more predictable the turn gap is, the easier the user perceives the network delay. Correspondingly, related studies in 2D reveals that conversation make users sensitive to the network delay in a turn-based task [?]. Thus, we define the second delay-sensitive level of tasks as (L2) *turn-based audiovisual tasks*. We regard (L3) *turn-based visual-only tasks* as the less delay-sensitive one.

Next, we introduce the theoretical background to further explain the problem. We refer local delay perception to *synchronous tasks*, Turn-Talking Model and Grounding Theory to *turn-based audiovisual tasks*, and Situation Awareness Theory to *turn-based visual-only tasks*.

Local delay perception. For local delay, 100 ms is an upper boundary for users to fell that the system is running instantaneously [62]. For a better performance, a local delay of 30

ms to 50 ms is needed [11]. In a *synchronous task*, the users are expected to act exactly at the same time. The users can perceive the network delay if their actions are obviously out of sync. We deem that users' ability to perceive the delay in this situation is similar to the perception of a local delay.

Bad case: the round-trip delay. Many *synchronous tasks* require one of the users to synchronize the actions by gesturing or saying something, e.g., "Three, two, one, go!". As [29] revealed, there is a difference between the delay perceptions of the *caller* and the other user *replier*. The *caller* is the user who try to synchronize the actions. As figure 1 shown, the *caller* starts a Rock-Paper-Scissors game. The *replier* just adjusts the timing of his motion to the *caller's* timing. Thus, he do not perceive any network delay. In contrast, the *caller* experiences the round-trip delay before seeing the partner's reaction. The round-trip delay may destroy the user experience in *synchronous tasks*. We should balance the perceived network delay of the the two users.

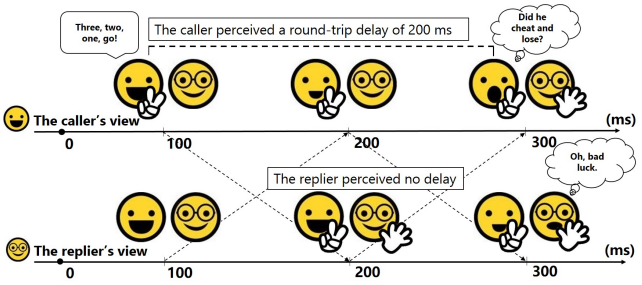

Figure 1: In a networked *Rock-Paper-Scissors* game with a delay of 100 ms, the *caller* is the player who try to synchronize the game, e.g., by saying "Three, two, one, go". He perceives a round-trip delay of 200 ms, while his partner perceives no delay.

Turn-Talking Model. Turn talking is a part of universal infrastructure for language [47]. In daily life we have learned to unconsciously manage a conversation by using the timing of the small pauses in speech [70]. Figure 2 shows the typical timing of the conversation. A Turn is 2 s in average. The language production system is slow: preparation before output begins takes 600 ms to 1500 ms [3, 27, 31]. However, switching of speakers is rapid, because the turn talking system relies on prediction [47]. The modal turn gap is around 200 ms [48].

As Turn-Talking Model suggests, the turn gap of a conversation is short and easy to predict because of our daily life training. As we explain above, the more predictable the turn gap is, the easier the user perceives the delay. Thus, conversation is also an important cue for delay perception. Previous studies have found that the users are sensitive to network

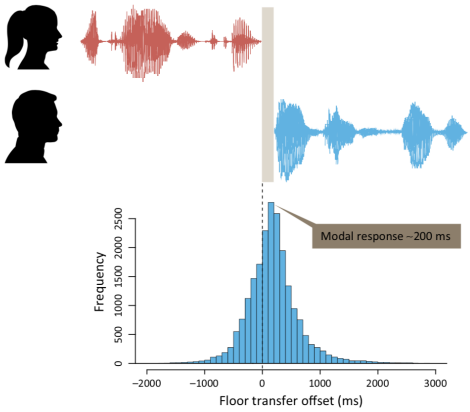

Figure 2: [注意，这是盗图] Turn Talking Model.

delay in a conversation. Delay of 100 ms is noticeable in audio communications [33, 69, 79]. Delay of 150 ms becomes industrial standard of the telephone network [69]. Delays greater than 450 ms can severely impact communication [24, 64, 77].

Situation Awareness Theory. Situation Awareness Theory holds that visual information helps pairs assess the current state of the task and plan future actions [20, 21]. A user need to maintain an on-going awareness of the partner' s actions and the status of the task objects. Through the observation of the turn gap, a user may perceive the network delay. However, the turn gap of an visual feedback is hard to predict. Thus, users are less sensitive to the network delay in a *turn-based visual-only task*. Previous works shows that the network delay requirement of the 2D *turn-based visual-only tasks* is from xxx ms to xxx ms [?, ?, ?].

Grounding Theory. Grounding Theory suggests that a successful communication relies on a foundation of mutual knowledge or common ground [12, 13]. Visual information is an important source of common ground that provides evidence of comprehension for communication [6, 44]. Sufficient visual information can assist the verbal communication [24] and reduce users' dependence on conversation. For example, a user can point at an object in the shared virtual space and refer to it using a simple pronoun "that". Because visual information reduces the dependence of conversation, we suggest that users is less sensitive to delay in 3D.

In conclusion, *synchronous tasks* require for the lowest network delay. The delay requirement of *turn-based tasks* are looser. Among *turn-based tasks*, the tasks with audio channel (conversation) require a lower network delay.

What is new in 3D?

There are two new features in 3D: the support of co-present and the enriched visual information. An advanced 3DTI system reconstructs the physical scene and render it in full 3D. It allows co-present, that is, the two users feel like exactly in the same virtual space. Meanwhile, 3DTI offer more visual information compared to the 2D systems, e.g., the user can perceive the distance to an object and observe it from different view points.

Co-present. Co-present is a critical feature in a full 3D tele-immersion system [44, 65, 80]. In previous 2D communications and even the 3DTI systems without HMDs, the interaction always occurs through a 'window' from one space into the other. Co-present allows the users to "touch" each other, to exchange their locations freely, to use the shared physical props [52] and so on. The significance of co-present is that 3DTI supports many impossible tasks in 2D or improve some existing tasks in a more natural manner. For example, as the users can "touch" each other in 3DTI, human can finally shake hands in distributed places.

Rich visual information. 3DTI renders the virtual space in 3D. A 3D environment offer richer visual information compared to the 2D situation. First, when the user move in the virtual space, his view dynamically changes according to his location. Second, the binocular visualization allows users to perceive distance in the virtual space. The significance of the rich visual information is twofold: first, it enables 3DTI to support more tasks; second, it further reduces users' dependence on conversation, which may lead to the less sensitiveness of network delay.

We give examples for each level of tasks. They show the possibility of 3DTI. Among the examples below, the tasks with the marker '*' is only supported by 3DTI:

- *Musical collaboration (L1)*: Development of audio transport over networked have supported professional-quality musical collaboration [8, 9]. The emerging 3DTI may support a vivid musical collaboration in the future.
- *Piano duet (L1)**: It is possible for a 3DTI system to fuse similar objects in both sides. Imaging there are two same pianos in two distributed rooms. 3DTI can fuse the two pianos into a virtual one, so that two distributed musicians can play piano duet.
- *Dancing together (L1)**: In the virtual space of 3DTI, two users can exchanges their locations freely. It allows them to dance together.
- *Shaking hands (L1)**: 3DTI allows the two users to "touch" with each other, so that they can shake hands.
- *The Rock-Paper-Scissors game (L1)*: The game requires the players to show a gesture at the same time. In physical world, a pair usually play in a very short distance. We deem that it is more natural in 3D communications.
- *Teleconference (L2)*: Teleconference is a typical tasks in 2D communications. It is a good example for remote conversation. 3DTI offers more visual information for the teleconference.
- *Building blocks (L2)*: It is a tele-collaboration task that a user imitates the other to build blocks. Though it is possible in 2D, it will be easier in 3D because the imitator can observe the remote blocks from different view points.
- *Remote interview (L2*: In an interview, it is important to preserve a dignified deportment. 3DTI offers more visual information about the interviewees.
- *Playing chess (L3)**: Each player interacted with a chessboard and chess pieces on his own side. 3DTI fuses the two scenes into a virtual one, so that each player could see both sides of chess pieces.
- *Surgery simulation (L3)**: The surgery simulation is between a nurse and a surgeon. For example, the nurse can prepare tools and materials in advanced by observing the actions of the surgeon.

More tasks fall into the most delay-sensitive level *synchronous tasks*

3DTI supports co-present and provides more visual information. As the examples shown above, 3DTI is possible to support more tasks or improve them in a more natural manner. The new supported tasks is mostly belongs to L1 and L3. The reason for the increasing L1 is that most of these tasks require for co-present. The increment of L3 is because visual information is usually not sufficient in 2D communication.

In particular, we have to pay attention to the increasing L1 (*synchronous tasks*) in 3D, because their network delay requirement is very challenging. Table xxx shows some previous studies on delay perception of *synchronous tasks*. We suggest that a network delay of about 50 ms is required.

————————

Musical collaboration [74] 30 ms ~ 50 ms
The Rock-Paper-Scissors game [29] 40 ms ~ 70 ms

————————

[表格xxx] The recommend network delay is the threshold that can provide a good experience. We summarize it according to the standard of 3.5 Mean Opinion Score (MOS) [19, 71] or the description of the paper.

The less sensitiveness and more tolerance to network delay in 3D conversation

In 2D, video weaken the negative impact of delay in remote interactions, because audiovisual interaction allows users to

see visual information [76]. As table xxx shown, there is a obvious tendency that tasks rely more on the audio channel require for a lower network delay.

————————

turn talking [41] 150 ms
The Rock-Paper-Scissors game [29] 40 ms ~ 70 ms
3D Visual Communication [82] 120 ms [??? bad case]
Video Group Discussion [72] 500 ms
Audiovisual telecommunication [76] 500 ms

————————

[表格xxx] The recommend network delay is the threshold that can provide a good experience. 在总结论文的时候，作者人工给这些tasks定性：纯语音、偏语音、语音视频都重要、偏视频或是纯视频。

We deem that this effect enhances in 3D. According to Turn-Talking Model and Grounding Theory, the users are more sensitive to the delay of conversation compared to visual feedback. Visual information reduce users' dependence on audio communication, so the sensitiveness of delay is decreased accordingly. As [44, 45] suggests, when all parties to the interaction are co-present, the users share a rich visual space. Thus, 3DTI systems with co-presence provide the richest visual information. We suggest that the users are less sensitive and more tolerable to the network delay in 3D.

We suggest that in 3D, the network delay requirement is 250 ms for (L2) *turn-based audiovisual tasks* and 300 ms for (L3) *turn-based visual-only tasks*.

Suggestions for network design

There is a space to improve the network design for a 3DTI system. Both the system service and the user experience in 3D are different from the 2D situation. On the one hand, the computation is tough in 3D, which leads to a generally longer computing time. The bandwidth requirement is also larger (1.5-fold of a 1080p video [?]); On the other hand, users' perception of network delay change a lot in 3D. The practitioners can suit their methods to the situation when designing the networked applications:

Zero-delay audio assistance (for L1). In *synchronous tasks*, the users are sensitive to the network delay. Moreover, as we explained above, the *caller* may perceive a round-trip network delay. To improve the user experience, we suggest to add a zero-delay audio prompt tone in the application. In the networked *Rock-Paper-Scissors* game, for example, we can add synchronized sounds of "tick, tick, tack" for both the users. The users can show their gesture when they hear the "tack" sound so that the actions can be exactly at the same time. The advantage of this strategy is twofold: first, to avoid the round-trip network delay; second, the psychological hint of fair game.

Though the network delay is unavoidable in the audiovisual transmission, it is possible to accurately synchronize the timing of two systems through the Network Time Protocol (NTP) [57]. Thus, this strategy is practicable. In our experiment, we validated that this strategy can significantly improve the user experience.

Trade bandwidth for time (for L1). In general, delay and bandwidth are a trade-off in network transmission. Compression is the key. It reduce the bandwidth requirement of a 3DTI network. Recent works on 3D data compression shows that a bandwidth of tens of megabits per second is possible to support a 3DTI system [14, 16]. However, compression is time-consuming. Some compression methods are based on inter prediction, which leads to several frames of latency.

The *synchronous tasks* require a network with delay of 50 ms. The service has almost no time for compression. In this situation, we can trade bandwidth for time, i.e., to use a lightweight compression method, or even transmit the raw data directly (about 500 Mbps). This strategy can reduce the network delay as well as maintain the highest quality. The price of this strategy is the very high requirement of network bandwidth. Thus, it is only possible if the practitioners have a dedicated network.

Buffer frames to recover lost packets and deal with jitter (for L2 and L3). Delay is not the only factor that affects the user experience in a networked communication. Jitters, delay spikes and network losses will degrade the user experience as well. Fortunately, the network requirement of turn-based tasks are looser in 3D. Compared to the 2D situation, we have extra 50 ms ~ 100 ms for recovering lost packets and smoothing data stream .

Trade time for bandwidth (for L2 and L3). As we explained above, delay and bandwidth are a trade-off. For turn-based tasks, we can trade time for bandwidth, i.e., to use a heavyweight compression method. The state-of-the-art 3D reconstruction pipelines [18, 65] track live 3D model based on temporal consistency. It provide convenience for the compression of data stream, i.e., we can leverage inter prediction in the transmission.

The audio can be earlier (for L2). In a *turn-based audiovisual task*, the transmission of audio is more lightweight than that of the video. Coincidentally, small delay can seriously disrupt the audio communication [43], while the delay requirement of the video is looser. However, most video conference systems synchronize video and audio by delaying audio, which reduces the responsiveness of the conversation [?, ?, ?, ?]. [32] suggests that we can transmit the audio a little bit faster than the video. Even if the gap between audio and video is noticeable, this strategy can somehow improve

the user experience. In the 3D situation, we suggest a gap between audio and video within 100 ms is acceptable.

3 SYSTEM OVERVIEW

Our 3DTI system fuses two distributed scenes into a same virtual space in full 3D. Figure 3 illustrates the functionality of our system. The end-to-end delay is 50 ms, i.e., the time interval between a user acts and his remote partner sees.

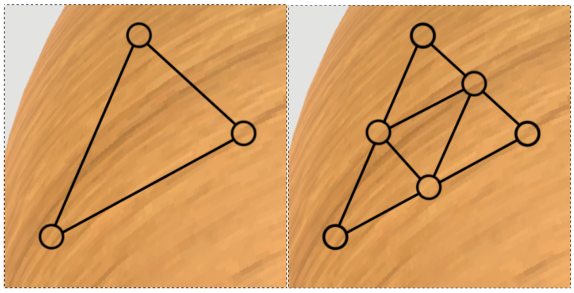

Figure 3: The functionality of our system.

In summary, a 3DTI system requires three processes: reconstruction, transmission and rendering [23]. For 3D reconstruction, we applied Truncated Signed Distance Function (TSDF) Volume [15] and Marching Cubes [51]. We did not focus on transmission as [4, 66] did, but used a 10 Gigabit Ethernet connection instead. For rendering, we applied HTC Vive (HMD) because it supports co-present interaction.

There are three features in our implementation. First, it reaches an end-to-end delay of 50 ms, which is responsive enough to support a delay perception research. Second, the system removes unnecessary background and retains only individuals and shared objects. Two shared objects in distributed places are fused into one virtual object. Third, the system is inexpensive and easy to build up. The price is about $ 7000 for two ends. The project is open-source [?]. We also provide a Unity plugin for the easy application development.

Hardware and Software Overview

Hardware. The system consists of two capture sites in the two distributed rooms. At each capture site, we had three depth cameras for capturing, a PC for computing and an HMD for rendering. Realsense D415 (depth cameras) were used to capture a volume of $2m \times 2m \times 2m$. The locating place of each camera and its contribution to 3D mesh are illustrated in Figure xx. Each PC had an Intel i7-7700k CPU and a GTX 1080Ti GPU. HTC Vive was used to present the fused reconstruction of both sides. Ten Gigabit network cards (Intel X520-SR2) were used to connect the two capture sites.

Software. OpenCV was used for camera calibration. CUDA was used for image processing and the kernel algorithm. Unity3D was used to implement the high-level application. It fetches live reconstruction from the kernel and renders it in HTC Vive. Python was used for audio transmission.

Calibration

Calibration between Cameras. The *camera calibration module* in OpenCV was used to calibrate the cameras. Each pair of cameras took ten snapshots (1080p color images) of a glass-made flat checkerboard. Then, OpenCV aligned their coordinates ($SD < 1pixel$).

Calibration between HMD and Cameras. The HTC Vive was calibrated by setting the original point in its software. We placed the original point of the camera coordinates at the same position by using the checkerboard. Hence, we aligned the HTC Vive with the cameras. This calibration is not necessarily accurate because the users can hardly perceive the error [?]. This step also aligned the coordinates of the two capture sites.

Preprocessing

Depth Processing. The cameras acquired depth images of 640×480 pixels at 30 FPS. The Realsense D415 is based on binocular disparity. Thus, disparity values (instead of depth values) were used in the processing to improve accuracy. We applied median filtering, spatial filtering, hole filling and temporary filtering on the depth images.

Color Processing. The cameras acquired color images of 960×540 pixels at 30 FPS. The exposure settings were manually adjusted. We used one RGB camera as a reference and warped the other cameras to this reference by white balancing and linear mapping.

Background Removal. In the calibration step, we recorded the background as RGBD images. At runtime, we removed pixels that are similar to the background based on thresholds.

3D Reconstruction

3D reconstruction is the kernel algorithm of a 3DTI system. We developed a real-time CUDA implementation of 3D reconstruction similar to KinectFusion [34]. First, the algorithm integrated depth images into a TSDF Volume [15]. Next, the 3D mesh was extracted from the TSDF Volume using Marching Cubes [51]. Then, the algorithm projected color images on the 3D mesh for colorization.

The resolution of TSDF volume was $256 \times 256 \times 256$ voxels. In the TSDF processing, we used a weighted average where $W = \frac{1}{Dist}$ on different cameras to minimize the error. In the colorization, we upsampled each triangle to four quartered parts to sample more colors (Figure 4). Because the users are more sensitive to the texture but not the shape [?].

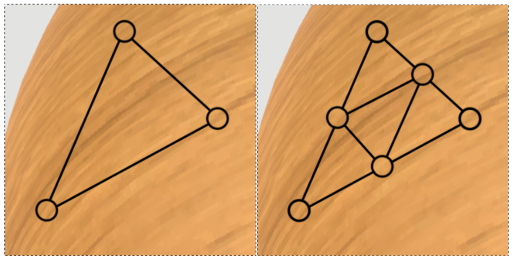

Figure 4: Left: Mesh without Supersampling; Right: Mesh with Supersampling.

We modified the TSDF algorithm to merge 3D meshes from the two capture sites. Figure 5 shows the weighted combination of the two profile from both sites. The merging rule is:

$$V_z = \min\{\frac{\sum_{local} W_{i,z} S_{i,z}}{\sum_{local} W_{i,z}}, \frac{\sum_{remote} W_{j,z} S_{j,z}}{\sum_{remote} W_{j,z}}\} \quad (1)$$

$$W_{i,z} = \frac{1}{dist(i,z)} \quad (2)$$

where $S_{i,z}$ is the Signed Distance Function (SDF) value of the zth volume from the ith camera. $dist(i,z)$ is the distance from the ith camera to the zth volume. V_z is the merged SDF value of zth volume after the combination.

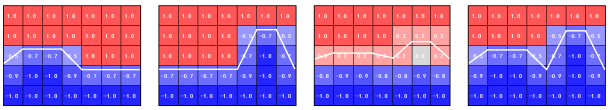

Figure 5: Left to right: 1)SDF value from camera 1; 2)Merged SDF value from camera 2; 3)SDF value by simple average; 4)Merged SDF value from our project.

End-to-end Delay

The lowest end-to-end delay of our system was 52 (± xx) ms. The frame rate of 3D reconstruction was 30 FPS. It was controlled by the frequency of the synchronized depth cameras. In average, a frame (33 ms) consisted of 19 ms processing and 14 ms idling. The remote images had one frame of latency. So the end-to-end delay was about $33+19 = 52$ ms. For artificial delay, we buffered remote data for frames. Figure 6 shows the pipeline in details. The rendering and the audio transmission were independent to the reconstruction pipeline. The frame rate of rendering reached 90 FPS so that the users do not feel dizzy. The audio channel was synchronized with the video channel by waiting.

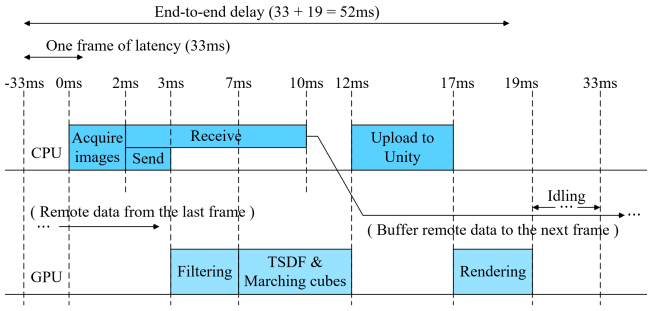

Figure 6

4 USER EXPERIMENT

The experiment has two parts. Part A is chess game tasks in two situation: with and without audio conversation. The goal of part A is to measure the network requirement of a *turn-based audiovisual task* and a *turn-based visual-only task* in 3D.

Part B is the Rock-Paper-Scissors game. It is a *synchronous task* that requires a low network delay. We tested the delay perception in two situation: with and without the zero-delay audio assistance. The goal of part B is to validate the effect of the assistance on user experience.

Part A: Playing Chess

Part A was to study the variety of delay perception in different synchronization level. The task was a chess game between pairs of participants in two rooms, with two conditions of cues: audiovisual mode and visual only mode.

Experimental design. We used a within-subjects experimental design. Each pair of participants played chess in two sessions of Communication Channel (CC): *audiovisual CC* and *visual only CC*, which correspond to the 2nd and 3rd synchronization level. The CC conditions were assigned to participant pairs in a Latin square design. Delay was another within-subjects factor with five conditions (50, 150, 250, 450, 750 ms in *audiovisual CC*; 150, 450, 1050, 1550, 2050 ms in *visual only CC*). In each session, we tested the five delay conditions in five trials. The delay conditions were assigned in a random order. In particular, we encouraged participants to chat in the *audiovisual CC*.

Task. In each trial, two distributed participants played chess "face-to-face" for three minutes. We adopted *Reversi* as the task. *Reversi* is simple enough that the participants can learn it in a short time. The game involves frequent interaction: when capturing, a participant should ask his partner to remove the captured chess. The participants have enough chances to perceive a noticeable delay.

In the physical world, each player interacted with a chessboard and chess pieces on his own side. The 3DTI system fused the two physical scenes into a virtual space. In the virtual space, each player could see not only his chess pieces but also his partner' s chess pieces.

Part B: Rock-Paper-Scissors

Part B was to evaluate the impact of delay in a high delay requirement situation. We used the Rock-Paper-Scissors game. We designed a synchronized audio cue to help users to synchronize with each other. The study assesses its effect on user experience.

Experimental design. This part was also a within-subjects design. Each pair played the Rock-Paper-Scissors game in two sessions: with and without the synchronized audio cue. We applied Latin square to the two sessions. In each session, we tested the delay of 50, 83, 117 and 150ms in four trials. The order of delay conditions was random. We also adopted Latin square on part A and part B.

Task. In each trial, a pair continuously drew the Rock-Paper-Scissors gestures until one of them won for ten times. There were two conditions to test: with and without the synchronized visual cue. In an actual network, it is possible to synchronize the time of two systems with almost zero milliseconds apart (the NTP protocol [?]).

Our 3DTI system provided zero-delay audio cues for both users to help them gesture exactly at the same time. The cue was an audio source of four seconds, with "tick" sounds at the 2nd, 3rd second and a "tack" sound at the 4th second. We told participants to gesture when their hear the "tack" sound.

Participants

We advertised our experiment on social media. Sixteen pairs of participants took part in our experiment (32 in total, 7 females). They all came from the campus, aged from 18 to 24. Participants were paid 150 yuan for the 90 minutes long study. The ten participants with the most conversation turn received extra 50 yuan.

Previous works have pointed out that the individual user differences affect study results of delay perception [?]. In our experiment, we control the source of participants carefully:

- *Relationship*: Each pair of participants are familiar with each other (friends, classmates or partners). This setting is to improve the conversation quality.
- *First language*: All the participants are native Chinese speakers. Their conversations are in Chinese in the experiment.
- *Experience in DIME*: Our participants have relatively high education levels. According to the self-report questionnaire, they are quite familiar with audiovisual multimedia (xx points in average, 5 for experts) and AR/VR (xx points in average).

Thus, our study results are rigorous but relatively low in the external validity. We recommend a larger amount of participants if the readers need a more general result.

Procedure

Before the experiment, we invited the participant pair to a room and explained our study. We explained the rule of *Reversi* and *Rock-Paper-Scissors*. Next, the pair had ten minutes to experience the physical interaction of these two games. We asked the participants to remember the feeling of physical interaction and regard it as a zero-delay experience. Then, we introduced our experimental procedure to the participants.

Part A had $2sessions \times 5trials = 10trials$, which lasted for 40 minutes. Part B had $2sessions \times 4trials = 8trials$ (20 minutes). In each trial, the participants experience the remote VR game. After each trial, participants filled in a short survey and rested for one minute. The questions in the survey are shown in Table 1.

Label	Question	Scale
quality	How do you feel during the experiment?	Bad <–> Excellent
noticeability	Can you perceive the delay in the connection?	Very much <–> Not at all
tolerance	To what extent where you annoyed by the delay?	Severe annoyance <–> No annoyance

Table 1: Questions and scale.

After each session, participants had a five minutes break. We conducted brief interviews with some subjective questions as followed:

- How do you notice the delay? What are the cues?
- What makes you annoyed in the task?
- Any other comments?

In particular, we explained to participants about the concept of network delay. We told the participants that we do not consider the local delay in the questionnaire. In the conversation, the users are forbidden to talk about the network delay directly.

Results

[NOTE from zsyzgu] I have several ideas about the result:

- In 3DTI, the users are much more tolerable than in audiovisual DIME. The immersive environment makes users focus on the interaction itself.

- A simple synchronized audio cue can help synchronization in tasks with high network requirement.
- Conversation is a stronger cue compared with visual feedback.
- Delay perception of 50 ms and 150 ms in audiovisual model have no significant difference.

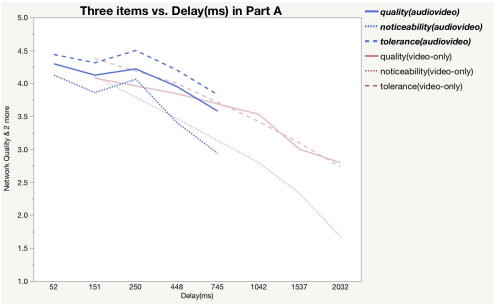

Figure 7: Effects of delay in part A(Playing Chess).

Figure 7 shows the results for the three questionnaire items in part A. We tried to find out when participants started to notice the delay in audio-video and video-only tasks, so we performed a pairwise comparison between the minimum delay and other delays in each mode.

In audiovideo mode, the analysis shows that network quality is not significantly different($p = 0.1914$) between 50ms and 150ms, then becomes significantly different($p = 0.042$) at 450ms. For noticeability, the results follow a similar pattern. Noticeability is not significantly different($p = 0.1592$) between 50ms and 150ms, then becomes significantly different($p = 0.0061$) at 450ms. For annoyance, the result is still not significant($p = 0.1012$) at 450ms and becomes significantly($p = 0.0058$) until 750ms.

In video-only mode, the analysis shows that network quality is not significantly different($p = 0.1546$) between 150ms and 450ms, then has a significant difference($p = 0.0079$) at 1050ms. For noticeability($p = 0.0114$) and annoyance($p = 0.0326$), the result is already significant at 450ms.

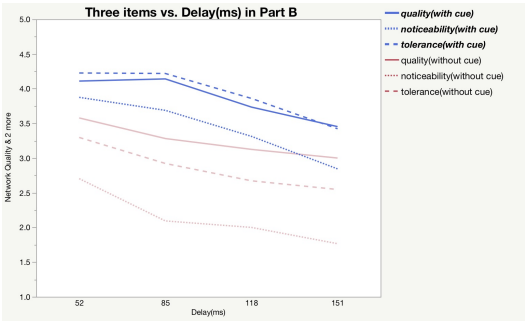

Figure 8: Effects of delay in part B(Rock-Paper-Scissors).

Figure 8 shows the results for the three questionnaire items in part B. We used the same method as Part A to analyze delay perception in two modes. When participants can hear synchronized audio, they gave three items significantly($p < 0.05$) lower score at 117ms, but not significant($p > 0.2$) at 87ms. Without the synchronized audio, noticeability score has significant($p = 0.0174$) drop at 87ms, but for network quality($p = 0.0451$) and annoyance($p = 0.0146$), the first significant drop shows at 117ms.

Through analysis we can get the following conclusions: in 3DTI, users have more tolerance to turn-based audiovisual tasks, they still don't notice the delay at 150ms in audiovideo mode, so the recommend network delay can be increased to 200 250ms. The recommend network delays of synchronized tasks and turn-based visual-only tasks appear to rise slightly. The synchronized audio cue can help synchronization in synchronized tasks, user experience is improved with the synchronized audio.

Observation and Interview

We summarized observation notes and transcripts of interviews. The main findings are as follow.

The effect of immersion in 3D. The 3DTI system support co-present and more visual information, which contribute to a higher level of immersion. Is the effect of immersion always positive? The answer is almost but not quite positive.

On the one hand, the immersion helps expression and comprehension in a communication. Most participants show body language to assist verbal communication, e.g., leaning forward before speaking; shrugging to express confusion. Some players often change their view point to observe the composition of the chess game.

On the other hand, the immersion confuse the users. Most participants confuse the "real" objects with the purely virtual objects. Even being informed of the virtuality of everything from the other end, most participants unintentionally tried to move other's chess pieces. "*I know it is no use to move the black chess pieces, but the behavior is just like a kind of instinctive reaction. The pieces are just being there, which seems to be of no difference with the chess in my hand.*" This observation enlighten us to design visual cues for distinguishing the objects from both ends.

How the participants perceive network delay? Though result shows a tendency that users are more sensitive to network delay with audio channel. Few participants reported that they perceive network delay by cues in the conversation. It indicates that the ability to perceive network delay from conversation is somehow unconscious.

In the Reversi, the expectation of removing captured pieces is reported as the most important cue for perceiving the delay. "*Until she removed the captured pieces, I repeated again*

and again, with my fingers pointing to the target chess. She couldn't have been thinking at that time. That must have been the delay, I bet." Some other cues are dependent on participants' relationship and experience. "*I said the funny word we invented together, and I knew she would have laughed out loud, but she didn't, until nearly one second latter.*"

In the Rock-Paper-Scissors game, a user can easily perceive the network delay by the observation that the partner throws slower then himself. "*It sounds ridiculous, but it does happen. I feel as if my partner throws paper after seeing I have thrown scissors.*"

The chess game sometimes becomes a synchronous task accidentally. We asked the participants not to talk about network delay directly. However, some participants were somehow foul because they misunderstood the goal of our experiment. In the Reversi, two pairs of participants consulted strategies to spy the network delay. a pair kept reversing the chess piece until the partner places the piece. The other pair appointed to keep their hands hovering on the chessboard until the partner places the piece. Both the two pairs give critical opinion scores to the network service. We finally removed their data from the result analysis. This observation shows that a task is easy to fall into the delay-sensitive level *synchronous task*.

The psychological hint of fair game with sync assistance. In the Rock-Paper-Scissors game, the threshold that provides good experience is xxx ms with sync assistance and xxx ms without the assistance. This gap is larger than one fold, while the gap between actual network delays is at most one fold. This phenomenon can not be explained by the network delay itself. We deem that it is because of the psychological hint of fair game with sync assistance. "*In the mode with sync assistance, I knew that the game is fair, so the network delay is tolerable for me.*"

5 RELATED WORK

In related work, we first review 3DTI techniques. The review shows that our system is in the mainstream, but sacrifices inter prediction for the responsiveness. Next, we review existing studies involving delay perception in 3DTI. These work do not exactly focus on delay perception. Moreover, they are limited in immature techniques (e.g., co-present not supported).

3D Tele-immersion

Optimal techniques toward 3DTI became clear in the last decade. Basically, a 3DTI system requires three processes: reconstruction, transmission and rendering [23]. For 3D reconstruction, the volumetric methods have become mainstream, e.g., TSDF Volume [15], Marching Cubes [51] and Fusion4D [18]. We do not focus on transmission as [4, 66] did. For rendering, we recommended Head-Mounted Display (HMD) because it supportd co-present and a high level of immersion.

3D Reconstruction. In early works, 3D reconstruction is either an off-line concept [15, 51] or some simple polygonal models that will look correct [22, 25, 38]. After 2000, researchers designed quasi real-time methods based on point cloud [28, 78] and triangulation [4, 46, 53, 67]. With the development of GPU, researchers achieved the real-time performance of high-quality reconstruction in the last decade [34, 50, 54]. Microsoft's KinectFusion [34] is the representative one. Numerous works improved 3D reconstruction within the framework of KinectFusion in regions of scale [10, 63], noise reduction [39, 60, 61] and so on.

In 2016, Microsoft proposed the state-of-the-art pipeline Fusion4D [18] and the matched system Holoportation [65]. The fourth dimension is time, indicating that it leverages inter prediction. However, Fusion4D is complex, not open-source and not responsive enough to study delay perception. Thus, we finally apply a 3D reconstruction method similar to KinectFusion.

3D Rendering. There are three categories rendering techniques in 3DTI: light field displays, Spatially Immersive Displays (SIDs) and HMDs. The light field displays [26, 35, 37, 40] suffers from low resolution because of its huge computation. SIDs are earlier. It becomes increasing significant around year 2000 [5, 25, 28, 28, 46, 78]. Recently, HMDs become popular. More 3DTI systems tend to apply HMDs for 3D rendering [49, 55, 65, 75].

Among HMDs, it is hard to give a choice between Augmented Reality (AR) and Virtual Reality (VR). In 2018, Microsoft proposed Remixed Reality [49]. This approach leverages both the benefits of AR and VR by rendering live 3D reconstruction of physical scene in VR. Finally, we applied VR (HTC Vive) for rendering.

Delay Perception in 3DTI

Most 3DTI works focus on algorithm and pipeline. Negative impacts of large delay are widely reported [4, 25, 46, 53, 68]. However, only a few works were conducted to study delay perception in 3DTI [30, 81, 82]. These works do not exactly focus on delay perception. Moreover, they are limited by single scenario and immature techniques, e.g., the 2D screen was used to display 3D scenes. Thus, we are the first to explore delay perception in an advanced 3DTI system, i.e., with reconstruction and rendering in full 3D.

6 LIMITATION AND FUTURE WORK

Our work has some limitations. First, the defects of our system implementation: while our system is very responsive,

the reconstruction quality is not state-of-the-art; the head-mounted display gets in the way of eye contact. Second, our experiment is low in external validity, e.g., the limited source of participants and the limited number of tasks. The experiment is an illustrating example that supports our framework, but can not validate the framework.

In future works, there are three directions. First, we can conduct more user studies to further investigate the delay perception in 3DTI, e.g., can the three levels be subdivided? Second, we can improve the reconstruction quality of our system. We deem that the effect of rendering quality on user experience is also a interesting research problem, e.g., to what extent is the texture more important than the shape? Third, we may try some low cost methods, e.g., can a 3D reconstruction based on point cloud provide good user experience?

7 CONCLUSION

[NOTE] In this paper, what we did? The framework give three levels, find the two tendencies of delay perception in 3D and give suggestions.

ACKNOWLEDGMENTS

We thank all the volunteers, and all publications support and staff, who wrote and provided helpful comments on previous versions of this document. Authors 1, 2, and 3 gratefully acknowledge the grant from NSF (#1234–2012–ABC). *This whole paragraph is just an example.*

REFERENCES

[1] David L Allen and Herold Williams. 1996. Teleconferencing method and system for providing face-to-face, non-animated teleconference environment. US Patent 5,572,248.

[2] Ignacio Avellino, Cédric Fleury, and Michel Beaudouin-Lafon. 2015. Accuracy of deictic gestures to support telepresence on wall-sized displays. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2393–2396.

[3] Elizabeth Bates, Simona D'Amico, Thomas Jacobsen, Anna Székely, Elena Andonova, Antonella Devescovi, Dan Herron, Ching Ching Lu, Thomas Pechmann, Csaba Pléh, et al. 2003. Timed picture naming in seven languages. *Psychonomic bulletin & review* 10, 2 (2003), 344–380.

[4] Stephan Beck, Andre Kunert, Alexander Kulik, and Bernd Froehlich. 2013. Immersive group-to-group telepresence. *IEEE Transactions on Visualization and Computer Graphics* 19, 4 (2013), 616–625.

[5] Hrvoje Benko, Ricardo Jota, and Andrew Wilson. 2012. MirageTable: freehand interaction on a projected augmented reality tabletop. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 199–208.

[6] Susan E Brennan. 2005. How conversation is shaped by visual and spoken evidence. *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (2005), 95–129.

[7] Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi, et al. 2013. Qualinet white paper on definitions of quality of experience. (2013).

[8] Alexander Carôt, Pedro Rebelo, and Alain Renaud. 2007. Networked music performance: State of the art. In *Audio engineering society conference: 30th international conference: intelligent audio environments*. Audio Engineering Society.

[9] Alexander Carôt and Christian Werner. 2007. Network music performance-problems, approaches and perspectives. In *Proceedings of the "Music in the Global Village"-Conference, Budapest, Hungary*, Vol. 162. 23–10.

[10] Jiawen Chen, Dennis Bautembach, and Shahram Izadi. 2013. Scalable real-time volumetric surface reconstruction. *ACM Transactions on Graphics (ToG)* 32, 4 (2013), 113.

[11] Jessie YC Chen and Jennifer E Thropp. 2007. Review of low frame rate effects on human performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 37, 6 (2007), 1063–1076.

[12] Herbert H Clark. 1981. Definite reference and mutual knowledge. *Elements of discourse understanding* (1981).

[13] Herbert H Clark and Deanna Wilkes-Gibbs. 1990. Referring as a collaborative process. *Intentions in communication* (1990), 463–493.

[14] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 69.

[15] Brian Curless and Marc Levoy. 1996. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 303–312.

[16] Ricardo L de Queiroz and Philip A Chou. 2016. Compression of 3d point clouds using a region-adaptive hierarchical transform. *IEEE Transactions on Image Processing* 25, 8 (2016), 3947–3956.

[17] Angus Donovan, Leila Alem, Weidong Huang, Ren Liu, and Mark Hedley. 2014. Understanding How Network Performance Affects User Experience of Remote Guidance. In *CYTED-RITOS International Workshop on Groupware*. Springer, 1–12.

[18] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. 2016. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 114.

[19] Thomas Enderes, Swee Chern Khoo, Clare A Somerville, and Kostas Samaras. 2002. Impact of statistical multiplexing on voice quality in cellular networks. *Mobile networks and applications* 7, 2 (2002), 153–161.

[20] Mica R Endsley. 2017. Toward a theory of situation awareness in dynamic systems. In *Situational Awareness*. Routledge, 9–42.

[21] Mica R Endsley and Daniel J Garland. 2000. *Situation awareness analysis and measurement*. CRC Press.

[22] Henry Fuchs, Gary Bishop, Kevin Arthur, Leonard McMillan, Ruzena Bajcsy, Sang Lee, Hany Farid, and Takeo Kanade. 1994. Virtual space teleconferencing using a sea of cameras. In *Proc. First International Conference on Medical Robotics and Computer Assisted Surgery*, Vol. 26.

[23] Henry Fuchs, Andrei State, and Jean-Charles Bazin. 2014. Immersive 3d telepresence. *Computer* 47, 7 (2014), 46–52.

[24] Darren Gergle, Robert E Kraut, and Susan R Fussell. 2006. The impact of delayed visual feedback on collaborative performance. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 1303–1312.

[25] Simon J Gibbs, Constantin Arapis, and Christian J Breiteneder. 1999. TELEPORT–Towards immersive copresence. *Multimedia Systems* 7, 3 (1999), 214–221.

[26] Daniel Gotsch, Xujing Zhang, Timothy Merritt, and Roel Vertegaal. 2018. TeleHuman2: A Cylindrical Light Field Teleconferencing System for Life-size 3D Human Telepresence. In *Proceedings of the 2018 CHI*

Conference on Human Factors in Computing Systems. ACM, 522.

[27] Zenzi M Griffin and Kathryn Bock. 2000. What the eyes say about speaking. *Psychological science* 11, 4 (2000), 274–279.

[28] Markus Gross, Stephan Würmlin, Martin Naef, Edouard Lamboray, Christian Spagno, Andreas Kunz, Esther Koller-Meier, Tomas Svoboda, Luc Van Gool, Silke Lang, et al. 2003. blue-c: a spatially immersive display and 3D video portal for telepresence. In *ACM Transactions on Graphics (TOG)*, Vol. 22. ACM, 819–827.

[29] Yousuke Hashimoto and Yutaka Ishibashi. 2006. Influences of network latency on interactivity in networked rock-paper-scissors. In *Proceedings of 5th ACM SIGCOMM workshop on Network and system support for games*. ACM, 23.

[30] Zixia Huang, Ahsan Arefin, Pooja Agarwal, Klara Nahrstedt, and Wanmin Wu. 2012. Towards the understanding of human perceptual quality in tele-immersive shared activity. In *Proceedings of the 3rd Multimedia Systems Conference*. ACM, 29–34.

[31] Peter Indefrey and Willem JM Levelt. 2004. The spatial and temporal signatures of word production components. *Cognition* 92, 1-2 (2004), 101–144.

[32] Ellen A Isaacs and John C Tang. 1994. What video can and cannot do for collaboration: a case study. *Multimedia systems* 2, 2 (1994), 63–73.

[33] T ITU. 2003. Recommendation G. 107 The E-model, a computational model for use in transmission planning. (2003).

[34] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. 2011. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 559–568.

[35] Andrew Jones, Ian McDowall, Hideshi Yamada, Mark Bolas, and Paul Debevec. 2007. Rendering for an interactive 360 light field display. *ACM Transactions on Graphics (TOG)* 26, 3 (2007), 40.

[36] Norman P Jouppi, Daniel J Scales, and Wayne Roy Mack. 2001. Robotic telepresence system. US Patent 6,292,713.

[37] Joel Jurik, Andrew Jones, Mark Bolas, and Paul Debevec. 2011. Prototyping a light field display involving direct observation of a video projector array. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*. IEEE, 15–20.

[38] Takeo Kanade, Peter Rander, and PJ Narayanan. 1997. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE multimedia* 4, 1 (1997), 34–47.

[39] Kourosh Khoshelham and Sander Oude Elberink. 2012. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors* 12, 2 (2012), 1437–1454.

[40] Kibum Kim, John Bolton, Audrey Girouard, Jeremy Cooperstock, and Roel Vertegaal. 2012. TeleHuman: effects of 3d perspective on gaze and pose estimation with a life-size cylindrical telepresence pod. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2531–2540.

[41] Itoh K Kitawaki N. 1991. Pure Delay Effect on Speech Quality in Telecommunications. *IEEE J. Sel. Areas Comm*, 586–593.

[42] Leonard Kleinrock. 1992. The latency/bandwidth tradeoff in gigabit networks. *IEEE Communications Magazine* 30, 4 (1992), 36–40.

[43] Robert M Krauss and Peter D Bricker. 1967. Effects of transmission delay and access delay on the efficiency of verbal communication. *The Journal of the Acoustical Society of America* 41, 2 (1967), 286–292.

[44] Robert E Kraut, Susan R Fussell, and Jane Siegel. 2003. Visual information as a conversational resource in collaborative physical tasks. *Human–Computer Interaction* 18, 1-2 (2003), 13–49.

[45] Robert E Kraut, Darren Gergle, and Susan R Fussell. 2002. The use of visual information in shared visual spaces: Informing the development of virtual co-presence. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. ACM, 31–40.

[46] Gregorij Kurillo, Ruzena Bajcsy, Klara Nahrsted, and Oliver Kreylos. 2008. Immersive 3d environment for remote collaboration and training of physical activities. In *Virtual Reality Conference, 2008. VR'08. IEEE*. IEEE, 269–270.

[47] Stephen C Levinson. 2016. Turn-taking in human communication–origins and implications for language processing. *Trends in cognitive sciences* 20, 1 (2016), 6–14.

[48] Stephen C Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology* 6 (2015), 731.

[49] David Lindlbauer and Andy D Wilson. 2018. Remixed Reality: Manipulating Space and Time in Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 129.

[50] Charles Loop, Cha Zhang, and Zhengyou Zhang. 2013. Real-time high-resolution sparse voxelization with application to image-based modeling. In *Proceedings of the 5th High-Performance Graphics Conference*. ACM, 73–79.

[51] William E Lorensen and Harvey E Cline. 1987. Marching cubes: A high resolution 3D surface construction algorithm. In *ACM siggraph computer graphics*, Vol. 21. ACM, 163–169.

[52] Paul Luff and Christian Heath. 1998. Mobility in collaboration. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work*. ACM, 305–314.

[53] Andrew Maimone and Henry Fuchs. 2011. Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras. In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*. IEEE, 137–146.

[54] Andrew Maimone and Henry Fuchs. 2012. Real-time volumetric 3D capture of room-sized scenes for telepresence. In *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2012*. IEEE, 1–4.

[55] Andrew Maimone, Xubo Yang, Nate Dierk, Andrei State, Mingsong Dou, and Henry Fuchs. 2013. General-purpose telepresence with head-worn optical see-through displays and projector-based lighting. In *Virtual Reality (VR), 2013 IEEE*. IEEE, 23–26.

[56] Jennifer Marlow, Scott Carter, Nathaniel Good, and Jung-Wei Chen. 2016. Beyond talking heads: multimedia artifact creation, use, and sharing in distributed meetings. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 1703–1715.

[57] David L Mills. 1991. Internet time synchronization: the network time protocol. *IEEE Transactions on communications* 39, 10 (1991), 1482–1493.

[58] Kana Misawa and Jun Rekimoto. 2015. ChameleonMask: Embodied physical and social telepresence using human surrogates. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 401–411.

[59] Carman Neustaedter, Gina Venolia, Jason Procyk, and Daniel Hawkins. 2016. To Beam or not to Beam: A study of remote telepresence attendance at an academic conference. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 418–431.

[60] Richard A Newcombe, Dieter Fox, and Steven M Seitz. 2015. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 343–352.

[61] Chuong V Nguyen, Shahram Izadi, and David Lovell. 2012. Modeling kinect sensor noise for improved 3d reconstruction and tracking. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*. IEEE, 524–530.

[62] Jakob Nielsen. 1993. Response times: the three important limits. *Usability Engineering* (1993).

[63] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. 2013. Real-time 3D reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)* 32, 6 (2013), 169.

[64] Brid O'CONAILL. 1997. Characterizing, predicting and measuring video-mediated communication: a conversational approach. *Video mediated communication* (1997).

[65] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. 2016. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 741–754.

[66] Fabrizio Pece, Jan Kautz, and Tim Weyrich. 2011. Adapting standard video codecs for depth streaming. In *Proceedings of the 17th Eurographics conference on Virtual Environments & Third Joint Virtual Reality*. Eurographics Association, 59–66.

[67] Benjamin Petit, Jean-Denis Lesage, Clément Menier, Jérémie Allard, Jean-Sébastien Franco, Bruno Raffin, Edmond Boyer, and François Faure. 2010. Multicamera real-time 3d modeling for telepresence and remote collaboration. *International journal of digital multimedia broadcasting* 2010 (2010).

[68] Suraj Raghuraman and Balakrishnan Prabhakaran. 2015. Distortion score based pose selection for 3D tele-immersion. In *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology*. ACM, 227–236.

[69] G Recommendation. 2003. 114-One-way transmission time ITU.

[70] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*. Elsevier, 7–55.

[71] Christian Schaefer, Thomas Enderes, Hartmut Ritter, and Marina Zitterbart. 2002. Subjective quality assessment for multiplayer real-time games. In *Proceedings of the 1st workshop on Network and system support for games*. ACM, 74–78.

[72] Marwin Schmitt, Simon Gunkel, Pablo Cesar, and Dick Bulterman. 2014. Asymmetric delay in video-mediated group discussions. In *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*. IEEE, 19–24.

[73] Marwin Schmitt, Simon Gunkel, Pablo Cesar, and Peter Hughes. 2013. A QoE testbed for socially-aware video-mediated group communication. In *Proceedings of the 2nd international workshop on Socially-aware multimedia*. ACM, 37–42.

[74] Nathan Schuett. 2002. The effects of latency on ensemble performance. *Bachelor Thesis, CCRMA Department of Music, Stanford University* (2002).

[75] Harrison Jesse Smith and Michael Neff. 2018. Communication Behavior in Embodied Virtual Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 289.

[76] Jennifer Tam, Elizbeth Carter, Sara Kiesler, and Jessica Hodgins. 2012. Video increases the perception of naturalness during remote interactions with latency. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2045–2050.

[77] John C Tang and Ellen Isaacs. 1992. Why do users like video? *Computer Supported Cooperative Work (CSCW)* 1, 3 (1992), 163–196.

[78] Herman Towles, Wei-Chao Chen, Ruigang Yang, Sang-Uok Kum, Henry Fuchs Nikhil Kelshikar, Jane Mulligan, Kostas Daniilidis, Henry Fuchs, Carolina Chapel Hill, Nikhil Kelshikar Jane Mulligan, et al. 2002. 3d tele-collaboration over internet2. In *In: International Workshop on Immersive Telepresence, Juan Les Pins*. Citeseer.

[79] Jian Wang, Fuzheng Yang, Zhiqing Xie, and Shuai Wan. 2010. Evaluation on perceptual audiovisual delay using average talkspurts and delay. In *Image and Signal Processing (CISP), 2010 3rd International Congress on*, Vol. 1. IEEE, 125–128.

[80] Steve Whittaker. 2003. Things to talk about when talking about things. *Human–Computer Interaction* 18, 1-2 (2003), 149–170.

[81] Wanmin Wu, Ahsan Arefin, Zixia Huang, Pooja Agarwal, Shu Shi, Raoul Rivas, and Klara Nahrstedt. 2010. " I'm the Jedi!"-A Case Study of User Experience in 3D Tele-immersive Gaming. In *Multimedia (ISM), 2010 IEEE International Symposium on*. IEEE, 220–227.

[82] Wanmin Wu, Ahsan Arefin, Raoul Rivas, Klara Nahrstedt, Renata Sheppard, and Zhenyu Yang. 2009. Quality of experience in distributed interactive multimedia environments: toward a theoretical framework. In *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 481–490.