A Study of Mediated Communication Enhancement Systems Employing Alignment Techniques between Remote Sites

内容記述

| 約
| 持続時間
| レベル |
|---|---|
| レベル |
| レベル |
A Study of Mediated Communication Enhancement Systems Employing Alignment Techniques between Remote Sites

Mamoun Nawahdah
Graduate School of Library, Information and Media Studies
University of Tsukuba

A dissertation submitted for the degree of
Philosophiae Doctor (PhD)
March 2013
Abstract

In the past, the communication circle among human beings was so limited in terms of distance, time, and people. Today, the communication circle is almost unlimited enhanced by information technologies. People can communicate with other people over the globe beyond barriers of distance, time and/or systems. This is a result of the large number of advanced communication systems realized with new media and technologies which are available these days. However, the network-based communication environment provided by the new technologies is different from the traditional face-to-face communication in many aspects, such as reality, affinity, sympathy, mutuality, simplicity, availability, usability, and efficiently. We can accept the current communication systems as it is and change our communication behavior in accordance with the systems, which is what we are doing even though it is not preferable for everyone. Alternatively, we can develop novel technologies to implement new systems to help people communicate over the network that are as good as or even better than face-to-face communication. Replicating face-to-face environment is a challenge considering limitation of the current technologies, the entire user behavior and huge application fields. Therefore, this growing research field is worth to investigate in order to enhance and implement new mediated communication systems to facilitate the people’s communication.

This dissertation discusses the empirical investigation and implementation of enhanced mediated communication systems between two sites, focusing on communication technologies between participants situated in different alignments, i.e., spatial alignment, surrounding alignment, and chronological alignment. Three experiments have been carried out to evaluate three alignment techniques in three different communications domains. The results showed that participants’ communicational behavior was significantly affected by the alignment, resulting in improved physical motion transference, enhanced conversation, and a heightened sense of other participants’ presence.

The experiment in the first communication domain was carried out as a learning task of physical body motion in a Mixed Reality environment. Mixed Reality refers to the process of merging computer generated graphics onto a real world scene to produce new environments and visualizations where physical and digital objects coexist and interact in real time. In this domain, the participants’ body movement and rotation is one important aspect of nonverbal communication. Conveying an appropriate physical body posture is very important to achieve a better communication. In this research, the author investigated the spatial alignment and the effects on learning outcome in the Mixed Reality environment. Spatial alignment
refers to the alignment of spatial aspects such as position, orientation, physical distances between people, body proxemics, body angling, etc. In this research, first, the author confirmed that the posture of a teacher body affects learning efficiency, when some teacher settings are more comfortable and easy to watch than others. A sequence of physical task learning experiments has been conducted using the Mixed Reality technology. The result suggested that the virtual teacher’s close side view is the optimal view for learning physical tasks that include significant single hand movements. However, when both hands are used, or body rotates around, a rotation angle adjustment becomes necessary. Based on these facts, the author proposed a novel automatic adjustment method “Motion Adaptive Virtual Teacher (MAVT)” governing the virtual teacher’s horizontal rotation angle, so that a learner can easily observe the important body motions. The method revealed effective for the motions that gradually reposition the most important moving part. This research activity indicates that spatial alignment of other person’s position and rotation is important for achieving optimal physical task learning.

The second communication domain is related to videoconferencing. In daily life, numerous factors affect the effectiveness of our communication. The surrounding is one of these factors. Common surrounding factors that affect communication are background, lighting, distractions, distance, viewing angle, noise, etc. Surrounding alignment refers to the alignment of surrounding aspects such as environment or setting, space, dimension, border, background, location, etc. In face-to-face communication, the shared surrounding plays an important role in smoothing the communication process. However, in mediated communication, surroundings of the participants are different, which may affect communication process. Aligning the surrounding may help enhancing mediated communication. In this research, the author investigated the background effect in video-conferencing communication. To achieve unified surrounding, he proposed a novel method, which is to superimpose the real-time image of a person at a remote site onto surrounding image of the local site. This technology makes a local user feel as though the other person at a remote site sits in the local site. The technology has been implemented as the “Being Here System (BHS)”. The system was evaluated through an experiment and it was revealed that the proposed technology significantly enhanced communication by improving the perceived presence of other person, in comparison with conventional videoconferencing. This research also investigated user’s communication behavior in the implemented high presence video-conferencing, in comparison with a conventional video-conferencing. User’s communication behavior, verbal and nonverbal, is likely to be affected. The analysis of recorded video of system evaluation experiment revealed that the proposed system significantly affected user’s communication behavior such as speaking turn taking, speech overlapping, total speech time, and number of individual gaze off, while no significant effects were found in terms of the percentage of individual user speech and percentage of gaze off. This research showed that surrounding alignment of the background view significantly enhances participants’ communication behavior.
The third communication domain is related to time shifted remote dining. Real time communication between people in remote sites requires that both persons are available at the same time and ready to communicate. This condition is not always easy to fulfill in the global networked environment because of time zone differences and other such contingent factors. One solution may lie in time shifted communication, or, in other words, delayed communication, which may affect fundamental communication process between people. In a time shifted environment, synchronization between participants is a challenge. This study proposed to minimize the synchronization gap through close chronological alignment. Chronological alignment refers to the alignment of chronological aspects that relate to events arrangement in the order of time. In this research, we investigated the effects of events synchrony in a time shifted tele-dining application. We proposed a system “KIZUNA” enabling people to enjoy a meal together in a virtual environment involving the transmission of recorded video messages. The system helps synchronize a video message with the dining pace of a participant who dines a meal watching the video message through synchronizing the start of dining sessions and adapting the displayed video’s playback speed to the difference in dining progress between local and remote person. This is likely to create a sort of real time behavioral mimicry which enhances communication and increases enjoyment while dining. A validation experiment revealed that the proposed KIZUNA adaptation method enhanced diners’ communication behavior and perceived presence of other person, in comparison with conventional time-shifted tele-dining. This result suggests a promising future for the KIZUNA system. This research activity shows that chronological alignment of events significantly enhances participants’ subjective impressions in communication.

The results of the research shown in this dissertation revealed significant enhancement of communication between spatially or temporally separated individuals, suggesting a promising new approach to the design of mediated communication systems. In particular, the results suggest that focusing on alignment between remote sites is of great value in enhancing communication.
I sincerely thank God, the most gracious and most merciful for enabling me to complete my Ph.D. successfully and for often putting so many good people in my way.

First and foremost, I wish to express my sincere thanks to my supervisor, Professor Inoue Tomoo, for the guidance, support, encouragement, and the great interest in this research. Without this supervision I would not have enjoyed and accomplished this work as much as I have.

Thanks must also go to my other supervisors, Professor Sugimoto Shigeo and Professor Nakayama Shin-Ichi, for their guidance and encouragement. I would like also to thank the members of my graduate committee, Professor Tanaka Kazuyo and Professor Kuzuoka Hideaki, for their remarks and suggestions.

I am grateful to all those people who took part in my experiments especially the students from my laboratory and to all my friends who have been waiting for this work to finish more than me.

Last but not least, no words are ever sufficient to express my everlasting gratitude, appreciation and thanks to my beloved parents, my wife, and my kids who have given me a lot of emotional support. Without their warm love, care, sincere prayers and support, it would have been impossible for me to be myself and to continue learning.

Thank you all.
# Contents

List of Figures iv

List of Tables vi

1 Introduction 1
   1.1 Background ................................. 1
   1.2 Motivation .................................. 2
   1.3 Problem and Approach ........................ 3
   1.4 Contributions ............................... 3
   1.5 Outline .................................... 4

2 Literature Review 5
   2.1 Communication Background .................... 5
   2.2 MC Enhancing .................................. 6
      2.2.1 Spatial Alignment ........................ 7
      2.2.2 Surrounding Alignment ..................... 9
      2.2.3 Chronological Alignment ................. 10
   2.3 Summary ..................................... 11

3 MAVT: A Study of the Effects of Spatial Alignment 12
   3.1 Introduction ................................ 13
   3.2 Related Works ................................ 14
   3.3 Determining the Best View of a Virtual Teacher 16
      3.3.1 Application Design ........................ 17
      3.3.2 Evaluation ................................ 21
   3.4 A Motion Adaptive Orientation Adjustment Method 26
      3.4.1 Application Design ........................ 28
      3.4.2 Evaluation ................................ 31
      3.4.3 Adjusting the Vertical View-Angle .......... 35
   3.5 Discussion .................................. 36
   3.6 Study Limitations ............................. 39
   3.7 Summary ..................................... 39
4 BHS: A Study of the Effects of Surrounding Alignment

4.1 Introduction ........................................... 42
4.2 Related Works ........................................... 43
4.3 Application Design ..................................... 46
  4.3.1 System Proposal ..................................... 46
  4.3.2 System Implementation .............................. 46
4.4 Evaluation .............................................. 50
  4.4.1 Setup ............................................. 50
  4.4.2 Conditions ......................................... 51
  4.4.3 Participants ........................................ 51
  4.4.4 Procedure ......................................... 52
  4.4.5 Questionnaire ...................................... 53
  4.4.6 Session Recording Setup ........................... 54
  4.4.7 Results ........................................... 54
4.5 Discussion ............................................ 59
4.6 Study Limitations ...................................... 62
4.7 Summary ............................................... 63

5 KIZUNA: A Study of the Effects of Chronological Alignment

5.1 Introduction ........................................... 65
5.2 Related Works ........................................... 66
5.3 Application Design ..................................... 68
  5.3.1 System Proposal ..................................... 68
  5.3.2 System Implementation .............................. 69
5.4 Evaluation .............................................. 72
  5.4.1 Recording Dining Video Messages ................. 72
  5.4.2 Dining Conditions .................................. 73
  5.4.3 Setup ............................................. 74
  5.4.4 Participants ........................................ 76
  5.4.5 Procedure ......................................... 76
  5.4.6 Questionnaire ...................................... 77
  5.4.7 Results ........................................... 78
5.5 Discussion ............................................ 79
5.6 Study Limitations ...................................... 80
5.7 Summary ............................................... 81

6 Discussion .............................................. 82

7 Conclusion ............................................... 86

References .................................................. 87
List of Figures

1.1 F2F model ................................................................. 2
1.2 MC model ................................................................. 2

3.1 Reality-virtuality continuum ........................................ 13
3.2 A user learns a physical-task while watching a virtual teacher through a HMD . 14
3.3 Learner-teacher model .................................................. 17
3.4 Examples of virtual teacher’s conditions .......................... 18
3.5 Physical workspace of physical task learning support system .......... 18
3.6 Physical-task learning support system’s configuration overview ........ 18
3.7 Computer-generated 3D virtual teacher’s appearance ................ 19
3.8 Marker locations on teacher’s body .................................. 20
3.9 Core scenes from one of teacher’s recording sessions ................ 21
3.10 120 evaluated conditions in first experiment ...................... 22
3.11 Average evaluation distribution of 120 conditions .................. 22
3.12 11 evaluated conditions in second experiment .................... 23
3.13 Button distribution on table .......................................... 24
3.14 Sample experimental physical task consisting of 10 motion units .... 24
3.15 Learner’s view through HMD ......................................... 25
3.16 Average accomplishment time and error rate per training session .... 25
3.17 Part 1: average accomplishment time and error rate per condition .... 26
3.18 Part 2: average error rate per condition ............................ 27
3.19 Part 2: average accomplishment time per condition ............... 27
3.20 Automatic adjustment processing flow chart ....................... 28
3.21 Virtual teacher’s environment divided into eight sectors ............ 29
3.22 Normal (fixed) learner’s view compared to adjusted view ........... 30
3.23 Physical workspace of physical-task learning-support system ....... 31
3.24 Button distribution on table .......................................... 32
3.25 3D virtual teacher’s appearance .................................... 32
3.26 Sample experimental physical task ................................... 33
3.27 Three fixed rotation-angle conditions: A) 180° B) 105° C) −105° .... 34
3.28 Virtual teacher’s appearance with: A) 180° B) 105° C) −105° ........ 34
3.29 Average error rate per condition ..................................... 35
3.30 Average vertical view-angle per number of sectors ................. 36
3.31 Percentage of experiment’s error types ............................ 37
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>A) Normal videoconferencing environment. B) Proposed superimpose videoconferencing environment</td>
</tr>
<tr>
<td>4.2</td>
<td>Physical workspace of BHS</td>
</tr>
<tr>
<td>4.3</td>
<td>Internal process flow diagram of BHS</td>
</tr>
<tr>
<td>4.4</td>
<td>Setup USB camera behind the display in the center</td>
</tr>
<tr>
<td>4.5</td>
<td>OpenNI scene analyzer view</td>
</tr>
<tr>
<td>4.6</td>
<td>The view after removing background</td>
</tr>
<tr>
<td>4.7</td>
<td>Superimposing other person’s figure on local front view background</td>
</tr>
<tr>
<td>4.8</td>
<td>Physical configuration of two experimental sites</td>
</tr>
<tr>
<td>4.9</td>
<td>Site B’s view, employing a 30” portrait display</td>
</tr>
<tr>
<td>4.10</td>
<td>Participant’s view of Normal videoconferencing mode</td>
</tr>
<tr>
<td>4.11</td>
<td>Participant’s view of Superimpose videoconferencing mode</td>
</tr>
<tr>
<td>4.12</td>
<td>Layout of used video cameras</td>
</tr>
<tr>
<td>4.13</td>
<td>Participants’ average sense of presence results. (Note: Q2’ result is positive form of original Q2)</td>
</tr>
<tr>
<td>4.14</td>
<td>Participants’ average response concerning the quality of presentation</td>
</tr>
<tr>
<td>4.15</td>
<td>Participants’ average response concerning the effect of display size</td>
</tr>
<tr>
<td>4.16</td>
<td>Screenshot of one of ELAN’s annotated video</td>
</tr>
<tr>
<td>4.17</td>
<td>Participants’ average number of turn-taking per minute</td>
</tr>
<tr>
<td>4.18</td>
<td>Participants’ average percentage of speech</td>
</tr>
<tr>
<td>4.19</td>
<td>Participants’ average percentage of speech overlapping</td>
</tr>
<tr>
<td>4.20</td>
<td>Participants’ average percentage of total speech</td>
</tr>
<tr>
<td>4.21</td>
<td>Participants’ average number of gaze-off per minute</td>
</tr>
<tr>
<td>4.22</td>
<td>Participants’ average percentage of gaze-off</td>
</tr>
<tr>
<td>5.1</td>
<td>A user dines while watching a recorded video message</td>
</tr>
<tr>
<td>5.2</td>
<td>KIZUNA system scenario</td>
</tr>
<tr>
<td>5.3</td>
<td>Physical workspace of KIZUNA</td>
</tr>
<tr>
<td>5.4</td>
<td>Procedure flowchart of KIZUNA</td>
</tr>
<tr>
<td>5.5</td>
<td>Internal folder structure of cloud storage</td>
</tr>
<tr>
<td>5.6</td>
<td>A participant’s eating progress compared to that of the recorded remote diner</td>
</tr>
<tr>
<td>5.7</td>
<td>A participant’s eating progress compared to that of the recorded remote diner, under the KIZUNA condition</td>
</tr>
<tr>
<td>5.8</td>
<td>Tele-dining site setup</td>
</tr>
<tr>
<td>5.9</td>
<td>Participants’ average communication-behavior results</td>
</tr>
<tr>
<td>5.10</td>
<td>Participants’ average sense of presence results</td>
</tr>
</tbody>
</table>
List of Tables

3.1 Sample motion-capture data (mm) ........................................... 20
3.2 Participant feedbacks after conducting physical-task learning in MR environment 39
4.1 Experimental session arrangements ........................................... 53
5.1 List of general questions and comments ..................................... 73
5.2 Experimental session arrangements ........................................... 76
Chapter 1

Introduction

1.1 Background

One important aspect of life is communication, where a good communication is one key for
a good life. Communication is a very important link with others, if we communicate well
with others, we are able to better understand what others around us want, need, expect of
us, and what they are able to do and likewise, they will understand what we want, need, etc.
The recent developments in information and communication technology, made the people lives
much accessible than ever. People now can run businesses from remote; learn what they want,
where they want, and when they want; they can maintain and establish relationships with other
people from around the globe. The wide ranges of communication mediums make the process
even much easier, affordable, and more efficient, where some communication mediums are more
efficient than others.

Communication can be carried out using both direct and indirect mediums of communication
such as face-to-face [F2F] interaction (Fig. 1.1), as well as mediated communication [MC] (Fig.
1.2). F2F communication type means the exchange of information, thoughts, and feelings
when the participants are exist in the same physical space at the same time. In this type of
communication, nonverbal cues (e.g., eye contacts, facial expressions, body movements, space,
time, distance, appearance, etc.) may influence the way the message is interpreted or decoded
by the receiver. MC type (also known as virtual communication) enabled new medium of
communication using the recent developments in technology. This type of communication has
many advantages for people given the increased globalization and the need for rapid knowledge
transfer across borders and time-zones. The mediated nature of MC allows greater manipulation
and more careful construction of information. The delayed nature in some MC gives participants
the opportunity to review, revise or cancel their communications before the information are
sent (1). However, the absence of nonverbal cues and tacit knowledge transfer may makes
communication difficult (2). These deficiencies reduce social presence and hinder relationship
formation, unity, and trust. Moreover, synchronous MC can be difficult to schedule in some
cases due to time-zones barriers (3). Despite the advanced communication systems we have
these days; they still can never fully replace the intimacy and immediacy of people conversing
in the same room.
1.2 Motivation

Effective communication between people requires that the communication exchange take place with respect to some level of common context (4). Context in this dissertation refers to the settings (around us and within us) in which all communication happens or develops. Settings could be physical, social, psychological, historical, and cultural. On the one hand, F2F offers a natural unified and mutual context between communication parties (i.e. sender and receiver), where both are exists in the same integrated and unified environment at the same time. On the other hand, in MC the integrated and unified context is not any more realizable given that the communication occurs across distance and/or time. Therefore, aligning the remote context approximate to the local context is believed to enhance the communication. Alignment here refers to the process of adjusting the MC contexts relative to each other to produce a proper connection or orientation for coordinated functioning. We assume that the closer the alignment is the greater the efficiency of the communication using MC technologies.

1.2 Motivation

In the past, the person’s communication circle was so limited in terms of distance, time, and people. Today, the communication circle is almost unlimited. People can communicate across distance and/or time with other people and/or systems from around the globe. This is a result of the large number of advanced communication mediums which are available these days. However, these communication systems are different than the traditional F2F communication in many aspects such as reality, affinity, sympathy, mutuality, simplicity, availability, usability, and efficiently. So it is either we accept the current communication systems as it is and change our ordinary communication behavior, or we come up with some novel techniques to implement new systems that are as good as or even better than F2F communication. Replicating F2F environment is a challenge considering the current technologies and considering the entire users’ behavior and huge application fields. Therefore, in this growing research field, we choose to
investigate this issue and to enhance or implement new MC systems in order to facilitate the people’s communication.

1.3 Problem and Approach

This dissertation focuses on a central problem which is how to achieve and implement enhanced MC systems through aligning the local and remote sites considering various alignment techniques. There are many alignment techniques that can be considered to enhance the communication. Implementing all these techniques might be difficult considering the current technologies. However, considering some specific techniques in a particular domain is a practical solution and this may enhance the communication process substantially. In particular, three communication scenarios were investigated in this dissertation based on different alignment techniques each.

To address this problem, we reviewed previous studies of MC and communication enhancement techniques. The literature review produced a clear view of the current knowledge in this field and the areas that require further investigation. For each area, a thoroughly analysis was conducted to identify the alignment technique that would enhance the communication and the factors that are needed to be considered when evaluating the enhancement. Three experimental systems were constructed and a set of experiments were conducted. The experiments’ results were carefully analyzed and published in related venues. The findings of the experiments were discussed in the context of other studies, and some conclusions were drawn out. Finally, the contribution of the research was examined and avenues of future work were identified.

1.4 Contributions

The following is a summary of the original contribution of this research:

MC enhancement can be achieved by considering various alignment techniques based on a specific communication task. This outcome is the result of three pilot studies and a set of experiments which was carried out to identify and study the alignment factor(s) that affect participants’ communication. The communication tasks were investigated using several objective and subjective measurements which provided some interesting and useful results. The three studies are:

- A study of the effects of spatial alignment to physical-task learning in mixed-reality environment (5, 6): In this study, we investigated the virtual teacher’s optimal position and rotation angle for the best learning outcomes. The outcomes were measured in terms of the required time to accomplish a physical-task and the number of committed errors. To achieve spatial alignment, we proposed a method for automatically adjusting the virtual teacher’s rotation angle when the virtual teacher is demonstrating physical-task motion. The method revealed to be effective and learning was significantly improved.

- A study of the effects of surrounding alignment in videoconferencing (7): In this study, we proposed a high-presence videoconferencing method and investigated the relative effect of surrounding background to participants’ perceived sense of other person presence and basic communication behavior. The method revealed to be effective where the perceived
• A study of the effects of chronological alignment in time-shifted tele-dining (8): In this study, we proposed a tele-dining system and investigated the relative effect of events synchrony to participants’ communication. The proposed system meant to enable distant people to enjoy a meal together in a virtual environment through chronological alignment. The alignment is based on the start and the progress of dining events. This created a sort of real-time behavioral mimicry which revealed to be effective while dining.

1.5 Outline

Chapter 2 covers the literature review. This chapter reviews the background work framing this dissertation, most specifically those related to MC in different aspects.

Chapter 3 covers the investigation of spatial alignment effects, in terms of position and rotation, over the learning outcome in mixed-reality environment. This chapter introduces the design, implementation, and evaluation of “Motion Adaptive Virtual Teacher (MAVT)” system.

Chapter 4 covers the investigation of surrounding alignment effects on communication, in term of the view’s background, in videoconferencing. This chapter introduces the design and implementation of “Being Here System (BHS)”. This chapter also presents the empirical results of evaluating the participants’ perceived sense of other person presence and the analysis of essential communication behavior.

Chapter 5 covers the investigation of chronological alignment effects, in term of events synchrony, on communication in time-shifted tele-dining. This chapter introduces the system design and implementation of a proposed time-shifted tele-dining system “KIZUNA”. This chapter also discusses the experiment results of evaluating the proposed system’s adaptation method on participants’ basic communication behavior.

Chapter 6 discusses the findings and reflection of the studies of this dissertation in the context of other work.

Finally, Chapter 7 summarizes the contributions of this dissertation.
Chapter 2

Literature Review

A rich and varied body of prior work related to the conducted research has been reviewed. This chapter reviews the background work framing this dissertation. Most specifically, those related to MC alignments in different aspects.

2.1 Communication Background

Within the domain of human interaction, there are several types of communication, each occurs in a different context (9):

- Intrapersonal communication: communication with oneself.
- Dyadic/Interpersonal communication: communication between two persons. Dyads are the most common communication settings.
- Small group communication: small groups are a common fixture of everyday life where every member can participate actively with the other members (e.g. family is a group).
- Public communication: occurs when a group becomes too large for all members to contribute.
- Mass communication: consists of messages that are transmitted to large, widespread audiences via electronic and print media.

Human interaction can also be classified based on the distance from others, known as body space or comfort zone. The following are the four social distances introduced by Edward Hall (10):

- Intimate distance (< 1.5 feet): for embracing, touching or whispering.
- Personal distance (1.5-4 feet): for interactions among good friends or family members.
- Social distance (4-12 feet): for interactions among acquaintances.
- Public distance (> 12 feet): used for public speaking.

Communication can be modeled as a simple linear process, where a sender encodes ideas and feelings into some sort of message and then conveys them through a channel to a receiver who decodes them (9). Communication channel is the method by which a message is conveyed between people (e.g. F2F contact, writing, MC, etc.). The word “mediated” reflects the fact
that messages are conveyed through some sort of communication medium. The channel we choose can make a big difference in the effect of the message. Another sophisticated model, “Transactional communication model”, considers simultaneous sending and receiving since most types of communication are two-way exchanges (11). Based on these models, we can conclude that communication requires a sender, a message, and a receiver. Although the receiver need not to be present or aware of the sender’s intention to communicate at the time of communication; hence communication can occur across vast distances in time and space (12).

Information exchanged can be through verbal (Oral), and/or nonverbal means. Some of nonverbal communication includes chronemics\(^1\), haptics, gesture, body language or posture, facial expression and eye contact, object communication such as clothing, hairstyles, architecture, and tone of voice as well as through an aggregate of the above (13). Nonverbal communication is distinct from verbal communication in that it is continuous and multi-channeled and it may be unintentional and ambiguous. The nonverbal part of the message is the primary conveyer of emotion (14). According to a study carried out by Albert Mehrabian, the total impact of a message breaks down like this: 7 percent verbal (words), 38 percent vocal (volume, pitch, rhythm, etc.), and 55 percent body movements (mostly facial expressions) (15).

Communication can be face-to-face [F2F] or mediated through a communication medium (i.e. MC). On the one hand, F2F communication is an indispensable activity in our lives. Traditionally, two or more people had to be physically present at the same time and same place in order to communicate with each other, also known as “Synchronous communication”. F2F occurs in real-time offers a unified context between sender and receiver. F2F communication is also considered rich because it includes nonverbal cues that give communicators cues about the meanings of one another’s words and offer hints about their feelings (16). People always communicate using nonverbal as well as verbal means (17). Researchers have estimated that in F2F communication as much as 60 percent of social meaning is a result of nonverbal behavior (18). Therefore, this situation will remain the preferable method of communication. On the other hand, MC offers an adequate solution for those cases where people are located far away from each other or cannot meet in person. MC has numerous benefits, among them travel cost/time reduction and convenience (19). However, the unified context which exists in F2F communication is not any more realizable in this type of communication. Moreover, when people are located far away from each other, this often means they are also living in different time-zones, and this time-difference is likely to present further difficulties for communication. The main difficulty lies in the misalignment of daily schedules between the two parties, and this is clearly apparent at some activities such as mealtimes. In these cases, time-shifted or delayed communication offers an acceptable solution, also known as “Asynchronous communication”.

### 2.2 MC Enhancing

All communications, intentional or unintentional, have some effect. The communication that involves achieving one’s goals is considered effective communication (20). We have to keep in mind that there is no ideal way to communicate. The type of communication that succeeds in one situation might fail in another. Numerous attempts have been made to enhance MC through various alignment techniques. In this regard, alignment here refers to the act/process of

\(^1\)Chronemics is the study of the use of time in nonverbal communication.
adjusting the MC contexts in relation (or relative) to each other to produce a proper relationship
(connection, rapport) or orientation for coordinated functioning. We assume that the closer
the alignment is the greater the efficiency of the communication using MC technologies.

A variety of alignment techniques was investigated across wide range of communication ap-
plications. The most dominant technique was through life-scale presentation of other person
(17, 21, 22). This way, the important nonverbal cues becomes clearer such as face expression,
gaze, gestures, etc. The distance approximation technique was investigated in (23). In this
research, the distance between the participants was maintained through mobile screen. Spatial
position was found significant in communication (24). In this research, the participants were
distributed around a table to preserve spatial cues. Similarly, seating arrangements was also
investigated (25, 26, 27), where some seating arrangements were found better for smoother
communication. 2D display hides many important nonverbal cues; accordingly the medium of
displaying the other person was investigated to enhance communication. Oyekoya et al. used
spherical display to present other person’s head, which reveled to be effective in identifying
other person and to tell where he/she is looking from any viewpoint (28). Kim et al. used
cylindrical display for presenting other person whole body. This research also reveled to be
effective for sense of social presence where participants were able to assess gaze and hand pointing
cues effectively. Physical embodied technique was found effective for social presence in commu-
nication as well. In one research, a dedicated display used to presents other person’s face and a
set of remote controlled cameras used to present other person’s sight (29). This technique was
enhanced by making display movable (30), using complete mobile robot (31), and even using
android “human-like robot” (32). Aligning remote events is important for communication that
involves physical motions. Synchronization was achieved using vibration sensors as in (23, 33).
Others used motion capture as in (21, 34, 35, 36, 37, 38). The surrounding environment has
important effect in creating a unified communication context. In particular, the different view’s
background effect was investigated. Morikawa et al. found that displaying both remote persons
over same background significantly enhanced communication (39). Gibbs et al. introduced a
technique for displaying other person’s figure over an extended virtual background (22). Har-
risson et al. created 3D background feeling (40). Being able to control something in remote
site was also investigated. In one implementation a user was able to move a plate and display
messages on a remote dining table (41). Controlling the view angle and displaying comments
remotely was also revealed effective (42). Time-difference alignment have been recently getting
more attention, some researches were focused on recorded video messages (43, 44, 45), other
involved recorded voices and gestures (46). Aligning events across distance and time was shown
to be effective in person daily behaviors (47, 48). Finally, haptic was also considered in convey-
ing nonverbal messages such as hand motions and hugs (49, 50, 51). All these related works are
samples of alignment techniques that were used to enhance MC across wide range of applica-
tions. The following subsections cover in more details the prior work related to the alignment
techniques that are covered in this dissertation.

2.2.1 Spatial Alignment

Spatial alignment in MC has been investigated across wide range of applications especially in
physical-tasks training and supporting. Spatial alignment refers to the alignment of spatial
aspects such as position, orientation, physical distances between people, body Proxemics, body
angling, etc. One of the most popular methods of learning new physical motions from remote is to watch a two dimensional [2D] video showing an expert performing the motions. However, some information cannot be conveyed well through 2D video images mainly when expert’s body move and/or rotate around. In such cases, a user cannot confirm how far the expert moved in any direction. Also when rotating, a user can not confirm expert’s front motion from behind. This triggered a large scale of investigation to find a way to align the other person’s body with respect to user through the usage of immersive virtual/mixed reality. In this regard, immersive virtual/mixed reality allows user to interact with other person and environment, as well as to perform novel functions such as sharing body space with other person(52).

An early study by Yang et al. tried to achieve spatial alignment through visualizing body of other person as a ghost that superimposed on user’s own body (35). The result revealed that this superimposing technique produced physical motion training as good as in the real world. Another study by Chua et al. considered multiple copies of other person’s body (21). In this study, they tried to achieve spatial alignment through tracking user’s body motion and display it concurrently besides other person body. They implemented several immersive techniques, such as providing multiple copies of other person body positioned around user and allowing user to superimpose his/her body directly over other person body. However, the results of the experiments showed that none of the used techniques proved to be significantly better than the basic technique when displaying one copy of other person in front of the user.

Movement alignment, in all directions, was investigated in a study by Nakamura et al. (23). In this study, they presented a multimodal information presentation method for basic dance training system. They tried to solve the movement problem by introducing an image display on a mobile robot. This way, user can easily tell the amount of required translation from robot’s movement. The experimental results showed the effectiveness of this method compared with typical fixed screen method. Using body acceleration data captured by myoelectric sensors, Hiyama et al. proposed an artisanship training system (33). The system displays egocentric visual and muscle activates of artisan which is not observed from outside.

Patel et al. tried to achieve spatial alignment by displaying an image of user him/her self rendered in the third person and other person from behind as well as a reflection of both images in a virtual mirror (53). The experimental result revealed that user learned more in this technique more than traditional 2D video system. Another study was conducted to find the optimal method to simplify the remote learning of various body movements by considering four basic visualization methods: a) teacher is facing student, b) like a) but teacher image is mirrored, c) teacher is facing same direction as student, d) like c) but here teacher is superimposed on student(34). The result revealed that d) method was most effective for partial movement repetition while other methods were effective for whole movements. The head physical orientation is one important nonverbal communication. From the head posture we can guess the target the other person is looking at. 2D view cannot convey such information; therefore, Oyekoya et al. proposed the usage of spherical displays to represent other person’s head (28). This way, it is possible to understand the identity of other person from any viewpoint; and also it is possible to tell where the person is looking from any viewpoint.

Controlling the view point is essential in enhancing the communication as well. Olmos et al. presented an interface allows a designated individual, the instructor, to provide the best viewing point to observe and execute a procedure, and simultaneously, offers the remote viewer the freedom to change viewpoints (54). The body posture was found to affect the spatial
patterns formed during F2F interactions between two or more people, known as ‘F-formation’. Kuzuoka et al. examined the effect of a robot rotating its body on the reconfiguration of the F-formation arrangement (55). The results showed that a robot can change the position of a visitor by rotating its body. They also confirmed that rotating the robot body is more effective than only rotating its head to reconfigure the F-formation arrangement.

2.2.2 Surrounding Alignment

Surrounding alignment refers to the alignment of surrounding aspects such as environment or setting, space, dimension, border, background, location, etc. In F2F communication, the shared surrounding physical context plays an important role in smoothing the communication process. However, in MC, two different environments are involved, which may affect the communication process. Observing the other environment may either be a distraction cause or be more engaging, giving a greater sense of other person’s environment (56). Unifying the environments may help enhancing MC. Thus, many studies introduced various methods to achieve this goal.

In virtual reality, surrounding alignment can be achieved by enabling all participants to feel as if they are sharing the same virtual space. A system called ‘HyperMirror’ designed by Morikawa et al. enabled both local and remote participants to appear together on a shared video wall (39). The experimental results showed that the participants sharing the screen tend to act as if they are in the same room. Similarly, Gibbs et al. presented an experimental teleconferencing system ‘TELEPORT’, where the goal was to provide greater realism similar to the F2F environment (22). The system achieved this goal by creating the illusion that other person, although actually distant, is present in local user’s physical space in life-scale by virtually extending user’s front view. The geometry, surface characteristics, and lighting of virtual extension were designed to closely match real room to which it is attached. This created very sophisticated surrounding environment integration. Using similar spaces is another popular technique to achieve comparable environments. Such technique was presented in a study by Luff et al. (57). In this study, they reported on some experiments with a high fidelity media space, ‘t-Room’, used in both sides to provide a consistent collaboration environment.

View’s background is considered the major surrounding physical environment. Chatting et al. investigated the view’s background to find out the effect on users conversation behavior when videoconferencing (56). The experimental results revealed that blurring background was preferred and more efficient than original remote view. Using 3D background effect in communication can also enrich the videoconferencing experience. Harrison et al. presented a method for producing a pseudo-3D experience out of 2D image using a single generic webcam at each end (40). To create a pseudo-3D view from 2D view, they first extract person’s figure from background environment, then the extracted figure and background composited as foreground and background layers with a small offset. This 3D effect is likely to convey real feeling which may enhance the communication.

Consistent furniture and seating arrangements across remote locations can minimize the effect of different environments in MC. A number of advanced videoconferencing systems, such as HP’s Halo1 and Cisco’s Telepresence2, creates the illusion that remote participants are in

1http://hphalo.org/
the same room together by a close configuration of large screens and consistent furniture arrangements across locations. A study by Broughton et al. revealed that the close configuration of cameras and displays along with the matching of color schemes and furniture makes the two sites appear blended into one which in turn facilitate a strong user experience of being collocated (58). Arranging the seating and the displays that represent other person(s) can solve the problematic behavior that users tend to fix their eyes on the display when videoconferencing in some line up seats (26). Another study described a side-by-side work environment that simulates a work environment where users sit next to each other (27).

Mixed reality environment can be used to present another person or virtual objects into local environment to enhance communication and minimize the effect of different environments. Kantonen et al. presented a mixed reality teleconferencing by aligning virtual avatars of remote meeting participants in real local physical space (59). Other studies presented actual other person’s figure into local environment (60, 61, 62). In real environment, presenting other person in form of an exciting real object was also confirmed effective (63). A study by Venolia et al. developed a telepresence device to represent other person locally (29). Lee et al. enhanced the previous idea by making the device movable and giving the other person the ability to control the movements remotely. Using android systems, human-like robot, enhanced the communication even further (32).

2.2.3 Chronological Alignment

Overall, simultaneous communication results better communication over nonsimultaneous communication. But this requires that both sides are available at the same time and/or ready to communicate. In many cases, this condition may be hard to fulfill. One solution may be through time-shifted or delayed communication, which may affect the communication process. However, through close chronological alignment, this effect can be minimized as indicated by some studies. Chronological alignment refers to the alignment of chronological aspects that relate to events arrangement in the order of time.

In physical motion training by watching a 2D video, the motion timing misalignment may affect the training performance. Nakamura et al. solved this issue by implementing an active device to direct action-starting cause with vibration (23). The experimental results showed the effectiveness of this method on concurrently mimicking the physical motion. Similarly, Hiyama et al. used vibration motors to synchronize learner’s and recorded teacher’s motions (33).

In asynchronous presentations, Lucero et al. introduced an interactive wall-mounted display tool that supports synchronization to convey the intended message or ideas (46). The tool records the presentation and organizes it into layers, which are first used to segment the presentation, and later to control the playback of presentation.

Tsujita et al. proposed a system which considers the time-difference between two locations by displaying recorded video of a remote activity after a time shift (44). The conversation was also recorded to form a type of delayed interactive communication as each person responds to the comments from the previous video. In meeting domains, Tang et al. proposed a system to enable participation in future meetings for a person who can not attend in real time (43). The alignment was achieved by uses the person’s recoded contributions to a meeting, which are played for other members during the meeting. The experimental results found that this video technique helped to integrate the absent person into the social context of the meeting.
Synchronizing remote appliances can also participate in unifying the communication environment. In this regard, Tsujita et al. proposed a system ‘SyncDecor’ that allow partners to remotely synchronize and provide awareness or cognizance about their partner - thereby creating a virtual “living together” feeling (47). Another system ‘MissU’ was implemented to share music and background sounds between two distant partners in order to enhance the intimacy and social presence (64).

In some cases, the physical body movement might be fast or hidden behind some objects. One solution is to delay physical movement so user can see it. This is likely to enhance communication. This technique was proposed in a recent study by Yamashita et al. in (65). A technique called “remote lag” was employed to alleviate the problems caused by the invisibility of remote physical motions in collaborative distributed tabletop environment. The technique provides people with instant playback of remote gestures to recover from the missed context of coordination. The experimental results show that remote lags effectively alleviated the invisibility problems.

2.3 Summary

This chapter covers the background work related to MC alignment techniques. The main goal behind such techniques is to reduce the effects that resulted from having different communication contexts and to build a sort of common grounding for communication. Each of these techniques was introduced to solve a specific issue concerning a particular communication task.
Chapter 3

MAVT: A Study of the Effects of Spatial Alignment

The participants’ body movement and rotation is one important aspect of nonverbal communication. Conveying an appropriate physical body posture is very important to achieve a better communication. In this study, we investigated the spatial alignment and the effects on learning outcome in mixed-reality environment. Spatial alignment refers to the alignment of spatial aspects such as position, orientation, physical distances between people, body Proxemics, body angling, etc. First, we confirmed that the virtual teacher body’s posture has effect on learning efficiency, when some teacher-settings are more comfortable and easy to watch than others. A sequence of physical-task learning experiments have been conducted using mixed-reality technology. The result suggested that the virtual-teacher’s close side-view is the optimal view for learning physical-tasks that include significant one-hand movements. However, when both hands are used, or body rotates around, a rotation-angle adjustment becomes necessary. Therefore, we proposed a novel automatic-adjustment method “Motion Adaptive Virtual Teacher (MAVT)” governing the virtual teacher’s horizontal rotation angle, so that the learner can easily observe the important body motions. The method revealed to be effective for motions that gradually reposition the most important moving part. This study is therefore considered likely to enhance communication when learning physical-task skills in mixed-reality environment.
3.1 Introduction

Mixed Reality [MR] refers to the process of merging computer-generated [CG] graphics onto a real-world scene to produce new environments and visualizations where physical and digital objects co-exist and interact in real time. A reality-virtuality continuum was proposed by Milgram et al., with the real environment at one end and the virtual environment at the other (66). Augmented reality [AR] and augmented virtuality [AV] are situated in between, depending on whether reality or virtuality is being modified (Fig. 3.1).

![Mixed Reality (MR)](image)

Figure 3.1: Reality-virtuality continuum

Physical-task learning that utilizes virtual reality and/or mixed reality technology has been actively researched. The use of purely synthetic scenarios in training systems reduces the authenticity of learning or training exercise (67), while the use of actual equipment in a real environment in physical-task learning is known to be very effective. MR technology supports this feature and allows exciting learning, in which learners can actively explore new tasks without the help of experienced instructors. Learners become actively involved in the learning process, and thus memorize more than without using MR (68). The use of MR also enhances users’ perception, and improves learners’ intuitive interaction with the real world (69). In light of this, a host of studies have investigated the support of physical-task learning in such an environment, using sensors and virtual reality (38, 70, 71). In the Computer-supported cooperative work (CSCW) context, collaborative physical tasks have been also investigated, and the use of deictic gestures is known to enhance task performance (72, 73). The results suggest that MR is suitable for supporting physical task learning. Thus, we have developed a physical-task learning-support system using MR (74, 75). The system visualizes a life-sized CG 3D virtual teacher model in front of the learner. Since appropriate feedback information is important for effective and smooth task learning (70), the developed system is also interactive, tracking the learner’s movements and providing basic feedback.

The motivated question was whether a virtual teacher’s location and rotation affects the physical-task learning outcome or not. We assumed that some teacher setups are better than others and accordingly this will affect the learning outcome in such environment. The virtual teacher setups that lead to optimum learning outcome were not determined before. Therefore, we conducted this research to accurately identify these setups and to study all the parameters that leads to better learning outcome when mimicking a virtual teacher using MR. Being able to accurately, safely and speedily perform the training is a major demands especially in the domains of medicine (76, 77), military (78), and industry (38). To perform such physical tasks, the learner must watch carefully and perform the same actions, in the same exact order, as are presented by the virtual teacher. Therefore, we first discussed the potential orientations of the virtual teacher model in a MR system, to determine the virtual teacher’s optimal position and rotation for physical-task learning. Optimalization was measured by the required time to accomplish the physical-task, and the number of committed errors. To investigate the effect of the virtual teacher’s body adjustment, two experiments were conducted; the first, to narrow
3.2 Related Works

down the large number of possible locations in which the virtual teacher may be presented; and the second, to determine the virtual teacher’s optimal position and rotation. To determine whether the virtual teacher’s solid appearance affected the results, experimental comparison with a semi-transparent virtual teacher was also conducted in the study.

In this study, we chose a simple, generic, push-button physical task as a learning model (Fig. 3.2). This task is considered very simple to perform, because the learner needs only move his/her hand and push one of the buttons. Despite its simplicity, however, the task involves the essential aspects of physical-task learning necessary to prove our hypothesis. To perform such a task, the learner must watch carefully and perform the same actions, in the same exact order, as presented by the virtual teacher. This kind of generic motion can apply to a wide variety of physical tasks, such as using some kinds of musical instruments and machines, performing simple dance movements and sports, building models from sub-models in a predefined order, and constructing new material by mixing sub-materials in a predefined order, etc.

![Figure 3.2: A user learns a physical-task while watching a virtual teacher through a HMD](image)

This study has shown that a virtual teacher-model’s orientation has significant effects on learning (6). The results show that the virtual teacher’s close side-view is the optimal view for physical task learning that involves one-hand motion. However, when the virtual teacher uses both his/her hands, or body rotates around, then rotation-angle adjustment becomes necessary. Therefore, a novel method “MAVT” of automatically adjusting the virtual teacher-model’s rotation angle during run-time was proposed by a previous research (79), and we implement and test it out (5). MAVT is based on the virtual teacher’s behavior, more specifically on his/her upper-body movements. The purpose of this method was to ensure that the virtual teacher’s most important moved body part in one motion segment is visible to the learner. This is likely to enhance the learning outcome and the learner may feel more comfortable and assured during learning. The outcome was measured in terms of the number of committed errors during a simple physical-task learning experiment.

3.2 Related Works

There have been various studies in various domains done on virtual reality and MR-based physical skill/task learning and training support and a number of systems have been developed (e.g.,
3.2 Related Works

in the industry domain: constructing machine-maintenance training system (80), metal inert gas welding training system (38), object assembly training system (81, 82), overhead crane training system (83), firefighting tactical training system (84), aesthetic industrial design (85), job training system for casting design (86); in the science and education domain: electrical experimental training system (87), application of geography experimental simulation (88), collaborative learning (89); in the medicine domain: ultrasound guided needle biopsy training system (76), baby feeding training system (90), endoscopic surgery simulation training system (77); in the tourist domain: tourist guide training system (91); in the military domain: missile maintenance training system (78); in the sports domain: Kung-Fu fight game (92), martial arts (21, 53, 93), physical education and athletic training (94), Golf swing learning system (95); in the dance domain: dance training system (23, 36), collaborative dancing (96); in the cooking and eating domain: augmented reality kitchen (97), augmented reality flavors (98), augmented perception of satiety (99); etc.). Many of these systems have employed a virtual teacher to perform the physical task in front of the learner (21, 35, 95, 100, 101). Some of these systems enhance the learning experience by virtually displaying related information and providing necessary feedback, which have been proved to be useful in the respective domains.

Horie et al., for example, proposed an interactive learning system for cooking in an MR environment, using video data extracted from TV cooking programs (102). The respective videos contain cooking experts performing cooking tasks, and the experts are displayed at a cooking table when needed in a fixed location. Another cooking-navigation system was proposed by Miyawaki and Sano, and in that, a virtual agent performing actions corresponding to the current cooking step is displayed in a fixed location at a table as well (103).

Regarding dance skills acquisition, Nakamura et al. developed a 3D dance model in the virtual world (100). The teacher and learner’s avatar were projected side by side on a projector screen. However, such video settings only allow the learner to watch the teacher, while immersive virtual reality allows the learner to interact with the teacher and the environment, as well as to perform novel functions such as sharing body space with the teacher. This capability has been introduced in a 3D immersive system developed by Patel et al., to teach moves from the Chinese martial art of ‘Tai Chi’ (53). In this system, the learner sees four stereoscopic human representations: an image of him and a teacher from behind, as well as a reflection of the front of these avatars in a virtual mirror. The image is displayed on a screen in front of the physical workspace. The results of this research showed that people learned more in the immersive virtual reality system in comparison to traditional 2D video systems.

In the aforementioned studies, the learner must look at a screen in front of him to see the virtual world. Multi-display systems, on the other hand, offer the learner a chance to conveniently view from arbitrary angles, and coordinate their body movements. This technique was implemented in a collaborative dancing system developed by Zhanyu et al. (96). Here, a 3D representation of the dancers is captured in real time, then streamed, and rendered in a shared virtual space. This system also features multi surrounding display to help the dancers conveniently view from arbitrary angle. For mobility and ease-of-watching, Chu et al. proposed a wireless virtual reality system for teaching Chinese ‘Tai Chi’ (21). The learner’s avatar and the teacher model were rendered in a generated virtual environment, and displayed via a light wireless head mounted display (HMD). Here, five interaction techniques were tested: one teacher, four surrounding teachers, four side by side, and two superimpositions. However, the results suggested that the techniques employed had no substantial effect on learning physical
3.3 Determining the Best View of a Virtual Teacher

In another study, by Kimura et al., four basic visualization methods were tested in a generic body-movement learning system: face-to-face, face-to-face with mirror effects, face-to-back, and superimposed (34). The results confirmed that the superimposed method is the most effective for the repetition of partial movements, while the others are effective for whole movements. All of these methods, except for the mirror format, were incorporated into our research. The mirror format was omitted because of the assumption that the mirror effect would cause learner uncertainty, and this would diminish learning performance.

Kirk et al. demonstrated how remote gestures influence the structure of collaborative discourse (72). The results suggested that the use of remote gesture technologies does influence the structure of language used by the collaborating parties. In the study, only helper hands’ view was projected into a fixed location on worker’s desk area. The worker can not move or control the projected view.

In conventional task learning with a real teacher, the teacher observes the learner and intervenes when the learner makes a mistake. To achieve such interactive information feedback for the learner, the ability to sense the learning task and its progress is built in to virtual reality-based learning support systems (70, 71, 100). Feedback information for the learner is also needed in MR-based task-learning support systems, and capturing the learner’s motion is very important in providing such feedback information (70). Such motion-capture technology is used in a dancing training system developed by Chan et al. (36). Here, the virtual teacher is projected on a wall screen, and the learner’s motions are captured and analyzed by the system, with feedback provided. A similar study, by Komura et al., proposed a martial arts training system based on motion capture (37). The learner wears a motion-capture suit and HMD. The virtual teacher appears, alone, in front of the user, through the HMD. This system analyzes the learner’s motion and offers suggestions and other feedback. In our research, the task learning and its progress are sensed using a motion capturing system as well. The captured data were analyzed, and the system provides the learner with basic tone feedback, notifying her/him whether her/his performed motion was correct or not.

On the other hand, our study focused on how virtual teacher should be aligned when it moves single or several body parts to enhance communication. For some simple motions a close side-view adjustment might be sufficient to clearly watch virtual teacher. But for other motions, a more flexible viewing adjustment has to be considered (5, 6, 75, 104). However, this problem has not been pointed out very often and the solution has not been provided. The method presented in this study provides a solution to this problem by automatically rotating virtual teacher’s body in appropriate horizontal angle. The vertical rotation angle adjustment has also considered for some specifics cases as well. As a result, the learners of physical-task movement can improve learning by this method.

3.3 Determining the Best View of a Virtual Teacher

In this section, we first discuss the potential setup of virtual teacher model in a MR system, to determine virtual teacher’s optimal position and rotation for physical-task learning. Optimization was measured by the required time to accomplish the physical-task, and the number of committed errors. To investigate the effect of virtual teacher’s setup, two experiments were conducted; the first, to narrow down the large number of possible locations in which virtual teacher may be presented; and the second, to determine virtual teacher’s optimal position and
3.3 Determining the Best View of a Virtual Teacher

To determine whether virtual teacher’s solid appearance affected the results, experimental comparisons with a semi-transparent virtual teacher were also conducted in the study.

3.3.1 Application Design

3.3.1.1 Learner-Virtual Teacher Model Specification

Figure 3.3 shows the learner-virtual teacher model employed in this research. In this model, \( d \) represents the distance between the centers of the learner and the virtual teacher model; \( \theta_1 \) represents the shifting angle between the front direction of the learner and the virtual teacher model; and \( \theta_2 \) represents the rotation angle of the virtual teacher model around himself.

![Figure 3.3: Learner-teacher model](image)

In this model, the learner will be located at a fixed real location while the virtual teacher will be relocated virtually according to the mentioned parameters (\( d \), \( \theta_1 \), and \( \theta_2 \)). We also fixed the virtual teacher’s size and height-level to mimic a normal person’s size, on the same level, according to the virtual distance between the real learner and the virtual teacher. Given that the maximum human horizontal viewing angle is about 200° (105), the natural constraint of \( \theta_1 \) range was set to ±100°. Examples of the virtual teacher model’s relative conditions are shown in Fig. 3.4.

3.3.1.2 Physical-Task Learning Support System Design Specifications

A physical-task learning support system was built for use in generic physical-task learning experiments (75). Fig. 3.5 shows the system’s physical workspace. The system consists of two subsystems: the motion-capture system and the mixed-reality system (Fig. 3.6). The motion-capture system is used to track and record a person’s motions, and save them to files, while the mixed-reality system is used to process the recorded motion files and prepare the respective task’s motion sequence for the learner to practice.

**The Physical-Task Learning Platform:** The physical-task learning platform contains a number of buttons placed on a table. The push-button task was adopted as a simple generic example of physical-task motion whose errors can be measured quantitatively. The virtual teacher appeared at the learner’s horizontal level. In such a setup the virtual teacher’s lower body movements could be ignored and all motions were carried out by the upper body, more...
3.3 Determining the Best View of a Virtual Teacher

Figure 3.4: Examples of virtual teacher’s conditions

Figure 3.5: Physical workspace of physical task learning support system

Figure 3.6: Physical-task learning support system’s configuration overview
3.3 Determining the Best View of a Virtual Teacher

specifically by the hands. The learner watched the teacher’s upper-body motion and performed a similar motion in real time. Displaying the body motion in such tasks might not be necessary in general. In our experiments, the body motion was not considered as long as the learner used the correct hand to push the correct button. However, displaying the buttons and the upper-body together had the effect of making the task’s instruction clearer and predictable (106). Thus, displaying the upper body motion in our experiments was considered appropriate.

**Motion-Capture System**: The motion-capture system is a computer system connected to six advanced NaturalPoint Optitrack™ (FLEX:V100) optical motion-tracking cameras through a hub (OptiHub). These cameras are used to detect the learner’s motion within the captured volume area by tracking visible reflective markers that are placed on the learner’s body. 7/16” diameter, premium reflective sphere markers are placed on the learner’s hand. Three markers placed at the vertices of a triangle are used to accurately capture the hand’s position and direction. A program written in C# receives the captured-motion data from the cameras and processes it. This program is used to determine which button the learner has pushed, and send this data to the mixed-reality system.

**Mixed-Reality System**: The mixed-reality system is connected to a webcam and HMD. The webcam is used to capture the learner’s view of the real world. The mixed-reality system is responsible for managing the 3D virtual teacher’s physical motion task, which is displayed to the learner through the HMD. The system has flexibility in displaying the virtual teacher, based on the learner-virtual teacher model described above. The distance between the learner and virtual teacher may be configured to 0, 1 or 2 m. The virtual teacher’s shifting angle can vary over a range of ±100°, and the virtual teacher’s rotation angle can vary over the full range of 360°. The system can also display the virtual teacher at two levels of opacity: solid and 50% transparent.

3.3.1.3 The Virtual Teacher’s Appearance and Motion

A recent study found that men’s decisions are strongly affected by certain aspects of the appearance of the virtual avatar, while women’s are not (107). Another study found that attractiveness (and gender) has an effect on the way that virtual interactions occur on both sides (108). Therefore, to minimize any effect of the virtual teacher model’s appearance on the task performance, a plain cylindrical computer-generated 3D model was used in our experiments, as shown in Fig. 3.7.

![Computer-generated 3D virtual teacher’s appearance](image)
The virtual teacher’s motion was randomly generated by combining basic-motion units during system run-time. Each of these motion units, which show the virtual teacher pushing one of the buttons, was prepared in advance by tracking and recording a real person’s motion while he performed these actions. This created a smooth and realistic motion.

To adequately capture and animate the real teacher’s upper-body motions, a minimum of 8 unique markers were placed on the teacher’s upper body, as shown in Fig. 3.8. The markers’ 3-dimensional coordinate data (X, Y, and Z) were recorded at a 100 frame-per-second rate. Table 3.1 shows samples of motion-capture data. Each line in the motion-data file represents one frame of motion data, and each frame contains the 8 markers’ location data. The motion-tracking software “OptiTrack® Rigid Body Toolkit” was used to capture the teacher’s motions. Core scenes from the teacher’s motion-unit recording sessions are shown in Fig. 3.9. The recorded motion data was animated by the free software RokDeBone

![Figure 3.8: Marker locations on teacher’s body](image)

### Table 3.1: Sample motion-capture data (mm)

<table>
<thead>
<tr>
<th>Frame #</th>
<th>Marker # 1: Head</th>
<th>Marker # 2: Chest</th>
<th>Marker # 3: Right Shoulder</th>
<th>Marker # 8: Left Hand</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>X</td>
</tr>
<tr>
<td>0</td>
<td>21.9</td>
<td>33.14</td>
<td>-16.19</td>
<td>3.4</td>
</tr>
<tr>
<td>1</td>
<td>22.0</td>
<td>33.15</td>
<td>-16.18</td>
<td>3.5</td>
</tr>
<tr>
<td>2</td>
<td>22.2</td>
<td>33.16</td>
<td>-16.20</td>
<td>3.6</td>
</tr>
<tr>
<td>108</td>
<td>22.3</td>
<td>33.17</td>
<td>-16.22</td>
<td>3.5</td>
</tr>
<tr>
<td>109</td>
<td>22.4</td>
<td>33.18</td>
<td>-16.23</td>
<td>3.4</td>
</tr>
</tbody>
</table>

1[http://www5d.biglobe.ne.jp/~ochikko/rokdebone.htm](http://www5d.biglobe.ne.jp/~ochikko/rokdebone.htm)
3.3 Determining the Best View of a Virtual Teacher

3.3.2 Evaluation

Two experiments were conducted to investigate the effect of the virtual teacher setup in MR physical-tasks learning. The first experiment was employed to narrow down the large sample number of possible conditions, while the second experiment was conducted to evaluate the top-rated conditions identified in the first experiment, as well as other conditions defined in a previous, related study.

3.3.2.1 Experiment 1: Narrowing Down the Virtual Teacher’s Possible Conditions

This experiment was conducted to investigate the range of virtual teacher positions and rotations most comfortable and instructive for the learner, during physical-task learning sessions.

**Specification of the Virtual Teacher’s Conditions:**

In this model, the virtual teacher’s positions and rotations can vary within the previously mentioned parameters ($d$, $\theta_1$, and $\theta_2$). These parameters produce an enormous number of conditions. Therefore, a systematic sampling method was used to reduce the number of evaluated conditions in this experiment. The distance $d$ was sampled at 0, 1, and 2 m. The shift angle $\theta_1$ was sampled at $0^\circ$, $\pm 30^\circ$, $\pm 45^\circ$, and $\pm 60^\circ$. The rotation angle $\theta_2$ was sampled at $0^\circ$, $\pm 45^\circ$, $\pm 90^\circ$, $\pm 135^\circ$, and $180^\circ$. This produced a total of 120 conditions (15 positions x 8 rotations).

Fig. 3.10 shows the 120 conditions where the learner is located at the origin.

**Participants:**

A total of 4 participants were hired to participate in this experiment. The participants were 4 male graduate students whose ages ranged from 23 to 28 years. All the participants were right-handed and had normal or corrected to normal vision.

**Procedure:**

The physical-task support learning system we developed was used to present a 3D virtual teacher model performing hand-movement tasks in each condition, through the HMD. Since we had many possible conditions to test, and the required result was a range of conditions, a subjective evaluation was employed. In this experiment, each condition was rated on a 7-point Likert scale, separately, by each participant, based on how comfortable and well-informed they felt in mimicking the virtual teacher’s physical-task of pushing the buttons. The participants were not required to finish each motion task and they were allowed to watch it as long as they need to evaluate the conditions. The evaluation scale was 1 “very confusing”, 2 “confusing”, 3 “somewhat confusing”, 4 “neither”, 5 “somewhat clear”, 6 “clear”, and 7 “very clear”. After
3.3 Determining the Best View of a Virtual Teacher

Figure 3.10: 120 evaluated conditions in first experiment

The evaluation process was completed, the average was calculated for each condition, and those conditions with a score of 4 or higher were considered good conditions.

Results:
The conditions’ average evaluation distribution is shown in Fig. 3.11, where each value represents the average value of the 4 evaluators’ results for each condition. Regarding the distance $d$, most of the top-rated conditions were close to the learner (1 m). Regarding the virtual teacher’s shifting angle $\theta_1$, most top-rated conditions were close to the learner’s center-view ($\pm 30^\circ$). And regarding the teacher’s rotation angle $\theta_2$, the side-view of the 3D virtual teacher model seemed to be preferred ($\pm 90^\circ$).

Figure 3.11: Average evaluation distribution of 120 conditions
3.3 Determining the Best View of a Virtual Teacher

### 3.3.2.2 Experiment 2: Determining the Virtual Teacher’s Optimal Condition

The objective of this experiment was to investigate and determine the virtual teacher’s optimal setup, for the highest learning performance in a mixed reality environment.

**Specification of the Virtual Teacher’s Conditions:**

Figure 3.12 shows the conditions that were tested in this experiment. The top-rated conditions defined by the first experiment (Section 3.3.2.1) were considered and reduced to eight conditions by merging similar location conditions. Conditions 1 to 4 were selected to investigate the effect of changing the virtual teacher’s shift angle. Conditions 5 and 6 were selected to investigate the effect of the distance from learner to teacher. Conditions 7 and 8 were selected to study the effect of teaching from the side view. Conditions 9 to 11 were considered as a result of previous, related research (21, 34, 35), in which they were positively evaluated in a virtual reality environment. Condition 9 represents a natural configuration in which the learner is located behind the teacher. Condition 10 represents another natural configuration, in which the learner is located in front of the teacher. Condition 11 represents a superimposed configuration, in which the teacher is virtually superimposed on the learner’s body.

![Figure 3.12: 11 evaluated conditions in second experiment](image)

**Experimental Physical-Task Specification:**

A physical task learning platform contains 5 buttons \([B_1 - B_5]\) placed on a table in two rows was used in this experiment (Fig. 3.13). The horizontal space between the buttons is 12 cm, and the vertical space between the two rows is 9 cm. The buttons were arranged in this way, so that the physical motions are distributed over the learner’s entire front space. This generates the kind of motions that cover a wide range of physical tasks.

Since there were 5 buttons, 5 recording sessions were conducted, to produce 5 motion units. An experimental virtual teacher’s physical-task motion was randomly generated by combining 10 basic-motion units during system run-time. Fig. 3.14 shows a sample experimental physical task.
3.3 Determining the Best View of a Virtual Teacher

![Button distribution on table](image1)

**Figure 3.13:** Button distribution on table

![Sample experimental physical task consisting of 10 motion units](image2)

**Figure 3.14:** Sample experimental physical task consisting of 10 motion units

**Participants:**
A total of 15 participants took part in this experiment, 7 females and 8 males. The participants’ ages ranged from 20 to 28 (mean=23, s.d.=3.1), and they were all undergraduate or graduate students. All the participants were right-handed and had normal or corrected to normal vision.

**Procedure:**
Because the participants were using this system for the first time, it was expected that they would become accustomed to the system after a while. To avoid this, training sessions involving the mimicking of physical task motions were first conducted. At the end of each session, the session’s time and errors were calculated. Based on these values, the experimenter decided whether the learner needed to conduct more training sessions or not.

To commence this experiment, the learner put on the HMD, and put the markers on his/her hand. When the system started up, it displayed the 3D virtual teacher model in one condition randomly, superimposed on the physical space, as shown in Fig. 3.15. The experiment was consisted of 11 sessions to evaluate the 11 conditions randomly. In each session one condition was evaluated. The learners were asked to complete the task of correctly mimicking the virtual teacher model’s motion as quickly and accurately as possible. The virtual teacher demonstrated randomly one of the recorded physical task motion units in front of the learner. The learner took a rest for 15 seconds between the experimental sessions. The sessions were recorded on tape. Afterward, the sessions were reviewed and the task’s error rate and accomplishment time were calculated for each condition. When the subject pushed a different button than the intended one, this was considered an error. The accomplishment time was measured from the start of each experimental session, until the subject successfully completed all the task’s motion units.
3.3 Determining the Best View of a Virtual Teacher

This experiment had two parts: the first was performed with a solid virtual teacher, by all the participants; and the second was performed with a 50% transparent virtual teacher, by 5 participants. The second part was conducted in order to determine whether the virtual teacher's transparency had any effect on the results for some of the tested conditions.

**Results:**

The training sessions’ results showed that learners became accustomed to the system after an average of 4 sessions, where no significant changes in the task’s accomplishment time, or the number of errors, were reported. Fig. 3.16 shows the average accomplishment time and average error rate per session.

**Part 1: Using a Solid Virtual Teacher:**

The experiment results show clearly that the virtual teacher’s setup has an effect on the learning outcome, both in terms of the required time to accomplish specific physical task learning, and in terms of the number of committed errors. Fig. 3.17 shows the average accomplishment time and error rate per condition.

Conditions 3 and 8 scored the lowest accomplishment time (mean = 15.1 sec.) and the lowest committed error rate (mean = 1.33%). To seek for any significant difference between the
tested conditions we used the t-test comparison test. First, we ran the test over the opposite side conditions of 3 and 8. Comparing the results of condition 3 to 4, we found no significant difference in the results [Time: \( t(14) = -1.3, p < 0.1 \), Errors: \( t(14) = -1.1, p < 0.1 \)]; and comparing condition 8 to 7, we also found no significant difference [Time: \( t(14) = 0.7, p < 0.1 \), Errors: \( t(14) = 1, p < 0.1 \)]. Next we ran the test over conditions 9, 10, and 11, with respect to condition 3. The results showed a significant difference between condition 3 and 9 [Time: \( t(14) = -6.9, p < 0.001 \), Errors: \( t(14) = -9.9, p < 0.001 \)], between condition 3 and 10 [Time: \( t(14) = -3.4, p < 0.01 \), Errors: \( t(14) = -2.8, p < 0.05 \)], and between condition 3 and 11 [Time: \( t(14) = -2.9, p < 0.01 \), Errors: \( t(14) = -1.2, p < 0.01 \)].

**Part 2: Using a 50% Transparent Virtual Teacher:**

The results of this part show no significant difference when using a semi-transparent instead of solid mode for the virtual teacher. Fig. 3.18 shows the resulting average error rate per condition, and Fig. 3.19 shows the average time per condition. We ran this second part specifically to investigate any effect on condition 9. The transparent mode scored better results (Time: mean=21.2 sec., Errors: mean=24%) in comparison with the solid mode (Time: mean=25sec., Errors: mean=36%); but this enhancement was not considered a significant difference according to the calculated t-test value [Time: \( t(4) = 0.93, p < 0.1 \), Errors: \( t(4) = 1, p < 0.1 \)].

### 3.4 A Motion Adaptive Orientation Adjustment Method

In this section, we introduce a novel method of automatically aligning the virtual teacher-model’s rotation angle during runtime. The method is based on the virtual teacher’s behavior, more specifically on his/her upper-body movements. The purpose of this method is to ensure
3.4 A Motion Adaptive Orientation Adjustment Method

Figure 3.18: Part 2: average error rate per condition

Figure 3.19: Part 2: average accomplishment time per condition
that the virtual teacher’s most important moved body part in one motion segment is visible to the learner.

### 3.4 A Motion Adaptive Orientation Adjustment Method

#### 3.4.1 Application Design

##### 3.4.1.1 Method Design

The automatic adjustment processing flow chart is shown in Fig. 3.20. The system is divided into two main processes: an initialization process and a run-time process. During the system initialization, the virtual teacher’s captured motion data is retrieved from a file system. Next, the task motion data is split into small fixed-duration segments. For each motion segment, the teacher’s optimal rotation angle is calculated. During system run time, the viewing angle of the each segmented teacher-task motion is automatically adjusted according to the pre-calculated angle, which is the side-view of the main virtual teacher’s movement, and displayed.

![Automatic adjustment processing flow chart](image)

**Figure 3.20:** Automatic adjustment processing flow chart

##### 3.4.1.2 The Virtual Teacher’s Rotation Angles

The method assumed that the virtual teacher is located (seated) at a specific fixed location and not moving; i.e., not completely moving from one location to another. Therefore, we fixed the virtual teacher’s location to the point of origin at the same learner’s horizontal level. To adequately assess the automatic adjustment method using generic physical-task motions, the virtual teacher’s environment must be divided into a sufficient number of sectors in such a way that the following motion scenarios are enacted:

- Having a virtual teacher’s physical motion move from a sector governed by the right-hand to another sector also governed by the right-hand; i.e., we need at least two sectors governed by the right hand in front of the learner. Similarly, we need at least two sectors governed by the left hand in front of the learner.
3.4 A Motion Adaptive Orientation Adjustment Method

- Having a virtual teacher’s physical motion move from a sector governed by the right-hand to a neighboring sector governed by the left-hand, and vice versa.

Based on these motion scenarios, the virtual teacher’s environment was divided into eight equal sectors as shown in Fig. 3.21. Each sector covers a 45° range, and each has an associated counter ($C_1 - C_8$). These counters were used to record the count of the virtual teacher’s maximum moved marker in each sector during the automatic adjustment process. The sector with maximum counter value is considered the sector that contains the most important movements. Accordingly, the virtual teacher is rotated to the sector’s predefined rotation angle ($\theta$). The sector’s predefined rotation angle ($\theta$) had been calculated so that the sector’s center angle faces the learner when selected using the following equation:

$$\theta = 360^\circ - SC$$

where $SC$ is the sector’s center angle.

![Figure 3.21: Virtual teacher’s environment divided into eight sectors](image)

3.4.1.3 Calculating the Optimal Segment’s Adjustment Rotation Angle

The automatic adjustment process starts by reading the segment’s motion data frame by frame. For each marker’s 3-dimensional coordinate data in the frame, the absolute marker’s movement amount $M_j$ in any direction is calculated based on the previous frame’s marker data:

$$M_j = \sqrt{(X_{cj} - X_{pj})^2 + (Y_{cj} - Y_{pj})^2 + (Z_{cj} - Z_{pj})^2}$$

where $j$ is the marker number ranging from 1 to 8; $X_{cj}$, $Y_{cj}$, and $Z_{cj}$ are the current frame $j$-marker’s position data; and $X_{pj}$, $Y_{pj}$, and $Z_{pj}$ are the previous frame $j$-marker’s position data.
3.4 A Motion Adaptive Orientation Adjustment Method

After calculating the frame’s eight markers’ absolute movement amounts, the maximum marker’s movement $MM_i$ is determined:

$$MM_i = \max (M_1, M_2, ..., M_8)$$  \hspace{1cm} (3.3)

where $i$ is the current frame number.

For this marker, which has the maximum absolute movement, we calculate the marker slope angle $O_i$ with respect to the XY plane:

$$O_i = \arctan \left( \frac{Y_i}{X_i} \right)$$ \hspace{1cm} (3.4)

Based on the calculated $O_i$ angle, the counter of the sector that includes this angle is increased by 1. Once all the segment’s frames are processed in the same manner, the maximum sector’s counter value $C_{\text{max}}$ is determined:

$$C_{\text{max}} = \max (C_1, C_2, ..., C_8)$$ \hspace{1cm} (3.5)

The resulting sector with $C_{\text{max}}$ is assumed to be that wherein the most important motion has occurred. Accordingly, the virtual teacher’s rotation angle in the entire segment will be set according to the selected sector’s predefined rotation angle ($\theta$).

Figure 3.22 shows the resulting views during the first two segments of a preliminary test of the method. The virtual teacher’s initial rotation angle was $180^\circ$. During the first segment, the virtual teacher used mostly his/her right hand over Sector 3. Therefore the virtual teacher’s rotation angle was automatically adjusted to $-112^\circ$. In the second segment, the virtual teacher used mostly his/her left hand over Sector 6, and in this case the virtual teacher’s rotation angle was adjusted to $112^\circ$. In each of these cases, the learner confirmed that he was able to clearly see the critical elements of the virtual teacher’s motion.

![Figure 3.22: Normal (fixed) learner’s view compared to adjusted view](image-url)
3.4.2 Evaluation

To evaluate if the automatic adjustment method produces a better view, a comparative generic physical-task learning experiment was conducted. The first part of this learning experiment was performed using three predefined and fixed virtual-teacher rotation angles. The second part was performed using the virtual teacher’s automatic adjustment method. The experiments were videotaped. A questionnaire was completed by the participants after each session. The error rates were compared and analyzed to find out any significant improvements between the conditions.

3.4.2.1 The Physical-Task Learning Platform

The MR learning-support system described in Sec. 3.3.1.2 was used to test our automatic adjustment method. The system physical workspace is shown in Fig. 3.23.

The physical-task learning platform contains eight buttons \([B_0 - B_7]\) of 85 mm in diameter and 10 mm in height, placed on a table, as shown in 3.24. The buttons diameter was determined according to the average person’s hand width, to minimize uncertainty in the view. For the learner to access the target buttons comfortably, the distance between each of the eight buttons and the learner’s hands was set to 250 mm. The buttons were arranged as seen in the figure so that the physical motions are distributed over the learner’s entire front space. In order to engage both the learner’s hands in the physical-task learning, four buttons \([B_0 - B_3]\) were operated by the learner’s right hand, and four other buttons \([B_4 - B_7]\) were operated by the learner’s left hand. This generates the kind of motions that cover a wide range of real physical tasks.

In the push-button physical-task learning platform, the learner was seated in a fixed location in front of the table. The virtual teacher appeared at the learner’s horizontal level. In such a setup the virtual teacher’s lower body movements could be ignored, and all motions were carried out by the upper body, more specifically by the hands. The learner watched the teacher’s upper-body motion and performed a similar motion in real time.
3.4 A Motion Adaptive Orientation Adjustment Method

3.4.2.2 The Virtual Teacher’s Appearance and Motion

Figure 3.25 shows the plain cylindrical 3D model that used in this experiment. Sub-motion units, which show the virtual teacher pushing one of the eight buttons, were prepared in advance by tracking and recording a real person’s motion while he performed these actions. This created a smooth and realistic computer graphic avatar movement when the motion data was animated. Since there were eight buttons in the physical-task learning platform, eight recording sessions were conducted to produce eight unique sub-motion units; four sub-motion units were right-handed motions, and the remaining four sub-motion units were left-handed motions.

In the MR system, the virtual teacher performs a gradual physical-motion task in front of the learner. Therefore a distinctive set of physical motion tasks had been produced at runtime. By using all the prepared basic sub-motion units, we systematically created a chain of sub-motions according to the following aims:

- We combined the prepared sub-motion units into variable-sized similar-motion blocks. In order to ensure that the user would not memorize the number of sub-motions within each block, we randomly employed a variable block size of 3, 5, or 7 sub-motion units.

- To avoid distracting the learner by frequently switching the used hand, we decided to create the physical-motion task out of two balanced parts. The first part was composed of four randomly-defined right-hand blocks, and the second part of four randomly-defined left-hand blocks.

Based on these aims, a total of 40 sub-motion units were combined to create a one-motion task. This produced a movie of 44 seconds’ length (Fig. 3.26). A) The motion task divided into
3.4 A Motion Adaptive Orientation Adjustment Method

eight blocks, where $n$ represents block size and has value of 3, 5, or 7. $B_R$ is one of right-handed sub-motion units. $B_L$ is one of left-handed sub-motion units. B) A sample motion task.

![Sample experimental physical task](image)

**Figure 3.26:** Sample experimental physical task

### 3.4.2.3 Participants

A total of 21 participants took part in this experiment as learners, 9 females and 12 males. The participants’ ages ranged from 20 to 33 (mean=24, s.d.=3.5), and they were mostly undergraduate or postgraduate students. The participants were divided into two groups. One group performed the first part of the experiment, while the other group performed the second part. There were 11 members in the first group, comprised of 6 males and 5 females; and 10 members in the second group, comprised of 6 males and 4 females. All the participants were right-handed and had normal or corrected-to-normal vision.

### 3.4.2.4 Fixed Rotation Conditions

It has been shown in the previous study that a virtual teacher-model’s position and rotation angle have significant effects on physical-task learning (6). That study suggested that the close side view of the virtual teacher is the optimal view for physical-task learning that involves one-hand motion. Based on this, we decided to assess the top three fixed rotation-angle conditions from that study (Fig. 3.27). The first condition has a $180^\circ$ rotation angle, the second condition a $105^\circ$ rotation angle, and the third a $-105^\circ$ rotation angle. The first condition represents a normal setup wherein the teacher is located in front of the learner, the second condition represents a teacher’s left-hand focused view, and the third condition represents a teacher’s right-hand focused view. In the three conditions, the virtual teacher was placed at one meter’s virtual distance away from the learner. Fig. 3.28 shows the resulting virtual-teacher view in the three fixed rotation-angle conditions.

Each participant in this part of the experiment performed three physical-task learning attempts by mimicking the virtual teacher’s motions. The virtual teacher appeared in front of the learner through the HMD with a fixed rotation-angle. Each learner performed the experiment in each of the three fixed rotation-angle conditions, one by one. The virtual teacher continuously performed one of the pre-generated motion tasks for 44 seconds in front of the learner. The learners were asked to watch and simultaneously push the correct button, and as many buttons as the virtual teacher pushed. The experimental sessions were recorded on tape. Afterward, the sessions were reviewed and the task’s error rate was calculated for each condition.
3.4 A Motion Adaptive Orientation Adjustment Method

3.4.2.5 Automatic Adjustment Condition

This part was similar to the fixed rotation-angle conditions experiment, except that here the participants performed one physical-task learning attempt only. In this part of the experiment, the virtual teacher’s rotation angle was automatically adjusted during the run time. The learners were asked to watch and simultaneously push the correct button and as many buttons as the virtual teacher pushed. The experimental sessions were recorded on tape. Afterward, the sessions were reviewed and the task’s error rate was calculated.

3.4.2.6 Results

Our primary goal was to find out whether or not the automatic adjustment method would minimize the number of errors when providing a better view. Minimizing the number of errors was assumed as one factor in improving physical-task learning. The statistical results of the two experimental groups were analyzed to determine whether using the automatic adjustment method significantly reduced the number of errors or not.

Figure 3.29 shows the average error rate per condition. The average error rate in each fixed condition was calculated to be: for the first condition (180°), 12.27% (s.d.=6.9%); for the second condition (105°), 12.5% (s.d.=4.6%); and for the third condition (−105°), 12.5% (s.d.=7.1%). First, we tested the error rate’s results of the three fixed rotation-angle conditions using ANOVA. The analysis confirmed no significant difference between the three conditions’ average error rate ($F(2,30) = -0.0047, p < 0.01$). Therefore, we summed up all the fixed-rotation conditions’ data to be used as the fixed condition’s data. This was to be used to compare with the automatic-adjustment condition’s data. The average error rate of automatic adjustment method was calculated to be 6.0% (s.d.=2.7%). The t-test (assuming unequal variances) was used to compare the means of the two conditions (the automatic adjustment
and the joined fixed rotation condition). We found that using the automatic adjustment method decreased the average error rate, and the average error rate was significantly different ($t(31) = 5.1, p < 0.01$).

![Figure 3.29: Average error rate per condition](image)

### 3.4.3 Adjusting the Vertical View-Angle

Our proposed method assumed that the learner will sit and see the virtual teacher in front of him at the same horizontal sight level as if in a real situation. Accordingly the method only controlled the view's horizontal rotation angle. The vertical rotation and orthogonal view were not considered in the previous experiments. Therefore, a final experiment was conducted to find out the effect of adjusting the vertical viewing angle on observing a clear motion while mimicking a physical-task motion.

**Experiment Design:**

The physical-task learning system was updated so that the learner can manually setup the vertical view-angle, while the horizontal view-angle was adjusted automatically using our proposed method. To find out the relationship between the physical motion with respect to the vertical view angle, 4 distinct motions were prepared in advance as follow:

- 1-Sector physical-task motion: the virtual teacher used his hand over one sector.
- 2-Sectors physical-task motion: the virtual teacher used his hands over 2 sectors.
- 3-Sectors physical-task motion: the virtual teacher used his hands over 3 sectors.
- 4-Sectors physical-task motion: the virtual teacher used his hands over 4 sectors.

The participants in this experiment were asked to mimic the virtual teacher’s physical motion. The vertical viewing angle was manually adjusted by the participants themselves to the degree that he/she felt more comfortable and the motion is the most clear. The vertical adjusting value ranged from $0^\circ$ to $90^\circ$. $0^\circ$ vertical view-angle represents the normal situation where the virtual teacher placed in front of the learner at the same horizontal sight level, while $90^\circ$ vertical view-angle represents the orthogonal view situation.

**Participants:**
A total of 5 participants were hired to participate in this experiment, 2 females and 3 males. All the participants were right-handed and had normal or corrected to normal vision.

**Results:**

Fig. 3.30 shows the average resulted vertical view-angle per the number of sectors. For 1-sector motion, the participants sit the vertical view-angle to $19^\circ$; for 2-sectors motion, a $28^\circ$ was selected; for 3-sectors motion, a $36^\circ$ was chose; finally for 4-sectors motion, a $43^\circ$ was best.

![Figure 3.30: Average vertical view-angle per number of sectors](image)

### 3.5 Discussion

Physical-task learning in mixed-reality environments becomes very popular in wide areas as seen in section 3.2. Without doubt, this training technique could become the first option in many domains if it designed well. The virtual teacher setup is one important aspect to achieve an effective training system. Many implemented training systems had considered a fixed and limited virtual teacher’s location(s) and rotation(s) (21, 34, 35, 53, 100). In contrast, we thoroughly investigated the virtual teacher setup and there effects on the sample physical-task learning.

The first experiment (Section 3.3.2.1) aimed to identify the ranges of the virtual teacher’s location and rotation most comfortable and instructive for the learner. The result shows that learner preferred front side-view of virtual teacher. Despite the small sample-size, the participants’ preference results were very consistent. This general result was expected as learner preferred to see the physical objects and virtual teacher before him/her in his/her spatial environment. This might be the reason why many of the implemented training systems had employed a virtual teacher in front of the learner such as (37, 53, 100, 103, 109).

The second experiment’s result verified that looking at the virtual teacher model from the side decreases both the time and error rate, and looking at the virtual teacher’s model from behind or in front increases the time and error rate significantly (Section 3.3.2.2). Regarding condition 9, the hand motion was completely or partially hidden by the virtual teacher’s body. Regarding condition 10, the opposite-hand view caused the learner to consume more time and commit more errors. A final interesting point: we found that conditions 1, 3, 5, and 8 scored better results, in comparison with conditions 2, 4, 6, and 7, even though all of these conditions represented a form of side-view. This is because our experiment’s physical motion task was
recorded using the right hand only, meaning that the physical motion was more visible from the right side-view.

For physical motion task where the virtual teacher uses both hands, or body rotates around, the proposed automatic adjustment method turned to be more effective compared with fixed rotation conditions (Section 3.4). The experiment result shows that the automatic adjustment method significantly decreased the average error rate. We found that the errors observed in the experiment could be categorized into three types, as follows (Note that the learner was supposed to watch the virtual teacher and simultaneously push the correct button in any manner he/she preferred as long as he/she used the correct hand; the learner’s body motion itself was not considered.):

- **Type A** error: When the learner pushes a different button than the intended one.
- **Type B** error: When the learner pushes a correct button but with the wrong hand.
- **Type C** error: When the total number of learner’s button pushes does not match the exact number performed by the virtual teacher. This covers the following two cases:
  - When the total number of learner’s pushes is more than the correct performed number. In this case, the extra pushes are considered errors.
  - When the total number of learner’s pushes is less than the correct performed number. In this case, the missing pushes are considered errors.

Figure 3.31 shows the error details. The Type A error, pushing the wrong button, was found to be the most common error across all the conditions with 83% of the total errors. This error typically seemed to occur when the learners could not see the virtual teacher’s motion clearly. The Type B error, using the wrong hand, made up 6% of the total errors. In this regard, we found that some of the learners tended to use their right hands more than their left hands. The Type C error, pushing more/less buttons, made up 11% of the total errors. In this type of error, most of the learners failed to push a button when they became confused and could not decide which one of the buttons was the correct one. On the other hand, few learners pushed the button extra times.

![Figure 3.31: Percentage of experiment’s error types](image)

A thorough analysis was conducted in order to determine what had caused some of the repeated errors in the experiment, and whether or not the automatic adjustment method had resolved those problems. In the fixed rotation-angle conditions, we noticed that some of the learners spent extra time at the beginning. This might be because they needed this time to
3.5 Discussion

figure out the experiment’s initial setup, and which hand they were supposed to use, despite the pre-session instructions, and the fact that the time before the first motion unit was displayed was the same in each session. Nonetheless, this may have caused some of them to miss the first motion unit in some cases. On the other hand, the automatic adjustment method provided a close and direct view of the initial virtual-teacher motion, which in turn minimized the confusion that occurred under the fixed-rotation conditions.

The generic experiment involved pushing the same button 3, 5, or 7 times. It was observed that the number of buttons pushed was sometimes one more than the correct number, when the correct numbers were 3 or 5. Six cases were found in the fixed rotation-angle conditions, and two cases were found in the automatic adjustment condition. Although the result was not statistically significant because of the small number of cases, the automatically adjusted view might alleviate this type of error.

The learners seemed to have some difficulty in recognizing the farthest two buttons in the view ($B_{0}$ and $B_{1}$) in the second fixed rotation-angle condition ($105^\circ$). The same difficulty was observed in the third fixed rotation-angle condition ($-105^\circ$), wherein the farthest two buttons were $B_{7}$ and $B_{6}$. The Type A error occurred 9 times in these conditions, and only 3 times in the automatic adjustment condition.

There was a case in which the current proposed method could not provide a good view. When the motion segment contained buttons from both the far ends ($B_{0}$ and $B_{7}$), the minority of motions suffered a bad view because the method gives the majority a good view.

The proposed method assumed that the learner will sit and see the virtual teacher in front of him at the same horizontal sight level as if a real situation. The method only controls the view’s horizontal rotation angle. The vertical rotation and orthogonal view was investigated in this study as well (Section 3.4.3). From the experiment of adjusting the vertical view-angle’s result, we concluded that the more sectors involved in the physical-tasks motion the more vertical view angles are wanted to see the whole motion clear. The method also assumed a gradual slow physical motion. To support fast motions more aspects would need to be considered, such as the segment length. In this evaluation we considered only fixed-length segments; a more dynamic, variable-segment length, based on the amount of motion, may improve the method outcome.

Participant Feedback:

A questionnaire was completed by the participants after each session. The questionnaire consists of the four questions shown in 3.2 as well as a free-feedback field. 3.2 shows the answers to the questions. Though the number of participants was 21, we excluded a few incomplete questionnaires and used 18 as the result.

We noted a number of tendencies in the answers, and confirmed the appropriate design of the experiment. Although only a three-point scale was used for answering the questions, and answers were not analyzed statistically given their small number.

Regarding the HMD, many participants felt it was somewhat troublesome to wear. They mentioned that the HMD’s weight was rather cumbersome, which is a common reaction to the HMD in general. Some participants noted that the HMD’s resolution was adequate but that they had expected better. The score here seemed somewhat less for the fixed rotation-angle conditions. This might be because the relatively crude resolution makes it more difficult to see the more distant, and thus smaller, motions of the virtual teacher in the fixed conditions. The majority reported some feelings of anxiety, as this was the first time they had used an
3.6 Study Limitations

This study has various limitations such as the chosen simple push-button physical task as a learning model. This task is considered simple to perform, because the learner needs only to move his/her hand and push one of the buttons. The learner’s body motion itself was not considered as long as the target button was correct. However, a real-life physical task (e.g. cooking task, sport task, dance task, etc.) may be better to accurately evaluate the proposed method. Next, the sample size is considered small which makes the generalization difficult. Even though the result shows significant differences; a larger balanced sample size is required to robustly adopt the results. This study did not deal with the gender and/or age differences, so the result obtained in this study may be limited to relatively young males. Taking those gender and age differences into account can be a future study issue. Finally, the 3D virtual teacher’s avatar was implemented using plain cylindrical model to minimize any effect of the virtual teacher model’s appearance. However, the virtual teacher’s shadow was not implemented at this stage. The effect of the shadow has been found very important in 3D feelings (92) and without it the motion and the distance may be falsely interpreted. Accordingly, shadow has to be considered in any future implementation.

3.7 Summary

In this study, we investigated the virtual teacher’s optimal setup for the best learning outcomes. More specifically, we investigated physical-task learning when mimicking a virtual teacher’s motion in a mixed reality environment. The outcomes were measured in terms of the required time to accomplish the physical task and the number of committed errors.

Table 3.2: Participant feedbacks after conducting physical-task learning in MR environment

<table>
<thead>
<tr>
<th>Question</th>
<th>Fixed (N=10)</th>
<th>Proposed (N=8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disagree</td>
<td>Age</td>
</tr>
<tr>
<td>Wearing the HMD and moving around was easy and comfortable</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>The virtual teacher’s motion was easy to follow</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>The virtual teacher’s motion was slow</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>The experiment’s session duration was short</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

augmented reality system. Regarding the virtual teacher’s view and motion, the participants seemed to feel somewhat easier in following the virtual teacher’s motion with the proposed method. Some participants felt that the motion task was too simple, but the motion speed was confirmed as appropriate for the task. Regarding the last question, on the session’s duration, it was confirmed that the experiment was not too long to have an effect on the result.
The virtual teacher’s optimal setup investigation involved two experiments. The first, preliminary experiment was conducted to narrow down the wide range of possible virtual teacher positions and rotations; and based on its results; the second experiment was conducted to determine the optimal virtual teacher position and rotation for the best learning outcome. The experimental results suggest that the 3D virtual teacher’s close side-view is the optimal view for such physical-task learning (which includes one-hand movement), and that displaying a semi-transparent virtual teacher has no significant effect on the results.

Afterward we proposed a method for automatically adjusting the virtual teacher’s rotation angle when the virtual teacher is demonstrating physical-task motion. This method will ensure that the learner sees most of the teacher’s motion from an optimal close-viewing angle. To determine whether the automatic adjustment method would produce a better view, comparative physical-task learning experiment was conducted. The first part of the learning experiment was performed using three predefined, fixed-rotation angles for the teacher view. The second part was performed using the teacher’s automatic adjustment method. The result showed that the automatic method scored a lesser error rate compared to the fixed-rotation angle method.

The proposed automatic adjustment method is significant for physical-task learning because such learning is mainly done by observation and nonverbal cues which plays important rule in conveying the correct message. This study shows that mutual body rotation adjustment is very important in communication where physical motion is involved.
Chapter 4

BHS: A Study of the Effects of Surrounding Alignment

In daily life, numerous factors affect how effective our communication is. The surrounding is one of these factors. The most common surrounding factors that affect communication are: background, lighting, distractions, distance, viewing angle, and noise. Surrounding alignment refers to the alignment of surrounding aspects such as environment or setting, space, dimension, border, background, location, etc. In F2F communication, the shared surrounding plays an important role in smoothing the communication process. However, in MC, two different contexts are involved, which may affect communication process. Aligning the surrounding help enhancing MC. In this study, we investigated the background effect in videoconferencing communication. To achieve unified surrounding, we proposed a novel method by the superimposition of other person’s extracted figure on local site front view. In this way, the user feels as though the other person is before him/her in his/her spatial environment. The method has been implemented in the “Being Here System (BHS)” accordingly. A system evaluation experiment revealed that the proposed method significantly enhanced communication by improving the perceived presence of other person, in comparison with conventional videoconferencing. In this study, we also investigated user’s communication behavior in the implemented high-presence videoconferencing, in comparison with a conventional videoconferencing. User’s communication behavior, verbal and nonverbal, is likely to be affected. The analysis of recorded video of system evaluation experiment revealed that the proposed system significantly affected user’s communication behavior such as speaking turn-taking, speech overlapping, total speech time, and number of individual gaze-off, while no significant effects were found in terms of the percentage of individual user speech and percentage of gaze-off.
4.1 Introduction

F2F communication is an indispensable activity in our lives. Traditionally, two or more people had to be physically present at the same time and place in order to talk to each other. This situation will remain the preferable method of communication. But in those cases where people cannot meet in person, videoconferencing offers an increasingly popular solution, with numerous benefits, among them travel cost/time reduction, convenience, and leveraging a global workforce (19). However, this solution may affect the communication process because a person may feel a lessened presence of other participants in communicating. Moreover he/she may fail to correctly and/or accurately interpret other people’s behavior. Therefore, one of the goals here is to create a videoconferencing system that is as close as possible to real F2F communication in terms of the felt presence of other people. ‘Presence’ or ‘sense of presence’ refers to “the user’s feeling of connection to the other person with whom they are interacting” (110).

Many studies have suggested that generating a life-sized view is likely to enhance the user’s sense of presence (17, 22, 26, 60, 111). The life-sized view makes it easy to read the other person’s fundamental behavior, such as eye movements, facial expressions, gestures, and postures, which are essential for smooth communication. Being able to do so is very important as it was indicated by many studies, where they estimated that 60% of conversations involves individual gaze and 30% involves mutual gaze (112). The gaze in this context serves at least 5 functions (113, 114):

- To regulate the flow of conversation.
- To provide feedback on how the communication is being perceived by the listener.
- To communicate emotions.
- To communicate the nature of the interpersonal relationship.
- To avoid excess information input.

Large displays can be used to achieve a life-sized view. However, typically, this will mean that a considerable region of the local person’s front view will be replaced by the remote site’s background, and this may decrease the user’s sense of presence, as there is no integration or continuity in the local person’s front view.

Another important issue is related to the amount of received view’s information from the other site. In many cases, the user might not be interested in seeing what goes around in the remote site. This might raise some privacy concerns for some people as well. Many studies had discussed the privacy issue in details as in (44, 115, 116, 117). Moreover, the remote site’s background in some environments might be ‘cluttered’ with static or movable objects as shown in figure 4.1.A. This may either be a distraction or be more engaging, giving a greater sense of the other person’s environment (56).

Therefore, we proposed a novel solution “BHS” to achieve high-presence videoconferencing through view’s background alignment (7). The system provides the communication environment, in which other person’s life-sized figure is visually situated in the local site (Fig. 4.1.B). The display shows site A’s front view, which would otherwise have been obstructed by the display, as a background. In this way, User A feels as if User B is present before him/her in the same room. BHS was evaluated by a questionnaire filled by users after performing a videoconferencing experiment. The results revealed that the BHS achieved a high sense of presence of other person, in comparison with a conventional videoconferencing system.
In this study, we further investigated the user’s behavior when communicating using BHS. The motivated question was whether BHS affects verbal and/or nonverbal communication structure. The considered verbal communication parameters in this study were: the person’s turn-taking, percentage of speech time, percentage of overlapping speech, and percentage of total speech time. Regarding the nonverbal parameters we considered the number and percentage of person’s gaze directions.

The user’s communication behavior analysis revealed that BHS had significantly affects number of person’s turn-taking, percentage of overlapping speech, percentage of total speech time, and number of individual gaze-off, while no effects were found in terms of the percentage of individual person speech and percentage of gaze-off.

4.2 Related Works

There have been various studies done on MC and media spaces, and a host of systems have been developed over time. Many of these studies have been devoted to proposing and/or implementing methods aimed at enhancing the sense of presence in videoconferencing.

An early system, introduced by Sellen et al., called “Hydra”, sought to enhance the sense of presence by supporting directional gaze cues and selective listening in 4-way videoconferencing (24). In this system, each participant was presented by a single monitor (8 cm diagonal), camera, and speaker. In this way, each participant occupied a stable and unique personal space, as if located in a real F2F meeting. However, given that the images of the participants in this system are small, it may be difficult for participants to read other participants’ facial expressions and gestures, which may affect the sense of presence. To overcome this shortcoming, a multi-party videoconferencing system called “MAJIC” was constructed by Okada et al. (17). In this system,

Figure 4.1: A) Normal videoconferencing environment. B) Proposed superimpose videoconferencing environment
4.2 Related Works

Life-sized video images of participants were projected onto a large curved transparent display. To support eye contact, each video camera was set up behind a transparent display at the center of a partner’s face. When user A turns his head to the right to look straight at user B, user B sees user A full-face and user C sees the left profile of user A. Despite such advances, this system suffered from significantly poor image quality. Because of the nature of the display employed, the image was neither bright nor clear, and the system’s sheer size made it impractical to install on site. Set up the camera at the center of user’s face to support direct gaze while the user facing the display is hard to achieve. Placing the camera over or below the display is proven to attain the same affects as long as the gaze deviation is less than 5 degrees (118). Acquiring large displays might be a costly choice in some cases. Some studies suggested using a tiled display instead to achieve the same objectives (119). Ishida et al. proposed a large scale tiled display environment to achieve high realistic sensation (111). A recent study by Bi et al. investigated the effects of interior bezels of tiled-monitor large displays (120). This study concluded that the number of interior bezels did not show any chief effect on either search time or error rate.

Another line of research focused on the seating arrangement in video-mediated meetings, in order to enhance the sense of presence. A videoconferencing system called “HERMES”, introduced by Inoue et al., sought to integrate F2F and video-mediated meetings (26). This study found that the video image should be placed where a viewer need to make no effort in order to see it. This finding is especially important when using normal displays such as those of computers or televisions, where users need a clear frontal view, as the picture is two dimensional.

A different approach to enhance the sense of presence was introduced by Morikawa et al. (39). In this study, a system called “HyperMirror” was constructed, in which all participants were meant to feel as if they were sharing the same virtual space. Both local and remote participants appeared together, life-sized, on a shared wall. This was done by extracting the participants’ figures, integrating these, and displaying the same resulting view to all participants. The implemented method’s resulting sense of presence was rated higher than that produced under normal conditions (i.e., when only the other participant was displayed).

To provide a greater sense of presence than had been achieved with conventional desktop videoconferencing, Gibbs et al. created the “TELEPORT” system (22), which was based on special rooms, called display rooms, in which one wall was a “view port” into a virtual extension. The geometry, surface characteristics, and lighting of the virtual extension were designed to closely match the real room to which it was attached. The viewing position of the local viewer was tracked, allowing imagery appearing on the wall display to be rendered from the viewer’s perspective. This system provided a shared physical context by superimposing other people’s life-sized figures onto a large virtual extension view. Our proposed system, on the other hand, provides the required shared physical context by superimposing other peoples’ figures onto the local person’s real front view using an appropriate display size.

It is natural to devote more attention to people present before one, since the felt presence of remote people is considerably weaker (121). To overcome this inclination, robotic means have been employed to convey the sense of presence in videoconferencing, enhancing the remote people’s felt presence. In this regard, a study by Sakamoto et al. investigated the effect of using a humanoid robot system as a telecommunication medium (32). The results confirmed that participants felt a stronger presence of the other person when he/she spoke through the humanoid robot than when he/she appeared on a video monitor. Another study, by Yankelovich
et al., introduced a system called “Porta-Person” to enhance the sense of social presence for remote-meeting participants (121). This goal was achieved by providing a high-fidelity audio connection and a remotely controlled telepresence display with video or animation. In the same manner, Venolia et al. developed a telepresence device, called “Embodied Social Proxy (ESP)”, which represented a remote coworker at roughly human-scale (29). In this system, they found that the physical presence of the ESP was a powerful reminder of the presence of the remote worker in the meetings.

The studies above focused primarily on creating a high-presence media space. Our study, in turn, makes its own contribution to this field. To mimic real situations, the other person’s figure should be presented locally, without his/her background. Typically, this can be achieved by using mixed-reality (MR) technology and special head-mounted display (HMD) equipment (59, 60, 61). Such a solution, however, implies that only one site will sense the realistic effects, while the other site will not. This setup is also likely to decrease the sense of presence. In contrast, our proposed method can be easily implemented in both sites, allowing both participants to experience the same effects.

Commercial Videoconferencing:

In commercial videoconferencing business firms, many solutions have been introduced under the name “Telepresence” technology for high presence feelings. Telepresence is defined as an illusion that a mediated experience is not mediated (110). In videoconferencing experience, telepresence gives you the feeling as if the remote participants are in the same room with you. To create the same-room illusion, some commercial telepresence solutions use a combination of technology elements, such as utilizing large displays for life-sized dimensions and hidden high-definition cameras strategically placed to create the appearance of a direct eye contact, and environmental design, such as consistent furniture arrangements across locations. The life-sized dimensions allow participants to see facial expressions, make eye contact, and read body language. Such solutions are: Cisco TelePresence TX9000 Series1, Polycom® RealPresence™ Immersive2, TANDBERG3, HP Halo4, PeopleLink TelePresence5, etc.

On the one hand, these solutions simulate high presence meeting environments as if the other people are sitting across the table in the same room. But on the other hand, these solutions are very expensive, require large-spaces, and have to be installed in a fixed environment with pre-installed matching furniture in both sides to achieve maximum telepresence feelings. In contrast, BHS can be implemented using an affordable equipments and can be installed easily almost anywhere.

Verbal and Nonverbal Communication Analysis:

It is well known that in F2F communication, people switch speaking and listening by using a complicated mechanism of verbal and nonverbal cues (122). A major nonverbal cue in speaking involves the use of eye contact (112). In F2F communication, failure to maintain eye contact is commonly considered to be a sign of deception, and leads to feelings of mistrust (122). Vertegaal et al. concluded that gaze is an excellent predictor of conversational attention in multiparty conversations (123). A study by Karmer et al. proposed a method of measuring people’s sense of presence in videoconferencing system based on linguistic features of their dialogues (124).

---

2http://www.polycom.com/products/telepresence_video/
3http://www.tandberg.com/
5http://www.peoplelink.in/telepresence.html
This study shows that 30% of the variance in self-reported presence can be accounted for by a small number of task-independent linguistic features.

The seating arrangements on group video communication affect participant’s behavior as well. A study by Inoue et al. presented a videoconferencing system “HERMES” that integrates F2F and video-mediated meetings (26). In this study they observed that participants tended to pay much attention to the monitor when using lined-up seating arrangement. This problematic behavior solved by the combination of round seat arrangement and multiple monitors. Another study by Yamashita et al. revealed that seating arrangements affect speaker switches without verbal indication of the next speaker (25). This study found that in some seating arrangement, the participants shared a higher sense of unity and reached a slightly better group solution.

Our study as well examined the proposed high-presence videoconferencing system for any verbal and/or nonverbal effects on communication comparing with a normal videoconferencing system.

4.3 Application Design

4.3.1 System Proposal

The following design requirements were identified as necessary to the proposed high-presence media space system:

- A life-sized video image of other person’s figure without his/her background.
- A shared physical work space that is integrated into real local site’s work space.
- Direct eye contact and gaze awareness.
- Use of a display capable of showing a life-sized image of an adult person’s upper body.
- Support communication within a social distance range: the social distance for interaction among acquaintances was assumed to be from 1.2 to 3.7 m (10).

Figure 4.1.B is a conceptual illustration of BHS scenario. The system provides a MC environment with other person, in which other person’s life-sized figure is visually situated in the local site. In this scenario, local User A is talking to other User B through the high-presence media space system. User B’s figure has been extracted from the remote site and superimposed on A’s local site large display. The display shows site A’s front view, which would otherwise have been obstructed by the display, as a background. In this way, User A will feel as if other User B is present before him/her in the same room. This should enhance A’s sense of presence of other person B.

4.3.2 System Implementation

An experimental media space system was constructed based on the system proposal, for people to communicate within a social distance. The constructed system consisted of two isolated sites, ‘Site A’ and ‘Site B’. The two sites were connected over a local network to permit the exchange of live video. The person in Site A could not see or hear the person in Site B without the system. Figure 4.2 shows the physical workspace of one site. Each site was equipped with a display installed upright 70 cm above the floor, a USB camera fixed behind the display, a Kinect™
4.3 Application Design

depth camera\(^1\), a computer connected to the network, two speakers and a microphone, and a chair. The user was seated at 1.2 m distance from the display as this was indicated to be the best distance for F2F meetings (125).

![Physical workspace of BHS](image)

The person’s figure and voice volume were so defined as to reflect the distance between the person and the display. The closer the person was to the display, the larger and louder they were in the other site, and vice versa. The site view was captured by the Kinect depth camera placed above the display, and the sound was captured by the microphone placed in front of the display. Two speakers were fixed on the sides of the display. Fig. 4.3 shows the internal process flow diagram of the two sites.

![Internal process flow diagram of BHS](image)

\(^1\)http://www.xbox.com/kinect
4.3 Application Design

4.3.2.1 Capturing the Local Site’s Front View

We used a high resolution USB camera to capture the local site’s front view, that is, the region concealed behind the display. The USB camera was placed behind the display in the center, and the camera’s angle and zoom were calibrated so that the area behind the display was exclusively captured (Fig. 4.4). This captured image was used as a background for the display.

![Figure 4.4: Setup USB camera behind the display in the center](https://example.com/figure4.4)

4.3.2.2 Extracting the User’s Figure from the Image

A popular technique to extract a person’s figure from its background involves placing the person in front of a uniform blue or green background. This technique is called “Chroma Keying”, and has been widely used in television and movies. But this technique requires a special environment and professional light setup which may be difficult to implement in some situations; and if implemented it is difficult to relocate from one site to another. Moreover, a chroma key subject must not wear clothing similar in color to the chroma key color(s), because the clothing may be replaced by the background. In BHS we used another technique, utilizing an affordable image depth camera to extract the user’s figure, which can be easily implemented anywhere.

To capture the site view and extract the user’s figure from it at run-time, we used a Kinect depth camera bar, OpenNI API\(^1\) version 1.3.2.3, PrimeSense’s NITE middleware\(^2\) version 1.4.1.2, and PrimeSense’s SensorKinect\(^3\) version 5.0.3.4. The Kinect sensor was placed over the display and focused on the person’s face in order to achieve the required direct gaze awareness. This is considered to be a best placement given the constraints of the environment (56). The person was seated 1.2 m from the sensor. This distance is within the social distance range required for the person to sense a connection with the other person. The OpenNI API provided the necessary means to analyze the Kinect image depth data, to distinguish between the foreground and background, and to identify the users in the scene. Fig. 4.5 shows the Kinect’s sensory depth data, analyzed by the OpenNI scene analyzer. The tracked user was colored

---

\(^1\)http://www.openni.org/
\(^2\)http://www.primesense.com/
\(^3\)https://github.com/avin2/SensorKinect
blue, and this image was used as a mask to remove the background from the original colored image and replace it with a transparent color (Fig. 4.6).

![Figure 4.5: OpenNI scene analyzer view](image)

![Figure 4.6: The view after removing background](image)

### 4.3 Application Design

#### 4.3.2.3 Transmitting the User’s Figure Images

A software module was developed to send/receive the extracted user’s figure images over a local network to/from the other site. Windows® Sockets 2 (Winsock)\(^1\) was utilized in our system for this purpose.

In order to optimize the process and increase the transmission speed for smooth broadcasting, a simple method was adopted, which involved the transmission of only the image data within the rectangle containing the user’s figure (Fig. 4.6). In this way, the image data was reduced by an average of 45%.

\(^1\)http://msdn.microsoft.com/en-us/library/windows/desktop/ms740673(v=vs.85).aspx
4.3.2.4 Superimposing the Received Other Person’s Figure on the Local Front View

The final step in the process was to superimpose the received person’s figure onto the local front view (Fig. 4.7). This was accomplished by merging other person’s figure and the background. The resulting image was acceptable, but the edges of other person’s figure were rough (jagged). In order to enhance the image, it was necessary to smooth other person figure’s edges. The Gaussian smooth function from the OpenCV\(^1\) library was used for this purpose. Finally, the resulting view was presented on a large display.

![Superimpose Remote User's Figure on Local Site View](image)

\*Figure 4.7: Superimposing other person’s figure on local front view background*

4.4 Evaluation

In this section, the evaluation experiment is discussed. The main objective of this experiment was to determine whether BHS would enhance the participants’ sense of other person’s presence, in comparison with conventional videoconferencing. Displaying other person’s life-sized figure alone, after removing his/her background, was expected to increase the sense of presence. The participants’ feelings and feedback were assessed by means of a questionnaire filled out by each of the participants separately, after each experimental session.

4.4.1 Setup

Using the constructed experimental media space, we conducted an evaluation of the proposed system, with recruited participants. Fig. 4.8 shows the two sites’ physical configuration. In Site A, a large flat-panel display (46 inches) was used, while in Site B, a 30 inches display was used. The 30 inches display was fixed in a vertical portrait position, presenting a life-sized image of

\(^1\)http://opencv.willowgarage.com
4.4 Evaluation

an adult’s upper body (Fig. 4.9). We used the portrait mode to study the effects of the display size on the sense of presence.

**Figure 4.8:** Physical configuration of two experimental sites

**Figure 4.9:** Site B’s view, employing a 30” portrait display

### 4.4.2 Conditions

Two videoconferencing modes were established to enable the participants to communicate with each other: “Normal mode”, in which the remote site view was displayed as is in the local site’s display (Fig. 4.10); and “Superimpose mode”, in which other person’s figure was extracted and superimposed on the local site’s front view (Figure 4.11).

In this experiment, we considered the following videoconferencing conditions, to evaluate the constructed system:

- **Large Superimpose**: superimpose mode via large display.
- **Portrait Superimpose**: superimpose mode via portrait display.
- **Large Normal**: normal mode via large display.
- **Portrait Normal**: normal mode via portrait display.

### 4.4.3 Participants

A total of 18 participants took part in the experiment, 7 females and 11 males. The participants’ ages ranged from 23 to 36 years old, most were undergraduate or graduate students from
the same university, and 17 participants had had previous experience using videoconferencing systems. Most used the videoconferencing principally to talk to a remote family member and/or a remote close friend. The participants were divided into groups of two. Within each group, we made sure that the participants were familiar with each other as we wanted the interaction to be as smooth as possible.

4.4.4 Procedure

In each experiment, a group of two participants were recruited to perform videoconferencing tasks with each other. One of the participants used the system at Site A, while the other used Site B. Before performing the videoconferencing tasks, the participants were asked to complete a basic demographic survey. After this, the researcher introduced the system to the participants. The experiment began with a familiarization session for five minutes. Each participant performed four videoconferencing sessions to test the conditions (two sessions at Site A, and two sessions at Site B). In each session, participants were instructed to talk about a selected general topic for an average of 10 minutes. After that, they were asked to complete the questionnaire about the system they experienced, independently of each other. The four general topics were:

- Study life in X city: discuss with the other person the pros and cons of studying in X city; how long you have been in X city; why you choose X university, compare X city with other cities you have been in, etc.
4.4 Evaluation

- Buying a new laptop: discuss the laptop’s specifications; the suggested shops; prices; usage; etc.
- Planning a trip: for the coming summer vacation, discuss the trip’s options; where to go; locally or abroad; cost; weather; attraction; etc.
- What your plans after graduation: discuss with the other person your plans after graduation, the possibility of pursuing a higher degree; work options, etc.

The conditions orders were counterbalanced across participants, to ensure that the order of the tested conditions would not affect the result (Table 4.1).

Table 4.1: Experimental session arrangements

<table>
<thead>
<tr>
<th>Participants</th>
<th>Session 1</th>
<th>Session 2</th>
<th>Session 3</th>
<th>Session 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1, P7, P13</td>
<td>Large</td>
<td>Large</td>
<td>Portrait</td>
<td>Portrait</td>
</tr>
<tr>
<td></td>
<td>Superimpose</td>
<td>Superimpose</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>P2, P8, P14</td>
<td>Portrait</td>
<td>Portrait</td>
<td>Large</td>
<td>Superimpose</td>
</tr>
<tr>
<td></td>
<td>Superimpose</td>
<td>Superimpose</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P3, P9, P15</td>
<td>Large</td>
<td>Large</td>
<td>Portrait</td>
<td>Portrait</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>Superimpose</td>
<td>Superimpose</td>
<td>Normal</td>
</tr>
<tr>
<td>P4, P10, P16</td>
<td>Portrait</td>
<td>Portrait</td>
<td>Large</td>
<td>Large</td>
</tr>
<tr>
<td></td>
<td>Superimpose</td>
<td>Normal</td>
<td>Superimpose</td>
<td>Normal</td>
</tr>
<tr>
<td>P5, P11, P17</td>
<td>Large</td>
<td>Portrait</td>
<td>Large</td>
<td>Portrait</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>Superimpose</td>
<td></td>
<td>Normal</td>
</tr>
<tr>
<td>P6, P12, P18</td>
<td>Portrait</td>
<td>Large</td>
<td>Portrait</td>
<td>Large</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>Superimpose</td>
<td>Normal</td>
<td>Normal</td>
</tr>
</tbody>
</table>

4.4.5 Questionnaire

In the questionnaire, we asked participants to evaluate each of the statements according to the feeling they experienced during the videoconferencing session. The principal aim of the study was to assess participants’ sense of the other person’s presence while using the constructed media space. The related works discussed in this chapter showed that a life-sized view was highly important to the correct recognition of the other person’s behavior and to the sense of their presence. In this study, we sought to determine whether our proposed method would enhance this sense or not using a common questions from related questionnaires from (126) (127) (125). To investigate the participants’ sense of presence in each condition, the following statements were used:

- Q1: “I felt as if the other person existed in the same room.”
- Q2: “I didn’t feel as if I were talking with the other person in the same room.”
- Q3: “I felt as if I were facing the other person in the same room.”

The feeling of spatial distance between the participants is one aspect of the sense of presence. In this research, our focus was limited to the social distance communication range, which is best for business meeting activities and social interactions. We assumed that participants who experienced the superimpose mode would feel as if the other person were much closer, almost as if ‘there’, in comparison with the normal mode. To evaluate this aspect, the following statement was used:
4.4 Evaluation

- Q4: “I felt that the distance between me and the other person was comfortable for chatting.”

In addition, we asked the participants to roughly estimate the distance between themselves and the other person while videoconferencing:

- Q5: “I felt that the distance between me and the other person was around: ”

A life-sized view is likely to improve a person’s ability to recognize the other person’s facial expressions and gestures. We assumed that the superimpose method would further improve this recognition, in comparison with the normal mode. The following statements were used to evaluate participants’ recognition:

- Q6: “The other person’s facial expressions were easy to recognize.”
- Q7: “The other person’s gaze direction was difficult to recognize.”
- Q8: “The other person’s gestures were easy to recognize.”

Regarding the quality of the presentation in each condition, the following statements were used:

- Q9: “The audio was clear enough to have a conversation.”
- Q10: “The live video was not clear enough.”

To investigate the effect of the display size concealing part of the user’s front view, the following statements were used:

- Q11: “I felt it was difficult to grasp the surroundings because of the display.”
- Q12: “I felt it was important to observe what was going on behind the display.”

All of these statements, except Q5, were rated on a 9-point Likert scale, where 1 = strongly disagree, 3 = disagree, 5 = neutral, 7 = agree, and 9 = strongly agree.

4.4.6 Session Recording Setup

Two cameras were used to record the experiment sessions at HD 720 resolution (1280 by 720 pixels). The first camera was placed over the display facing the participant in order to capture his/her facial expressions, gestures, and postures. The second camera was installed upright 1 m above the floor beside participant in order to capture him/her from the left side and the display content. Fig. 4.12 shows the cameras setup layout.

4.4.7 Results

4.4.7.1 Statistical Results

Figure 4.13 shows the average results of the participants’ sense of the other person’s presence while videoconferencing, under the four conditions. A comparison was done using a two-factor ANOVA test. The first factor is the videoconferencing mode (i.e. Normal and Superimpose). The second factor is the used display (i.e. Large and Portrait). We found a main effect of videoconferencing mode over the participants’ sense of other person’s presence as if in the same room, (Q1: $F(1,68) = 55.26, p < 0.01$), (Q2: $F(1,68) = 14.08, p < 0.05$), and (Q3:
4.4 Evaluation

Figure 4.12: Layout of used video cameras

\[ F(1, 68) = 31.71, p < 0.01 \] This indicates that the superimposed videoconferencing mode enhanced the presence feelings more than the normal videoconferencing mode. On the other hand, the results shows no main effect of the used display over the participants’ sense of other person’s presence as if in the same room, (Q1: \( F(1, 68) = 0.31 \)), (Q2’: \( F(1, 68) = 0.0 \)), and (Q3: \( F(1, 68) = 0.06 \)). The result also shows that there is no interaction between the used mode and display over the presence feelings, (Q1: \( F(1, 68) = 2.31 \)), (Q2’: \( F(1, 68) = 2.73 \)), and (Q3: \( F(1, 68) = 1.34 \)).

Figure 4.13: Participants’ average sense of presence results. (Note: Q2’ result is positive form of original Q2)

Moreover, we found a main effect of videoconferencing mode over the feeling of comfortable distance between the user and the other person (Q4: \( F(1, 68) = 7.14, p < 0.01 \)), while no main effect of the used display was reported (Q4: \( F(1, 68) = 0.89 \)). The result also shows that there is no interaction between the used mode and display over the distance, (Q4: \( F(1, 68) = 0.62 \)). This indicates that the superimposed videoconferencing mode enhanced the feeling of comfortable distance between the user and the other person. In addition, participants who used the superimpose videoconferencing mode were able to estimate the distance more accurately. The average estimated distance was as follows:

- Large Superimpose: 1.3 m (s.d. = 0.6).
4.4 Evaluation

- Portrait Superimpose: 1.3 m (s.d. = 0.6).
- Large Normal: 2.2 m (s.d. = 1.2).
- Portrait Normal: 1.8 m (s.d. = 1.0).

(The actual distance between the participant and the display was 1.2 m).

Figure 4.14 shows the average results of the participants’ feedback concerning the quality of the presentation, for each condition. The statistical results revealed that there were no significant differences among the conditions with respect to the participants’ ability to recognize the other person’s facial expressions (Q6: $F(2, 34) = 0.53, \text{not significant}$), gaze direction (Q7: $F(2, 34) = 3.92, \text{not significant}$), and gestures (Q8: $F(2, 34) = 1, \text{not significant}$). Moreover, there were no significant differences among the conditions with respect to the audio quality (Q9: $F(2, 34) = 0.12, \text{not significant}$), and video (Q10: $F(2, 34) = 0.74, \text{not significant}$).

![Figure 4.14: Participants’ average response concerning the quality of presentation](image)

Figure 4.15 shows the average results of the participants’ feedback concerning the effect of the display size concealing part of the user’s front view. The statistical results showed a significant difference in participants’ sense of difficulty in grasping the surroundings because of the display (Q11: $F(2, 34) = 6.32, p < 0.05$). Furthermore, the results showed that participants preferred to see what was concealed by the display, regardless of the display size or the videoconferencing mode (Q12: mean = 6.3).

![Figure 4.15: Participants’ average response concerning the effect of display size](image)
4.4 Evaluation

4.4.7.2 Analysis of Video Data

ELAN\(^1\) tool was used to annotate the recorded video of experiment sessions. A total of 36 recorded video (9 groups by 4 conditions) were annotated for user’s communication behavior such as talking and gaze. Only the middle 2 minutes of each session were analyzed at this stage of analysis (a total of 72 minutes of data). Fig. 4.16 shows a screenshot of one of the ELAN’s annotated video.

Figure 4.16: Screenshot of one of ELAN’s annotated video

The following terms were used in annotating and analyzing the recorded video sessions:

- **Talking**: happens when a person speaks for at least 1.5 seconds (128).
- **Turn-taking**: In conversation analysis, turn-taking term is defined as the manner in which orderly conversation normally takes place. The principles of turn-taking were first described by sociologists Sacks et al. in (129). In this study, we adopted the same turn definition from (130) as the person’s number of continuous segment of speech between silent intervals for at least 1.5 seconds.
- **Overlapping**: is a simultaneous speech by two persons. This might happen as speaking turn change or as simultaneous reply to other person’s speech while talking.
- **Gaze**: happens during a conversation when two people look at one another (112).
- **Gaze-off**: this term is defined for the analysis in this paper. It happens when the person avert his/her gaze from other person.

**Talk Analysis:**

Figure 4.17 shows the average results of the participants’ number of turn-taking while videoconferencing, under the four videoconferencing modes. A comparison was done under the four conditions using a one-way repeated-measures ANOVA test. We found a significant difference in number of turn-taking \(F(3,51) = 6.49, p < 0.05\). A Tukey’s HSD post-hoc test was performed in order to determine which condition’s mean was different from the others. For this aspect, we found that the superimpose conditions were significantly different from the normal conditions.

\(^1\)http://www.lat-mpi.eu/tools/elan/
4.4 Evaluation

Figure 4.17: Participants’ average number of turn-taking per minute

Figure 4.18 shows the average results of each participant’s percentage of speech while videoconferencing. For this aspect, we found no significant difference between the tested conditions ($F(3, 51) = 0.48$).

Figure 4.18: Participants’ average percentage of speech

Figure 4.19 shows the average results of the participants’ percentage of overlapping talk while videoconferencing. We found a significant difference in percentage of overlapping ($F(3, 51) = 11.69, p < 0.01$). For this aspect, we found that both the superimpose conditions were significantly different from the normal conditions.

Finally, Fig. 4.20 shows the average results of the both participants’ percentage of total speech while videoconferencing. We found a significant difference in the percentage of total
4.5 Discussion

Figure 4.19: Participants’ average percentage of speech overlapping

speech \( F(3, 51) = 2.28, p < 0.1 \). For this aspect, we found that both the superimpose conditions were significantly different from the normal conditions.

Gaze Analysis:

Figure 4.21 shows the average results of the participants’ number of gaze-off while videoconferencing. We found a significant difference in number of gaze-off \( F(3, 51) = 4.83, p < 0.05 \). For this aspect, we found that both the superimpose conditions were significantly different from the normal conditions.

Figure 4.22 shows the average results of participants’ percentage of gaze-off while videoconferencing. For this aspect, we found no significant difference between the conditions \( F(3, 51) = 1.1 \).

4.5 Discussion

Presence Feelings:

The constructed superimpose videoconferencing mode, using either large or portrait display, improved participants’ sense of presence of the other person, in comparison with the normal mode. This result supports our assumption that background alignment through displaying the remote site’s life-sized figure alone (after removing the remote site background) would increase the realism of the view and the sense of the other person’s presence. The portrait display was introduced to study the effect of the displayed background size if any. The results showed that no difference between the conditions across the same videoconferencing mode (superimpose and normal). One participant mentioned that the portrait display’s wide border consumed a considerable amount of the front view compared with the large display. A study by Bi et al. on the effects of bezels of large tiled display revealed that bezels affect tunnel steering (120).
4.5 Discussion

Figure 4.20: Participants’ average percentage of total speech

Figure 4.21: Participants’ average number of gaze-off per minute
The superimpose mode simulated an actual F2F communication configuration. Some participants noted that the superimposition of the other person’s figure onto the local front view made it appear as if they were seeing the other person in three dimensions. Some of the participants enjoyed the superimposing mode, with some attempting to shake hands with each other through the system, and others moving and rotating and asking the other person to do likewise. In the normal mode, nothing curious in participants’ behavior was noticed.

Regarding the estimation of the distance between the participants during the videoconferencing sessions, the superimpose mode tended to result in a better distance assessment, compared to the normal mode, possibly owing to the fact that participants managed to compare the relative distance of other person’s figure with respect to the local site’s front view shown in the display.

On the other hand, we expected that our method would improve the recognition of the other person’s essential behavior; however, the results showed no significant differences between the conditions, perhaps because the life-sized view made it easy to recognize the behavior despite the videoconferencing conditions.

The superimpose videoconferencing mode, using either large or portrait display, reduced participants’ sense of difficulty in grasping the surroundings, principally because of the display size used, in comparison with the normal mode. This result supported our assumption that displaying the region concealed by the display would enhance the person’s awareness of the surroundings. Some participants noted this effect, especially when someone walked behind the display. They said it was very interesting to see the walking person even when he went behind the display.

Finally, we expected that our method would reduce the need to observe what was going on behind the display, since the concealed region was shown in the display as background. The result showed no significant differences between the conditions in this regard. Despite this result, the total average result showed that participants felt that it was pleasant to see what was concealed by the display.

**Verbal/Nonverbal Communication:**
The high presence feelings affected significantly the participants’ verbal communication. The results revealed that the proposed system drew more turn-taking comparing with normal videoconferencing system (Fig. 4.17). The superimpose mode increased the number of turn-taking by 1.4 more than the normal mode. This result was expected as some related research found that video-mediated condition tends to increase turn length relative to F2F conversations (131, 132). The participants’ average percentage of speech was not affected by the tested conditions (Fig. 4.18). The result shows that each participant talked on average 48% of the session time. This result is consistent with a related research by Sellen et al. (130). However, we found that the percentage of speech overlapping in the superimpose conditions were twice more than the normal conditions (Fig. 4.19). In Sellen et al. study they found that F2F imposes more simultaneous talk compared with video conditions. From this we can conclude that our proposed superimpose mode was closer to the F2F in this aspect. Finally, the result shows that the total speech time in the superimpose conditions were higher than normal conditions (Fig. 4.20). The participants while using the superimpose mode felt more engaging in the conversation as if in the same room and accordingly spoke more. A study by Mantei et al. reported that people tended to talk more often with others in the same office than with others in a remote office, when a media space connected two offices (133).

In our study, we investigated the gaze-off (avert) aspect while communicating under the tested conditions. The result shows that the participants tended to avert there gazes more when they used the superimposed compared to the normal videoconferencing conditions (Fig. 4.21). In F2F conversations, people use more gaze when they are further apart (112). The participants whom used the superimposed conditions felt more close to the other person, and accordingly they averted there gaze more as if in F2F.

4.6 Study Limitations

In this study, the hidden area behind the display was captured using a USB camera fixed behind the display. The captured view was initially clipped and calibrated to provide good front view integration for the user who supposed to set and see the display from a specific location. In real, this assumption cannot be guaranteed since the person will keep adjust his posture while communicating. In this case, the user’s front view suffered some view misalignment between the real front view and what displayed in the screen. This in turn may affect the communication. Keep adjusting the displayed view based on the user’s location may solve this issue. But this requires detecting the users exact view angle which might be a difficult process. A better way is by using a transparent display instead of normal display and USB camera. This may provide a better and naturally adjustable integrated view which in turn may enhance the communication.

The user’s figure extraction was achieved by a simple extraction technique using depth camera. The resulting image was acceptable, but the edges of the user’s figure were rough (jagged). We tried to enhance the image through smoothing the edges using Gaussian smooth function. This enhanced the image but the jagged edges were still noticeable. This captured the participants’ attention which in turn might affected the communication. Finally, in this study we only compared our method with the traditional videoconferencing method. Other studies considered other background effects such as blurring the background as in (56). To better evaluate our method, a comparison with other background alignment techniques has to be considered.
4.7 Summary

In this study, we investigated the surrounding alignment effects on communication. More specifically, we investigated other person’s presented background effect on videoconferencing communication. To achieve unified surrounding physical context, we proposed a novel method involving superimposition other person’s figure on local site front view. In this way, the local person felt as if other person were present before him/her in his/her spatial environment. The design and construction of a high-presence media space called “BHS” was presented, along with the method, results and discussion of a system evaluation experiment. BHS effectively reduced the psychological distance between the remote participants. Also we investigated some verbal and nonverbal user communication behavior while using the system comparing with a normal videoconferencing system. The users’ communication behavior analysis revealed that using a high presence media space affects some verbal and nonverbal aspects. In the verbal communication, we found that the number of turns-taking, speech overlapping, and total speech time were significantly higher in the proposed high presence media space. Furthermore, the nonverbal gaze-off aspect was significantly higher as well.
Simultaneous communication with a remote person requires that both persons are available at the same time and/or ready to communicate. Because of time-zone differences and other such contingent factors, this condition can often be hard to fulfill. One solution may lie in time-shifted or delayed communication, which may affect communication process. In a time-shifted environment, achieving synchronization is a challenge. However, through close chronological alignment, this effect can be minimized. Chronological alignment refers to the alignment of chronological aspects that relate to events arrangement in the order of time. In this study, we investigated the effects of events synchrony in time-shifted tele-dining application. We proposed a system “KIZUNA” enabling people to enjoy a meal together in a virtual environment involving the transmission of recorded video messages. The system achieves enhancement through synchronizing the start of dining sessions and adapting the displayed video’s playback speed to the difference in dining progress between local and remote person. This is likely will create a sort of real-time behavioral mimicry which enhances communication and increase enjoyment while dining. A validation experiment revealed that the proposed KIZUNA adaptation method enhanced diners’ communication behavior and perceived presence of other person, in comparison with conventional time-shifted tele-dining. This result suggests a promising future for the KIZUNA system.
5.1 Introduction

During meals, families have a chance to catch up on what has happened in their lives, and to strengthen the bonds that hold the family together. The busy lifestyles of our time, however, often make such family dining very difficult; especially with older people living independently, younger people living alone, and working people traveling more and/or working from afar (134). In these cases, family members must often eat alone, which is a possible source of loneliness and unhappiness (135)(136), in contrast to the pleasure and longer duration typical of social dining (137).

A variety of inexpensive videoconferencing technologies offer a decent solution for distant family members to maintain a sense of social connectedness, which is defined as “positive emotional appraisal, characterized by a feeling of staying in touch within ongoing social relationships” (138). But when family members are located far away from each other, this often means they are also living in different time-zones, and this difference is likely to present further difficulties for communication between distant family members. The main difficulty lies in the misalignment of daily schedules between the two parties (139), and this is clearly apparent at mealtimes.

In real-time group dining, participants tend to wait each other to start dining together. The dining starts when the food is placed on the dining table and all the involved participants are seated around table. When this happen, the dining may start automatically or often after saying a specific phrase depends on the culture. During dining session, many researches had demonstrated that eating behavior is greatly affected by social influences (140). Several studies showed that the presence of others affects the amount of consumed food in a meal, where people tend to eat more in the presence of others than alone (141, 142). Similarly, a person’s eating can be modified by partner’s eating behavior, where a person tends to eat as much or as little as do those with whom he/she eat (143, 144, 145). A possible explanation for this behavior is that the eating participants’ food intake becomes synchronized through processes of behavioral mimicry (140). Watching a partner taking a bite activates, within the person, the same type of action, which may have made him/her more likely to take a bite as well. This natural dining behavior makes dining process goes smooth where all the participants enjoy the meal, conversation, and finish dining more or less together. The difference between the participants’ communication behavior with and without meal was revealed. For instance, the body motion synchrony of table talk was investigated in (146). In this study, the body motion synchrony in three-party table talk was analyzed, where the body motion synchrony in the head was measured. The results revealed that there is a high mutual body motion synchrony between participants. In time-shifted tele-dining environment, achieving this kind of synchronization is not easy therefore it is been researched.

In this study, we propose “KIZUNA” system to overcome the problem of eating alone. The system is principally based on time-shifted communication involving the transmission of recorded video messages (Fig. 5.1). To achieve a sort of real dining synchronization between local dining and the displayed recorded video, KIZUNA controls the time to display recorded video and adapts video’s playback speed to the difference in dining progress (DDP) between the local and remote user. The KIZUNA’s adaptation algorithm was validated by an actual dining experiment using the Wizard of Oz (WoZ) system simulation technique. Utilizing WoZ, we replicated an envisaged future scenario for the KIZUNA system, and the experimental
results revealed that the proposed KIZUNA adaptation method enhanced diners’ communication behavior and significantly enhanced the perceived presence of other person, in comparison with conventional time-shifted tele-dining.

5.2 Related Works

Numerous studies have investigated ways to utilize information and communication technology in minimizing the gap between people living apart. In this section, we briefly review research related to people dining together despite being located in remote places and/or different time-zones.

Synchronous Dining:

Family members and close friends tend to have meals together, but lack of time or proximity often makes this hard to achieve. To overcome this situation, the international consulting firm Accenture introduced a tele-dining prototype called the “Virtual Family Dinner” that would allow families to dine together in a virtual environment\(^1\). This prototype was essentially a videoconferencing system targeting people with limited knowledge of technology, such as the elderly. Thus, Accenture made it highly automated and easy to operate. The system monitors the site and when it detects a meal dish on the table, it goes through a list of contacts, trying to reach one who is available for a dinnertime chat.

The advertising agency Wieden+Kennedy’s Amsterdam office produced a website called Virtual Holiday Dinner\(^2\), enabling scattered friends and family to have a dinner party of up to five people via Skype\(^{TM}\). Guests can call into the dinner, and their faces are shown on the displays as the heads of models physically sitting around a dining table. The models are equipped with video cameras fitted with facial tracking software, so each guest can look around the dining table, ‘from’ the respective model, by moving his/her head.

The sense of coexistence among family members is likely to be affected when living apart. To maintain a heightened sense of coexistence among family members, a system called CoDine was introduced (41). This system consists of a dining table embedded with interactive subsystems.

---

\(^1\)http://gizmodo.com/accenture-virtual-family-dinner/
\(^2\)http://www.virtualholidaydinner.com/
that augment and transmit the experience of communal family dining. CoDine connects people in different locations through shared dining activities, such as gesture-based screen interaction, mutual food serving, ambient pictures on an animated tablecloth, and the transmission of edible messages.

Dining together is also an enjoyable experience. In some cultures, meals are typically consumed in groups, usually made up of family members or close friends. Examples of popular group meals are the Chinese Hot-Pot, the Japanese Nabe, and the Arabic Mansaf. In this regard, remote group-meal communication was investigated to render such meals more enjoyable (147). In the study (of the Chinese Hot-Pot meal), three factors essential to the group-meal experience were identified: interacting with food as a group, a central shared hot-pot, and a feeling that others are nearby.

All such previous research primarily focused on synchronous dining among a group of people living apart, assuming that all participants are available at the same time. Our research, in contrast, focuses on the cases where only one participant is currently available for dining.

**Asynchronous Dining:**

When all participants are available at the same time, dining together can be achieved by a variety of videoconferencing systems. However when only one participant is available, things become more difficult. In this situation, asynchronous (time-shifting) collaboration techniques offer a general solution. Because users can receive the information they want when it is most convenient for them, such techniques have been widely employed in work settings, but they may lack a sense of immediacy. For example, in the distant-learning domain, Ocker et al. revealed that asynchronous collaboration is as effective as F2F collaboration in terms of learning, quality of solution, solution content, and satisfaction with the solution quality (148). However, students were less satisfied with the asynchronous learning experience in terms of both the group interaction process and the quality of group discussion. This suggests that asynchronous interaction is satisfactory for some requirements, but not for others, such as affective satisfaction or social connectedness. Asynchronous video messages have also been employed to support interpersonal relationships in separated families (149). Here, the recipient views the video messages asynchronously, creating a non-stressful, continuous line of communication. This is believed to enhance the connectedness and intimacy between separated family members. Inkpen et al. showed that despite enjoying F2F interaction, close friends used asynchronous video communication tools extensively to augment their existing relationships (45). Time-shifting has also been considered for meetings in the workplace. Tang et al. introduced a system enabling a distant person to contribute to a meeting by pre-recording comments to be played during the meeting when needed (43). The conducted field experiment showed that most of the recorded messages were played in the meetings; however, a lesser percentage of the messages generated in the meeting were reviewed by the distant person.

In the dining domain, Tsujiita et al. proposed the system “CU-Later”, to be used in time-shifted dining (44). This system plays a recorded video of remote dining after a specific time-shift, when the local user is in front of a display placed on a dining table, enabling the local user to watch the video automatically when he/she is eating. As the video is played, the system records the local user’s session as well, so that the remote user can watch the local user’s session later on—rather like a video mail exchange with automatic playback and recording.
Social dining communication means more than just information exchange; it typically also involves affective satisfaction and social connectedness, and may thus be seen as a form of ‘consummatory communication’ (which involves the sharing of experiences, emotions, knowledge, and/or opinions (150)). To support this type of communication, a sense of co-presence is often crucial. Thus, we proposed system addresses these more complex needs, and further explores the potential of video exchange, but is more focused on the social connectedness between remote users, based on the assumption that aligning the dining activities of remote participants may significantly enhance the users’ dining experience.

5.3 Application Design

5.3.1 System Proposal

Figure 5.2 shows a conceptual illustration of the KIZUNA (‘bond’ in Japanese) system scenario. In this scenario two persons, local user A and remote user B, are dining together in a virtual environment, although they are physically dining in different spaces and/or at different times. First (Fig. 5.2 Session 1), a recorded video message from the remote user B is created while dining.

The system automatically starts working at site A when the local person A gets seated and some dishes are placed on the dining table to simulate real-time dining (Fig. 5.2 Session 2). This increases the system-usage portability, especially for people with limited knowledge of the
5.3 Application Design

The system starts playing the video of remote user B (recorder earlier) on a large display in front of the local dining table, at life-size scale. It also starts recording the local user A’s dining and his/her reaction to user B’s video message. Similarly, the system automatically starts working at site B when the remote person B gets seated and some dishes are placed on the dining table (Fig. 5.2 Session 3). The system then starts playing the video of user A, and the same process is repeated.

This automatic playback and recording of videos provides the illusion of co-dining for the users, though they may be dining at different times and/or in different places. We assume that merely watching the other user’s video may not, by itself, arouse a sense of co-dining; but that the synchronization of both dining activities (not only the start and end time but also the entire process) will create a sort of real-time behavioral mimicry. Specifically, this synchronization can be achieved by controlling the video playback speed according to the DDP between the local and remote user. In this way, the local user can enjoy the company of the remote user throughout his/her meal. Also in this way, we ensure that the local user can see the entire remote dining session while he/she is eating. These factors appear likely to enhance the dining experience.

**System Requirements:**

Based on the system proposal, we defined the following set of design requirements to implement the KIZUNA system:

**Dining detections:** dining status and progress are necessary information to control the system. This includes the following:
- User detection: the system must be highly automated and easy to operate. User presence at the dining table should activate the system and display the remote participant’s recorded dining video. This will ensure high system usability and portability, especially for people with limited knowledge of the technology.
- Food detection: detecting a user at the dining table is not sufficient to ensure that a dining session is about to begin; the system must also detect whether food has been placed on the dining table or not.
- Amount of food detection: based on the above information, the system can assess the user’s dining behavior, particularly the user’s dining progress.

**Video manipulation:** in the proposed KIZUNA system, users can communicate through the exchange of recorded video messages. Therefore, the following video manipulations are required:
- Video recording: to record the local dining session.
- Video streaming and controlling: to control the displayed video’s playback speed according to the DDP.

### 5.3.2 System Implementation

An experimental time-shifted tele-dining system is under construction based on the KIZUNA system proposal and relevant design requirements. The experimental site is equipped with the following: dining table, chair, large flat-panel display, two USB cameras, speakers, microphone, motion detection sensor, two spotlights, and computer. Figure 5.3 shows the preliminary system’s physical workspace.

Figure 5.4 shows the system procedure flowchart, comprised of the following major modules:
5.3 Application Design

Figure 5.3: Physical workspace of KIZUNA

Figure 5.4: Procedure flowchart of KIZUNA
5.3 Application Design

**User detection:** The system utilizes a generic motion sensor to detect any human motions within a specified interest volume. The human motion sensor “AT Watch NET IR mini”\(^1\) is used for this purpose. This sensor is fixed above the dining table at a height of 230 cm. This way, the sensor is able to detect any motion within a ground radius of 230 cm.

**Food detection:** This module detects any food placed on the dining table, in two steps; first step: detecting dishes on the dining table; second step: detecting the total amount of food in the detected dishes. In this study, we assumed that all the food would be served ‘upfront’, with the food placed directly onto the plates and brought to the table at once (151).

- **Dish detection:** this is performed using shape and color recognition. A previously devised method (152) is used to detect the dishes. A USB camera is fixed above the dining table at a height of 200 cm, to exclusively capture the dining table from above. At this stage of construction, we considered white circular dishes only. In addition, we used dishes with solid colored rims to easily distinguish the dishes. The free OpenCV library is utilized to perform this task, and to determine the dishes’ position on the dining table.

- **Amount of food detection:** this is achieved using image processing techniques. A previously devised method (152) is used to estimate the remaining food from a 2D image of the dishes. At the beginning of the dining session, all the detected amount of food in all the dishes is summed together to produce a total initial amount of food. Periodically, this detection is repeated to determine the amount of consumed food compared to the total initial amount. The process ends when the user is no longer present at the dining table and/or the remaining amount of food becomes zero.

**Video recording:** Recording of a dining session starts when actual dining activity is detected. This includes a person sitting at the dining table (User detection), and food placed on that table (Food detection). Accordingly, the beginning of a dining session is indicated by the diner moving his/her hand over one of the dishes. This is achieved using hand detection by the background subtraction method. A previously devised method (152) is used to detect the user’s hands on the dining table.

As soon as a dining session starts, a USB camera placed over the large display in front of the person starts capturing images, and a pin microphone starts capturing sounds. The Microsoft DirectShow application programming interface (API) is used to create a media file (.avi) from the captured images and sounds. During the dining session, periodic dining progress information is collected and saved to a designated metadata file. This information includes a time stamp, a meal progress percentage, and whether or not user motion was detected at that time.

At the end of each dining session, the recorded video and the metadata file are uploaded to cloud storage in a designated user’s folder. Figure 5.5 shows the internal folder structure of the cloud storage. In this structure, each user has a folder containing all the recorded dining sessions. Each dining session consists of a single video file and a designated metadata file. Dining sessions seen by other users are automatically archived in the user’s old dining sessions folder.

**Video streaming and controlling:** As soon as an actual dining session starts, the most recent recorded remote dining video is streamed and displayed on a large display in front of the local diner. The libVLC API\(^2\) is used to play and control the streamed video. In order to ensure

---

1http://www.ntt-at.co.jp/page.jsp?id=1793&content_id=886
2http://www.videolan.org/
that the displayed recorded remote dining progress status matches the local dining progress status, the displayed video playback speed is controlled. The system accesses information from the metadata file and compares it to the current detected local dining progress. Based on this comparison (remote percentage eating progress minus local percentage eating progress), the system sets the displayed video playback speed. Through experimentation, we found that diners can watch a video with adjusted playback speed in a range from 0.7X to 1.5X without losing any information or sensing any distortion. Thus, the playback speed adaptation method is based on this range. If the percentage DDP is less than a specific amount, the playback-speed is set to slow-speed mode (0.7X). If the percentage DDP is greater than a specific amount, the playback speed is set to fast-speed mode (1.5X). Otherwise, the playback speed is set to normal-speed mode (1X).

5.4 Evaluation

In this section, the experimental validation of the KIZUNA’s adaptation method is discussed. The main objective of the experiment was to determine whether using KIZUNA’s time adaptation method would enhance a given diner’s communication and sense of presence of other diner, in comparison with conventional time-shifted tele-dining. Synchronizing the recorded remote dining activity with the local dining progress is expected to enhance these factors. The participants’ feelings and feedback were assessed by means of a questionnaire filled out after each tele-dining experimental session.

5.4.1 Recording Dining Video Messages

In real-life social dining, a person usually enjoys the company of familiar people, such as family members, friends or co-workers. This way, the person feels more relaxed, communicating easily and without constraint (137). In the validation experiment, we achieved this by inviting potential participants to dine while watching a recording of another familiar person (or ‘actor’). This limited the pool of participants. To extend the pool of participants, we recorded three dining sessions, each with a different actor, in a different language (Japanese, Chinese and Arabic). To achieve consistent eating progress among actors, we recruited three adult male graduate students (A1: Japanese, A2: Chinese, and A3: Arab). The actors were instructed to dine at a
normal, relaxed speed, and provided with a list of general questions and comments to be delivered in their native languages, at specified times, while eating. Table 5.1 shows the list of these general questions and comments, and the time at which they are to be delivered. The questions and comments were collected and extracted from a typical basic dialog between two persons conversing at a distance. The time between each question and comment was determined based on the expected average length of reply.

Table 5.1: List of general questions and comments

<table>
<thead>
<tr>
<th>Time</th>
<th>Question or Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:45</td>
<td>Q: Hello, how are you today?</td>
</tr>
<tr>
<td></td>
<td>Begin</td>
</tr>
<tr>
<td>01:30</td>
<td>Q: Do you like your meal?</td>
</tr>
<tr>
<td>02:15</td>
<td>Delicious, I like cury rice</td>
</tr>
<tr>
<td>03:00</td>
<td>Q: By the way, What’s your favorite food?</td>
</tr>
<tr>
<td>03:45</td>
<td>Personally, I like the (Italian) food a lot.</td>
</tr>
<tr>
<td>04:30</td>
<td>Q: Where do you live?</td>
</tr>
<tr>
<td>05:15</td>
<td>I like (Tsukuba) city. It’s safe, clean and the people are so friendly.</td>
</tr>
<tr>
<td>06:00</td>
<td>Q: Do you have any plan for the summer vacation?</td>
</tr>
<tr>
<td>06:45</td>
<td>I like the sea a lot, so most probably I will go to a beach and have some relaxed time.</td>
</tr>
<tr>
<td>07:30</td>
<td>Q: Which country would you like to visit?</td>
</tr>
<tr>
<td>08:15</td>
<td>Nice, I like to visit (Italy) I want to go there to eat (Italian) food.</td>
</tr>
<tr>
<td>End</td>
<td>Thank you. I am looking forward to meeting you again in the next video.</td>
</tr>
</tbody>
</table>

In the recording sessions, we used 400 g of curried rice and a glass of juice as the meal. This amount is considered sufficient for an adult person. We chose curried rice as a widely enjoyed meal. Moreover, this kind of food is often eaten with a spoon, while other kinds of food might involve the use of different utensils, such as forks or chopsticks, and this, we found through experimentation, might affect the dining progress. Analysis of the actor’s recorded video showed that this amount of food (400 g) was consumed in nine minutes on average.

5.4.2 Dining Conditions

In the experiment, we considered the following two time-shifted tele-dining conditions:

- **Normal condition**: participant dines while watching a recorded video of a familiar person eating sometime earlier.
- **KIZUNA condition**: same as normal condition, but with the video playback speed adjusted according to the DDP.

Using the same amount of food in the experiment as in the recording sessions caused small variations in terms of the DDP. Furthermore, 400 g may represent either a large or small amount of food to different participants, which may cause them to feel uncomfortable while dining. Accordingly, after surveying meals sold in the market, as well as participant preferences, we decided to use the following meal amounts in the experiment, according to each participant’s preference:

- Small meal: 300 g of curried rice (25% less than the amount used in the recording sessions)
- Large meal: 500 g of curried rice (25% more than the amount used in the recording sessions)
5.4 Evaluation

We performed a preliminarily dining experiment, and ran a simulation, to investigate the playback-speed adaptation method, and to determine the adaptation values. Figure 5.6 shows one participant’s eating progress (large meal) compared to that of the displayed remote diner (Note: the numbers above the chart data point indicate the percentage difference in DDP between the sessions). The overall DDP was around 3 minutes, with an average of 14% DDP over the length of the meal.

![Figure 5.6: A participant’s eating progress compared to that of the recorded remote diner](image)

To achieve fine synchronization using the specified meal amounts, the following adaptation values were determined:

- More than -5% DDP: the local diner is considered to be eating slower than the displayed video, and the video playback speed should be changed to slow-speed mode (0.7X).
- More than 5% DDP: the local diner is considered to be eating faster than the displayed video, and the video playback speed should be changed to fast-speed mode (1.5X).
- Between ±5% DDP: the local diner is considered to be eating at a normal speed compared to the remote diner, and the video playback speed should be set at normal speed (1X).

Figure 5.7 shows a participant’s eating progress compared to the displayed remote diner progress under the KIZUNA condition. After 2 minutes of dining, the DDP exceeded the limit (-5%), so the playback speed was changed to slow-speed mode (0.7X). At the 6th minute, the DDP decreased to -2%, so the playback speed was changed to normal-speed mode (1X). At the 7th minute, the DDP increased to -5%, so the playback speed was again changed to slow-speed mode (0.7X). Under this condition, the overall DDP was reduced to only 30 seconds, with an average of -4% DDP over the length of the meal, because of the playback-speed adaptation.

5.4.3 Setup

The validation experiment for the proposed adaptation dining method took place in a custom-built time-shifted tele-dining booth. The booth was enclosed by a curtain to isolate it from the lab environment. Figure 5.8 shows the booth’s internal setup. In the tele-dining site, we used a USB camera, placed over the display, to record the participant’s facial expressions and hand gestures. A small flat cooking scale was placed under the plate to accurately measure the food amount. A second USB camera was used to capture and record the cooking scale reading.
5.4 Evaluation

Figure 5.7: A participant’s eating progress compared to that of the recorded remote diner, under the KIZUNA condition

Figure 5.8: Tele-dining site setup
5.4 Evaluation

For KIZUNA validation, the assessment of the participant’s eating progress was performed by employing the WoZ system simulation technique. The second USB camera was connected to a display placed outside the tele-dining site. The researcher monitored the DDP and periodically altered the displayed video playback speed.

5.4.4 Participants

A total of 22 healthy participants took part in the experiment (12 females and 10 males). The participants’ ages ranged from 21 to 32 years old (average = 24.1 years, s.d. = 2.7); most were undergraduate or graduate students; and 13 participants had had previous experience of videoconferencing systems. Most had used videoconferencing mainly to talk to remote family members and/or friends. The participants were divided into two groups: Group 1 participants \((d1 - d11)\) performed under the Normal condition, while Group 2 participants \((d12 - d22)\) performed under the KIZUNA condition. All the participants experienced each condition for the first time.

Participants were given their choice of the amount of food (small or large meal) they would consume in the experiment, and 10 participants chose the small meal, while the other 12 chose the large meal. Participants were also given the choice of dining at lunch or dinner time, and 17 participants chose lunch, while the other 5 chose dinner. Table 5.2 shows the participant-video arrangements, the diners’ relationship, and the chosen meals, for the two groups; Where \(V1, V2,\) and \(V3\) are recording of \(A1, A2,\) and \(A3\) dining sessions respectively.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Sex</th>
<th>Meal</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d1)</td>
<td>M</td>
<td>300</td>
</tr>
<tr>
<td>(d2)</td>
<td>F</td>
<td>V3</td>
</tr>
<tr>
<td>(d3)</td>
<td>F</td>
<td>Friend</td>
</tr>
<tr>
<td>(d4)</td>
<td>M</td>
<td>V1</td>
</tr>
<tr>
<td>(d5)</td>
<td>M</td>
<td>V1</td>
</tr>
<tr>
<td>(d6)</td>
<td>F</td>
<td>V1</td>
</tr>
<tr>
<td>(d7)</td>
<td>F</td>
<td>V3</td>
</tr>
<tr>
<td>(d8)</td>
<td>F</td>
<td>V3</td>
</tr>
<tr>
<td>(d9)</td>
<td>F</td>
<td>V1</td>
</tr>
<tr>
<td>(d10)</td>
<td>M</td>
<td>V1</td>
</tr>
<tr>
<td>(d11)</td>
<td>M</td>
<td>V1</td>
</tr>
</tbody>
</table>

5.4.5 Procedure

Each tele-dining experiment began with the researcher asking the participant to sign the experiment consent form, and to complete a basic demographic survey form. Following this step, the researcher guided the participant to the tele-dining site and introduced the system environment. We asked the participant to imagine him-/herself dining and talking with a family member or friend who lives in a different time-zone. At this point, we displayed the familiar partner’s static image on a display in front of the dining table. We explained to the participant that he/she would be dining while watching a previously recorded video of his/her partner. The participant was told that the partner would talk to him/her while dining and then later
5.4 Evaluation

review the current recorded session. The participant was instructed to dine as he/she normally
does. The participant was not apprised of the respective experimental dining condition (normal
or KIZUNA). The researcher then left the dining booth and activated the recorded video as
an indication of the beginning of the dining session. After each dining session, the participant
was instructed to complete a questionnaire about the session he/she had just experienced. The
dining sessions were recorded on tape. Afterward, the sessions were reviewed, and the DDP
calculated for each condition.

Under the KIZUNA condition, the researcher played the role of WoZ, by monitoring the
participant’s dining progress and controlling the displayed video playback speed accordingly.
The alterations were performed each minute, according to the adaptation values noted above,
to achieve fine synchronization.

The experimental design was between-subject, and the analysis was performed on the results
of the participants’ first-ever experience of tele-dining (whether in Group 1 or Group 2).

5.4.6 Questionnaire

In the questionnaire, we asked the participants to evaluate a series of statements according
to the feelings they experienced during the time-shifted dining session. The principal aim of
the study was to assess participants’ communication with their partners, and to determine
whether synchronizing the dining sessions would affect communication or not, using a selection
of straightforward questions from related questionnaires (26)(130). The following statements
were used for this purpose:

- C1) “I wanted to talk to the partner.”
- C2) “I enjoyed talking with the partner while eating.”
- C3) “The partner’s talking distracted me from my meal.”
- C4) “The content of the conversation was natural.”
- C5) “The timing of the partner’s delivery was natural.”
- C6) “I could communicate with the partner naturally.”

The sense of the partner’s presence while tele-dining is another important aspect to consider
in assessing the proposed system; and synchronizing the dining sessions is believed to affect this
sense of the partner’s presence. In this study, we sought to determine whether our proposed
system would enhance this sense or not, using common questions from related questionnaires
(126)(127)(125). To investigate the participants’ sense of presence under each condition, the
following statements were used:

- P1) “I felt as if the partner and I were eating together in the same room.”
- P2) “I felt distant from the partner.”
- P3) “The partner’s facial expressions were easy to recognize.”
- P4) “The partner’s gaze direction was easy to recognize.”
- P5) “I was able to make eye-contact with the partner.”
- P6) “The partner’s gestures were easy to recognize.”

All of the above statements were rated on a 9-point Likert scale, where 1 = strongly disagree,
3 = disagree, 5 = neutral, 7 = agree, and 9 = strongly agree.
5.4 Evaluation

5.4.7 Results

Figure 5.9 shows the average results of participants’ communication-related feedback on their dining experience, under the two conditions. Comparison was done using a between-groups t-test, to examine whether the means under the two conditions differed significantly from one another. We found significant differences in participants’ feelings of distraction caused by the partner talking (C3: \( t(20) = 2.62, p < 0.05 \)) and in the timing of the partner’s delivery (C5: \( t(20) = 3.66, p < 0.01 \)); while no significant differences were found in participants’ willingness to talk with the partner (C1: \( t(20) = -0.68 \)), enjoyment of talking with the partner while eating (C2: \( t(20) = -1.6 \)), content of the conversation (C4: \( t(20) = -1.4 \)), or ability to communicate with the partner naturally (C6: \( t(20) = -0.81 \)).

![Figure 5.9: Participants’ average communication-behavior results](image)

Figure 5.10 shows the average results of participants’ felt sense of the partner’s presence while dining, under the two conditions. We found significant differences in participants’ sense of the partner’s presence as if in the same room (P1: \( t(20) = -3.48, p < 0.01 \)), and participants’ felt distance from the partner (P2: \( t(20) = 2.07, p < 0.1 \)); while no significant differences were found in recognizing the partner’s facial expression (P3: \( t(20) = -0.11 \)), the partner’s gaze direction (P4: \( t(20) = -0.61 \)), or the partner’s gestures (P6: \( t(20) = -0.51 \)), or in the ability to make eye-contact with the partner (P5: \( t(20) = 0.28 \)).

The exact deployment of the V1 to V3 videos was not identical in both conditions, although all the actors were instructed to behave in the same manner, according to the scenario in Table 5.1: the V1 video was used 7 times, the V2 video 2 times and the V3 video 2 times under the Normal condition; while the V1 video was used 5 times, the V2 video 4 times and the V3 video 2 times under the KIZUNA condition. To determine whether the videos had any effect on the ratings, correlations between the videos and the ratings were examined for P1, P2, C3, and C5, which indicated the differences between the conditions. We found a weak correlation in Normal P2 (\( \rho = 0.41 \)) and no correlation for the others. However, this Normal P2 used more V1 videos than under the KIZUNA condition, and the V1 video actually had relatively lower ratings on average. This suggests that the marginal significant difference in P2 is not caused by the video effect.
5.5 Discussion

The proposed adaptation algorithm meant to synchronize the participant’s and partner’s dining progress by keeping the DDP within a specific range (±5%). This in turn reduced the overall DDP. For the Normal sessions, the average DDP was 117 seconds (s.d. = 48.8), while for the KIZUNA sessions, the average DDP was only 19 seconds (s.d. = 15.9), and this suggests a significant difference in the DDP (t(20) = 6.07, p < 0.001).

5.5 Discussion

The questionnaire results revealed that the proposed time-shifted tele-dining method improved participants’ communication while dining, in comparison with the Normal mode. This result supports our assumption that synchronizing the dining sessions (local and remote recorded dining sessions) would increase the realism of virtual social dining. In particular, participants found that the partner’s talk in KIZUNA dining was less distracting than in Normal dining (Fig. 5.9 - C3). This may be a result of the close alignment of the dining progress. Moreover, participants found the partner’s conversational delivery timing more natural under KIZUNA dining conditions than under Normal dining conditions (Fig. 5.9 - C5).

Since a sense of other participant’s presence affects the local diner’s communication, we tried to enhance this sense of presence by depicting the latter’s life-sized image on a large display (60)(111)(17), and by matching the furniture (22) of the local and remote dining environments. The large display was placed in front of the local dining table, at the same level; hence the remote dining table appeared as if it was an extension of the local dining table (Fig. 5.1). We also assumed that synchronizing both dining sessions would further enhance the sense of presence. The questionnaire results support our assumption, confirming that participants felt as if the partner and they were eating together in the same room during KIZUNA dining (Fig. 5.10 - P1), and that participants felt less distance from the partner during the KIZUNA dining sessions (Fig. 5.10 - P2).

Though participants were aware that they were watching a recorded video, we noticed that some communicated more naturally with their partners during KIZUNA dining sessions, and many asked more questions in responding to the recorded video (with an average of 1.02...
questions asked, compared to 0 in the Normal dining sessions). Many participants noted that it felt strange not to receive a direct response to their questions. This reaction is very similar to the “uncanny valley” telepresence reaction caused by humanoid robots that are highly similar to humans (153). In the experiment, other person’s life-sized representation, combined with the matching furniture and dining utensils, misled local viewers into trying to interact with other person in real time.

The proposed system achieved better communication and sense of presence by merely controlling the playback speed of the large displayed video. Our proposed algorithm had three playback-speed modes (normal, fast and slow). The fast and slow modes were carefully chosen so that viewers would not sense any distortion in the picture or sound. In this regard, we asked participants to judge their partner’s dining speed. Despite the varying playback-speed mode, most participants said that the partner’s dining speed was normal in general. However, a few participants noted that they felt the partner’s motion was slightly slower in some sections, and attributed this to a problem with the video player. No comments were made, however, regarding the fast mode, which may indicate that this mode was acceptable.

It is natural for a person to look at their dinner partner while dining, and to observe their dining progress. In this regard, some participants noted that they adopted the strategy of watching their partner’s plate, which gave them an indication of their partner’s dining progress, by which they estimated how much time remained for the dining session. We also noticed a common behavior among participants, to echo their partner’s actions, most noticeably the action of drinking. Whenever the partner performed a drinking action, the local participant performed a similar action, either synchronously or once he/she had finished their current action. We found no difference between the two conditions in this regard.

The controlled dialog during the experimental dining sessions was meant to be general and uniform for all participants. The dialog consisted of a series of questions and comments, delivered at specified times based on the dining progress. The preliminary conversational analysis showed that almost all the partner’s questions were answered by participants, though only half of the partner’s comments were responded to, usually in the form of a new question or agreement (or disagreement) with the respective comment. And again, no differences were found between the two conditions. When we reviewed the recorded experimental dining sessions, we found that the dialog was typically natural, especially when the partner’s follow-up question or comment was a natural reply to the participant’s utterance. This tended to engage participants more in the dialog, as noted by one participant: “I was surprised when the partner answered my question immediately, as if we were talking F2F.” However, some participants found the pre-set question/answer dialog style abnormal, preferring a dialogue style based on a story or the latest news.

5.6 Study Limitations

Despite the achieved results, this study has many limitations. For example, the validation experiment was performed in a lab environment and the number of participants was quite small (3 actors and 22 participants). Moreover, the experimental setup assumed that all the food would be placed on the dining table beforehand and served on one plate only. In addition, our current system is designed for two diners only; in the future, we will study how KIZUNA usage would scale if more than two diners wished to dine together. Moreover, there were other
ecological shortcomings, such as participants’ age, location, bond, nationality, and culture. Our study was limited to young student participants and actors (from 21 to 32 years old), and the participants and actors were located in the same city, which meant that they see each other frequently. The relationship between the participants and actors was mainly one of friendship (Table 5.2). We would like to validate and extend the current findings with qualitative studies, such as longitudinal observations and investigations of dining behavior.

At this stage, we implemented a simple means of adapting the video playback speed by comparing the local and recorded dining progress. This can be further enhanced by analyzing the video content and locating ‘rich’ sections in the recorded dining sessions; for example, those that include an abundance of conversation and/or dining actions. Finally, the method assumed discrete levels of video playback speeds (slow, normal and fast), and this may affect the diner’s viewing experience. A recent study found that a person preferred gradual, rather than sudden or dramatic alterations, in playback speed (154). Thus, a gradual continuous video playback speed may enhance the communication even further.

5.7 Summary

In this study, we investigated the effect of chronological alignment on communication. More specifically, we investigated the synchronization of dining events between local diner and a recorded video message of a remote diner in time-shifted tele-dining. The design of a tele-dining system “KIZUNA” and the validation of proposed adaptation algorithm are presented. The KIZUNA system is meant to enhance communication and sense of connectedness between family members and friends who live apart in different time-zones. The system involves recording the remote person’s dining sessions and then displaying these sessions for the local person when he/she starts eating, and vice versa. Based on the difference in dining progress, the system adapts the playback speed of the displayed remote recorded dining session to achieve a sort of real-time dining behavioral mimicry. A validation experiment was conducted to thoroughly investigate the proposed adaptation method. The results revealed that the KIZUNA system effectively enhanced communication and sense of other person’s presence, by synchronizing the dining sessions.
Chapter 6

Discussion

Nonverbal cues are considered a key aspect in communication. Being able to convey the necessary nonverbal cues is essential in enhancing MC. One factor to do so is the size of other person’s image, which we consider as the core alignment in visual communication. If the image was small, we may feel as if other person is far away (17). Furthermore, reading other person’s facial expressions or gestures might be difficult. Various researches have confirmed that the sense of copresence is increased when the image is life-sized (22, 39, 155, 156). This aspect, in particular, has been widely adopted by many commercial videoconferencing systems these days. This motivated the conveyance of life-scale images across the performed studies. In MAVT study, the other person ‘Virtual Teacher’ was presented in life-scale at the same height-level as the user through a HMD (Section 3.3.1); in BHS study, the other person’s figure was displayed in life-sized without his/her background on a large display (Section 4.3.1); and in KIZUNA study, the recorded video was displayed in life-size scale on a large display as well (Section 5.3.1).

Proxemics is a subcategory of the study of nonverbal communication. Various researches have emphasized the impact of proxemic behavior (the use of space) on interpersonal communication (157, 158). People like to keep distance from others and there are very specific social rules about how close they can go to others in particular situations (Section 2.1). In F2F communication, maintaining a suitable distance from other while communicating is significant in achieving enhanced communication. MC is no different in this regard (159). The sensed distance in visual MC, also known as “Virtual Distance”, is influenced by a number of factors include spatial distance from display, size and video image quality, background, and voice fidelity (17). Various virtual distances were considered in the performed studies in this dissertation according to the communication application. In MAVT study, three virtual distances were investigated (i.e. 0m, 1m, and 2m) (Section 3.3.2). The users’ preferences as well as the experiment results showed that 1m virtual distance was the preferred while conducting mimicking tasks. In BHS study, we considered a virtual social distance (i.e 1.2m), which allows participants to feel as though they are talking at social distance when videoconferencing (Section 4.3.1). Finally, in KIZUNA study, we considered a virtual personal distance (i.e. 0.5m), which is a natural distance when dining with a companion around a dining table.

The people’s body movement is one important aspect of nonverbal communication. Conveying an appropriate body posture is very important to achieve better communication. On
the one hand, in F2F communication where two persons are involved, two natural and mutual behaviors occurred to enhance communication process. The first behavior concerning other person, where he/she tends to adjust his/her body posture and movements to show the user the target and to make certain nothing block the view. This obliged the other person to keep adjusting his/her posture in corresponding to his/her own physical movements and to the user’s location and view angle. The second behavior is concerning user him/her self. The user tends to watch the other person’s physical movements and target from a good view angle. This obliged the user to update his/her posture and view angle in corresponding to the other person’s physical movements. On the other hand, in visual MC environment, the other person and the user’s position and rotation are typically fixed. This setup may harm communication process as the needed spatial alignment in some applications is missing.

Achieving spatial alignment involves capturing the details of other person’s physical movements and location as well as the user’s (21, 53). This way, it is possible to align the view of both sites based on the performed body movements by each or both sites. As we saw in the literature review, many studies had investigated this issue by proposing fixed and limited methods to capture and visualize the other site’s movements as in (37, 53, 100, 103, 109). In contrast, we thoroughly investigated the virtual teacher’s setup and its effects on the sample physical-task learning (Section 3.3.2).

The MAVT study revealed that considering spatial alignment is significant for enhancing remote learning communication (Chapter 3). The learner’s head movement was captured and based on it the remote view was updated. This simulated the real-life situation from learner perspective. The teacher’s body movement was captured as well and based on it the teacher’s rotation angle was adjusted to provide an easy view to the learner. This simulated the real-life situation from teacher perspective. The optimal teacher’s rotation angle was not confirmed before this study. Our investigation showed that the close side view is the optimal for learning physical-tasks motions (Section 3.3). However, the teacher’s natural body movement might not be limited; he/she might use both hands or rotate around. In such cases the view of the teacher has to be adjusted according to his/her physical body movements. This led us to propose a novel method to automatically adjust the view of teacher based on teacher’s physical body movement (Section 3.4). The method revealed to be effective for body movement that gradually repositions the most important moving part. This study is therefore considered likely to enhance communication when learning physical-task skills in mixed-reality environment.

The surrounding is one of the factors that affect how effective our communication is. The most common surrounding factors are: background, lighting, distractions, distance, viewing angle, and noise. On the one hand, in F2F communication, the two parties share the same surrounding environment. The first party’s view is the integration of the other party’s view. This surrounding integration has been shown in many studies as a major factor in facilitating the communication process (73). A study conducted by Fussell et al. demonstrated the value of the shared view of the work environment for remote collaboration on physical tasks (160). In this setup, referring to anything in the surrounding would be straightforward (39), and any conducted action needs no or minimum verbal explanation (161). Moreover, any surrounding effect, either distraction cause or engaging cause, would affect both sides equally. On the other hand, in MC, each side is naturally surrounded by a different environment. This in turn might harm the communication process as one of the needed conversational grounding is missing (4) and may also makes the users feel more distant. Another important issue is related to the
amount of received information from other site. In many cases, the user might not be concerns in seeing what is happening in the other site. This may raises some privacy concerns for some people as well. Many studies had discussed the effect of privacy and communication as in (115, 117). From our early investigation, we found that privacy is greatly depends on whom you are communicating with (i.e. family, close friend, business, etc.). In a preliminary semi-structure interview with some participants, we found that they preferred to see the other side’s background when they are video chatting with a family or a close friend. However, when video chatting with foreigners or for business, they preferred to hide their own background.

Achieving complete surrounding alignment in MC might be difficult as this requires replicating all the details which is might be hard to be captured and presented using the existing technology. However, some partial environment alignment techniques help minimizing the differences between communication sides as presented in the study in (Chapter 4). Various simple techniques such as blurring the background in (56) was found effective in communication. Another technique by using matching furniture across the sites was also found effective(22). This technique, in particular, has been widely adopted by many commercial videoconferencing systems, such as HP’s Halo and Cisco’s Telepresence systems. Another dominant technique is by unifying the surrounding by changing the view background for both sides(39). Quite similar to these techniques; we proposed a yet simple but effective alignment method by extracting other person’s figure and superimpose him/her into local captured front view. This simulated F2F communication, where a user can manage to see his/her own environment and the other person’s figure at the same time. This technique revealed to be effective in communication where the perceived sense of other person presence was improved and the conversation behavior was enhanced. The proposed system also effectively reduced the psychological distance between remote persons and guarantee full local surrounding awareness.

In F2F, the communication process occurred in real-time ‘Synchronous’. When the user conveys a message, the receiver receives it instantaneously and he/she can react and adjust to any nonverbal cues. Any response or feedback will also be executed instantaneously. This creates a feeling of community because we are better able to socialize and interact with one another. However, by definition, this kind of communication requires same-time participation, any different time-zones and/or conflicting schedules can create communication challenges. On the other hand, time-shifted or delayed communication ‘Asynchronous’ enables communication over a period of time through a “different time different place” mode. This kind of communication allows people to connect together at each person’s own convenience and own schedule and makes the involvement of people from multiple time-zones much easier. However, many studies revealed that this is considered as one of the major factors that affect communication(162). In some cases, messages get misinterpreted and a sense of personal connection is never truly established or maintained. Despite any drawbacks, there are many cases where time-shifted communication is the only visible mean of communication. For those cases, minimizing the time difference gap becomes a necessity.

There have been a number of studies in various domains done on chronological alignment in communication. Some of these studies focused on synchronizing the start point of communication as in (44). Other studies tried to achieve chronological alignment through showing related content on demand as in (43). These techniques may enhance communication process, but in some social communication, the communication means more than just information exchange; it typically also involves affective satisfaction and social connectedness. To support this type of
communication, a sense of co-presence is often crucial. Thus, we proposed a system addresses these more complex needs, and further explores the potential of video exchange, but is more focused on the social connectedness between remote users, based on the assumption that aligning the activities of remote participants may significantly enhance MC. In the conducted study (Chapter 5), the proposed system achieved better communication through making it highly automated and by merely controlling the playback speed of the received time-shifted video message. This mimicked the natural synchronization found when eating with a companion (140). In this study, the displayed recorded video speed is adapted according to user’s dining progress. This technique revealed to be effective in communication where the perceived sense of other person presence was significantly improved and the conversation behavior was enhanced as well.
Chapter 7

Conclusion

The main problem that was addressed in this dissertation is how to implement enhanced MC systems by aligning the local and remote sites. To this aim, we synthesized results from researches on MC and communication enhancing to identify the alignment technique(s) that would enhance the communication and the factor(s) that is needed to be considered when evaluating the enhancement. Three comprehensive studies were conducted to validate the alignment concept with respect to spatial, surrounding, and chronological. The results of the conducted studies confirmed that applying such alignment could enhance the communication significantly.

The spatial alignment investigation involved two parts. In the first part, we investigated the optimal other person’s setup in terms of position and rotation for the best learning communication outcome. The experimental results suggested that the other person’s close side-view is the optimal view for physical-task learning which includes one-hand movement. In the second part, we proposed a method to automatically adjusting the other person’s rotation angle when he/she is demonstrating physical-task motion. This method ensured that the user sees most of other person’s motion from an optimal close-viewing angle. This study revealed that considering spatial alignment can indeed enhance communication in physical-task learning domain.

The surrounding alignment investigation considered the view background effect in video-conferencing. In a high-presence media space, users have to feel as if other person is present in their local environment. This setup was achieved by background alignment through extracting the other person’s figure from remote site view and superimposing it onto the local front view. In this way, the local user felt as if other person were present before him/her in his/her spatial environment. This study revealed that considering surrounding alignment can also enhance communication and affect conversation behavior.

Finally, the chronological alignment investigation considered the events synchronization in time-shifted tele-dining. The alignment involves recording remote person’s dining sessions and then automatically displaying these sessions for the local user when he/she starts eating, and vice versa. Based on the difference in dining progress, the system adapts the playback speed of the displayed remote recorded dining session to simulate the natural synchronization that found when eating with a companion in real. The experimental results revealed that chronological alignment effectively enhanced communication and the perceived sense of other person presence.
References


REFERENCES


REFERENCES


REFERENCES


REFERENCES


[81] Dawei Jia, Asim Bhatti, and Saeid Nahavandi. Design and evaluation of a haptically enabled virtual environment for object assembly training. In IEEE International Workshop on Haptic Audio visual Environments and Games, HAVE’09, pages 75–80, Nov. 2009. 15


Jung Sang Hack and Bajcsy Ruzena. Learning Physical Activities in Immersive Virtual Environments. Fourth IEEE International Conference on Computer Vision Systems (ICVS’06), pages 5–5, Jan 2006. 15


Matthew Lombard and Theresa Ditton. At the Heart of It All: The Concept of Presence. J. Computer-Mediated Communication, 3(2), 1997. 42, 45


94
REFERENCES


[121] Nicole Yankelovich, Jonathan Kaplan, Joe Provino, Mike Wessler, and Joan Morris DiMicco. Improving audio conferencing: are two ears better than one? In Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, CSCW ’06, pages 333–342, 2006. 44, 45


REFERENCES


[141] John Castro and Marie Brewer. The amount eaten in meals by humans is a power function of the number of people present. Physiology and Behavior, 51(1):121–125, 1992. 65


REFERENCES


