Title
Generating, Evaluating and Recognizing Embodied Communication

Author
Abdullah, Ahsan

Publication Date
2022

Peer reviewed|Thesis/dissertation

Generating, Evaluating and Recognizing Embodied Communication

By

AHSAN ABDULLAH

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Michael Neff, Chair

Nina Amenta

Hao-Chuan Wang

Committee in Charge

2022

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my sincerest gratitude and regard for my advisor, Professor Michael Neff, who provided amazing guidance for all my projects, helped shape my research and encouraged me through challenging times with incredible patience and compassion. Working with him, I learned all key aspects of the research process and have always been inspired by his work ethic. He always supported me and exposed me to different industrial and academic opportunities that have improved and shaped my professional skill set. I would also like to extend my thanks to my committee members Professor Nina Amenta and Professor Hao-Chuan Wang, who served on my qualifier committee, provided thought provoking insights on my work and worked with me on my final dissertation as well.

The love, encouragement and emotional support of my parents and wife made this time much easier and enjoyable. I would like to thank my amazing family in Pakistan and US, whose support was an impeccable force during the whole journey. I would like to thank all my friends from Davis, Bay Area, Seattle, New York, London, Pakistan, and around the world who created an incredible support system for me.

I would like to thank my lab mates at Davis, Mohammad Adil, Gabriel Castillo and Nicholas Toothman. A special thanks to Jesse Smith, for his thoughtful insights on our professional and personal discussions and always bringing positive energy to our conversations.

This is just an incomplete list of people, to whom I am eternally grateful for their love, compassion, support, friendship, mentoring and more.

Abstract

Generating, Evaluating and Recognizing Embodied Communication

Designing embodied virtual environments where humans can naturally communicate through verbal and non-verbal channels is a challenging problem. It requires a deep understanding of human communication behaviors and patterns, and sophisticated models that can project this knowledge onto digital avatars. Non-verbal gestures and cues are the dominant channel for conveying information in interpersonal interactions. In this dissertation, we explore sub-problems in the domain of understanding human communication patterns, generating digital avatar behaviors and studying how humans communication patterns vary in digitally embodied environments.

We first present a study comparing group interactions in Virtual Reality and Videoconferencing settings. During interactions, people were able to achieve similar performance across tasks, however their gaze and other nonverbal behavior patterns varied in VR and VC settings. Findings of this study inform how sharing an embodied 3D environment impacts our ways of communicating information. Significant behavioral differences are observed. These include increased activity in videoconference related to maintaining the social connection: more person directed gaze and increased verbal and nonverbal backchannel behavior. Videoconference also had reduced conversational overlap, increased self-adaptor gestures and reduced deictic gestures as compared with embodied VR.

We then explore the design of a pedagogical virtual agent, that takes students through a discovery-based learning environment, the Mathematical Imagery Trainer for Proportionality (MITp). The agent helps foster insights about concepts of ratios and proportions by helping students through well crafted set of tutorials. It is a challenging task for agent technology as the amount of concrete feedback from the learner is very limited, here restricted to the location of two markers on the screen. A Dynamic Decision Network is used to automatically determine agent behavior, based on a deep understanding of the tutorial protocol. A pilot evaluation showed that all participants developed movement schemes supporting proto-

proportional reasoning. They were also able to provide verbal proto-proportional expressions for one of the taught strategies.

As a sub-problem in understanding gesture based human communication, we also explore the Sign Language Recognition (SLR) problem. Isolated Sign Language Recognition (ISLR) is an important constituent task in developing SLR systems in which videos of individual, word-level signs are correctly identified. We develop a novel model for ISLR based on a unique G3D-Attend module that further uses spatial, temporal and channel self-attentions to contextualize aggregated spatial and temporal dependencies. We augment the datasets with 2D and 3D skeleton data, which is used along with RGB data in an ensemble-based approach to achieve state-of-the-art recognition rates. We then extend the approach to create weak sign spotting labels. Spotting involves identifying individual signs in multi-sign sentences and is a significantly more difficult task due to co-articulation effects, differences in signing speeds and the influence of contextual information on sign production. Generating sign labels in continuous domain manually is a laborious task, and this approach can provide a set of labels that can then be used to provide weak supervision for future machine learning applications.

Chapter 1

Introduction: Embodied Communication

Aristotle, the Greek philosopher, wrote [30], "Man is by nature a social animal; an individual who is unsocial naturally and not accidentally is either beneath our notice or more than human. Society is something that precedes the individual. Anyone who either cannot lead the common life or is so self-sufficient as not to need to, and therefore does not partake of society, is either a beast or a god". His words highlight how social interactions are important to us as humans, and how we strive to be a part of communities. A key component of our social life is how we effectively communicate with each other, convey our thoughts, understand others and share moments with each other.

Communication is one of the fundamental skills that we naturally learn as humans. In this complex multi-modal realm, we commonly use a combination of speech and gestures to convey the message. Over time we have established sophisticated modes of communication such as spoken languages, that are able to convey complex concepts in words or sentences. Similarly there are sign languages, used mostly by deaf communities around the world, that are entirely gesture based languages. We communicate through reading and writing, a popular setting where we do not have to socially share the space but both sides must agree

on a language. In face-to-face communication, our body language and gestures can help us communicate to a large extent, even if we do not understand each other verbally. Face-to-face communication is one example of embodied communication where nonverbal cues from physical presence play the key communication role.

1.1 Verbal and Nonverbal Communication

Over time we have developed sophisticated languages with a defined vocabulary set that could be written, preserved and widely understood. Modern day languages have evolved and adapted over centuries. Early humans used bodily gestures and tone of voice to convey the message. These gestures were also limited to survival concepts like hunting, fire, water, danger etc. It was tough to converge on a single definition for a wider range of communities, as each group would have their own signs [116]. Pictorial languages [153] were an early solution to this problem, where symbols for everyday concepts were drawn on big rocks and people would learn and remember these symbols to communicate further. Symbols for phonetics were also developed and over time it led to languages that larger groups could understand and communicate complex concepts with. Today there are around 6900 languages used around the world[9]. These languages differ immensely in their vocabulary and phonetics, but the nonverbal cues that accompany these languages are almost all globally understood.

1.2 Role of Embodiment in Communication

Spoken languages have helped us evolve, survive and collaborate more efficiently over time. However studies show that humans rely on the accompanying nonverbal gestures more than spoken words. Albert Mehrabian [107] presented a 7-38-55 rule that says our words only matter 7% in our everyday communication. Our voice and tone of speech conveys 38%, whereas the majority 55% is expressed through our gestures and body language. This rule further implies that in cases where there is a contradiction between speech and body lan-

guage, people tend to believe the body language more. The exact numbers might vary in specific settings but the general idea of the nonverbal channel of communication being dominant in face-to-face settings prevails. In another study [124] on recorded sales negotiations, researchers also found that body language was the dominant force of impact in negotiating. They found that people with strongest arguments would win negotiations over phone more often than in face-to-face settings. It shows how a person's body language can influence their efficacy of delivering the message and the other party's decision-making.

Video conferences are preferred over audio conferences all around the world for similar reasons. People prefer seeing the other person during discussions so they can establish mutual trust and communicate effectively so that ideas can flow more effortlessly. A similar pattern was shown by Smith et al. [144] where participants interacted with each other with and without embodiment in VR, and social presence in embodied setting was found to be pretty similar to face-to-face situation and no embodiment was construed as lonely and a degraded communication experience.

1.3 Technology as means of Communication

Technological advancements in recent times and their adaptation to communication based applications have entirely changed our landscape of social interactions. To begin with, it has taken away the requirement of being in a shared physical space to be able to visually communicate with each other. For some time this communication was limited to telephone calls, fax and emails, but in the past few decades people in different parts of the world can easily chat, play, work and spend quality time together. Covid-19, the current pandemic, stress tested our ability to communicate and collaborate remotely at every level and these technologies stood up to the challenge. Among other use cases, advances in remote communication enabled us to perform our jobs from our homes in this challenging time. Remote work has been gaining popularity as studies show that it had little to no negative, and in many cases a positive impact on productivity of employees [26, 93].

Apart from working and collaborating via technology, round the clock we are sharing stories of our daily life with loved ones and responding with emojis to an awesome sunset they saw thousands of miles away. Advancement in various fields such as semi-conductors, cameras, internet and machine intelligence have minimized the impact of physical distance on our social life. We are now moving towards virtual environments, where people can remotely share a virtual world with each other, hangout, travel and do all sorts of activities all while physically sitting at their home. These virtual worlds focused on social connection are also called Metaverses. Early adaptation of these concepts can already be seen in games like Grand Theft Auto V (GTA V) and Minecraft among others. GTA V had peak concurrent users at 220 thousand, whereas Minecraft currently has daily concurrent users ranging between 2.8 - 3.6 million. In these games you are represented as a digital avatar with features of your choice, and your movements and actions are controlled via keyboard or joystick controls. A few natural extensions to the current state of metaverse experience are realistic avatars that look, move, talk and behave like the individual behind that avatar and being able to interact with each other and the natural environment more naturally. A major portion of advances in virtual, augmented and mixed Reality are geared towards this idea.

1.4 Avatar based interactions

We refer to autonomous digital humans as virtual agents. In future, we might not be able to differentiate them from real humans in virtual and augmented environments. The term 'Avatar' refers to visual appearance and embodiment of humans or agents in virtual environments. Avatars add the element of embodiment in our remote social interactions. Autonomous and intelligent virtual agents have been designed as independent entities working with each other and real humans in virtual environments. Some virtual agents such as Ellie [44] and Maria [1](Chapter 3) remotely interact with humans in real-time and guide them through varying experiences. For example, Ellie [44] conducts a personalized interview with

a patient and picks up signs of depression and post-traumatic stress disorder. Maria [1] is a pedagogical agent that guides students through an exploratory experience to help them build intuition for the mathematical concept of ratios.

Designing human-like intelligent virtual agents is a complicated task. Apart from looking realistic, the avatar has to mimic natural movement, stylistic variations in body gestures, plausible gaze patterns, facial expressions, synchronization of audio with lips and rest of the body pose. All these traits paint an idea of an avatar's personality in our mind and develop a sense of trust with that virtual human in our shared virtual space. Even the slightest anomalies in avatar's behavior can cause discomfort. The task also involves gathering information from the shared environment and responding to changes in environment while interacting with the human or other virtual characters. Modeling the behavior pattern that captures this dynamic nature of communication requires processing complex multimodal data. Avatar's behavior must also seem natural and that requires understanding how humans behave in those settings. Hence we must first learn communication patterns from real human data, and then try to impart those behaviors to virtual agents.

1.5 Dissertation Breakdown

This dissertation focuses on sub-problems in mapping human-like natural communication behaviors and routines to virtual avatars. It's a broad domain that involves deep understanding of human communication patterns, perception of virtual humans, and how to incorporate those human communication traits in virtual embodied communication. In the scope of this dissertation we explore challenges in understanding embodied communication patterns from actual human data, design behavioral models for intelligent virtual agents and understanding the impact of virtual avatar embodiment on human communication in virtual reality settings. Summaries for each project are given below.

Videoconference and Embodied VR: Communication Patterns Across Task and Medium

In this project we compare communication patterns during task performance in embodied VR against videoconferencing. Videoconferencing is a widely used medium for conducting remote meetings, and most of us are fairly comfortable with it. This study is designed to evaluate how the two media compare on the communication and task performance fronts. Thirty five groups of three participants went through a set of four tasks in VC setting, and exactly the same circumstances were replicated for embodied VR. Embodiment in VR used state-of-the-art hand, body, face and eye tracking on avatars that matched participant's basic profile. Participants performed similarly in terms of task performance across VC and VR but behavioral differences in their communication patterns were observed. These include increased activity in videoconference related to maintaining the social connection: more person directed gaze and increased verbal and nonverbal backchannel behavior. Videoconference also had reduced conversational overlap, increased self-adaptor gestures and reduced deictic gestures as compared with embodied VR.

Pedagogical Agents to support Embodied, Discovery-based Learning

This project focuses on the behavioral model design for Maria. Maria is a virtual agent that teaches students the mathematical concept of ratios and proportions by guiding them through the Mathematical Imagery Trainer for proportionality (MITp) system. MITp is a discovery based learning environment. Discovery-based learning guides students through a set of activities designed to foster particular insights. The only input participants provide is touch coordinates of their right and left hand on the screen. Our agent, Maria, explains the protocol, how the participant must interact with the screen, provides performance feedback, leads participants to have different experiences and provides remedial instruction when re-

quired. A Dynamic Decision Network is used to automatically determine agent behavior, based on a deep understanding of the tutorial protocol. A pilot evaluation shows results on what insights and reasoning participating students were able to develop at the end of the experience.

Sign Recognition in Isolation and in Context for American Sign Language

In this project we tackle the problem of recognizing individual signs in videos of American Sign Language. Sign Language Recognition (SLR) from video could enable a range of beneficial applications for the deaf community. Sign depends on the intricacies of finger motion, along with non-manual features, the learning architecture must be able to learn spatial and temporal dependencies for micro and macro movements. Isolated Sign Language Recognition (ISLR) is an important constituent task in developing SLR systems in which videos of individual, word-level signs are correctly identified. We develop a novel model for ISLR based on a unique G3D-Attend module that further uses spatial, temporal and channel self-attentions to contextualize aggregated spatial and temporal dependencies. The approach is tested on two large, public datasets of labeled, isolated sign videos of American Sign Language (ASL), WLASL and MSASL. We augment the datasets with 2D and 3D skeleton data, which is used along with RGB data in an ensemble-based approach to achieve state-of-the-art recognition rates. We then extend the approach using SVMs to create weak sign spotting labels. Spotting involves identifying individual signs in multi-sign sentences and is a significantly more difficult task due to co-articulation effects, differences in signing speeds and the influence of contextual information on sign production. Our sign spotting approach is tested using the How2Sign video dataset. Generating sign labels manually is a laborious task, and this approach can provide a set of labels that can then be used to provide weak supervision for future machine learning applications.

Chapter 2

Videoconference and Embodied VR: Communication Patterns Across Task and Medium

2.1 Introduction

Videoconference, the dominant medium for remote meetings, uses video cameras to provide remote participants with a 2D, screen-based visual connection. Embodied VR, an emerging alternative, uses motion tracking and VR headsets to place participants in a shared 3D environment. This immersive 3D experience is more similar to face-to-face interaction, although at a lower fidelity, and it is important to understand how it impacts the collaborative experience. This paper describes a study comparing people's *behavioral patterns* across the two media and over a set of four tasks spanning different elements of workplace meetings. It builds on a long standing interest in how the affordances of communication media support various tasks [37, 53, 149]. Understanding the behavior engendered by the different affordances has important ramifications for the design of remote collaboration systems. As but one timely example, recent work postulates the exhaustion people feel from videoconfer-

encing, so called "Zoom Fatigue", may result in part from the behavioral pattern of people receiving too much gaze [13].

The study employs state-of-the-art, embodied VR technology that includes body tracking, face tracking and finger tracking to drive the movement of semi-realistic avatars (Figure 2.1, accompanying video), which provides a compelling interaction experience. Great care was taken to ensure that the videoconference (VC) and virtual reality (VR) conditions were as evenly matched as possible in the experiment, for example by employing a recommended videoconference framing that shows the upper body so that gestures read clearly, providing a shared mouse interface so that people could still point at shared artifacts in VC like they can in VR, and having all participants maintain a fixed, seated position in both conditions. Differences remain, however. The model based avatars have lower fidelity than video and do not fully reveal a person's identity (gender and ethnicity were matched). Conversely, the avatars allow people to be located in a shared 3D space, while videoconference remains screen based.

Participants worked in groups of three to complete a warm up and four experiment tasks that were designed to replicate different types of activities that might occur during meetings. An *intellective task* required them to come up with answers to questions where there was a correct answer. A *decision making task* required them to reach consensus when there was not a single correct answer. Two *mixed-motive tasks* required them to negotiate where each team member had different desires. The second mixed-motive task introduced a floor plan to visually ground the task and explore how this impacted nonverbal behavior. The experiment was run with Medium as a between subjects condition and Task as a within subjects condition. In other words, every participant was assigned to only one Medium, but completed all four Tasks. In total, 210 people participated in the study, 35 groups of three in each medium.

An analysis of performance and subjective measures (e.g. social presence) did not reveal notable differences between the two media. For instance, post task surveys derived from [149, 119, 148, 25, 61] showed high ratings without significant differences between media for the

Figure 2.1: Frames from the video log showing participants interacting over video conference (left) and in embodied virtual reality (right) while discussing an apartment floor plan.

scales *Satisfaction with Medium* (mean 6.30 for VC and 6.41 for VR on seven-point Likert scales), *Co-presence* (mean of 6.48 for VC and 6.53 for VR) *Mutual Understanding* (mean of 6.12 in VC and 6.24 in VR) and *Clear Communication of Affect* (mean of 5.42 for VC and 5.37 for VR) This paper focuses on the marked differences in behavioral measures.

Problem Statement: This work seeks to understand the behavioral differences that arise from people's use of either embodied VR or videoconference as a medium for conducting work meetings and if these behavioral differences are based on the nature of the work task. As the current default remote meeting option, VC provides an important comparison point for evaluating the behaviors induced by embodied VR. It is important to understand the potential and impacts of embodied VR ahead of potential widespread adoption, and these are partially contained in the behavioral patterns the media encourages.

Contributions: As far as we are aware, this paper reports on the first large scale comparative study of behavioral patterns of triad interaction across videoconference and embodied VR. The study included 210 participants with diverse demographics. To illuminate the role played by the different technologies, the basic meeting configuration is kept as similar as possible between the two media. Different types of tasks are contrasted to explore if

behavior changes as a function of task. Evaluated behavioral measures include conversational turns, gaze patterns and nonverbal behavior. *The paper contributes clear evidence of marked behavioral differences across task and media. Media differences include that VC participants spent much more time looking at interlocutors, especially their faces, they provided more verbal and nonverbal backchannels, performed more self-adaptors, fewer deictic gestures, and in some cases, had longer conversational turns. A potential explanation for these behavior changes is that participants in videoconference exerted greater effort to maintain a social connection than participants in embodied VR.* This suggests a differential in exertion required to use the two media that will likely impact users and warrants further in depth study.

2.2 Background

Theory

Media Affordances and Nonverbal Communication: There has been a long term interest in studying how the affordances of different media impact conversational interaction (e.g. [37]). Conversation is a collaborative process in which meaning is incrementally constructed together. It relies on both coordination and communication across verbal and nonverbal channels. It is made more efficient through *grounding*, a process through which interlocutors develop a shared understanding, and this coordination can be easier with a shared environment [159].

Conversational turn management indicates when it is another person's turn to speak and is largely done nonverbally. Head nods, gaze, and gesture all mediate turn taking [159]. People use more words and turns in audio telephony than face-to-face communication, and most notable for this study, turn taking in video conferencing tends to be more formal compared to face-to-face communications and is more similar to that seen in telephone calls [54, 117]. It is postulated that the increased verbal communication is a compensation for visual grounding that is less effective than in face-to-face settings [54].

Backchannels are feedback that listeners provide to speakers indicating that they are paying attention, have understood what is being said, etc. They are often nonverbal and include actions like head nods [23] and also phrases like "mmm" and "mhmm". Backchannels lead to smoother communication between the speaker and listener. O'Conaill et al. [117] found more auditory backchannels in face-to-face meetings than an early, high quality videoconference system, and by far the fewest in a low quality, single duplex videoconference system.

Gaze plays a rich set of functions, including expressing intimacy, exercising social control, regulating interaction and providing information [87]. It communicates a person's attention. In face-to-face communication, it allows people to tell who is staring at whom [53]. Gaze duration and looking at another's face are powerful cues [159].

Deictic gestures establish reference by pointing at objects and assist with grounding. Nonverbal deixis can increase the efficiency of communication [159]. Gestures can make representations of objects (iconic gestures) or ideas (metaphoric) [106]. Gestures are also used to regulate turn taking. They can be used to indicate emphasis, tone and subtext [85].

Finally, nonverbal communication performs a range of social functions, including: impression formation, person perception, communication of emotions and interpersonal attitudes [23] Bente et al. [23] argue that "[t]he effects emerge from implicit dynamic qualities, which rarely pass the threshold of conscious registration."

Task: Several taxonomies of group work have been put forward (e.g. [59]) and we rely on that of McGrath in designing our tasks [105, 149]. Their circumplex model consists of two dimensions, one runs from Conceptual to Behavioral; the second relates to the degree and nature of interdependence in three levels: collaboration, coordination, conflict resolution [149]. Following [149], social context cues are relatively unimportant when there is a demonstrably correct answer. The communication medium is more likely to have an effect when tasks require coordination, expression and perception of emotion, and persuasion or reaching consensus [149]. Whittaker argues "[f]or social tasks, there are clearly differences between mediated and face-to-face interaction, but for many cognitive tasks (especially those that do not require access to a shared physical environment), outcomes may

not be different." [159] Tasks involving interdependence or uncertainty require substantial amounts of interpersonal communication to be successful [91], which in turn places demands on the quality of the communication medium. Our tasks were designed to span this range (Sec. 2.3).

Related Work

There is a large related literature that is briefly sampled here. Influential for this work is the study of Strauss and McGrath [149] that focused on how medium (online chat system or face-to-face) interacts with task type (idea generation, an intellective task, and a judgment task). Results showed very similar quality output across the two interfaces, but face-to-face was more efficient. In the judgment task, people were less productive for the mediated interface and responded more negatively to the medium and task.

An early ethnographic study of videoconferencing [73] showed it had advantages over audio only in factors like showing understanding, expressing attitudes and nonverbal communication, but performed worse than face-to-face for peripheral cues, controlling the floor and pointing to objects, with the lack of correct eye contact seen as distancing. Dong and Fu [47] found videoconference more successful than audio or text for negotiation and attributed the difference to exchanging information in small pieces. Hauber et al. [64] found that spatial interfaces based on using multiple video screens to create a 3D environment positively influenced social presence and copresence measures in comparison to 2D, but the task measures favored the two dimensional interface. Other work shows a benefit of adding spatial video to an audio conference [72]. Nguyen et al. [115] compared videoconferencing systems that had a single camera for 2-3 remote participants with ones that had dedicated cameras and projectors directed for each participant. The non-directed video condition showed significantly less co-operative behavior than either directional video or face-to-face. Follow-up work showed greater empathy for an upper-body video framing than head-only [114], so that wider framing was adopted in this study. Other work has explored video projections for

two person remote interaction [125]. Schroeder [135] suggests that avatars could provide the spatial component missing in video, but they suggest a concern that the representation of the person may not be authentic. Wong and Gutwin [161] found that pointing in collaborative virtual environments benefited from being able to observe the prepatory arm motion, a direct connection between the gesturer and referent, and awareness of others views.

Early avatar research used a mixed head mounted displays with participants at workstations, and found a positive relationship between presence and co-presence, with accord increasing with presence [143]. Dodds et al. [45] found that a gesturing avatar led to more words guessed correctly in a game scenario and more gestures than a static avatar. Bente et al. [23] conducted an early, large scale study of avatar representations in which pairs of participants selected job applicants using one of six interfaces: text, voice, video conferencing, low-fidelity avatar (cartoon-style) and high-fidelity avatar (3D character). Text performed worse than all conditions on perceived intimateness, co-presence and emotionally based trust. Most dependent variables did not show a difference between video and avatar conditions. Notably, their avatars were displayed on 2D screens rather than the shared 3D environments used in this study. Steptoe et al. introduced one of the first avatar gaze tracking systems and provide preliminary evidence that it improves communication [145]. They later found that realistic eye movement increases participants ability to detect truth and deception [146].

A preliminary evaluation of the Holoportation AR avatar system suggests participants experienced spatial and social presence, and appreciated being able to control their point of view [120].

Pan and Steed showed that people asked for less advice from a key-frame animated, 2D projected avatar than a video or a robot expert, but would always prefer the more expert agent [122]. Smith and Neff [144] showed similar social presence and behavioral patterns for faceless, motion tracked avatars and face-to-face communication, but lower presence and a shift in communication patterns when avatars were not present in VR. Other work found no differences for avatar rendering style, but some differences for the amount of the body that was motion tracked [164] and a preference for full body avatar motion [65]. Jo et al. [78]

found that video performed worse than avatars on a measure that included spatial and social presence questions.

This study employs model-based avatars. A pre-rigged character model is used for each avatar, much like would be used in videogames, its movements driven by live tracking. An alternative avatar technology employs depth cameras or other optical techniques to create a point-cloud model of the person in real time, also known as 3D video (e.g.. [120, 58, 16] and hybrid approaches [84]). 3D video has the potential advantage of better preserving the person's identity, but tends to suffer from visual artifacts such as tear out (holes in the mesh), pixelation and issues with occlusion, as well as requiring complicated capture setups. It is also difficult to place multiple avatars in a shared 3D environment with this technique without also rendering the head mounted displays on the avatars, which blocks the face and limits communication. This paper is not focused on the particular avatar technology and we will simply note that at this time, model-based approaches offer more consistent visual quality and easier immersion in 3D.

The authors of impressive recent work suggest that it is the first to feature avatars with live tracking of the body, gaze and the lower portion of the face [131] and develop methods to augment avatar behavior beyond participants' actual motions. Our work tracks the same features, and also tracks hands and the full face, but our focus is on studying interaction patterns relative to a videoconference baseline, so we do not intentionally modify participant behavior. Other work has also explored the potential of adaptation, focusing on facial expressions in VR [62]. These studies point to the additional potential VR offers for modified or augmented social interaction.

2.3 Method

Experiment Design

Participants attended a single session, during which they interacted with two other participants in technologically mediated social interaction. The experiment design is mixed, with a between groups factor *mediation interface* at two levels: embodied VR and VC (see sec. 2.3), and a within group factor of *task type* at four levels (see sec. 2.3). The order of *task* was randomized. In short, groups of three completed all tasks with one of the two mediated interfaces. A range of behavioral measures were calculated from live measurement and analysis of the session recordings (Sec. 2.3).

Tasks

After a short warm-up task (a version of the Desert Survival Game, [94]), participants completed the following four tasks which were selected to cover different task types on McGrath's circumplex [105], which models different forms of group interaction. Task details and instructions are included in the appendix.

Estimation: An "intellective task" (Type 3 on McGrath's circumplex) involves solving problems with correct answers. Participants were asked to determine answers as a group to questions that require them to make statistical estimates. For example, "How many times does an average person blink in a day?"

Bribery: A "decision making task" (Type 4) involves coming to agreement on a matter that does not have a demonstrably correct answer. The experiment employed a moral judgment task in which participants act as a tribunal on a whistle blower case set in the workplace and must decide on appropriate punishment. A top salesperson has accepted an expensive paid trip from a client without reporting it. The salesperson's boss heard about this from a whistleblower, but failed to report it. The group needed to decide on a punishment for both while considering various stakeholders within the company.

Party Planning: A "Mixed-motive Negotiation Task" (Type 6) involves people coming to agreement when participants have different motives. Both mixed-motive tasks used a form of multi-issue bargaining, in which participants must agree on several different issues [57]. The scenario required the participants to agree on terms for a company party: the number of security guards to hire, the end time of the party, the price for guest tickets and how many knife jugglers to hire for entertainment. One participant was made Head of Security, one Head of Finance and the third Head of Social Planning, giving them conflicting interests. Each participant has a different points-table reward structure based on how each issue is settled, with conflicting and complimentary goals. They had time to study this before the session.

Floor Plan: The final task was also a mixed-motive negotiation task, but it introduced an artifact - a floor plan - to visually ground the discussion. Participants were told they are roommates that will be sharing an apartment. They need to decide on the room allocation and how to split the rent, with conflicting and complementary desires for rooms and additional features such as an extra closet. These were represented in a points table. In VR, the floor plan was placed on the table in front of participants. In VC, the floor plan was displayed on the same monitor with the remote participants and each participant had a different colored mouse they could use to point to items on the floor plan. This allowed a form of gesturing in both media to maintain an equivalent task.

Procedure

Upon arrival, participants were kept physically separated to ensure that all interactions between participants only occurred through the mediated interface. The participants were lead to three separate but similarly arranged areas (small conference rooms for VC, and partitioned areas for VR). To familiarize participants with the system and each other, the warm-up task was completed first, followed in randomized order by the four experiment tasks. Instructions were provided at the start of each task. Participants were told they could receive

a reward of up to $11 per task based on their performance to incentivize engagement. All participants were paid the full reward at the end of the experiment.Each task could last up to 15 minutes, and the total time for an entire session was up to 195 minutes. Following each task, the remote connection was stopped temporarily for the participants to complete post-task surveys and to take a brief break. The experimenter's role was only to initiate the start and end of each task and to answer questions regarding task instructions.

Participants

There were 70 groups of three, a total of 210 participants, divided evenly between the two media conditions. Five participants did not provide demographic information. The remaining 205 were diverse in terms of age (M = 32.82, SD = 8.07), race (34 Black/African American, 39 Asian/Asian American, 4 Pacific Islander/Native Hawaiin, 84 White/Caucasian, 29 Latin/Hispanic, 2 American Indian/ Alaska Native, and 6 Other/Prefer not to say), gender (96 females, 105 males, 2 non-binary/third gender, 2 other), and education (15 completed high school, 48 completed some college, 23 with an Associate's degree, 93 with a Bachelor's, and 26 participants with a graduate degree). VR participants were diverse in terms of their experience in VR (35 with no prior experience, 43 with some experience with VR, 14 who had experienced VR several times, and 3 who own their own VR headset. 10 did not state their experience with VR).

Participants were recruited from a community participant pool and a research recruitment vendor. They were paid $200 for the 3 hour and 15 minute session, plus the $44 bonus. All participants were 18-45, without eye conditions that would impact tracking and comfortable speaking English. All groups were strangers, except one VC group that had two distant acquaintances. An effort was made to gender balance across media. The gender composition for video conferencing was 6 all-female groups, 7 all-male groups, 19 mixed groups, and 3 groups where we do not have enough information to classify. For the virtual reality condition, we had 8 all-female groups, 8 all-male groups, 16 mixed groups, and 3 groups where we do

not have enough information to classify. One VC session was dropped due to failed eye tracking and one VR condition due to failed data recording.

Apparatus

Each participant used one of three identical *stations* featuring one of the two mediation interfaces:

VC

Each VC setup consisted of a participant looking at the other two participants on a 55" screen at a distance of roughly 5.5 feet. We used Zoom for videoconferencing with settings to keep participants in the same sized window throughout the task and their self view turned off. We followed the camera placement recommendations of Chen [35]. Eye Tracking was done with Tobii Pro Nanos. We used Microsoft LifeCam for the participant video streams. Interactions were recorded using OBS desktop capture. In addition, the video stream with the gaze overlay data from the Tobii Pro Software was recorded for later analysis.

Each participant saw two other participants horizontally laid out on screen for the Estimation and Bribery tasks. For mixed motive negotiations, the right half of the screen was vertically divided to show participants. The left half showed shared visual artifacts (the floor plan) and/or private points tables. A software called "UseTogether" allowed participants to each move a mouse on a shared floor plan.

VR

Participants wore a modified Oculus Rift Head Mounted Display (HMD) to view the VR scene of an office meeting room. They had their body, finger and face movement tracked in order to project them into the scene as avatars. Body tracking was performed using a single Kinect sensor as input and a custom motion solver that estimated the participant's skeleton pose. Body landmarks are inferred and an IK algorithm solves for skeleton joint angles by

minimizing the squared distance of the observed landmarks and the attachment positions on the skeleton. Finger pose was calculated using HMD mounted cameras and a custom solver based on [60]. Face tracking was performed with cameras placed inside and outside of the HMD to view the participant's eyes and mouth. These cameras provide direct gaze tracking and were also used to estimate gains for a set of facial blendshapes in order to track facial deformation. Each station used two computers that each housed a 12 core Intel Xeon processor, with 64 GB memory and either two or three Nvidia GTX 2080 graphics cards to perform tracking and stream combined motion into the VR scene. Overall, this provides direct tracking (not "head and hand" tracking) for a ¿100 DOF skeleton along with a 70 blend shape facial model, yielding very nuanced nonverbal behavior. All data is broadcast to a local network shared by the three setups. The system runs between 55-75fps.

A custom set of 36 avatars was built that included 3 male and 3 female avatars for each of 6 racial groups (Caucasian, East Asian, South Asian, African, Middle Eastern and Hispanic). This allows for basic matching between participants and their avatars. We ended up using 31 of those during the VR study sessions. During testing, it was found that extreme facial expressions on some avatars would create distracting artifacts, such as the eyeball penetrating the eyelid. To avoid these ever appearing during interaction, the range of the avatar expressions was reduced. A side effect of this was that overall expressiveness of facial motion was reduced. While speech activity was clearly visible on the animation of the mouth and lips, facial expressions were present, but damped.

Measures

Behavioral Measures were tabulated for conversational turn-taking (described in detail in Sec. 2.4), gaze (Sec. 2.5) and gesture (Sec. 2.6) behaviour. The analysis of the gaze data from the Tobii eye trackers in VC and internal HMD cameras in VR was largely automated and is described in detail in the Appendix. The other behavioral measures rely on annotations of the video logs of the sessions (Fig. 2.1). Videos were divided into one minute segments

for annotation and annotators coded a single speaker at a time. Annotation was done by two remote annotation teams that were trained for this work. They were given detailed instructions that were then reviewed together with the research team. During training for both gestures and conversational turn annotation, annotators were given examples of correct annotations from the researchers. The final task for training was for each annotator to complete 20 annotations and for a researcher to manually evaluate and approve that they completed training successfully. During an initial test phase, annotators annotated clips for which a gold standard annotation had already been produced to check quality. After any issues were addressed, annotators proceeded to the main data. When they had questions on any part of the annotation, these were addressed by the research team. Annotations were spot-checked to ensure accuracy and annotators were encouraged to seek clarification throughout. Annotation was done using a customized annotation tool for remote video annotation.

Gesture annotation was completed by 4 annotators who were blind to the research goals and were not involved in data collection for the study. The annotators completed annotation using a predefined list of gestures and were asked to mark the start and stop of each gesture, the gesture label, and the reference label. The predefined list of gestures is detailed in the Appendix. Each 1-minute clip was independently annotated by 2 annotators, and if there was a mismatch between their annotations, then a 3rd annotator would review the annotations and arbitrate to produce a single annotation per gesture for our final analysis. Gesture annotation with a large label set is a challenging task given the subjective quality of co-verbal gesture – our annotators would initially agree on about 77% of gesture labels and 84% of reference labels – so the arbitration process provides a realistic method to achieve a high quality annotation. Spot checks were conducted daily by the researchers checking 20% of the annotations.

Conversational turn annotation was completed by 10 annotators who did not work on gesture annotation nor data collection, and were also not familiar with the research goals. Each 1-minute video clip was annotated once. Initial tests showed that this was a straight-

forward task that could be done accurately by a single person. Spot checks were performed by both an annotation manager and the researcher daily. After annotators completed annotations, the annotation manager reviewed 20% of their jobs and had any issues addressed. Once the annotation manager's checks were completed, the researcher reviewed about 10% of all jobs submitted for each day, which included both those that went through the annotation manager's review and those that did not.

Statistical Tests Used

To avoid cluttering the discussion, statistical methods will be summarized here. Distributions were checked for normality. When normal, a linear mixed-effects model was fit to the data using the lmer() function in R. Linear mixed effect models are used to predict the dependence of a response variable (i.e. the item being measured, such as gaze duration) on one or more covariates (e.g. the Medium). They include both fixed and random effect terms, where a repeatable factor, such as Medium, is fixed and a non-repeatable factor, such as participant, is modeled with a random-effect term. Further details and information on the lme4 package which implements lmer() and glmer() can be found in [18, 19]. For non-normal distributions, a generalized linear mixed-effects model was used with either the glmer() or glmmPQL() [157] function and a log normal or Gamma distribution, depending on the data, as these provide a more accurate fit of the data. Significance of main effects and interactions was calculated using Anova, which performs Wald tests. Post-hoc tests were performed using estimated marginal means (emmeans() [136, 133]) which can be used with mixed effect models to compute pairwise comparisons which applies the Tukey method for correction. In some cases, a Wilcoxon rank sum test with continuity correction was used to compare two non-normal distributions and Bonferroni correction was applied as needed.

2.4 RESULTS: Conversational Turns

Analysis in this section focuses on three types of activities.

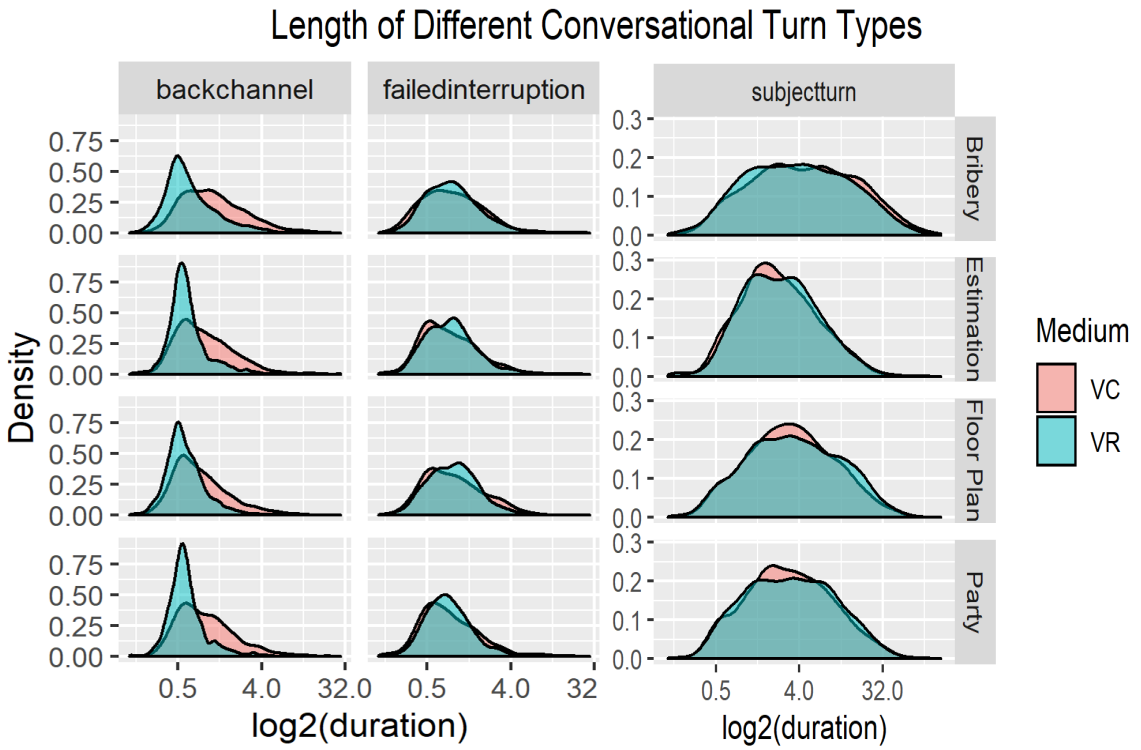

Figure 2.2: Ratio of session time spent on all turns.

Turn Duration

A conversational turn (*speaking turn or subject turn*) is the period someone holds the floor while talking before yielding to another participant. Occasionally, nonverbal cues can be used to hold the turn, such as holding a hand out during a pause to indicate that you are not done speaking. Conversational turn length and interruptions provide an indication of the fluidity of conversation. For example, longer turns occur when people do not perceive that others want to speak[117, 159]. A *Backchannel* occurs when the listener provides acknowledgement, such as saying "mhmm" or nodding their head. *Failed interruptions* occur when someone tries to take the conversational floor, but the person speaking does not yield.

We find that the length of speaking turns (Figure 2.2) is impacted by task, with an

ordering from longest of: Bribery (median 4.08s , mean 8.94s, sd 12.5), Floor Plan (median 3.24, mean 6.14, sd 7.71), Party Planning (median 3.09s, mean 5.63s, sd 7.24) and Estimation (median 2.28s, mean 3.88s, sd 4.75). In all cases, these differences are significant for both media at $p < .0001$ except for the differences between the Floor Plan and Party Planning tasks. The latter are significant for VR (t.ratio = 2.762, p= 0.029), but not VC (t.ratio = 2.24, p= 0.11).

The impact of medium is reflected in significant interactions. The largest difference is for Bribery, where VC turns are significantly longer, over 20% on average (t.ratio = 9.88 , p¡.0001; VC median 4.41s, mean 9.86s, sd 13.8; VR median 3.73s, mean 8.13s, sd 11.04). Turns are significantly shorter for VC in Floor Plan (t.ratio = -2.46, p =0.014; VC median, 3.17s mean 5.76s, sd 7.22; VR median 3.38s, mean 6.53s, sd 8.19) and Party Planning (t.ratio = -2.023, p =0.043; VC median 3.00, mean 5.50, sd 7.20; VR median 3.18, mean 5.80, sd 7.29), but these differences are somewhat less marked, averaging 13 and 5 percent respectively. Durations were not significantly different for Estimation.

There is no significant difference in the length of failed interruptions.

The duration of backchannel turns is significantly longer for VC than VR (Chisq 738.45, Df 1, p ¡ 2e-16; VC median 0.91s, mean 1.42s, sd 1.62; VR median 0.58s, mean 0.77s, sd 0.87) and this relationship remains significant across all Tasks. There is also an effect of Task (Chisq 278.71, Df 3, p¡ 2e-16). Post-hoc analysis reveals that backchannels are longer for Bribery than all other tasks and this relationship holds for both VC and VR (Bribery median 0.82s, mean 1.36s, sd 1.66; Estimation median 0.65s, mean 1.01s, sd 1.21; Floor Plan median 0.64s, mean 0.96s, sd 1.05; Party Planning median 0.68s, mean 1.08s, sd 1.64)

Turn Frequency

The duration of the turn provides one characterization of conversation. Frequency of turn types is an important complement (Figure 2.3). Examining backchannels per minute shows a significant main effect for Medium (Chisq 7.4957, Df 1, p = 0.0062) and Task (Chisq

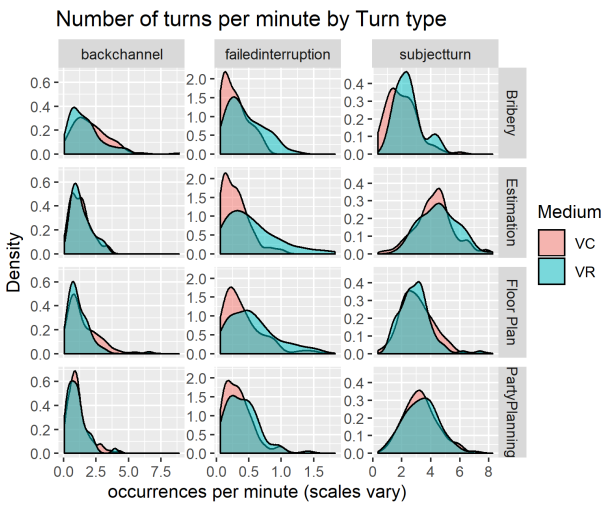

Figure 2.3: Turns per minute.

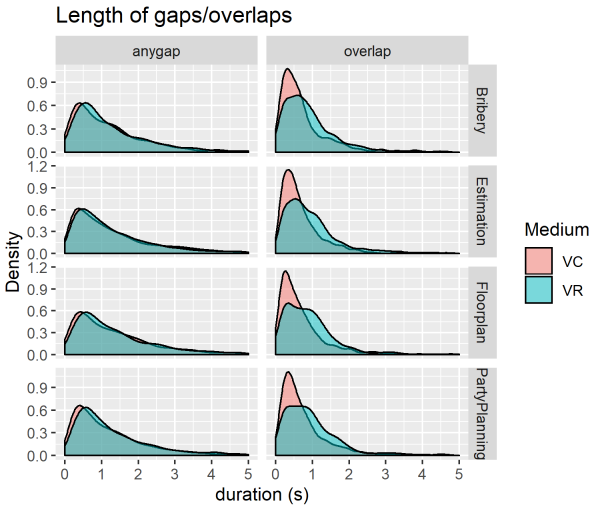

Figure 2.4: Gaps and overlap.

71.2297, Df 3, p¡.0001), but no significant interaction (Chisq 4.2161, Df 3, p = 0.24). VC backchannels are more frequent (VC median 1.18 per min., mean 1.49, sd 1.12; VR median 1.06, mean 1.30, sd .95). Post-hoc analysis shows that backchannels are more frequent in Bribery than all other tasks (all p¡.0001). The frequency was also significantly higher in Estimation than Party Planning (z.ratio -3.184, p = 0.0079. The overall frequencies by Task are Bribery (median 1.60, mean 1.89, sd 1.33); Estimation (median 1.16, mean 1.33 sd .79); Floor Plan (median .98, mean 1.30, sd 1.00); Party Planning (median .92, mean 1.06, sd .78).

An analysis of the frequency of failed interruptions showed a significant main effect of Medium (Chisq 45.55, Df 1, p¡.0001) and of Task (Chisq 13.75, Df 3, p=0.0033) and a tendential interaction (Chisq 6.85, Df 3, p=.077). Failed interruptions are more frequent in VR (VC median .29 per min., mean .33, sd .23; VR median .43, mean .49, sd .34). Post-hoc analysis suggests that they were more frequent in Floor Plan than Bribery (z.ratio 3.182, p=0.0080) and Party Planning (z.ratio -2.962, p=0.0161) (In order by mean occurrences per minute: Floor Plan median .42, mean .49, sd .34 Estimation median .36 mean .43, sd .34

Bribery median .31 mean .38, sd .25 Party Planning median .32, mean .37, sd .24.)

Turning to the frequency of speaking turns, there was no main effect for Medium (Chisq 2.0485, Df 1, p=0.15), but there was a main effect for Task (Chisq 300.15, Df 3, p ¡ 2e-16) and a significant interaction (Chisq 10.7463, Df 3, p = 0.013). There is a significant interaction between Medium and Task for Bribery (z.ratio 3.332, p = 0.0009), with VC being less frequent (VC 2.02 vs. VR 2.42). The differences between all Tasks are significant for both Media, except Floor Plan and PartyPlanning, which are significant for neither (VC z.ratio=1.341, p=0.5370; VR z.ratio=2.496 p=0.061). Tasks by order are: {Bribery median 2.13, mean 2.21, sd 1.01}, {Floor Plan median 2.95, mean 2.99, sd 1.11, PartyPlanning median 3.28, mean 3.32, sd 1.16}, {Estimation median 4.48, mean 4.38, sd 1.31}.

Conversational Gaps and Overlap

We looked at both the length of the gap between speaker turns and the amount turns overlap (Fig. 2.4). The gap duration data did not show a consistent pattern, but there is a clear difference for overlap. There is only a significant main effect of Medium (chisq 104.64, Df 1, p¡2e-16). This shows that overlaps are significantly longer in VR (VC median .52, mean .71, sd .62; VR .77, mean .93, sd .72).

Discussion

Bribery is clearly distinguished from the other tasks. It had the longest speaking turns, the longest backchannels, the least frequent speaking turns and the most frequent backchannels. *In short, people spoke longer but less frequently, and they both provided more backchannels and these had a longer average duration.* The Bribery task is social, subjective and collaborative. The longer speaking turns could be explained by the need to make more complex, and hence longer, arguments, given the subjective material. The more frequent and longer backchannels would seem to reflect the need for increased coordination as there was a need to reach consensus on a more subjective and emotional task. By contrast, during Estimation

people were often sharing short facts or guesses, which could explain the Estimation speaking turns being the shortest and most frequent.

The most striking difference between media occurs around backchannels, which were both longer and more frequent in VC than VR. Backchannels serve to maintain the social connection, acknowledging that the other person is heard and understood. People felt a need to do more of this "connection maintenance" in VC. Another possible explanation is that people were simply less connected or more tuned out in VR, but social presence surveys run during the experiment showed similar levels across the media, speaking against this explanation.

Previous work has shown that turn taking in video conferencing (VC) is more formal than in face-to-face communication [54, 117]. It appears that it is also more formal in VC than embodied VR. Failed interruptions were more frequent in VR and there were longer overlaps of speaking turns in VR. Both suggest less careful adherence to strict turn taking behavior. It is important to note that in our coding, failed interruptions included all times a person interrupted or interjected, but did not obtain the floor, so they are not necessarily negative. Listening to the sessions, it appears that people were more comfortable and more able to effectively talk over each other in VR by providing brief interjections of helpful content, whereas they followed a more strict turn taking approach in VC. This is also consistent with the increased overlap in VR.

One could argue that this overlap occurred because people receive less clear signals in VR that they are speaking over top of another and it took them longer to detect this "error". However, the fact that the duration of failed interruptions was similar across media speaks against this. It suggests both media provided sufficient clues that the interruption would not be successful (or neither media made it clear to the speaker that someone was trying to interrupt), and suggests that people gave up on attempts to interrupt after a similar effort. People appear to be more comfortable overlapping dialogue in VR.

The length of speaking turns is longer in VC for Bribery, but shorter in VC for Floor Plan and Party Planning. For the latter two tasks, the payoff tables in VR were displayed with a virtual touch interface that users found a bit difficult to use, which might have led to longer

turns. The differences are much larger in Bribery. Shorter turns tend to occur when there are more clear nonverbal signals for turn taking [159], which might explain the difference in Bribery, but conclusions should be drawn with caution as VR only outperformed in this one task.

2.5 RESULTS: Gaze

Technical details on gaze tracking analysis are contained in the appendix.

Categories of Gaze

A first analysis considers the distribution of gaze by task and condition. Gaze is broken into three broad categories: *Body,* which includes gaze at any other participant; *Task,* which includes gaze at task artifacts when they exist (the payoff tables or floor plan); and *Elsewhere* which includes all remaining gaze.

Figure 2.5 shows the portion of time participants looked at various body parts in VR and VC. This data was best fit to a generalized linear mixed effects model with a Gamma distribution, containing all main effects and interactions. Type II Wald chisquare tests show no main effect for Medium ($\chi^2(1) = .0271$, $p = 0.8692$), but significant main effects for Task ($\chi^2(3) = 1241.9$, $p < 2.2e-16$) and Category ($\chi^2(2) = 1471.0$, $p < 2.2e-16$), as well as all interactions being significant: Medium:Task ($\chi^2(3) = 33.37$, $p = 2.697e-07$), Medium:Category ($\chi^2(2) = 344.2$, $p < 2.2e-16$), Task:Category ($\chi^2(4) = 148.4$, $p < 2.2e-16$), Medium:Task:Category ($\chi^2(4) = 208.4$, $p < 2.2e-16$) . Table 2.1 shows the means for VC and VR in each Task and gaze target category, along with post-hoc statistics computed using the Tukey method and emmeans(). Participants in VC spent significantly more time looking at their interaction partners than in VR across all tasks, on average, *56% more time*. Participants in VR spent significantly more time looking at Task artifacts than in VC for the Party Planning task.

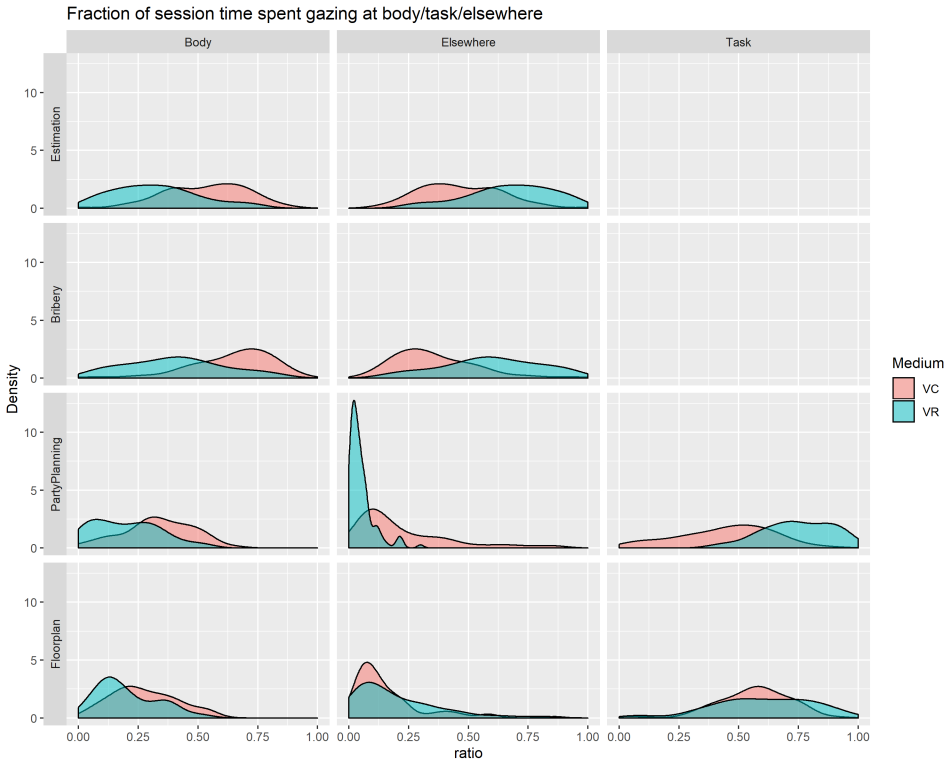

Figure 2.5: Ratio of gaze at different targets across Task and Medium.

	Body				Elsewhere				Task Item			
Task	VC	VR	z ratio	p	VC	VR	z ratio	p	VC	VR	z ratio	p
Estimation	.532	.325	-8.429	$< .0001$	.468	.675	7.601	$< .0001$	-	-	-	-
Bribery	.642	.406	-8.973	$< .0001$	.358	.594	9.173	$< .0001$	-	-	-	-
PartyPlanning	.334	.201	-6.216	$< .0001$	.222	.0517	-8.875	$< .0001$	.444	.747	11.210	$< .0001$
Floor Plan	.275	.202	-3.366	0.0008	.162	.197	1.725	0.0846	.563	.600	1.344	0.1789

Table 2.1: Proportion of time spent looking at other participants (Body), at something not task related (Elsewhere) or a task artifact. When differences are significant by medium, they are color coded pale red for the more frequent, blue for less.

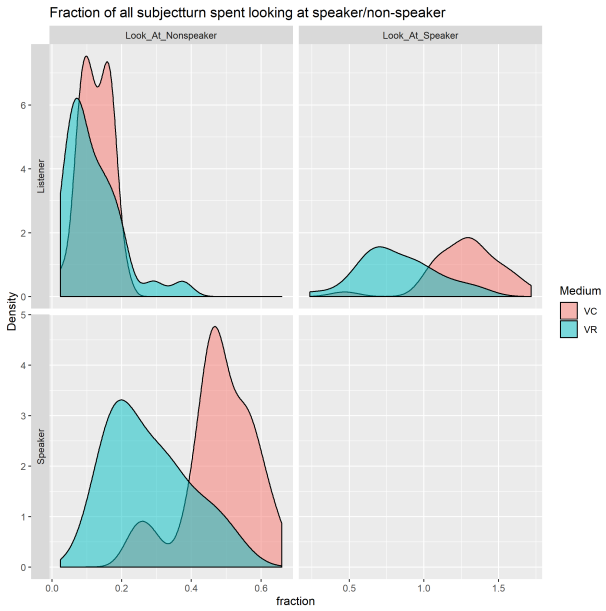

Figure 2.6: The proportion of time people spend looking at speakers and listeners during conversational turns.

Ordering by the mean proportion of time, people spent the most time looking at other participants in Bribery, followed by Estimation, Party Planning and Floor Plan. The difference between Task was significant in all cases except the difference between Party Planning and Floor Plan in VR.

Gaze and Body Areas

The time participants spent looking at another participant is broken down by the areas of the body they gazed at in Figure 2.7. If participants only briefly gazed at another participant, these ratios may not have stabilized. Therefore any *participant sessions* (a single participant on a single task) for which there were less than 45 seconds of body gaze were dropped from the analysis. This left 677 participant sessions for analysis.

Models were fit to the data by progressively adding the factors Category, Medium and Task. The best model fit was obtained using a Laplace approximation for a generalized

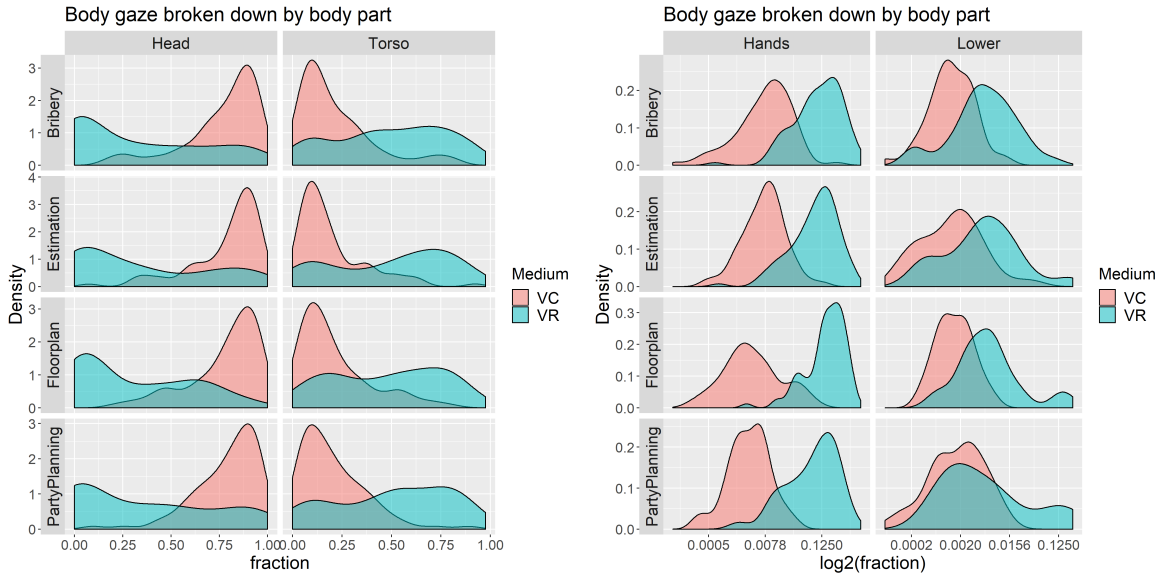

Figure 2.7: These figures show the distribution of time looking at different body parts when a participant is gazing at the body. Note that the Head and Torso figure on the left is on a linear scale and the Hand and Lower body figure on the right is on a log scale

linear mixed effect model with a Gamma distribution for gaze Category and Medium. This indicates that Task did not have a significant effect on gaze distribution, which is consistent with Figure 2.7. There was a significant main effect for gaze Category ($\chi^2(3) = 4903.3658$, $p < 2e-16$), but not Medium ($\chi^2(1) = 0.12$, $p = 0.73$). There was a significant interaction between Body and Medium ($\chi^2(3) = 1329.88$, $p < 2e-16$). The means and significant differences for Medium are summarized in Table 2.2, which shows that Medium had a significant impact on every Category but Lower body. Differences between Categories were always significant except for the difference between Hands and Lower Body for VC.

Head and Torso are the main gaze targets overall. For VC, the Head is by far the dominant focus of attention (78%), with the Torso receiving the bulk of remaining attention (20%). Participant gaze in VR is more dispersed across the body. The Torso was the most significant category (49.2%), followed by Head (34.3%) and Hands emerge as an important target (16.0%). The Lower body receives little direct attention in either medium. It was

	Head				Torso				Hands				Lower Body			
Task	VC	VR	z ratio	p	VC	VR	z ratio	p	VC	VR	z ratio	p	VC	VR	z ratio	p
All	.78	.34	-27.07	***	.21	.49	20.42	***	.010	.16	13.43	***	.0013	0.0049	0.35	0.73

Table 2.2: Proportion of time spent looking at different body parts, averaged across Task because Task did not lead to significant variation. The table shows means for VC and VR. When differences are significant by medium, they are color coded pale red for the more frequent, blue for less. *** indicates p< .0001

also largely obscured by the edge of the screen in VC and table in VR. It is worth noting that these statistics only account for direct gaze. Participants may of course still perceive movements of other body parts through their peripheral vision.

Gaze During Conversational Turns

Figure 2.6 shows how frequently participants look at the speaker or listeners during conversational turns. Note that for one speaker, there are two listeners and this figure sums results over the total number of people in a role (i.e. if both listeners look at the speaker throughout, this would produce 200% listener-at-speaker gaze). Since the data is not balanced and there is no anticipated relationship between the categories, separate tests were used to compare the impact of Medium on each condition. A speaker gazed at a listener an average of 48.1% of the time in VC and 28.0% in VR. The distributions are significantly different according to a Welch Two Sample t-test (t = 7.0329, df = 54.843, p = 3.38e-09). The listeners gazed at a speaker on average 128.4% of the time in VC and 82.5% of the time in VR. This difference is statistically significant according to a Welch Two Sample t-test (t = 6.9751, df = 55.547, p = 3.95e-09). Listeners would gaze at non-speakers on average 12.4% of the time in VC and 11.9% of the time in VR. Since the VR distribution is not normal, a t-test based on various robust estimators was used and indicated no significant difference (pb2gen, Test statistic: 0.0211, p = 0.27).

Discussion

The two media are clearly distinguished by differences in gaze behavior. People spend much more time looking at each other in VC than VR, on average 56% more time, and when people do look at each other, they are much more likely to look at the head in VC (78%) than in VR (49%), where gaze is more distributed, with the Torso (34%) and Hands (16%) also receiving notable attention. This behavior occurs for both speakers and listeners. Speakers in VC look at a listener 48% of the time, compared to only 28% in VR. Listeners look at speakers on average 64% of the time in VC versus 41% of the time in VR. Interestingly, listeners look at other listeners about the same in either medium. This seems to suggest that the increased gaze is therefore about the speaker-listener interaction or connection, and not about maintaining a connection to everyone in the triad.

We postulate that this additional gaze felt necessary in order to maintain the social connection in VC. This is based on a reflection on the many functions of gaze. One function is to both give and show attention [87]. People are clearly giving more attention to the people at the other end of the speaker-listener interaction in VC. This could reflect that people are for some reason more interesting in VC, perhaps because their facial features are rendered with higher fidelity. If this was so, however, it seems reasonable that there would be some difference reflected in subjective measures of social presence, and those were not seen. Conversely, it is possible that the VR environment was more interesting, but both the VR and the VC environments were quite plain offices. Perhaps a more likely explanation is that this increased attention feels necessary to maintain connection with someone who is remote and located in a different 3D space. This need could be driven by two sources: people might feel that they need to look at the other person so that they are not distracted by other items in their own space which the remote participant cannot in general see or perhaps they feel a need to show to the other person that they are paying attention by directing their gaze at them. We expect the latter may be dominant, but this requires further investigation. When people are co-located in the same space in embodied VR, a lower level of gaze may have felt

sufficient to maintain the connection.

Another function of gaze is to regulate turn taking [159]. If other methods of turn taking regulation, such as gesture, are less effective in VC, this might lead to increased reliance on gaze cues, and hence increased gaze, although Beattie suggests gaze becomes a less effective turn taking cue as the baseline level of gaze increases [22].

There are also costs associated with gaze. It is a powerful cue [159] that can express intimacy and exercise social control [87]. It is known that there is social pressure associated with receiving directed gaze, and this has even been exploited in VR by artificially increasing gaze to increase listener attention [15, 14]. There may also be costs with *giving* gaze. It is worth considering what burden this increased gaze may place on users.

Explanations for the more distributed attention in VR could include that people gathered more information from other body parts in VR, perhaps because gesture was more effective, or again that they felt more of a need in VC to look directly at a speaker as a way to show attention. It may also be that the face in VR provided a less reliable signal with the current state of technology, so people relied more on other signals. Another possibility is that the people in VC occupied less of people's field of view, so they could see the whole person by staring at their head, but gaze needed to be more distributed in VR to fully observe the person.

Increased gaze during Bribery could occur because this is the most subjective, social task and so relies heavily on people being able to read social signals. Party Planning and Floor Plan both featured Task objects that garnered significant attention. They may have looked at these frequently to remember the payoffs or to intentionally avoid gaze in a competitive task. The increased task gaze in VR Party Planning may be due to delay caused by the increased difficulty of using a virtual touch interface compared with a mouse.

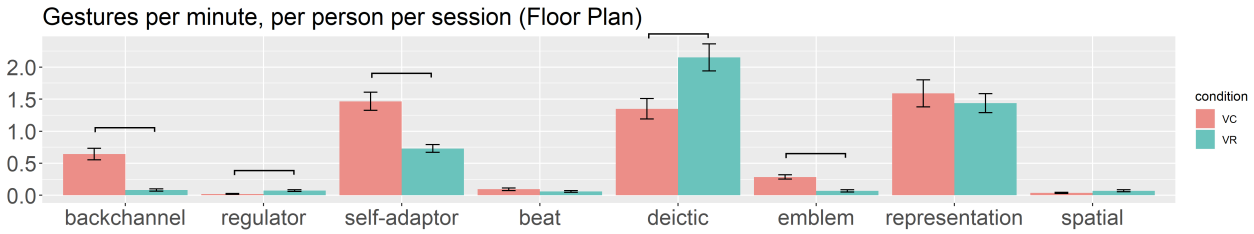

Figure 2.8: A comparison of gesture rates between VC and embodied VR for the visual aided out FloorPlan task. Significant differences are marked.

2.6 RESULTS: Nonverbal Analysis

A detailed annotation (details in Appendix) was completed of the gesture behavior for participants in the Floor Plan task. Due to the high effort involved in such an annotation, we were only able to complete this for a single task and the Floor Plan task offers the richest visual grounding, so is particularly appropriate for a gesture analysis.

Gesture Frequency

Figure 2.8 shows the frequency of different gesture types across media. Each gesture type was analyzed separately to see if it varied based on Medium. Applying Bonferroni correction on eight tests yields an $\alpha = .00625$ threshold. Since the distributions are not normal, Wilcoxon rank sum tests are used throughout.

Backchannels are significantly more frequent in videoconference (W = 5723.5, p ¡ .0001, VC median .30 per minute, mean .64, sd .81; VR median 0, mean .083, sd .18). Self-adaptors are also significantly more frequent in videoconference (W = 4882, p-value = 0.00015, VC median 1.19, mean 1.47,sd 1.29; VR median .57, mean .73, sd .58). Regulators, gestures that explicitly turn over the conversational turn, were quite rare with only 73 occurrences in the corpus, but significantly more common in VR (W = 2860.5, p-value = 0.0018; VC median 0, mean .021, sd .049; VR median 0, mean .075, sd .13).

There was not a significant difference in the rate of Beats (W = 4182, p-value = 0.048).

Deictics were significantly more frequent in VR (W = 2644, p-value = 0.0018; VC median .88, mean 1.35, sd 1.46; VR median 1.74, mean 2.15, sd 1.96). Emblems were uncommon, but significantly more common in videoconference (W = 5281, p-value = 4.5e-08; VC median .21, mean .29, sd .31; VR median 0, mean .071, sd .16). The rate of Representational gestures did not vary significantly based on medium (W = 3469, p-value = 0.57). Spatial gestures were also relatively uncommon and did not differ significantly by Medium (W = 3365, p-value = 0.26).

In terms of novel information in gestures, 5.3% of gestures were marked reference pronoun (definitions in Appendix, Table 2.4) in videoconference compared to 15.8% in VR. Other unique information occurred with 6.65% of gestures in videoconference and 4.80% in VR. Overall, 12.0% of videoconference gestures and 20.6% of VR gestures contained unique information. All these differences are significant using a 2-sample test for equality of proportions with continuity correction ($\chi^2(1) = 220.24$, $p =< 2.2e-16$, $\chi^2(1) = 11.955$, $p = 0.00054$; $\chi^2(1) = 103.5$, $p < 2.2e-16$)

Discussion

As with conversational turns, nonverbal-only backchannels are more frequent in videoconference. *Assuming that people are not more agreeable in one condition than the other, this may show a greater felt need to actively maintain the social connection in videoconference and is consistent with behavioral changes seen for conversational turns and gaze.* Self-adaptors are associated with anxiety [158], so their increased prevalence in videoconference could suggest less comfort.

One explanation for increased person-directed gaze in VC is that this is required for turn regulation. The increased use of regulators in VR could reflect the greater ease of passing the turn in a 3D environment using gesture, compared with having two interlocutors on a screen, which would be consistent with the discussion related to gaze. However, these were still rare events, so do not suggest a major change in behavior between media, and hence,

turn regulation is unlikely the dominant reason for gaze differences. There are also multiple syntactic and paralinguistic cues for turn taking that could be deployed [22].

Even though every effort was made to provide a shared mouse interface that facilitated pointing, it is still likely easier to point in a shared 3D environment, so it is not surprising that deictics were higher in VR. The increased emblem use in videoconference could be because people found them easier to use in this medium if hand tracking was not perfect, gestures like deictics were less accessible so people switched to emblems or perhaps there was more use of "thumbs up" and similar gestures in an effort to maintain a social connection. Again, these were relatively infrequent, so strong inferences should not be drawn.

There was a marked overall increase in unique information conveyed by gestures in the VR condition, although videoconference had more unique information that was not associated with a reference pronoun. The increase in reference pronoun use in VR exceeds the increase in deictic gestures alone. This likely suggests that it was more fluid in VR to point at things, so people were more likely to use that pointing as a substitute for speech.

2.7 Discussion

This section will relate our work to previous research and offer interpretations for our findings.

Task: Considering the Task factor first, *the behavior during Bribery is clearly differentiated from the other tasks. It had the longest speaking turns, the longest and most frequent backchannels, and the most person-directed gaze. The Bribery task requires a high level of coordination between the participants.* It is a subjective task where the group must strive for mutual understanding and to reach consensus. Estimation also requires consensus, but on factual questions so it lacks the subjective nature of Bribery. Both Floor Plan and Party Planning are negotiations that can be viewed, at least in part, as adversarial and do not involve a sensitive discussion of moral norms nor group consensus, like Bribery. The longer turns in Bribery may suggest more elaborate arguments. Longer backchannels and increased gaze seem to reflect attentive listening, which is consistent with the task. This increase in

prosocial behavior provides evidence related to theoretical arguments such as information richness theory, that suggest that richer (more embodied) media are better when equivocality is higher [41] and there are needs for group coordination [149]. People are at least making more use of additional affordances on the most equivocal task, whether or not this actually improves their performance. The increased gaze and backchannel behavior is also consistent with research suggesting embodiment is more important for social, interpersonal communication, whereas speech is more task-oriented [159].

Performance: Turning to the impact of Medium, while our detailed performance analysis is not included in this paper, *it is perhaps unsurprising that we did not observe performance differences between VR and VC.* Much previous work comparing audio and video interfaces has failed to show a critical impact of visibility on completion time or quality of work [54] and given that audio lacks any visual component, it is arguably a more different interface than the two visual media explored here. Evidence for the benefit of video tend to be for physical manipulation tasks [54], which were not explored here. Even comparing text and face-to-face, Strauss and McGrath [149] showed similar quality output across the two interfaces, but face-to-face was more efficient and participants were more negative towards using text for their judgment task, which is similar to our bribery task, a difference we did not see for the media in this study.

Considering the limited related work on avatars, Bente et al. [23] also did not see notable performance differences between videoconference and avatar based interaction. Jo et al. [78] found that video performed worse than avatars on a job interview task, but for a measure that combined spatial and social presence questions, making it difficult to interpret. *Our contribution to this literature is to show that there are strong behavioral differences between embodied avatars and videoconference, even if performance is similar.*

Conversational Turns: Previous work based on conversational game analysis found that face-to-face interaction was more efficient than audio-only because participants would rely on visual indicators to confirm understanding, rather than requiring verbal feedback, but this efficiency improvement did not hold for video when compared to audio [46]. *A key*

finding of our work is that backchannels are longer and more frequent in VC than VR. This appears to reflect a greater need to align and affirm in the medium, as was seen with audio and VC in the past.

The more formal nature of turn taking in audio and VC has been discussed above. For example, one study found people interrupted each other around twice as much in face-to-face as VC (12.6 to 6.5 interruptions per dialogue) [46]. *Our work shows a less formal interaction style in embodied VR, with more failed interruptions and more overlapping speech. It thus appears more similar to face-to-face.*

Gaze: *Previous work found the same increase in gaze for VC as compared to face-to-face that we observed for VC compared to VR.* A central finding of this paper is that people gaze at each other much more in VC than embodied VR, and this gaze is more focused on the head. Previous work has found higher rates of gaze in VC than face-to-face [118, 46, 27]. The study reported on in O'Malley et al. [118] and Doherty-Sneddon et al. [46] shows that subjects using video gazed at each other 56% more than subjects who were face-to-face. This is exactly the same increase that we saw for gaze in VC vs. embodied VR, suggesting VR is more in line with face-to-face. They refer to the VC behavior as "overgaze," since it represents an unnatural increase over the face-to-face baseline. Interestingly, they found co-present subjects would gaze much more when speaking than listening, but there were no such differences in video [118], whereas we found listeners gazed more at the speaker in both media, as was also reported in recent research on in person conversation [104]. Setlock et al. [138] reported much lower gaze levels in a video-only study that used the desert survival game as a task, our warm up exercise. They found gaze never occurred more than 25% of the time, whereas our mean for VC was 45% for speakers gazing at listeners and 64% for listeners gazing at speakers. It may be that this decrease in gaze occurred because they gave participants a paper printout of the desert survival items which they could place in front of them during the exchange and they felt free to look at. Other differences may also be important, such as the closeness of the screen or changes in use of VC that may have occurred as it became prevalent. Finally, we studied triads and group size has been shown

to impact conversational gaze [104]

Gesture: *Several findings suggest nonverbal behavior may have been more effective in VR, especially when it had a spatial component.* Participants had longer conversational turns in VC for Bribery, which could suggest they were less aware of attempts to take the floor. This effect was reduced but opposite for Floor Plan and Party Planning. Both of these tasks featured a more difficult to operate virtual touch interface, so it is difficult to say if these cases offer counter evidence or the duration is related to people trying to display their points table. Deictic gestures were much more frequent in VR and people were much more likely to replace a word with a reference, likely suggesting that pointing was easier. Previous work has suggested that deictic gestures may increase efficiency of communication [159], but we saw no clear evidence of that. Other forms of unique gesture content were slightly higher in VC, but overall, information that was only available in gesture was much higher in VR (20.6% of gestures compared with 12% for VC). Regulators that manage the conversational turn and are often directed towards a speaker were uncommon, but more likely in VR. Emblematics were also uncommon, but show the opposite trend, being more likely in VC. They generally have no spatial information.

Interpretation: *People were making more effort in VC, in terms of increased prosocial behavior like backchannels and gaze, in order to maintain a similar level of connection as reported on the subjective surveys.* The subjective surveys conducted after each task suggest that participants were experiencing a high level of social connection in both media, and most importantly, there were not notable differences between VR and VC. It is therefore difficult to attribute the observed behavioral changes to differences in social experiences of the media and other explanations must be sought.

To understand what might be underlying the behavioral changes, it is helpful to clarify the differences between our VC and embodied VR manifestations by comparing them on the factors of the decompositional model of copresence developed by Kraut et al. [92] and summarized in [54]. *Field of view* relates to whether people can see what entities other people are oriented towards. This is easier in our VR setup as people can freely adjust their

orientation and other interlocutors can tell what they are oriented towards because they can see the same items. The *spatial perspective* is similar in both media as participants are seated and hence can see their own view, but cannot reorient to share their interlocutors perspective (something that would be easy in VR in general, but we intentionally restricted). *Display symmetry* differs as in VR, interlocutors can all see essentially the same environment, but in VC, much of people's local environment is off camera and not visible to their remote interlocutors. The *dimensionality* is 3D in VR and 2D in VC. *Spatial resolution* is arguably higher in VC as people's facial details are rendered with more fidelity. *Temporal delay* is low and comparable in both. *In short, the critical differences between media arise from users being in a shared, 3D environment in VR, combined with a secondary factor of arguably higher visual fidelity of VC.*

We postulate that the behavioral differences across media arose because people are making a greater effort in VC in order to actively manage the social connection. Gaze performs multiple social functions in conversations, including providing information such as cues for attention and competence, regulating the interaction, expressing intimacy and exercising social control[87]. Higher gaze leads to increased compliance [137] and gaze aversion has shown consistently negative effects [29]. While these findings suggest why people might feel the need to maintain gaze in general, they do not address the differential between VC and VR. *Given that subjective experience was similar, we postulate that these behavioral differences arise from the main difference in the media: that VR places people in a shared, 3D space. People's substantially increased tendency to look at each other in VC could reflect that such visual attention was needed for them to remain engaged, or perhaps more likely, they fear that not showing such attention would be viewed as rude or inattentive.* The gaze differences may arise from people wanting to show that they are in the same environment as their interlocutors, rather than looking at items only they can see in their local environment. This is not an issue in VR as people are de facto in the same virtual space.

Further evidence is consistent with the hypothesis that people felt a greater need to actively maintain the social connection in VC. Backchannels are a social feedback mechanism,

so their increased use in VC suggests effort to manage the connection. This connection maintenance explanation is also consistent with Bailenson's hypothesis for explaining Zoom fatigue whereby he suggested people experience increased cognitive load in Zoom because of increased effort to send and receive nonverbal signals, including attention [13]. Zoom often creates increased intimacy by placing people close to a large image of another's face [13]. Interestingly, we intentionally took a wider viewing angle and greater distance, giving people more latitude to look elsewhere on the screen, but they *chose* to look at each others' face a high proportion of the time. By contrast, people looked at each other less in VR, and when they did, gaze was more distributed between the head, torso and hands. Finally, it is worth noting that people displayed more self adaptors in VC, which are related to anxiety [158], possibly suggesting less comfort, which could actually be triggered by the increased gaze and/or effort to maintain the connection.

It is important to also consider other potential explanations. In trying to explain similar gaze differences, O'Malley et al. "suggest that when speakers are not physically co-present they are less confident in general that they have mutual understanding, even though they can see their interlocutors, and therefore over-compensate by increasing the level of both verbal and nonverbal information" [118]. They also suggest that to gain as much information as they would in face-to-face interaction, more gaze may be necessary, or that people felt their communication was less effective, so compensated with more gaze [46]. These are also plausible explanations for the differences seen with embodied VR, worth considering, although their exact mechanism of action is unclear. They can be contrasted with the much lower gaze rates observed in Setlock et al. [138], which may be more easily explained within the framework of a social connection hypothesis. The gaze reduction could have occurred because both participants were given a paper artifact by the experimenter and it was clearly part of the task, so they did not feel rude looking at it and it felt socially acceptable to reduce gaze. If gaze was needed for understanding, it is less clear why it would have been lower. Explanations of interface novelty that were offered in the past to possibly explain increased gaze in VC [46] seem less likely now that it has become a common form of interaction, and

embodied VR is likely more novel for all participants.

Although we did not observe performance differences, there may be negative consequences of these behavioral changes. O'Malley et al. [118] raise the concern that "overgaze" may increase cognitive load and inhibit verbal processing, since people tend to avoid gaze during periods of increased cognitive load [21]. Bailenson raises concerns about fatigue in videoconferencing [13]. Beattie [22] argues that evidence from face to face studies suggest that gaze may be most effective for turn management when its background level is low, so it may become a less effective cue in VC.

When considering overlapping speech, failed interruptions and gaze, the behavioral differences observed between the media indicate that embodied VR is closer to what has been observed for face-to-face interaction [118, 46] and suggest less formal interaction.

Relying on a definition from Clark and Brennan [37], Fussell and Setlock [54] view copresence as sharing the same physical environment and argue that it is particularly important for collaborative physical tasks, such as repairing an item. *Our work shows that co-presence, as achieved by being in a shared 3D environment in VR, impacts behavioral patterns even when people are not working on such physical manipulation tasks.*

There are also additional factors worth considering. Although people were told their avatar looked like them, they could see the avatars of other people were of moderate fidelity, so this likely provided some sense of anonymity compared with video. This could have influenced their interaction. Facial motion was reduced due to limitations in the avatars, but they did exhibit facial motion, body motion and finger motion that tracked that of their users, likely providing one of the highest quality avatar experiences in a study of this scale. To explore an effort hypothesis, it would also be valuable in follow up work to measure energy expenditure or similar qualities to see if the behavioral changes took a toll on participants.

2.8 Conclusion

This paper discusses a study comparing behavior during remote, multiparty meetings

across two conditions. The first condition was Medium. Participants interacted through one of videoconference or embodied virtual reality. A second, within subject condition was Task type, that included four different tasks that match the types of activities people often perform during meetings: an intellective task (Estimation), a decision making task where groups needed to reach consensus on a problem without a clearly correct answer (Bribery), and two mixed motive negotiation tasks, one with visual grounding of the conversation on a map (Floor Plan) and one without (Party Planning). *Notable behavioral differences were observed. People showed longer conversational turns, more gaze and more backchannels for the Bribery task that required a high level of coordination and subjective discussion. Interaction in VC showed more frequent and longer backchannels, more formal turn taking, much higher rates of gaze, more gaze at the face, more self adaptors and less unique information in gestures, among other differences. In comparison with previous work, the levels of VR behaviors often appear closer to face-to-face interaction. To explain some of these differences, we postulate that people needed to make increased effort in VC to maintain a similar level of social connection.*

We feel that embodied VR has great potential to be a beneficial technology for remote collaborations. We find it encouraging that some of the behavioral patterns observed here are closer to what has been observed in face-to-face interaction than those produced through VC interactions. While much more work is needed to understand the impact of these behavioral differences, and certainly VR is a long way from matching the richness of face-to-face interaction, we hope that this provides early evidence of the potential of VR to support more natural remote interactions. Reducing the environmental impact of travel alone makes this a worthwhile goal.

In closing, it is worth noting that VR is a far more flexible technology than was explored here. For the purpose of this study, people were required to remain seated in our Embodied VR condition to maintain equivalence with VC, as well as replicating a common meeting scenario, but this is in no way a technological requirement. One of the strengths of VR is that people can walk around in their space, they have full control of their viewing direction,

and they can potentially even group up and have side conversations, as people do in real life. There is evidence that sharing a common viewpoint is particularly useful in supporting physical manipulation tasks [54] and we would like to explore the appropriateness of VR for such applications in the future. In addition, it is important to undertake longitudinal studies of extended work collaborations conducted through VR, these would include both much longer individual sessions and work that continued over days or months. VR also offers great potential for visual and aural displays than was investigated here. People could start a meeting around a conference table, transfer to a construction site and then switch to a design studio, all while remaining in the same physical space. Much work is needed to understand how to make best use of these affordances. Finally, it would be interesting in future work to see if any of the findings of this study change when the visual fidelity of VR avatars reaches that of video.

This work would not have been possible without the support of Ronald Mallet, Reena Philip and Elif Albuz. It was informed by useful discussions with Ronit Kassis. Sohail Shafii, Carsten Stoll, Michael Ranieri and David Altman contributed to software development and technical support. Jeremy Schichtel provided support with technology and experiment execution. Aaron Ferguson, Victor Knai and Thierry DiDonna assisted with art assets. Many others contributed to the underlying technologies.

Appendix

Processing Gaze Data

Eye tracking in VR was performed with additional cameras mounted in the headset. For VC, gaze was recorded with a Tobii eye-tracker mounted on a stand in front of participants.

VR Each participant's avatar in the VR condition had colliders assigned to the different avatar body parts. For the FloorPlan and PartyPlanning tasks, the "task objects" (i.e. virtual scoreboard and floor plan on table) were also assigned colliders.

Using logged pose data, we reconstructed each recorded frame from each experiment session. This pose reconstruction sets the eyeball orientation. A sphere-cast is performed from each eye, which sends a sphere along a ray in the view direction and logs all intersecting collider objects and their distance. To confirm integrity of the technique, visual inspections of a sampling of data were performed (e.g. Figure 2.9). During this process, some instances were noticed of gaze rays barely missing the body colliders, especially around the neck area, even though it looked like participants were focusing on those body parts. To reduce this error, we enlarged the colliders between 10 and 20%.

VC Tobii screen captures with associated eye-tracking data were exported for automatic labeling using an external tool developed for this purpose. As additional input for this tool, we first manually defined rectangular regions of interest (ROIs) for each screen capture. These ROIs describe in screen-capture space where each participant's video streams were, and where task-related information was shown. Next, our tool automatically annotated the participant ROIs with body part segmentation masks using the detectron2/densepose network [162, 129]. This mask assigns a label to each pixel with either *background* or with the body part inferred by the network. Body part categories matched those used in VR.

Figure 2.9: Visualization of gaze tracking rays.

[1]. Next, the Tobii eye-tracking coordinates in screen-capture space are associated with each frame and their intersection with body parts masks are determined. To be consistent with the sphere cast technique used in VR labeling, a circle of comparable size was placed at the gaze location and the label was assigned based on a majority vote of the pixels in this circle.

Post-processing: Every gaze sample was labeled with 1) who it was from 2) whether the gaze was at the `Body`, a `Task` object or `Elsewhere`, and 3) body gaze was broken down to `Head, Torso, Hands and Lower` categories. For VR, there are two samples for each pose frame (one for each eye) and the one with smallest depth was selected.

Equalizing Gaze Data Across Media: Because VR gaze is tracked in the 3D scene, data is available wherever the person looks. For VC, however, gaze is only tracked when they look at the screens. No data for VC gaze means either that the participant is looking off-screen or the gaze tracker failed. To understand which event is likely, a sampling of 120 gaps of varied duration with no tracking data were randomly selected and manually coded based on a video review with a 1 if the participant was gazing off-screen and 0 if the participant was looking at the screen, but tracking had failed. Gaps $\leq 0.5s$ generally correspond to blinks and short

[1]Given the amount of data to be processed, and this process being computationally expensive, the body part segmentation masks were only updated every fifth frame.

term tracking failures, so were filled in with the surrounding tracking labels. Gaps of longer than 100s never consisted of participants looking off-screen, so these were considered failed tracking and discarded. Segments in between the two extremes were a mix of participants looking off-screen or tracking failures (often caused by participants shifting in their seats). Longer segments were more likely to be failed tracking, but there were cases of both off screen gaze and failed tracking throughout the range. To account for this, a regression line was fit to the 1 or 0 labels and used to indicate the probability that sample was valid off-screen data (below the line) or failed tracking (above). The equation of the line was $y = 0.8605 + -0.007278x$ where x represents the duration of the segment. Gaze marked off-screen was included in the "Elsewhere" category. While an individual case may be wrong, this provides a statistically reasonable categorization of gaze.

Unlike VC, VR gaze tracking can detect when a person is looking at their own body. A common VR gaze target was the users own hand (9.3% of gaze), which sometimes occurred when they were using a virtual task interface. To further make VR and VC more comparable, we opted to ignore self-hand gaze collisions and instead take the next collision behind the hand. This was a Task target 30-40% of the time, and "Elsewhere" the rest.

Nonverbal Annotation Process

After the study sessions, remote annotators annotated the groups' nonverbal behavior during the floor plan negotiation task using an audio and video feed that displayed all participants. Annotators labeled the type of every gesture performed, with types shown in Table 2.3. In addition, they could indicate if the gesture contained unique information with the options in Table 2.4. Mouse movement in VC was included in the annotation as a substitute for manual gesture (e.g. deixis and spatial gestures were often performed with the mouse).

All annotations for the project were completed by a team of four annotators who saw the videos in random order. Every video was independently annotated by two different annotators. If there were mismatches between their ratings, they were resolved by a third

Gesture Type	Description
Backchannel	Acknowledgments of interlocutor, including head nods and manual gestures.
Beat	Small movements of the hand in rhythm with the spoken prosody.
Deictic	Pointing gesture.
Emblem	Well defined gestures used as a substitute for words, such as a thumbs up.
Glitch (in VR)	Tracking failure generating an unnatural movement.
Regulator	A gesture that passes the conversational floor to the person who should speak next.
Representation	Metaphoric and iconic hand movements, illustrative of an idea (but not fitting in "Spatial or Distance").
Self-adaptor	Self-manipulations not designed to communicate, such as nose scratches.
Spatial or Distance	Gestures conveying more complex spatial or distance information, such as a path through the apartment.

Table 2.3: Annotators applied the most appropriate label to each observed gesture.

Novel Content	Description
Reference Pronoun	Use of pronouns that must be disambiguated by the gesture, such as saying 'this' while pointing.
Unique Information	Information other than a reference pronoun that is only provided through the gesture
Default	Applied if no other option was selected.

Table 2.4: Annotators could apply additional metadata about each gesture.

annotator. The research team performed quality checks throughout the annotation process.

Task Instructions

Estimation

In this task, you will be asked a series of questions where you will need to estimate some real world statistics, for example, what the median age is in the U.S. Your bonus of up to $11 for this task will depend on how many questions you answer correctly. This task will last 15 minutes. After each question, please discuss your answer with the group. When you've agreed, report your answer by saying "Final Answer" and let me know what the answer is. You only have one chance to give an answer for each question. If you discuss a question for more than 3 minutes, I will give you a 15 second warning and then move on to the next question. I will not be answering any questions during the task, so please ask questions now if you are unclear about anything. Any questions?

Ok. Let's do a quick check for understanding.

- How will your bonus be determined for this task?
- Does everyone need to agree with the answer?

Bribery

In this scenario, you are all members of a personnel discipline committee at a major company. An employee we'll call "Salesperson" works in sales. Salesperson's boss we'll call "Boss" and is head of the sales department. Salesperson is one of the top performers in the company, ranking in the top 5% of sales. Last year Boss found out that Salesperson had accepted an expensive trip from one of the company's clients, against company policy. Salesperson apologized, promised to never do it again and pleaded to avoid punishment. Company policy requires Boss to report this offence to his or her boss. Afraid of losing a top producer, Boss did not report Salesperson and let Salesperson off with a warning. A whistle blower has reported the issue and it has been turned over to your committee. In your deliberations, please consider three groups with potentially conflicting interests:

- The sales team, that wants to keep their star.
- The CEO, who is worried about insubordination and the negative impact an ethics violation would have if this reached the press.
- The overall company culture.

You need to decide on four issues:

- The salesperson's punishment.
- Whether the salesperson can keep their job
- The boss's punishment
- Whether the boss can keep their job

Examples of punishment include:

- No penalty
- Stiff warning
- Stiff warning and probation
- Fine equal to the cost of the trip
- Fine double the cost of the trip
- Loss of job

You must collectively decide on the resolution for each issue and you must agree on a justification for each decision. Your bonus of up to $11 will depend on the quality of the plan you develop as a group, as well as how much you contributed to the discussion. Your group must come to an agreement in order to receive a bonus. This task will last 15 minutes and I will give you a 2 minute warning towards the end. Do you have any questions?

Ok. Let's do a quick check for understanding.

- What did Salesperson do wrong?
- What did Boss do wrong?
- What three groups with potentially conflicting interests should you consider?
- What are some example punishments you might consider?
- Are you allowed to consider other punishments?
- What are the four issues you must resolve?

Party Planning

You are each senior managers at a major company. One of you is Head of Finance, one of you is Head of Security, and one of you is Head of Social Events. You have been given the task of planning the end of year party for 500 people and you need to reach a collective decision on each of four issues:

- How many knife jugglers should you hire to perform at the party? Your options are 1, 2, 3 or 4.
- How many security guards should you hire for the party? Your options are 5, 10, 15 or 20.
- How much should you charge employees to bring a guest? Your options are $0, $10, $20 or $30.
- What time should the party end? Your options are 8pm, 9pm, 10pm or 11pm.

Each of you have preferences for each of these items, based on the work unit you represent. We will give you your preferences separately. Your job on this task is to negotiate with each other to reach a final decision on each of these items. Your bonus for this task will be up to $11 and will depend on how many of the total points you receive. Once the task begins,

you will have 15 minutes to negotiate and I will give you a three minute warning towards the end. If you do not reach an agreement, no one will receive any bonus. Do you have any questions?

Ok. Let's do a quick check for understanding.

- What are the four topics you are negotiating on?
- What are your options for the amount of knife jugglers you can have at the party?
- What are your options for the amount of security guards to hire at the party?
- What are your options for the amount for the ticket cost for guests?
- What are your options for what time the party should end?
- What does your bonus depend on?
- What is your bonus if you do not reach agreement?

Floorplan

The three of you have agreed to be roommates. The floorplan shows the apartment you have agreed to rent. Study it closely to understand the different features of the apartment. There are three bedrooms, two with large windows, two bathrooms, a living room and a kitchen. Two bedrooms have large windows, one of the two also has a view. You need to decide who gets each bedroom and which room will be a living room. In addition, there is a closet in the common space that can be assigned to one of the roommates. The total rent is $3,000. You need to decide how to fairly split the rent and best share the space.

As in most situations, each of you value different things. We'll give you a reward table that summarizes what's important to you. Do not share or talk about this table. You should negotiate with each other to decide on the room allocation and what portion of the rent each of you should pay. Try to make convincing arguments. Your potential bonus of up to $11 for this task will depend on how well you meet your goals.

You each have been given your personal reward tables. You should not directly reveal the numbers in this table. You have two minutes to study the information and come up with the best arguments you can for why you should get the features you want. You are not allowed to look at these papers during the task. If you forget the information, you will be able to click on a button to reveal it while negotiating. Any questions?

Ok. Let's do a quick check for understanding.

- What is the total rent of the apartment?
- What are the four issues you be negotiating?
- How is your bonus determined for this task?
- What happens if you do not come to an agreement by the time is up?

Chapter 3

Pedagogical Agents to Support Embodied, Discovery-based Learning

3.1 Introduction

Discovery-based learning is an educational activity paradigm whereby students are led through well-specified experiences that are designed to foster particular insights relevant to curricular objectives. It differs from many of the applications in which pedagogical agents have traditionally been used, in that the knowledge desired for the student is never explicitly stated in the experience. Rather, the child discovers it on her or his own. Testing also differs, as the goal is a deeper conceptual understanding, not easily measured by right or wrong answers. Although discovery-based learning has been a major approach in reform-oriented pedagogy for over a century in classrooms, only recently has begun to be incorporated into interactive technology. The broadest objective of the current paper is to highlight an approach, along with challenges and responses, for building autonomous pedagogical agents for discovery-based interactive learning.

This work explores the application of pedagogical agents to the experiential goal of discovery-based learning. In particular, we add a pedagogical agent to an embodied math

learning environment, MITp, designed to teach children proportion. MITp is described in detail in Sec. 3.2. The basic idea is that a child is encouraged to move two markers on a screen with her fingers. As she does this, the screen changes color. If the height of the two markers is in a particular ratio, say 1:2, the screen will go green, and as the ratio varies away from this it will go to yellow, and then red. The child is never told anything about ratios or proportion or how to make the screen green. Rather she is guided to discover different ways to create a green screen, and by doing so, begins to build an understanding of proportion. This system has been used extensively in learning research with a human tutor guiding students. In this work, we seek to understand what is required to make an animated pedagogical agent effective in this tutoring role.

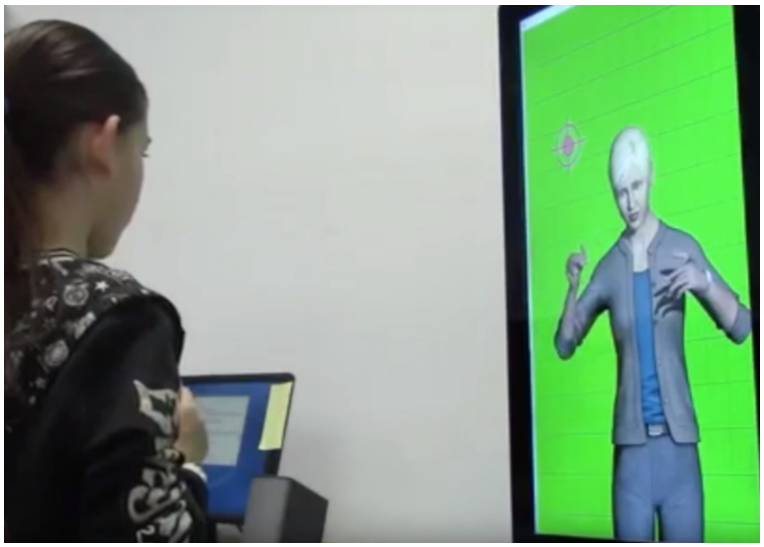

Figure 3.1: A child listens to the pedagogical agent explain concepts within the MITp learning environment.

MITp is an embodied learning experience where the child learns through performing physical movements. Embodied pedagogical agents are particularly useful in this setting because of the engagement they engender and, most importantly, because they can enact virtual actions and gestures they wish the learner to perform.

This type of learning application creates unique computational challenges. Chief among them, it is very difficult to measure the student's progress when it is not possible to ask questions with right or wrong answers that are easy for a computer to grade. Our design process employed a deep analysis of the process used by human tutors, including reviewing many hours of video recorded interactions. We identified the key stages in the tutorial process, the types of actions tutors took and when they took them. From this analysis, we identified the following activity types the agent must engage in: *instructing* the child what to do; *valorizing* success; *waiting* so the child can explore on her own; *providing remedial training* when the child is blocked and *advancing* the child through the tutorial process.

While the human tutor sits beside the learner, we placed our animated agent on the other side of the screen from the child, with access to the same touch surface as the learner (Figure 3.1). Dubbed *Maria*, our agent can execute a sequence of action blocks. Each block consists of any subset of spoken audio, facial animation, lip syncing and body animation, including arm and finger gestures. There are well over 100 actions that the agent can perform. We use Dynamic Decision Networks to decide when the agent should perform an action block and which action to perform.

We have conducted a pilot evaluation of the system. It showed that students were able to effectively explore the screen to find greens and enact a particular search strategy taught by the system. Students provided proto-mathematical descriptions of one solution strategy, although not the other. The results demonstrate good progress and also illuminate potential directions for future work. In essence, this chapter presents a pedagogical agent designed to support students in an embodied, discovery-based learning environment. Discovery-based learning guides students through a set of activities designed to foster particular insights. In this case, the animated agent explains how to use the Mathematical Imagery Trainer for Proportionality, provides performance feedback, leads students to have different experiences and provides remedial instruction when required. It is a challenging task for agent technology as the amount of concrete feedback from the learner is very limited, here restricted to the location of two markers on the screen. A Dynamic Decision Network is used to automat-

ically determine agent behavior, based on a deep understanding of the tutorial protocol. A pilot evaluation showed that all participants developed movement schemes supporting proto-proportional reasoning. They were able to provide verbal proto-proportional expressions for one of the taught strategies, but not the other.

3.2 MITp Learning Environment and Tutorial Protocol

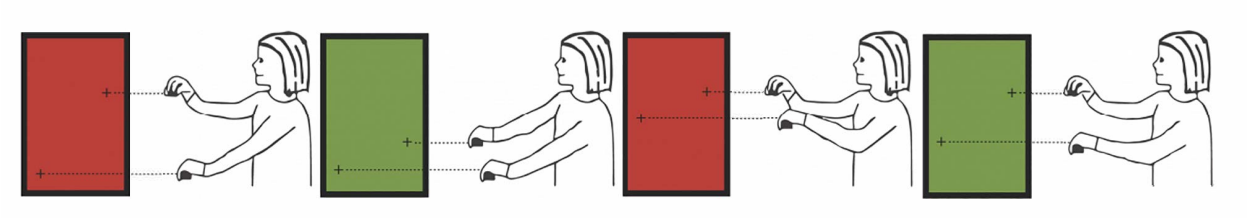

Figure 3.2: The MITp environment. The screen is green when the hands' heights match a pre-programmed ratio.

The current study builds upon an earlier educational-research effort to design and evaluate embodied-interaction technology for mathematics instruction. Specifically, the pedagogical agent described herein was integrated into an existing activity design architecture called the Mathematical Imagery Trainer for Proportionality (MITp) [3, 2, 68], which we now present.

Proportional reasoning is important yet difficult for many students. It involves understanding multiplicative part-whole relations between rational quantities; a change in one quantity is always accompanied by a change in the other, and these changes are related by a constant multiplier [28, 96, 127].

Our MITp approach to support students in developing multiplicative understanding of proportions draws on embodiment theory, which views the mind as extending dynamically

through the body into the natural-cultural ecology. Thus human reasoning emerges through, and is expressed as, situated sensorimotor interactions [8, 95]. Educational researchers informed by these theories have created technologies to foster content learning through embodied interaction (e.g., [11, 49]).

The MITp system (Figure 3.2) poses the physical challenge of moving two hands on a touch screen to make it green, a result which occurs when the ratio of hand heights matches the pre-programmed ratio of 1:2. Through this process, students can develop presymbolic mathematical understanding by engaging in this embodied activity and building particular movement schemes related to proportions. By introducing specific tools into the environment, here a grid and numbers, students are given progressively more mathematical tools with which to express those strategies.

The MITp system has been extensively tested for its educational effectiveness. Using qualitative analyses, the researchers demonstrated the variety of manipulation strategies students developed as their means of accomplishing the task objective of moving their hands while keeping the screen green [68]. Moreover, it was shown that students engaged in deep mathematical reflection as they were guided to compare across the strategies [2]. The studies have presented empirical data of students shifting from naive manipulation to mathematical reasoning as they engage the frames of reference introduced into the problem space and the tutor's critical role in facilitating this shift [3].

Interview Protocol

Maria is programmed to lead students through a series of activities on the MITp touchscreen, each supporting the development of particular movement strategies deemed relevant to proportional reasoning. Broadly, the two main phases are exploration, targeting the strategy "Higher-Bigger," and "a-per-b." In "Higher-Bigger," participants meet Maria on a screen with a red background, which is later overlaid with a grid. Participants are instructed to move the cursors up and down to make the screen green. At each green, Maria valorizes

their work and asks them to make another green, either higher, lower, or elsewhere. Other than moving the cursors up and down, participants receive little guidance on particular movement strategies. With time, the grid is overlaid on the screen. The goal of this stage is for students to notice that when they make a green higher on the screen, the gap between their hands is bigger ("Higher-Bigger"). Next, in the "a-per-b" phase, participants are given instructions to start at the bottom of the screen, move their left hand up one grid unit, and then place their right hand to make green. Finally, the grid is supplemented with numerals. Participants are periodically asked to reflect on their rule for making green. Though Maria does not (yet) recognize speech, these reflections promote verbal description through which the developers can assess the participants' proto-proportional understanding. Participants took about 20 minutes on average to complete the task.

While participants interact with Maria, the presiding human interviewers try to minimize human-to-human interaction, albeit occasionally they respond to participant queries, confusion, or frustration.

3.3 Related Work on Pedagogical Agents

Our work finds its roots in previous work on Pedagogical agents and Intelligent Tutoring Systems. Intelligent tutoring systems are computer softwares designed to simulate a human tutor. Pedagogical agents aid the process by adding a human-like character to the learning process. Research over the past few decades [79] has validated the positive impact of having an embodied presence in virtual learning environment. They have been a success primarily because they add emotional and non-verbal feedback to the learning environment [80]. More expressive pedagogical agents tend to improve the learning experience [63].

Intelligent tutoring systems (ITS) have been developed for a wide range of topics. Cognitive Tutors [88] have been adapted to teach students mathematical and other scientific concepts like genetics. The Andes Physics tutor [154] focuses on helping students in introductory Physics courses at college level. Writing Pal [130] and iStart [75] help students in

developing writing and reading strategies respectively. Decision theoretic tutoring systems have also been very successful and range from generic frameworks like DT Tutor [112] to domain specific systems such as Leibniz [55]. The feedback and learning mechanism behind all these activities revolves around the tutor provided instructions or some sort of rule specification, followed by student responses, given as either text or multiple choice selections, to posed challenge questions. Our discovery based learning methodology differs fundamentally as the system never describes how to achieve the desired goal, and the student response has to be gauged in real-time based solely on the touch screen coordinates. There are not questions that can be used to directly gauge progress.

Pedagogical agents can interact with the student in various roles, such as interactive demonstrators, virtual teammates and teachers. Steve [81] is an early example of a demonstration based pedagogical agent to train people to operate ship engines. INOTS [32] teaches communication and leadership skills to naval officers. The agent is questioned about a case by officers during training, and their performance is evaluated by rest of the class watching this interaction. AutoTutor [56] is a modern system used to teach concepts in Science and Mathematics. The student agent works with the human student to solve the problems in different ways. Adele [139] and Herman the Bug [97] are two classic pedagogical agents designed to teach medicine and botanical anatomy respectively. Decision Networks have also previously been used in the development of adaptive pedagogical agents such as [38] and [128] and for narrative planning [111]. These decision theoretic agents work on concrete feedback from the user in form of biological signals or responses to questions. In fact, all the above mentioned agents use the standard teach and test framework in order to gauge student's performance, allowing them to focus on specific concepts in the learning that the student is struggling with. Our pedagogical agent operates in a quite different, discovery-based learning paradigm. Some previous pedagogical agents have also targeted more open-ended learning environments. For example, a system designed for children with ASD allows children to interact with the system by telling stories, control the agent by selecting pre-defined responses or author new responses to create new stories [150]. Related work has sought to use agents

not as instructors, but as virtual peers [86, 51].

3.4 System

System Overview and Architecture

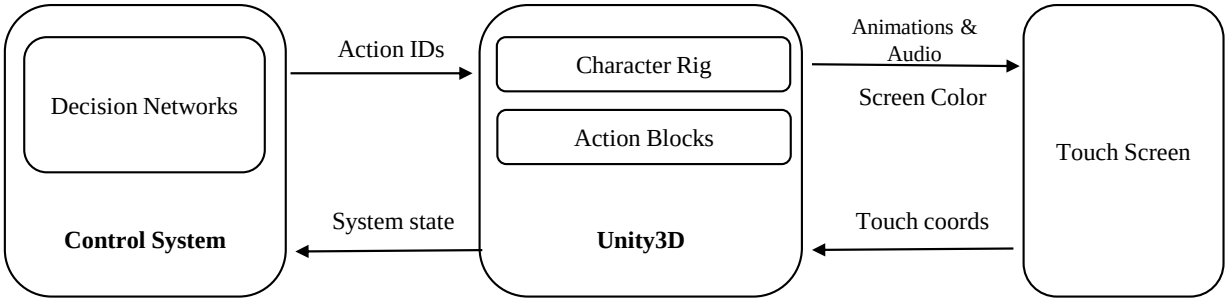

Figure 3.3: System Architecture

An overview of our MITp autonomous agent system architecture is shown in Figure 3.3. It consists of a control system and Unity3D front end. Students interact with our agent using a touch screen. The screen is virtually divided into left and right halves around the agent, designating two large tracks where the learner can move the markers. Maria is standing in the middle as shown in Figure 3.1 and can reach most of the screen. The Unity client sends the system state consisting of the two touch locations to the control system, which then instructs the agent to perform particular actions by specifying Action IDs.

The control system employs dynamic decision networks [42] to model the behavior of our pedagogical agent. The decision networks are updated based on the evidence received from Unity and history maintained throughout the interaction. They may decide to do nothing or have the agent perform one of many pre-designed actions, depending on what appears most efficacious in the current context. Triggered actions are sent to Unity as action block

IDs. Each action block consists of an audio file, a facial animation and a body animation. Our database contains 115 different action blocks.

Generating meaningfully labelled data for learning agent behavior is a challenge in a discovery-based setting. At this point, we do not make assumptions about student's learning during interaction and instead rely on a robust tutorial process which can adapt and provide remedial instruction if the student is struggling. Modeling student's learning on the fly requires mapping from patterns of finger movement to their mental state, which remains future work. Due to these issues, learning based techniques such as [74] are not a good fit for the current problem. DDNs allow us to leverage our strong understanding of the tutorial process by pre-encoding it into system parameters.

Decision Networks

Decision networks find their roots in *Bayesian Networks* [123], which are graphical models consisting of chance nodes representing a set of random variables. Random variables are events that could possibly occur in the given world. Chance nodes, drawn as ovals in the graph, can take any type of discrete value, such as a boolean or an integer. These values are generally finite and mutually exclusive. Arcs between these nodes capture conditional dependencies. Cycles are not allowed. Given the conditional probabilities, prior and evidence, inferences can be made about any random variable in the network.

Decision networks [67] extend Bayesian Networks by using Bayesian inference to determine a decision that maximizes expected utility. Decision networks add decision and utility nodes, represented by a rectangle and diamond respectively. Decision nodes represent all choices the system can make while the utility node captures our preferences under different factors impacting the decision. This concept can be further extended to *Dynamic Decision Networks* (DDNs) [42]. DDNs consist of time varying attributes where there are conditional dependencies between nodes at different time steps. Decisions made in previous time steps can impact the probability distribution or state of the network in future steps. DDNs provide

a useful way of modeling evolving beliefs about the world and changing user preferences.

Building Decision Networks for Experiential Learning

Discovery-based learning is challenging compared to other learning settings because the pedagogical agent must make decisions with very impoverished information as there is no continuous stream of concrete verbal or q&a based feedback that the agent can use to assess the student's understanding. For example, in our case, we only have student's touch coordinates on the screen as input. Our agent guides the learner through the discovery process using this input and a deep understanding of the tutorial process, encoded in the DDNs. Per the tutorial analysis, the agent must *instruct*, *valorize*, *wait*, *provide remedial training* and *advance* the child to the next stage. To decide if the agent should *wait*, the child must have a notion of the passage of time, which can be combined with the child's activity pattern to determine if allowing time for exploration is appropriate. *Remedial* actions are triggered when the child is not executing desired movement patterns after repeated instructions and ample trial time. An example remedial strategy involves the agent displaying a marked location for the child's one hand and then encouraging them to find green by only moving the other hand. Based on their performance, the agent may ask students to repeat activities to deepen their learning.

The learning experience progresses through multiple activities tied to particular interaction strategies: "Higher-Bigger" exploration without and with the grid, "a-per-b" without and with numbers, and an optional "speed" activity. All the activities are modeled using dynamic decision networks. Space limitations preclude discussion of each, but we will explain our approach using the "Higher-Bigger" task as an exemplar.

Exploring "Higher-Bigger"

is the introductory activity students go through, as outlined in Sec. 3.2. The activity guides the student to find greens at several locations on the screen, with the goal of having the

child realize that the separation between her hands is larger for greens higher on the screen than greens that are lower. Figure 3.4 and 3.5 show the decision network which governs the behavior of our pedagogical agent for this activity at two different points during an interaction.

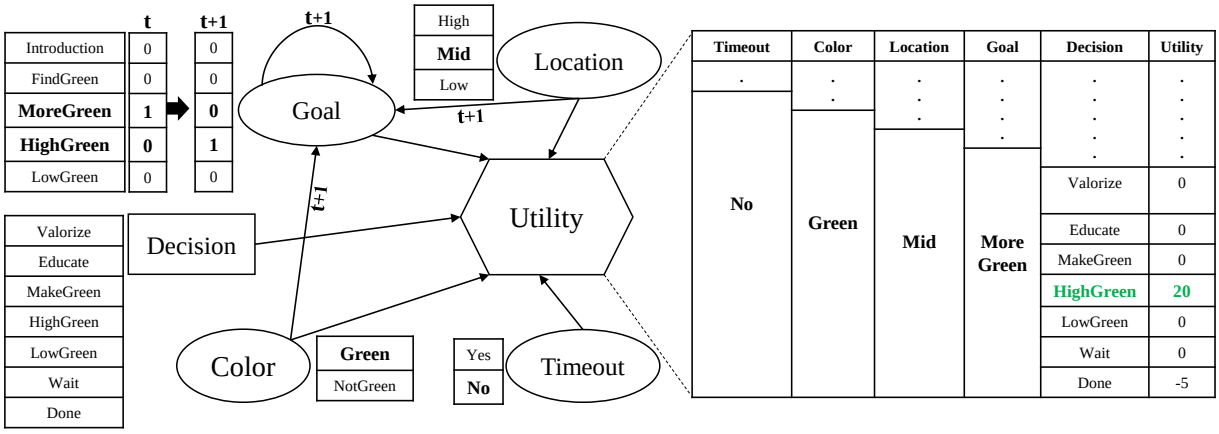

Figure 3.4: Decision Network for the exploration task at time step t. This figure shows an example query and update mechanism of our decision network. The arcs labeled t+1 indicate that the state of Goal node at time $t+1$ depends on the state of Location, Color and Goal at time t. At time t, the agent's goal is for the student to find *MoreGreen* on screen. The student finds green somewhere in the middle of the screen, and the evidence for nodes Location, Color and Timeout (bold words) is set. We now query the network to give us a decision with maximum utility given the circumstances represented by the network state. As shown in green, the agent decides to guide the student to the new task of finding green in the upper portion of the screen.

The decision network is updated multiple times each second to ensure real time responses. The network encodes our agent's belief about the state of tutorial process and student's interaction at each time step. Factors impacting the agent's decision for the network shown in Fig. 3.4 and 3.5 are:

Goal : Models agent's temporally evolving expectations for the student. During this activity the student is first expected to find a couple of greens anywhere on the screen, followed by greens in specific portions of the screen. Possible node states are shown in the network

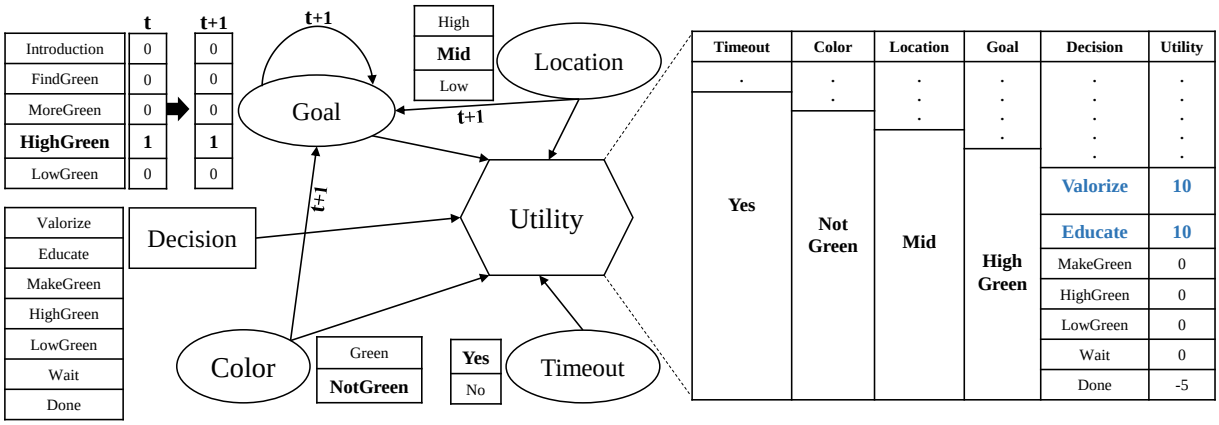

Figure 3.5: At this point, the child is either actively exploring, but failing to satisfy the goal, or has stopped interacting with the system. The agent queries the network for the optimal decision to make after setting evidence for Timeout, screen Color and current exploration Location on the screen (bold words). The decision network suggests Valorization and Educating the child about the task as actions of equal and optimal utility (shown in blue). We choose either action randomly. Remedial activities override the decision network and are triggered if there have been multiple continuous timeouts and student hasn't been able to achieve a goal for an extended time.

above.

Timeout : Models the time elapsed since important events. It includes time since the last agent action, time since last touch by the student, time since the last green and time since last achieved goal. These factors, both individually and in combination, are critical in behavioral modeling when the student is struggling in finding patterns on the screen.

Location : Captures the portion of the screen that the student is currently exploring, discretized into {*High, Medium, Low, NA*}.

Screen Color : Models the current screen background color as a binary node that can be either be green or non-green.

Decision : Decision node that contains all the high level decisions the agent can take. For this activity, the agent can choose to instruct the student about the current task, valorize them, prompt them to explore different areas for green, provide location specifications for

finding a green or stay quiet to give the student time to think and explore.

At each time step, the network is updated with available evidence and provides a decision with maximum expected utility. Directed arcs in the network above show conditional dependencies. Those labeled as 't+1' show temporal dependencies, such that the state of node *Goal* at time t+1 depends on the goal, screen color and finger touch coordinates of the user at time t.

3.5 Evaluation

This section describes our first explorative experimental evaluation, methods, results and discusses implications for future pedagogical agent technology.

Experiment Description

The agent followed the protocol outlined in Sec. 3.2.

Participants included 10 children (5 male, 5 female) aged 9 - 12 years old.

Data Gathering: As participants worked with Maria, three sets of data were collected simultaneously. One researcher used an objective-observation instrument, detailed below, to note the occurrence and frequency of key participant movements and expressions. A second researcher took qualitative field notes of the participant's interaction with Maria. Finally, the interview session was video and audio recorded. The qualitative field notes and video records were used to verify key moments indicated in the observation instruments.

Note that these data sources are consistent with qualitative methods. As this system represents a new genre of agent-facilitated, discovery-based learning, we first seek to understand the types of interactions and insights that emerge as children work with the system. Such qualitative work is not consistent with statistical analyses or pre/post comparison, which pre-suppose which ideas will emerge as salient for participants and considers only those insights that can be quantitatively evaluated. With a qualitative understanding in hand, later

work can focus the system on insights deemed most mathematically productive, at which stage a quantitative evaluation would be appropriate. This progression is well established within grounded theory [40] and design based research [10].

Observation Instrument: A one-sheet observation instrument was used during each interview to code, in real time, when each participant exhibited particular benchmark movements and expressions. These movements and expressions were determined in consultation with researchers familiar with analogous, human-run interviews. For example, a participant expression of "the higher I go on the screen, the bigger the distance must be between my hands to make green" is a proto-proportional expression that reflects the changing additive difference between numerator and denominator. Additionally, a movement scheme of raising the right hand 2 units for every 1 unit increase in the left hand reflects the constant multiplicative relationship between equivalent ratio (e.g., 4:8 and 5:10). These two strategies are termed "Higher-Bigger" and "a-per-b," respectively.

Results

Subject	Greens by Screen Region			"a-per-b"
	Low	Middle	High	
1	14	4	5	1
2	12	7	7	8
3	6	9	3	4
4	10	16	7	1
5	11	4	6	7
6	11	15	7	9
7	9	11	5	4
8	7	5	5	5
9	8	15	8	5
10	2	7	10	4

Frequency of greens, by region, and "a-per-b" performance.

"Higher-Bigger"	"a-per-b"	Other Insights
0	1	1
0	1	3
0	2	2
0	0	5
0	2	1
0	0	4
0	2	2
0	1	3
0	2	2
0	0	3

Participant expressions by category.

The observation instrument was designed to measure study participants' physical movements and/or verbal and gestural utterances that would imply they are engaging in "the higher, the bigger" or "a-per-b" strategies as their means of solving the bimanual manipulation problem. Performance of a particular movement pattern suggests that the participant is enacting a perceptuo-motor strategy that could result in conceptual learning. Verbal description of those movements in proto-mathematical terms indicates further progress along that learning pathway per our instructional design.

As indicated in Table 3.5, all participants produced green low down, in the middle, and high up on the screen. Interestingly, no participant described the changing distance between their hands at these various regions (Table 3.5). Additionally, all participants performed

the "a-per-b" strategy (Table 3.5 indicates the number of times this was observed), and 7 participants verbally described it, such as "for every 1 my left goes, my right goes 2." Notably, all participants developed other insights into the system's functioning, which fell into 1 of 4 categories: observational ("Generally, my right hand has to be on top"), feedback-based ("If you see the red screen flash [to green], move back to where you were"), memorization ("4 and 7 or 8 will make green. 6 and 11 or 12 will make green. 2 and 3 makes green."), and procedural ("Keep one hand in one spot and move your other hand around").

Unanticipated was that researchers were obliged to interact with participants on average twice per interview. In 7 of the 10 interviews, this interaction involved researchers restating the "a-per-b" instructions with similar phrasing to Maria's.

Discussion

We are encouraged by the widely adopted movement strategies, both for making green in all regions of the screen and in adopting the "a-per-b" strategy. Instructed by Maria, and largely independent of human intervention, all participants developed movement schemes supporting proto-proportional reasoning.

This work surfaced a gap, however, between participants' performed movements and their descriptions thereof. Participants adeptly made green in all regions of the screen and performed the "a-per-b" strategy, yet they did not develop proto-mathematical descriptions of "Higher-Bigger," only of "a-per-b." Comparing the conditions of "a-per-b" work with the conditions for "Higher-Bigger" work suggests sources of this disparity. In the "a-per-b" phase, participants received verbal instructions for the target movement strategy as well as a visual grid and numerals. The verbal instructions highlighted discrete, alternative hand movements (laying the "_-per-_" foundation) while the grid and numerals drew attention to unit quantities (supporting specifically "*1*-per-*2*"). In contrast, instructions for the "Higher-Bigger" phase did not explicitly mention the distance between participants' hands. Additionally, students were not instructed to follow a particular green progression, for example

making green low down, in the middle, and up high, that would facilitate noticing a growing distance. Future efforts should focus on understanding how to embed in the pedagogical agent's models and actions the nuances of human-tutor actions that have led students to attend to the interval between their hands.

Patterns in researcher intervention suggest another area for design iteration. Researchers consistently interacted with participants by repeating the "*a*-per-*b*" instructions after participants worked unsuccessfully for 3 or more minutes. During this time, participants developed a host of less effective movement patterns - alternating left and right but moving each 1 unit or raising the left hand to the top of the screen then raising the right hand. Though Maria repeated fragments of her original instruction, the timing and particular fragment selected often did not correct the participant's movement. And while researchers tried to mimic Maria's exact instructional language, their choice of *when* to give those instructions and *which words to emphasize* gave more information than Maria is programmed to do. Further work is required to analyze these researcher interventions and convert their decisions into procedures for the autonomous agent.

Overall, we find this work to tentatively support the added value of a virtual agent in discovery-based learning environments. The agent provides feedback on student work and suggests corrective or novel movement strategies that would likely not arise in agent-free work. In particular, the agent draws the student's attention to multiple parametric regions of the problem space, such as particular spatial locations on the monitor, that had not occurred to the student in their free exploration, and the agent suggests new spatial-temporal interaction schemes, such as introducing a sequential bimanual manipulation regime where the student was trying only simultaneous actions. The agent also provides encouragement, validating the students' efforts and encouraging them to explore in new ways. As none of the participants noticed the "Higher-Bigger" relationship, this first prototype was somewhat less successful than human tutors. However, this is not surprising. Human tutors perceive a wider range of student behaviors (posture, oral expressions, facial expressions) contemporaneous with their on-screen actions, giving more information upon which to determine the

pacing and content of guidance. Additionally, human tutors enjoy the full range of their gestural and verbal vocabularies in responding to and guiding participants. Consequently, we did not expect that the virtual agent would perform to the same benchmark as human tutors. Nevertheless, we see the results of this work as a success, then, in that all participants performed, and almost all expressed, the proportional "a-per-b" relationship under the virtual agent's guidance.

3.6 Conclusion

The MITp system presents a very challenging application for pedagogical agents as they must determine appropriate actions based on very little feedback from the learner, in this case, only the location of two markers on the screen. Such constraints are typical of discovery-based learning, where the asking of concrete questions is limited, and the learner is given freedom to explore. The system performed quite well on the task, leading to appropriate movement patterns in all cases and desired verbal expressions for one of the two movement strategies taught. This suggests that the potential for pedagogical agents in discovery-based learning is high and that DDNs represent an effective control strategy.

The system was effective due to a very thorough understanding of the tutoring protocol that was then encoded in the DDNs. In our case, this was based on an analysis of human-led tutoring of the same task. Two significant shortcomings were noted in the study, students failed to verbally explain the "Higher-Bigger" pattern and some amount of human intervention was required, generally when students failed to progress later in the last task. Both of these suggest the need to further refine the protocol encoded in the DDNs. Future work should consider using verbal input from the learners.

Chapter 4

Sign Recognition in Isolation and in Context for American Sign Language

4.1 Introduction

Approximately 70 million people across the globe, and half-a-million in the United States, use sign language as their primary form of communication. For majority deaf communities in United States and Canada, ASL is considered the mother tongue. Studies have also reported ASL as the fourth most used language in the United States[109]. The inherent difficulties in spoken language acquisition for the the deaf community lead to disadvantages when they are restricted to word-based communication media, as for example, they have below par reading comprehension [152].

To improve the sign language community's access to modern technological infrastructure, systems need to be developed that support communication in sign. Research in the domain of Sign Language Recognition, Translation (SLT) and Production (SLP) is designed to bridge this communication gap for deaf communities. SLR is further categorized into word-level recognition (WSLR), also known as isolated (ISLR) sign recognition, and sentence-level recognition, also known as continuous sign recognition. In this paper we focus on the ISLR

problem for publicly available modern day ASL datasets, namely WLASL [99] and MSASL [82]. We further investigate how trained ISLR models can be used to create weakly supervised training datasets for the sign spotting task on a recently introduced How2Sign[48] dataset. Sign spotting is a localization problem of detecting isolated signs in continuous videos where longer sentences are signed. How2Sign[48] is the first continuous sign language dataset of its scale for ASL.

While related to the action recognition problem, SLR presents unique challenges. Signs are performed in a limited physical space around the signer, and motion is generally limited to subtle upper body and finger motion. In this confined space, simply changing the placement of hands while performing an otherwise identical gesture could alter the meaning of a sign. Stokoe [147] and Battison [20] quantified the structural and temporal properties of sign language. Principal parameters that define a sign according to their work are hand motion path, hand pose shape, palm orientation choice and area of the body near/around which the sign is performed, a.k.a the point of interest. Unlike SLR, finger motion is generally not an important source of information for everyday human action recognition. There are also non-manual properties of a sign, such as movement of head, eye brows, eyes and other facial movement. Capturing and learning all these properties is a challenging computer vision task. Sign spotting in continuous videos introduces additional complexities of co-articulation, contextual dependencies and signing speeds. Sign spotting can help expand ISLR datasets by localizing samples from continuous in the wild videos. It can also be used to search sign language videos with video based action queries, and potentially in the future could be used to help train signers.

Recognition approaches can be broadly split into RGB-D image-based [167, 34, 98] and skeleton-based [156, 76, 142, 99, 82]. Most recent work in action recognition, and specifically for SLR, proposes a late fusion, ensemble based approach combining image-based and skeletal modalities. Image-based approaches use pixel information directly from video frames, whereas skeleton-based approaches expect body keypoints and skeleton connectivity information [163, 156, 76]. Skeleton based approaches have become popular due to the availability

of skeleton-pose recovery software [33, 141, 39] that does not require motion capture hardware. These pose extractors work well at high frame rates and resolutions, but struggle with artifacts such as hand motion blur that occur in ASL datasets [99, 82] that rely on old, lower quality videos.

In this paper, we focus on the challenging WLASL and MSASL datasets and achieve state of the art ISLR accuracy for both. We introduce 3D keypoint pose data, along with 2D joints and bones, and evaluate the utility of each modality. With 3D pose estimation we recover a view independent posture that can help in cases where signers are not consistently aligned with the camera. Finger data is reconstructed using a state of the art hand pose recovery method [168] to achieve a 3D hand pose directly from video frames. Upper body 3D pose complementing the 3D hands is recovered using using XNect [108]. To aid future research, we will make the 3D reconstructed data publicly available.

To summarize, Isolated Sign Language Recognition (ISLR) is an important constituent task in developing SLR systems in which videos of individual, word-level signs are correctly identified. We develop a novel model for ISLR based on a unique G3D-Attend module that further uses spatial, temporal and channel self-attentions to contextualize aggregated spatial and temporal dependencies. The approach is tested on two large, public datasets of labeled, isolated sign videos of American Sign Language (ASL), WLASL and MSASL. We augment the datasets with 2D and 3D skeleton data, which is used along with RGB data in an ensemble-based approach to achieve state-of-the-art recognition rates. We then extend the approach using SVMs to create weak sign spotting labels. Spotting involves identifying individual signs in multi-sign sentences and is a significantly more difficult task due to co-articulation effects, differences in signing speeds and the influence of contextual information on sign production. Our sign spotting approach is tested using the How2Sign video dataset. Generating sign labels manually is a laborious task, and this approach can provide a set of labels that can then be used to provide weak supervision for future machine learning applications.

4.2 Previous Work

Research in sign language recognition has advanced with the availability of large sign language datasets and consistent improvements in deep learning approaches for action recognition. In this section we discuss the literature on ASL datasets, sign recognition and spotting.

ASL Datasets

Sign language datasets can be categorized into continuous (sentence level) and isolated (word level) datasets. Continuous datasets [48, 52, 5] contain signs for complete sentences and these are mostly different from a concatenation of isolated signs. It is hard to localize individual words in continuous signs, because of co-articulation effects. The starting point of a word sign in a sentence might be dependent on the preceding word and the end-point be dependent on the next one. NCSLGR [113], RWTH-Boston-104/400 [166] and How2Sign [48] are the three publicly available, continuous ASL datasets with How2Sign [48] being the most thorough with a 16k vocabulary size, 11 signers and 6 different data modalities.

There are five publicly released word-level sign language datasets. RWTH-BOSTON-50 [165] was published in 2005 with 483 samples across 50 different glosses (written approximation of a sign) and 3 signers. The dataset was captured under a controlled environment in a lab, however 75% of the videos are grayscale and the dataset was limited in terms of number of samples and gloss. Purdue RVL-SLLL [160], published in 2006, is a unique isolated sign database that is broken down into different components of the sign instead of a complete sign corresponding to verbal English vocabulary. Boston-ASLLVD [12], published in 2008 had a bigger vocabulary size of 3300 glosses with a total of 9800 samples by 6 signers [6]. With less than 3 samples per gloss on average the dataset is not well suited for state of the art deep learning architectures.

Almost a decade later, MSASL [82] (2018) and WLASL [99] (2020) were published. Both these datasets consist of short videos available publicly online. The dataset comes as a list of web links to rgb videos (e.g. youtube), associated gloss and signer ids. This leads to its own

issues as will be discussed in Section 4.3. MSASL consists of 25,513 videos across 1000 glosses by 222 signers. WLASL has 21,083 videos for 2000 glosses, signed by 119 signers. These datasets are much better suited to deep learning based frameworks. Both these datasets also include most publicly available videos from previous datasets. It makes these datasets a super set for their predecessors, however the video quality within these datasets is not consistent. Some videos were captured back in late 90s and early 2000s when consumer grade cameras were lower quality.

SLR Algorithms

SLR algorithms can broadly be categorized into two broad categories: (a) RGB-D image based approaches, that take in 2D images as input, with or without a depth map, and (b) Skeleton-based methods that take skeleton joint keypoints and bone connectivity information as input.

RGB-D Based Approaches

Earlier work in the domain extracted hand-crafted features from images, followed by some temporal modeling leading to classification or recognition. The Scale Invariant Feature Transform (SIFT)[126] and Histogram of Oriented Gradients (HOG)[103] have been leading choices for initial feature extraction for visual recognition tasks. These features are then fed into Support Vector Machines (SVMs) or a temporal sequence model like Hidden Markov Models (HMMs)[83] for classification purposes. This methodology is excessively reliant on the robustness of the initial set of features, hence researchers started working with segmented hand regions and hand motion trajectories for creating the feature set. Later, 2D Convolutional Neural Networks (CNNs)[89] generating spatial dependencies, Recurrent Neural Networks (RNNs) modeling temporal connections and 3D CNNs [24] simultaneously representing spatio-temporal dependencies were learned for the task. LSTMs[90] and GRUs, being more efficient RNNs for modeling long-term temporal dependencies, also saw some suc-

cess. The most widely accepted architecture out of these has been 3D-CNN based inflated-3DCNN (I3D) [34], further optimized in the separable 3D-CNN (S3D) [167]. It has been used as a baseline performance metric and key ensemble component for recent action and sign language recognition models [36, 102, 163]. ResNet(2+1)D [151] has recently seen some success too.

Skeleton Based Approaches

Skeleton-based approaches try to recognize gestures, signs and actions based on skeleton keypoint data. These approaches require joint or bone data along with skeleton hierarchy information to capture their spatial layout. RNNs, LSTMs and CNNs were early deep learning approaches applied to human motion captured data for recognition tasks[71]. These approaches capture the temporal trajectory of individual keypoints, but are unable to capture the spatial structure and jointly optimize based on the spatio-temporal relationship between keypoints. PoTion [36] tried to address this issue by stacking joint trajectories on top of each other to create a spatio-temporal map. This issue was effectively addressed when Graph Convolution Networks (GCNs) were adapted into Spatial Temporal Graph Convolutional Networks (STGCNs) by Yan [163]. Although STGCNs work well for human action recognition tasks, they performed rather poorly on SLR tasks [7]. Their limitation seemed to be a fixed skeleton graph that only captures local dependencies between joints and misses out on implicit joint/keypoint correlations. These problems were addressed in follow up works AS-GCNs [100], 2s-AGCNs [140] and MS-G3D [102]. In order, these approaches capture richer dependencies by using an encoder-decoder structure, a two stream framework to capture first and second order information together and using a multi-scale aggregation scheme to disentangle the importance of nodes in specific neighborhoods. Manuel [156] adapted [102] to SLR and SAM-SLR [76] used a similar late ensemble fusion based approach for the problem.

Sign Spotting

Sign spotting has not been studied on ASL, largely because of the lack of large-scale continuous datasets along with gloss annotations in video clips. The recent release of How2Sign[48] provides a large continuous ASL dataset, but it does not provide word level sign segmentations. Li [98] reported a single result on a self gathered and annotated news dataset, that is not available anymore. Most of the sign spotting work has been done on British Sign Language (BSL)[5, 155, 110] and German Sign Language(DGS)[77]. Both these sign languages have large datasets with word level annotations available on continuous sign datasets. Isolated glosses in BSL Sign Bank dataset [50] for example, can match against annotations on continuous signs in BSLCORPUS[134]. Similarly, RWTH-Phoenix-Weather dataset [31] for DGS has spotting annotations. For BSL, Albanie [4] introduced automatic annotation for co-articulated signs using mouthing cues. However mouthing and non-manual features are very region dependent. We observe limited mouthing in ASL datasets [99, 82, 48] and literature has shown limited mouthing in native ASL speakers[121, 17]. This paper employs ASL datasets to create weak labels for sign spotting on continuous How2Sign dataset clips using ISLR models.

4.3 WLASL and MSASL Datasets

Although MSASL [82] came out earlier, WLASL [99] was published without having access to MSASL dataset. As a result, there is considerable overlap between the two with 82.4% of MSASL glosses/words/classes also being a part of WLASL. It may be worthwhile to merge the two datasets in future, and come up with a larger and richer dataset, however for the scope of this paper we keep them as-is. Both datasets feature a single signer, performing a single sign, per video sample with upper body clearly visible within the frame.

MSASL is further split into four subsets containing 100, 200, 500 and 1000 glosses. MSASL100 contains the most frequent 100 glosses only, MSASL200 has the most frequent

200 and so on. Similarly WLASL is split into WLASL100, WLASL300, WLASL1000 and WLASL2000. Splitting into smaller datasets makes evaluating new algorithms on representative subsets easier. Data distribution among classes is also skewed, so it helps in evaluating how the model scales with more glosses and unbalanced data. These training and testing splits are provided in the original papers [99, 82].

Both these datasets consist of publicly available video links from various sources. Users are expected to download and process the dataset based on the provided configuration files. This format is good for easy access and future scalability of the datasets, but the provided web links change over time and have to be continuously updated to keep the datasets viable and available. Similar to [98], we were able to gather all samples for WLASL, but could not access 6,164 out of 25,513 samples for MSASL (24%) due to dead links. For the rest of this paper, we will refer to our recovered portion of MSASL as MSASL_R.

Data Pre-processing

All videos are rescaled down to 256x256 resolution for further processing. These videos are used as-is in I3D. For body and hand pose estimation, all videos are processed with Openpose [33, 141]. Extracted 2D keypoints and corresponding RGB frames are then fed into 3D pose recovery pipeline (below). We obtain four data modalities from these keypoints, namely: 2D Joints, 2D Bones, 3D Joints and 3D Bones.

2D to 3D Pose Estimation

Estimating the correct 3D body and hand pose information of the signer from monocular videos is a challenging task. This information is also missing from most of the prior sign language datasets since accurate 3D body pose is hard to obtain without using the visually obtrusive markers and uncomfortable body suits of the standard marker-based motion capture systems. Recent deep learning-based approaches have shown promising results towards reliable 3D human body and pose estimation from monocular images. These methods rely

on the availability of large scale datasets containing 3D human pose annotations tracked using motion capture systems. Fortunately, most of the sign language gesture datasets are recorded from a front facing pose which are commonly found in 3D human pose datasets, thus lowering the chance of such estimators to fail.

We use two separate approaches for body and hand tracking. Since part of the signer's body is typically not visible in the videos, we employ XNect [108], a state-of-the-art multi-person 3D human pose estimation which can perform reliable estimation under partially occluded settings. To this end, we use the per-frame estimation result from the *Stage II* of the method to obtain 13 upper body joint predictions. Note that the prediction for each frame is relative to the 3D position of the pelvis.

3D pose information of the hands is obtained with the real-time capable monocular hand tracking method of Zhou [168]. In addition to regressing the hand joint position, this method also consists of a kinematic regression module that allows the model to improve the initial regression by leveraging prior hand pose information. Since this method assumes the input to be tightly cropped around each hand, we first crop the source image around a 2D bounding box which we estimate from the 2D hand pose estimation result of OpenPose [141]. Once predicted, we collect the information for the 21 joints of each hand alongside the prediction confidence value for each joint. In contrast to the body, hand pose estimation is more challenging since hands are more likely to be occluded or of lower quality due to the motion blur. To alleviate these issues, we discard the 3D hand prediction result if the prediction confidence value falls below a certain threshold. We then fill-in this missing information using an off-the-shelf cubic interpolation method.

4.4 Method

The overall approach combines predictions from two types of networks (I3D and MSG3D-Attend) using an ensembling approach.

RGB Based - Inflated 3D

I3D [34] has been widely used as a baseline for numerous activity and action recognition papers. The original WLASL [99] and MSASL [82] papers that introduced the datasets also reported I3D results. 3D-CNN based approaches have proven to be fairly efficient at capturing spatio-temporal relationships between the inputs, which is a much needed requirement for the task at hand. I3D is a 3D-CNN based feature extraction method that starts with a 2D base architecture and inflates its filters and pooling kernels to add the additional temporal dimension. These 2D filters can be taken from already trained, large scale datasets like ImageNet [132], which adds to the efficacy of the model. Carreira further fine tuned the model on the Kinetics dataset [34]. To cater for the varying nature of videos in datasets and dynamic nature of actions, kernels that are non-symmetric across temporal dimensions are used.

In our experimentation, we used the WLASL I3D implementation provided by the authors [99] and used pre-trained model weights on a British Sign Language dataset BSL-1K [5] that has shown to improve the performance on both WLASL and MSASL. Although the BSL-1K dataset is not yet publicly available, pre-trained model weights can be downloaded from the project website. Our training on a BSL-1K base resulted in 49.57% per instance accuracy (2.75% greater than [5]) on WLASL and 63.8% (0.91% less than [5]) on MSASL_R.

Skeleton Based - MSG3D-Attend

MSG3D-Attend builds on Liu 's MSG3D [102]. MSG3D jointly models dependencies across space and time with a G3D module. While other approaches rely solely on a factorized formulation where graph convolutions (GCNs) and temporal convolutions (TCNs) are computed independently and later merged together, this architecture utilizes both factorized and jointly modeled pathways. Each bulding block (STGC) of the MSG3D model sends data along two parallel pathways, the G3D and a factorized pathway that consists of one GCN and two TCNs layered vertically. The G3D pathway uses cross space-time edges that

create unbiased, direct and data driven dependencies in data at multiple scales. It creates spatio-temporal windows and performs multi-scale graph convolutions on them. In each iteration G3D aggregates responses from sliding windows at different strides and dilations and sends it back to STGC block where the aggregated response is merged with the outcome of factorized pathway, and passed out of the block via an MS-TCN [102] layer.

MSG3D-Attend adds a self attention layer at the end of G3D module that contextualizes the aggregated information before merging with the factorized pathway. The model architecture is shown in Figure 4.1. The proposed self-attention layer has spatial, temporal and channel attention modules connected in a series similar to STC module in [76]. Spatial and Temporal attentions are 1D convolutions in respective dimension followed by a sigmoid activation. Channel attention consists of chained linear layers with ReLU activation after the first layer and sigmoid after the second. Along the factorized pathway, both MS-TCN layers use a stride of 2. We add together the output along both pathways, and apply a ReLU activation on top as a final output of the block.

Ensembling

The idea of ensembling is to merge prediction scores coming from each individual data stream in order to make the results more robust and accurate. We use standard multi-stream ensembling techniques to produce the final SLR predictions[76, 156]. Ensembles are evaluated for skeleton-based MSG3D-Attend data modalities and a combination of MSG3D-Attend with RGB based I3D.

Prediction score vectors can be added without any priority/weighting factor associated with individual stream's vector, known as un-weighted ensembling, or individual weights can be assigned to each stream, called weighted ensembling. Our experiments employ the weighted ensembling approach. Multiple methods have been proposed for finding the best weights. We found optimal weights using Nelder-Mead [43] optimization available in Python Scipy optimization package. The optimization objective was to max out the Top-1 accuracy

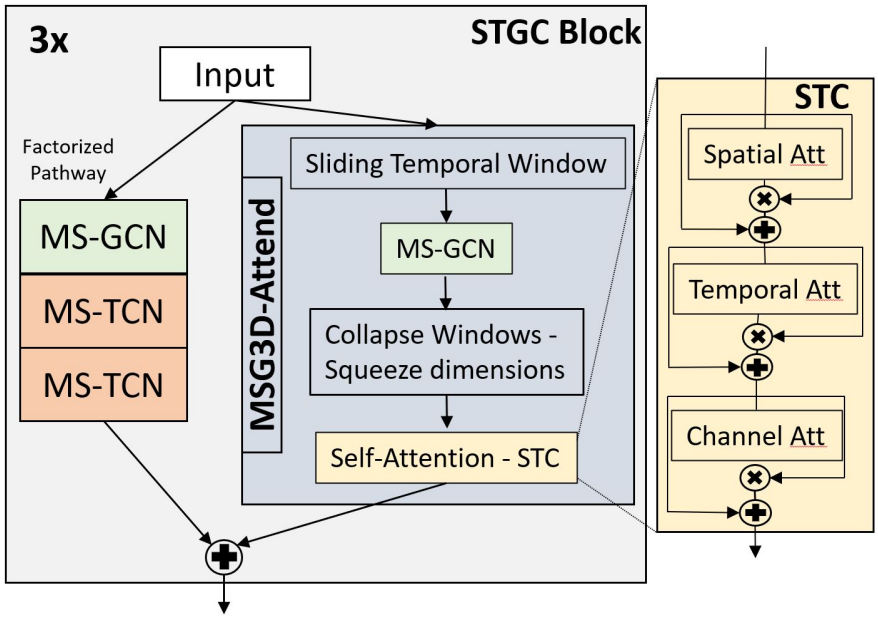

Figure 4.1: MSG3D-Attend model architecture. STC block represents layered spatial, temporal and channel attentions layered on top of eachother. '3x' on top left corner of the figure shows three STGC blocks serially connected in our model. Keyword MS- stands for Multi-Scale, GCN stands for Graph Convolutional Networks and TCN are Temporal Convolutional Networks. Input data is split across two pathways, MSG3D-Attend and a factorized pathway stacking spatial-only and temporal only convolutional layers. The output of both pathways is added together and passed through a ReLU activation.

for each data subset evaluated in Tables 4.1 and 4.2.

4.5 Experiments and Results

All experiments were run using pytorch. Standard data splits provided with the datasets are used. For I3D runs, we used the Adam optimiser with binary cross entropy loss, base learning rate of 0.001 and plateau patience of 5 steps before changing the learning rate by a factor of 0.3. 64 consecutive frames were selected from each clip randomly in each iteration. At training time, videos were randomly mirrored horizontally and a randomly cropped 224x224 region was returned. All datasets were rescaled down to a resolution of 256x256 beforehand.

During testing, the middle 224x224 region was cropped and returned. Pretrained weights on BSL-1K[5] were used for training.

For MSG3D and MSG3D-Attend training, 50 consecutive frames were randomly chosen. Sequences shorter than 50 frames were padded by randomly repeating the first or last frame. For MSG3D multi-scale learning, 2 G3D scales and 8 GCN scales were chosen. 3 STGC blocks were stacked on top of each other with 48, 96 and 192 feature channels respectively. Stochastic Gradient Descent with Nesterov's accelerated gradient (momentum: 0.9, weight decay: 0.0003) was used. Batch size of 32 was used across the board. An initial learning rate of 0.1 was decayed by a factor of 0.1 at epochs 45, 65 and 85 with 95 max epochs. MSG3D results with all configurations the same as the original paper were slightly lower than reported here.

Tables 4.1 and 4.2 show SLR results for WLASL and MSASL_R respectively. Each table is divided into three sections. The first section refers to results reported in previous work. The middle section reports results of our runs on MSG3D [102, 101] for WLASL and MSASL_R datasets. The third section reports results of our proposed MSG3D-Attend model and final ensemble numbers. The second last row, labeled 'MSG3D-A - Ensemble' refers to an ensemble for all four skeleton based modalities predicted through our proposed MSG3D-Attend model. Similarly, the last row reports our final results ensembling four skeleton based streams and one rgb based I3D stream.

Discussion

For WLASL, our approach out performs current state of the art [70] by **at least 5%** for Top-1 accuracy consistently across all subsets. For MSASL_R, we outperform [70] on Top-1 accuracy numbers for both the smallest and largest subset and slightly under perform (0.2%) on MSASL200_R subset. It must be considered that similar to [98], results reported on MSASL are indicative, as we were not able to download all videos for that dataset.

Our modified MSG3D-Attend model also out-performs MSG3D on all skeleton based data

	$WLASL100$			$WLASL300$			$WLASL2000$		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
Li [98]	77.52	91.08	-	68.56	89.52	-	-	-	-
SignBert - Pose[70]	79.07	93.80	-	70.36	88.92	-	47.46	83.32	-
SignBert - RGB[70]	82.56	94.96	-	74.40	91.32	-	54.69	52.08	-
BSL-1K[5]	-	-	-	-	-	-	46.82	79.36	-
Fusion-3[66]	75.67	86.00	90.16	68.30	83.19	86.22	38.84	67.58	75.71
Hu [69]	-	-	-	-	-	-	51.39	86.34	-
$MSG3D_{2D}^{Joints}$ [102]	68.89	88.19	93.7	59.91	83.96	90.17	46.0	76.62	81.01
$MSG3D_{2D}^{Bones}$ [102]	64.57	89.34	94.88	63.24	85.02	91.08	45.9	76.13	81.33
$MSG3D_{3D}^{Joints}$ [102]	72.44	89.76	93.7	63.39	87.14	91.23	41.15	74.02	78.95
$MSG3D_{3D}^{Bones}$ [102]	65.75	87.80	93.31	59.76	85.48	90.62	42.32	75.23	80.01
I3D - Pretrained $BSL\text{-}1K_{OurRuns}$	83.72	93.80	95.83	72.57	86.78	91.14	49.57	81.42	87.29
Our - $MSG3D\text{-}A_{2D}^{Joints}$	75.39	92.13	95.87	69.59	89.03	93.04	48.62	80.631	87.23
Our - $MSG3D\text{-}A_{2D}^{Bones}$	72.83	89.96	93.50	63.31	85.48	91.91	48.14	80.86	86.52
Our - $MSG3D\text{-}A_{3D}^{Joints}$	73.43	91.14	93.50	65.66	88.05	92.89	44.80	77.83	84.66
Our - $MSG3D\text{-}A_{3D}^{Bones}$	76.57	92.13	94.09	63.31	87.82	92.81	43.28	77.21	83.95
MSG3D-A - Ensemble	85.43	95.67	96.06	75.49	92.59	96.67	54.78	87.19	91.93
Our - I3D + MSG3D-A - Ensemble	**90.55**	**96.85**	**97.24**	**83.24**	**95.61**	**97.73**	**59.66**	**90.73**	**94.44**

Table 4.1: WLASL Recognition Results. Highest accuracy numbers (micro) are shown in bold format. Our ensembles combining MSG3D-Attend skeleton-based modalities and our run of I3D pre-trained on BSL-1K report the best performance. Results are reported as average per-instance accuracy also known as micro-average accuracy.

modalities, across datasets and data subsets. On average there is a 3 to 4% improvement per data modality using MSG3D-Attend.

One major difference between the two evaluated ASL datasets is that $MSASL_R$[82] has half the gloss labels and a much larger average number of samples per class (19.3) compared to WLASL (10.3)[99]. I3D performing better on $MSASL_R$ than WLASL, both in terms of accuracy numbers and scaling with dataset size, can be attributed to this fact. On the other hand, videos in WLASL are comparatively much more consistent and cleaner in terms of signer placement in front of the camera, viewing angles, zoom level and clip timing segmentation. These factors result in better skeletal pose estimation for WLASL dataset, and might

	$MSASL100_R$			$MSASL200_R$			$MSASL1000_R$		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
Li [98]	83.04	93.46	-	80.31	91.82	-	-	-	-
SignBert - Pose [70]	81.37	93.66	-	77.34	91.10	-	59.80	81.86	-
SignBert - RGB [70]	89.56	**97.36**	-	**86.98**	**96.39**	-	71.24	89.12	-
BSL-1K [5]	-	-	-	-	-	-	64.71	85.59	-
Hu [69]	87.45	96.30	-	85.21	94.41	-	69.39	87.42	-
$MSG3D_{2D}^{Joints}$ [102]	62.5	84.26	89.56	62.36	82.21	86.66	43.27	68.35	73.15
$MSG3D_{2D}^{Bones}$ [102]	63.86	83.82	89.85	64.99	82.37	87.56	44.32	69.84	75.01
$MSG3D_{3D}^{Joints}$ [102]	66.47	88.24	92.94	63.51	82.87	87.64	41.01	66.01	71.78
$MSG3D_{3D}^{Bones}$ [102]	61.62	85.44	91.32	59.64	81.47	87.23	40.35	65.51	70.56
I3D - Pretrained $BSL\text{-}1K_{OurRuns}$	84.61	95.80	97.23	82.24	91.78	93.14	63.8	82.49	87.29
Our - $MSG3D\text{-}A_{2D}^{Joints}$	68.6	86.91	90.15	67.83	84.31	88.59	49.44	71.05	78.15
Our - $MSG3D\text{-}A_{2D}^{Bones}$	67.5	83.9	89.26	65.03	82.58	87.19	48.5	70.66	76.97
Our - $MSG3D\text{-}A_{3D}^{Joints}$	68.68	88.38	93.53	64.58	84.84	89.37	43.77	67.43	75.42
Our - $MSG3D\text{-}A_{3D}^{Bones}$	67.06	86.62	91.62	63.1	80.81	86.49	43.68	67.21	74.85
MSG3D-A - Ensemble	79.68	92.21	95.00	76.45	88.55	91.85	59.35	78.58	83.7
Our - I3D + MSG3D-A - Ensemble	**89.97**	96.51	**97.88**	86.78	95.79	**97.79**	**72.03**	**90.76**	**93.34**

Table 4.2: MSASL Recognition Results. Highest accuracy numbers (micro) are shown in bold format. Our ensembles combining MSG3D-Attend skeleton-based modalities and our run of I3D pre-trained on BSL-1K report the best performance. Results are reported as average per-instance accuracy also known as micro-average accuracy.

be the reason behind drop of performance for skeletal data modalities for $MSASL_R$. This could also explain why skeleton based MSG3D-Attend ensembles consistently out-perform I3D on WLASL, but the pattern gets reversed for $MSASL_R$. The only other skeleton-based benchmark from previous work is Pose-Based SignBert[70] and it's performance is at the bottom (excluding MSG3D [102] results) for MSASL. Comparatively, on WLASL Pose SignBert performs better than [98]. It is possible that owing to the nature of videos in dataset, skeleton pose estimation on MSASL [82] results in lower quality data.

No single modality appears to dominate the results, with each of the five modalities having the highest recognition results for at least one test case. I3D often has the highest individual score, especially on the MSASL dataset. Considering the skeletal modalities, 3D joint data

often scores highest for the original MSG3D model and the MSG3D-Attend model for the MSASL dataset. However, 2D joints often perform best for the MSG3D-Attend model on the WLASL dataset. Some further insight on the importance of the various modalities can be gained from examining the optimal ensemble weights. For example, for WLASL2000, the weights by modality are: 0.1589 (2D joints), 0.2318 (2D Bones), 0.1507 (3D Joints), 0.1310 (3D Bones) and 0.3274 (I3D-RGB). This shows that I3D is indeed given greatest weight, but each modality is providing useful input, with the lowest weight at 0.13. One might expect 3D data to become even more valuable for in-the-wild videos in which the camera angle is less carefully controlled.

4.6 Sign Spotting

In this section we explore the feasibility of using our designed models to create sign spotting labels in continuous clips from the How2Sign[48] dataset. Labels for component signs are unavailable on any sentence-level ASL dataset and would need to be annotated by hand, so sign spotting offers an appealing alternative to manual labeling as a way to build labeled datasets. The process does not involve any training on How2Sign or any other continuous domain dataset. We strictly train on ISLR videos from WLASL and test on a subset of How2Sign clips. These labels can be used as a stepping stone for future labeling efforts or weakly-supervised learning for the problem. Our generated labels are available upon request for future work.

Dataset Preparation

We used clips from the publicly available, continuous sign language dataset How2Sign[48] as a test set. The dataset contains roughly 79 hours of footage with sentence level English language translation annotations and sentence-level segmentations. The dataset, however, does not provide sign spotting labels. Clips on the project website are a mixture of 24 and

30 fps videos. All videos used in the experiments were at 24fps. We test the approach by generating weak spotting labels for How2Sign clips with our model trained on WLASL100 dataset and evaluate it on a subset of 120 randomly chosen and manually labeled clips from How2Sign. Chosen clips had at least one word from the WLASL100 labels in the provided continuous sentence translations and no label appeared more than twice in the testing set. We excluded ambiguous words like 'all', 'but', 'now', 'can', 'like' etc.

Training ISLR Models for Sign Spotting

New versions of the 2D joints MSG3D-Attend and I3D model were trained to operate on a smaller window needed for the spotting task. The average length of our manually created spotting annotations is 16 seconds with a median of 14.4, so both models were trained with 15 frame inputs. Other than the standard data augmentation techniques for training mentioned in Section 4.5, every other frame in videos was sampled at random training instances. Hence the model was trained at different signing speeds for the same sign. All other settings were kept the same. Our I3D model trained for spotting reported 64.7% Top-1 accuracy on WLASL100 test set and MSG3D-Attend reported 68.11%. Once the models are trained, we remove their final fully connected layers and use the rest of the model as a feature extractor. For I3D, this provides a 1024-dimensional feature F^I for each input 15 frame sample. MSG3D-Attend results in a 192-dimensional feature F^G.

SVMs for Sign Spotting The sign spotting approach assumes access to subtitles or sentence level translations for the clip. A word is then chosen from the annotated sentence and the network attempts to localize its corresponding Sign S in the video clip C. The network annotates every 15 frame segment of the clip with a binary label, identifying whether or not it contains the desired sign. It is thus a binary classification problem.

Support Vector Machines (SVMs) are employed for the task. To train the SVMs, F^I and F^G feature vectors are calculated for all training samples in WLASL100. When spotting a sign S in clip C, we train an SVM for S against K randomly chosen signs from the training

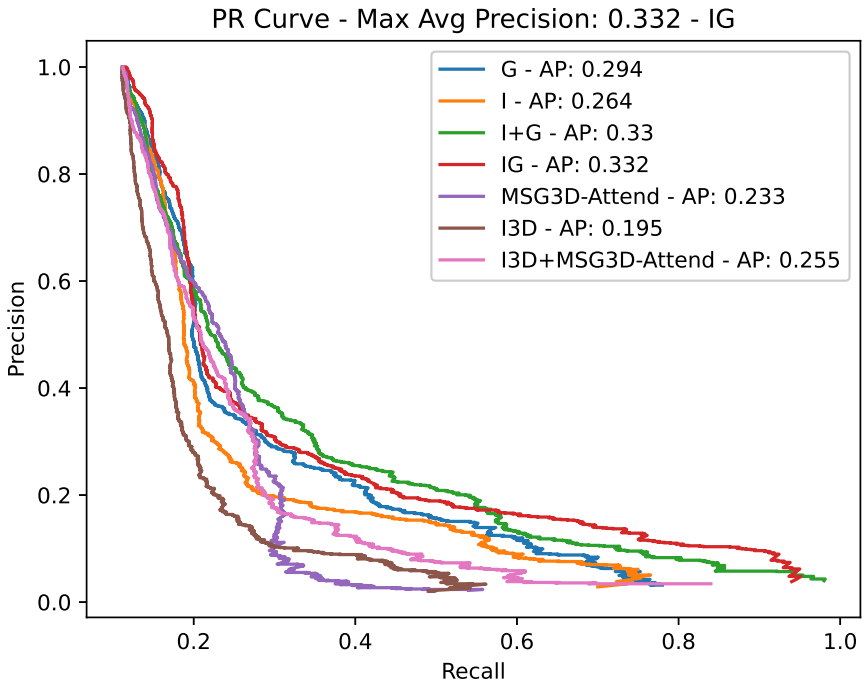

Figure 4.2: Precision-Recall Curve for SVM based SVM^G (G), SVM^I(I), SVM^I + SVM^G(I+G), SVM^{IG}(IG), MSG3D-Attend, I3D and I3D+MSG3D-Attend. Average Precision (AP) numbers for each model are also shown. SVM^{IG} outperforms others on this metric, while I3D performs the worst.

dataset. Binary decision SVMs are trained where $F^G(S)$ and $F^I(S)$ are labelled as positive and all other training samples for K classes are marked negative. N such SVMs are trained, each against K other classes, and the prediction scores are averaged to increase robustness of the predictions. If a single SVM against all other training classes for WLASL100 is to be trained, we will use N=1 and K=99. Binary SVMs are trained on F^I, F^G and F^{IG} features respectively where F^{IG} is a concatenation of F^I with F^G to create a 1216-dimensional representation. Corresponding SVMs are denoted SVM^I,SVM^G,SVM^{IG}.

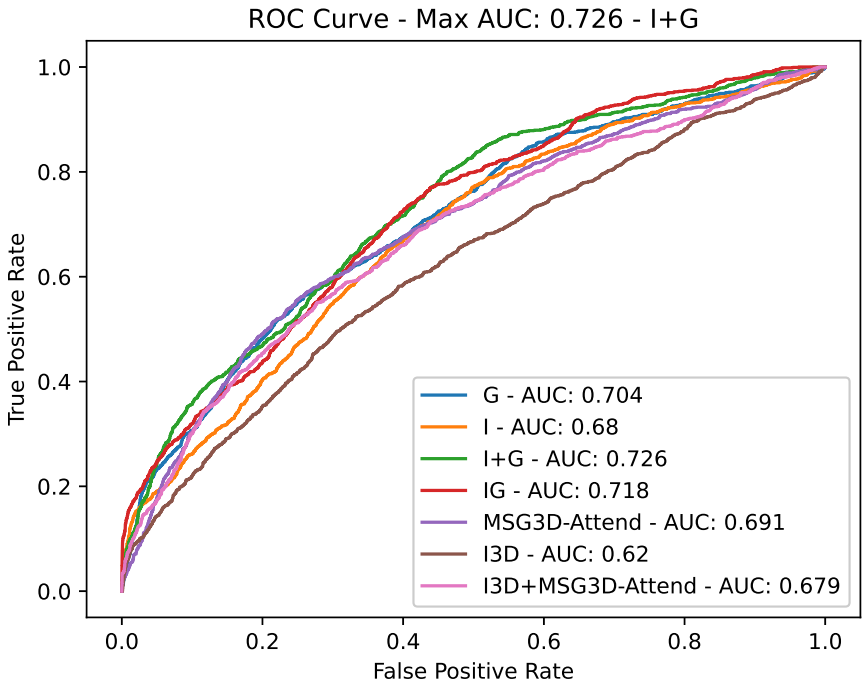

Figure 4.3: ROC Curve for SVM based SVM^G (G), SVM^I(I), SVM^I + SVM^G(I+G), SVM^{IG}(IG), MSG3D-Attend, I3D and I3D+MSG3D-Attend. Area Under the Curve (AUC) numbers for each model are also shown. SVM^I + SVM^G outperforms others on this metric, while I3D performs the worst.

Evaluation

Sliding windows of 15 frames are tested for sign spotting. If a window is predicted as a positive spotting instance and it overlaps with a true label by 5 or more frames, it is marked as a true positive instance. Figures 4.2 and 4.3 compare the precision-recall and ROC curves for 7 classification models, namely SVM^G labeled as G in the figures, SVM^I(I), SVM^I + SVM^G(I+G), SVM^{IG}(IG), MSG3D-Attend with the final logits layer, I3D and I3D+MSG3D-Attend. I3D performs worst on both metrics while SVM^I + SVM^G and SVM^{IG} lead both the charts with highest AUC score on ROC curve and highest average precision.

Table 4.3 compares the F-1 harmonic mean score for the four SVM variants. SVM^{IG}

	K=1	K=3	K=6	K=9	K=12
SVM^G	0.234	0.272	0.274	0.235	0.198
SVM^I	0.202	0.151	0.152	0.136	0.109
SVM^I+ SVM^G	0.220	0.266	0.198	0.105	0.110
SVM^{IG}	**0.262**	**0.298**	**0.314**	**0.286**	**0.261**

Table 4.3: F1 score table. Row-wise the table shows results for SVMs trained on features from MSG3D-Attend (F^G), I3D (F^I), combination of prediction scores from the former two models, and concatenating features from both (F^{IG}). We show the F1-score trend at varying K.

outperforms the other variants on all values of K. Higher value of K leads to higher precision and lower recall as we train discriminatory SVMs against a higher number of classes, so the model gives out positive labels more and more conservatively. F-1 scores peak at 0.314 at K=6 for SVM^{IG}. F1-score for baseline models directly getting results from I3D is 0.165, MSG3D-Attend is 0.178 and I3D+MSG3D-Attend is 0.209.

4.7 Conclusion

This paper presents a new method for isolated sign recognition (ISLR). Using an ensemble approach, it combines prediction networks using RGB data, as well as 2D and 3D skeletal data. Recognition on the skeletal data is performed using a novel MSG3D-Attend model. State-of-the-art performance is obtained both for the skeletal modalities and the overall ensemble. This is also the first work we are aware of to explore the application of 3D skeletal data to the sign recognition problem.

A novel extension of the technique is explored for *sign spotting* – recognizing signs in longer sentences. Initial results are promising and the technique could be useful in boot strapping annotation work on sign utterances, among other applications. The paper will also mark the release of two large, 3D skeleton ASL datasets derived from the MSASL and WLASL datasets. This is likely to be useful in much future research in the area.

Sign recognition in longer utterances remains a very challenging problem. Using semi-supervised approaches to help build annotated corpora, such as the sign spotting work explored in the second part of this paper, may be a profitable way to create large, labeled datasets, and in turn build more robust sentence-level recognition models.

Chapter 5

Conclusion and Future Work

Embodiment is a key component of interpersonal interactions. This dissertation covers topics that inform the design and modeling of natural communication patterns for avatars in virtual environments. We seek to learn from real life examples of human signs and gestures, design models for autonomous agent decision making and study how humans perceive their peers when they are embodied as avatars in virtual environments.

We first present a study comparing group interactions in Virtual Reality and Videoconferencing settings. Embodied experience with participants sitting in chairs is provided for both conditions. VR was accompanied with full body, hands, face and eye tracking. Each group of three people were given carefully designed tasks to perform. Evaluation metrics showed that people were able to able to achieve similar performance across tasks, however their gaze and other nonverbal behavior patterns varied throughout. These observations suggest that when in a shared 3D space (even if it is virtual), people tend to rely on different nonverbal modes of communication. Direct gaze was more dominant in VC compared to VR i.e. people would look directly at the speaker's face and vice versa a lot more. One possible explanation is that people gather more information from other cues when in a shared embodied setting than VC. VC also had more formal turn taking in speech, longer backchannels and less unique gestures. This postulates that people have to put in more effort in VC to maintain similar

social connection as in VR. These findings are also consistent with studies comparing VC to face-to-face interactions. It is worth noting that in this study VR interactions were limited to people sitting in their seats to keep it comparable with VC. Settings where people can freely move in VR, actively adjust their distance during social and professional interactions, etc. could potentially have a large impact on communication patterns. In the future, studies without the mobility restrictions can help us improve our understanding further.

For autonomous agents, we present a behavioral design model for a pedagogical agent, Maria, that takes students through a discovery based learning environment. Discovery-based learning is an educational activity paradigm whereby students are led through well-specified experiences that are designed to foster particular insights relevant to curricular objectives. The students are never explicitly told what they are expected to learn, but are rather taken through a series of exercises and asked about their observations and insights. Maria takes students through the Mathematical Imagery Trainer for proportionality, where students build intuition for the concepts of ratios and proportions. Students are asked to touch the screen with a finger from both hands and depending on relative height of their fingers, the screen color changes. They are tasked with making the screen green and must explain their observations on how to turn the screen green. In this experience Maria is able to take students through different exercises to develop the strategies "a-per-b" and "Higher-Bigger" to find embodied proportions that make the screen green. A user study showed that students were able to foster insights on one of the strategies. One future direct is to incorporate more teaching material in the agent tutorials. This can improve the interaction between Maria and the students, so students provide more than just screen coordinates to analyze their understanding and Maria can accordingly provide custom experiences for students.

We also explore the problem of sign language recognition and spotting. Sign Language is the primary mode of communication for deaf communities around the world. Sign language is also differentiated into word-level or isolated signs and continuous sign language. Isolated signs represent a single word or gloss, whereas continuous sign language is equivalent to uttering larger sentences. Sign Language Translation is a broad problem that targets translating

sign language videos into verbal language, and could bridge the communication gap between deaf and hearing communities. Sign Language Recognition is a sub-problem where we try to recognize individual signs. We present a state of the art model for recognition on isolated American Sign Language (ASL) recognition datasets using Multi-scale Graph Convolution Networks and then present a base framework of weak labelling for Sign Spotting task. Sign Spotting is when you try to localize isolated signs in continuous signs. The continuous sign language problem for ASL has been under explored due to a lack of labelled data. With new datasets like How2Sign [48], and algorithms to automatically annotate these datasets, progress can be made on this front and the communication gap can be reduced for deaf communities.

Bibliography

[1] Ahsan Abdullah et al. "Pedagogical agents to support embodied, discovery-based learning". In: *International Conference on Intelligent Virtual Agents*. Springer. 2017, pp. 1–14.

[2] Dor Abrahamson et al. "Coordinating visualizations of polysemous action: Values added for grounding proportion". In: *ZDM* 46.1 (2014), pp. 79–93.

[3] Dor Abrahamson et al. "Fostering hooks and shifts: Tutorial tactics for guided mathematical discovery". In: *Technology, Knowledge and Learning* 17.1-2 (2012), pp. 61–86.

[4] Samuel Albanie et al. "BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues". In: *European Conference on Computer Vision*. 2020.

[5] Samuel Albanie et al. "BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues". In: *CoRR* abs/2007.12131 (2020). arXiv: 2007.12131. URL: https://arxiv.org/abs/2007.12131.

[6] *American Sign Language Lexicon Video Dataset (ASLLVD)*. http://www.bu.edu/asllrp/av/dai-asllvd.html.

[7] Cleison Correia de Amorim, David Macêdo, and Cleber Zanchettin. "Spatial-temporal graph convolutional networks for sign language recognition". In: *International Conference on Artificial Neural Networks*. Springer. 2019, pp. 646–657.

[8] Michael L Anderson, Michael J Richardson, and Anthony Chemero. "Eroding the boundaries of cognition: Implications of embodiment". In: *Topics in Cognitive Science* 4.4 (2012), pp. 717–730.

[9] Stephen R Anderson. "How many languages are there in the world". In: *Linguistic Society of America* (2010).

[10] Terry Anderson and Julie Shattuck. "Design-based research: A decade of progress in education research?" In: *Educational researcher* 41.1 (2012), pp. 16–25.

[11] Alissa N Antle. "Research opportunities: Embodied child–computer interaction". In: *International Journal of Child-Computer Interaction* 1.1 (2013), pp. 30–36.

[12] Vassilis Athitsos et al. "The american sign language lexicon video dataset". In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. 2008, pp. 1–8.

[13] Jeremy N Bailenson. "Nonverbal overload: A theoretical argument for the causes of Zoom fatigue". In: *Technology, Mind, and Behavior* 2.1 (2021).

[14] Jeremy N Bailenson et al. "The use of immersive virtual reality in the learning sciences: Digital transformations of teachers, students, and social context". In: *The Journal of the Learning Sciences* 17.1 (2008), pp. 102–141.

[15] Jeremy N Bailenson et al. "Transformed social interaction, augmented gaze, and social influence in immersive virtual environments". In: *Human communication research* 31.4 (2005), pp. 511–537.

[16] H Harlyn Baker et al. "Understanding performance in coliseum, an immersive videoconferencing system". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 1.2 (2005), pp. 190–210.

[17] Charlotte Lee Baker-Shenk. *A microanalysis of the nonmanual components of questions in American Sign Language*. University of California, Berkeley, 1983.

[18] Douglas Bates et al. “Fitting linear mixed models in R”. In: *R news* 5.1 (2005), pp. 27–30.

[19] Douglas Bates et al. “Fitting linear mixed-effects models using lme4”. In: *arXiv preprint arXiv:1406.5823* (2014).

[20] Robbin Battison. “Lexical borrowing in American sign language.” In: (1978).

[21] Geoffrey W Beattie. “Contextual constraints on the floor-apportionment function of speaker-gaze in dyadic conversations.” In: *British Journal of Social & Clinical Psychology* (1979).

[22] Geoffrey W Beattie. “The regulation of speaker turns in face-to-face conversation: Some implications for conversation in sound-only communication channels”. In: (1981).

[23] Gary Bente et al. “Avatar-mediated networking: Increasing social presence and interpersonal trust in net-based collaborations”. In: *Human communication research* 34.2 (2008), pp. 287–318.

[24] Yunus Can Bilge, Nazli Ikizler-Cinbis, and Ramazan Gokberk Cinbis. “Zero-shot sign language recognition: Can textual data uncover sign languages?” In: *arXiv preprint arXiv:1907.10292* (2019).

[25] Frank Biocca, Chad Harms, and Jenn Gregg. “The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity”. In: *4th annual international workshop on presence, Philadelphia, PA*. 2001, pp. 1–9.

[26] Nicholas Bloom et al. “Does working from home work? Evidence from a Chinese experiment”. In: *The Quarterly Journal of Economics* 130.1 (2015), pp. 165–218.

[27] Leanne S Bohannon. “Effects of video-conferencing on gaze behavior and communication”. PhD thesis. Rochester Institute of Technology, 2010.

[28] Ty W Boyer and Susan C Levine. “Prompting children to reason proportionally: Processing discrete units as continuous amounts.” In: *Developmental psychology* 51.5 (2015), p. 615.

[29] Judee K Burgoon, Deborah A Coker, and Ray A Coker. "Communicative effects of gaze behavior: A test of two contrasting explanations". In: *Human Communication Research* 12.4 (1986), pp. 495–524.

[30] John Burnet et al. *Aristotle on education: being extracts from the Ethics and Politics*. At the University Press, 1905.

[31] Necati Cihan Camgoz et al. "Neural sign language translation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7784–7793.

[32] Julia Campbell et al. "Developing INOTS to support interpersonal skills practice". In: *Aerospace Conference, 2011 IEEE*. IEEE. 2011, pp. 1–14.

[33] Zhe Cao et al. "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". In: *CVPR*. 2017.

[34] Joao Carreira and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset". In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.

[35] Milton Chen. "Leveraging the asymmetric sensitivity of eye contact for videoconference". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2002, pp. 49–56.

[36] Vasileios Choutas et al. "Potion: Pose motion representation for action recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7024–7033.

[37] Herbert H Clark and Susan E Brennan. "Grounding in communication." In: (1991).

[38] Cristina Conati. "Probabilistic assessment of user's emotions in educational games". In: *Applied Artificial Intelligence* 16.7-8 (2002), pp. 555–575.

[39] MMPose Contributors. *OpenMMLab Pose Estimation Toolbox and Benchmark*. `https://github.com/open-mmlab/mmpose`. 2020.

[40] John W Creswell. *Qualitative inquiry and research design: Choosing among five approaches*. Sage Publications, 2012.

[41] Richard L Daft and Robert H Lengel. "Organizational information requirements, media richness and structural design". In: *Management science* 32.5 (1986), pp. 554–571.

[42] Thomas Dean and Keiji Kanazawa. "A model for reasoning about persistence and causation". In: *Computational intelligence* 5.2 (1989), pp. 142–150.

[43] John E Dennis Jr and Daniel J Woods. *Optimization on microcomputers. the nelder-mead simplex algorithm*. Tech. rep. 1985.

[44] David DeVault et al. "SimSensei Kiosk: A virtual human interviewer for healthcare decision support". In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 2014, pp. 1061–1068.

[45] Trevor J Dodds, Betty J Mohler, and Heinrich H Bülthoff. "Talk to the virtual hands: Self-animated avatars improve communication in head-mounted display virtual environments". In: *PloS one* 6.10 (2011), e25759.

[46] Gwyneth Doherty-Sneddon et al. "Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance." In: *Journal of experimental psychology: applied* 3.2 (1997), p. 105.

[47] Wei Dong and Wai-Tat Fu. "One piece at a time: why video-based communication is better for negotiation and conflict resolution". In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 2012, pp. 167–176.

[48] Amanda Duarte et al. "How2Sign: A Large-Scale Multimodal Dataset for Continuous American Sign Language". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 2735–2744.

[49] Carolien ACG Duijzer et al. "Touchscreen tablets: Coordinating action and perception for mathematical cognition". In: *Frontiers in Psychology* 8 (2017).

[50] Jordan Fenlon, Kearsy Cormier, and Adam Schembri. "Building BSL SignBank: The lemma dilemma revisited". In: *International Journal of Lexicography* 28.2 (2015), pp. 169–206.

[51] Samantha Finkelstein et al. "The effects of culturally congruent educational technologies on student achievement". In: *International Conference on Artificial Intelligence in Education*. Springer. 2013, pp. 493–502.

[52] Jens Forster et al. "RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus." In: *LREC*. Vol. 9. 2012, pp. 3785–3789.

[53] Henry Fuchs, Andrei State, and Jean-Charles Bazin. "Immersive 3d telepresence". In: *Computer* 47.7 (2014), pp. 46–52.

[54] Susan R Fussell and Leslie D Setlock. "Computer-mediated communication." In: (2014).

[55] João Carlos Gluz et al. "Helping students of introductory calculus classes: the Leibniz pedagogical agent". In: ().

[56] Arthur C Graesser et al. "AutoTutor: A tutor with dialogue in natural language". In: *Behavior Research Methods* 36.2 (2004), pp. 180–192.

[57] Jonathan Gratch et al. "Negotiation as a challenge problem for virtual humans". In: *International Conference on Intelligent Virtual Agents*. Springer. 2015, pp. 201–215.

[58] Simon NB Gunkel et al. "Virtual Reality Conferencing: Multi-user immersive VR experiences on the web". In: *Proceedings of the 9th ACM Multimedia Systems Conference*. 2018, pp. 498–501.

[59] J Richard Hackman. "Effects of task characteristics on group products". In: *Journal of Experimental Social Psychology* 4.2 (1968), pp. 162–187.

[60] Shangchen Han et al. "MEgATrack: monochrome egocentric articulated hand-tracking for virtual reality". In: *ACM Transactions on Graphics (TOG)* 39.4 (2020), pp. 87–1.

[61] Chad Harms and Frank Biocca. “Internal consistency and reliability of the networked minds measure of social presence”. In: (2004).

[62] Jonathon D Hart et al. “Emotion sharing and augmentation in cooperative virtual reality games”. In: *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. 2018, pp. 453–460.

[63] Dai Hasegawa et al. “The Effect of Metaphoric Gestures on Schematic Understanding of Instruction Performed by a Pedagogical Conversational Agent”. In: *Learning and Collaboration Technologies: Second International Conference, LCT 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings*. Ed. by Panayiotis Zaphiris and Andri Ioannou. Cham: Springer International Publishing, 2015, pp. 361–371. ISBN: 978-3-319-20609-7. DOI: 10.1007/978-3-319-20609-7_34. URL: http://dx.doi.org/10.1007/978-3-319-20609-7_34.

[64] Jörg Hauber et al. “Spatiality in videoconferencing: trade-offs between efficiency and social presence”. In: *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. 2006, pp. 413–422.

[65] Paul Heidicker, Eike Langbehn, and Frank Steinicke. “Influence of avatar appearance on presence in social VR”. In: *2017 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE. 2017, pp. 233–234.

[66] Al Amin Hosain et al. “Hand Pose Guided 3D Pooling for Word-Level Sign Language Recognition”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 3429–3439.

[67] Ronald A Howard. *Readings on the principles and applications of decision analysis*. Vol. 1. Strategic Decisions Group, 1983.

[68] Mark Howison et al. “The Mathematical Imagery Trainer: From embodied interaction to conceptual learning”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2011, pp. 1989–1998.

[69] Hezhen Hu, Wengang Zhou, and Houqiang Li. "Hand-Model-Aware Sign Language Recognition". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 2. 2021, pp. 1558–1566.

[70] Hezhen Hu et al. "SignBERT: Pre-Training of Hand-Model-Aware Representation for Sign Language Recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11087–11096.

[71] Earnest Paul Ijjina and C Krishna Mohan. "Human action recognition based on mocap information using convolution neural networks". In: *2014 13th international conference on machine learning and applications*. IEEE. 2014, pp. 159–164.

[72] Kori Inkpen et al. "Exploring spatialized audio & video for distributed conversations". In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 2010, pp. 95–98.

[73] Ellen A Isaacs and John C Tang. "What video can and cannot do for collaboration: a case study". In: *Multimedia systems* 2.2 (1994), pp. 63–73.

[74] Tommi Jaakkola, Satinder P Singh, and Michael I Jordan. "Reinforcement learning algorithm for partially observable Markov decision problems". In: *Advances in neural information processing systems*. 1995, pp. 345–352.

[75] G Tanner Jackson, Chutima Boonthum, and Danielle S McNAMARA. "iSTART-ME: Situating extended learning within a game-based environment". In: *Proceedings of the workshop on intelligent educational games at the 14th annual conference on artificial intelligence in education*. AIED Brighton. UK. 2009, pp. 59–68.

[76] Songyao Jiang et al. "Skeleton aware multi-modal sign language recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 3413–3423.

[77] Tao Jiang, Necati Cihan Camgoz, and Richard Bowden. "Looking for the Signs: Identifying Isolated Sign Instances in Continuous Video Footage". In: *arXiv preprint arXiv:2108.04229* (2021).

[78] Dongsik Jo, Ki-Hong Kim, and Gerard Jounghyun Kim. "Effects of avatar and background types on users' co-presence and trust for mixed reality-based teleconference systems". In: *In Proceedings the 30th Conference on Computer Animation and Social Agents*. 2017, pp. 27–36.

[79] W Lewis Johnson and James C Lester. "Face-to-face interaction with pedagogical agents, twenty years later". In: *International Journal of Artificial Intelligence in Education* 26.1 (2016), pp. 25–36.

[80] W Lewis Johnson, Jeff W Rickel, and James C Lester. "Animated pedagogical agents: Face-to-face interaction in interactive learning environments". In: *International Journal of Artificial intelligence in education* 11.1 (2000), pp. 47–78.

[81] W. L. Johnson and J. Rickel. "Steve: An Animated Pedagogical Agent for Procedural Training in Virtual Environments". In: *SIGART Bull.* 8.1-4 (Dec. 1997), pp. 16–21. ISSN: 0163-5719. DOI: 10.1145/272874.272877. URL: http://doi.acm.org/10.1145/272874.272877.

[82] Hamid Reza Vaezi Joze and Oscar Koller. "Ms-asl: A large-scale data set and benchmark for understanding american sign language". In: *arXiv preprint arXiv:1812.01053* (2018).

[83] Timor Kadir et al. "Minimal Training, Large Lexicon, Unconstrained Sign Language Recognition." In: *BMVC*. 2004, pp. 1–10.

[84] Tuomas Kantonen, Charles Woodward, and Neil Katz. "Mixed reality in virtual world teleconferencing". In: *2010 IEEE Virtual Reality Conference (VR)*. IEEE. 2010, pp. 179–182.

[85] Sara Kiesler. *The hidden messages in computer networks*. Harvard Business Review Case Services, 1986.

[86] Yanghee Kim and Amy L Baylor. "A social-cognitive framework for pedagogical agents as learning companions". In: *Educational Technology Research and Development* 54.6 (2006), pp. 569–596.

[87] Chris L Kleinke. "Gaze and eye contact: a research review." In: *Psychological bulletin* 100.1 (1986), p. 78.

[88] Kenneth R Koedinger, Albert Corbett, et al. *Cognitive tutors: Technology bringing learning sciences to the classroom*. 2006.

[89] Oscar Koller, Hermann Ney, and Richard Bowden. "Deep learning of mouth shapes for sign language". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2015, pp. 85–91.

[90] Oscar Koller et al. "Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos". In: *IEEE transactions on pattern analysis and machine intelligence* 42.9 (2019), pp. 2306–2320.

[91] Robert E Kraut. "Applying social psychological theory to the problems of group work". In: *HCI models, theories and frameworks: Toward a multidisciplinary science* (2003), pp. 325–356.

[92] Robert E Kraut et al. "Understanding effects of proximity on collaboration: Implications for technologies to support remote collaborative work". In: *Distributed work* (2002), pp. 137–162.

[93] OWL Labs. "State of Remote Work 2021". In: (). URL: https://resources.owllabs.com/hubfs/SORW/SORW_2021/owl-labs_state-of-remote-work-2021_report-final.pdf.

[94] J. C. Lafferty, Eady, and J. Elmers. *The desert survival problem*. Plymouth, Michigan: Experimental Learning Methods, 1974.

[95] George Lakoff and Rafael E Núñez. *Where mathematics comes from: How the embodied mind brings mathematics into being*. Basic books, 2000.

[96] Susan J Lamon. "Rational numbers and proportional reasoning: Toward a theoretical framework for research". In: *Second handbook of research on mathematics teaching and learning* 1 (2007), pp. 629–667.

[97] James C Lester, Brian A Stone, and Gary D Stelling. "Lifelike pedagogical agents for mixed-initiative problem solving in constructivist learning environments". In: *User modeling and user-adapted interaction* 9.1 (1999), pp. 1–44.

[98] Dongxu Li et al. "Transferring cross-domain knowledge for video sign language recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6205–6214.

[99] Dongxu Li et al. "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2020, pp. 1459–1469.

[100] Maosen Li et al. "Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition". In: *CoRR* abs/1904.12659 (2019). arXiv: 1904.12659. URL: http://arxiv.org/abs/1904.12659.

[101] Ken Ziyu Liu. *MSG3D*. https://github.com/kenziyuliu/MS-G3D. 2020.

[102] Ziyu Liu et al. "Disentangling and unifying graph convolutions for skeleton-based action recognition". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 143–152.

[103] Stephan Liwicki and Mark Everingham. "Automatic recognition of fingerspelled words in british sign language". In: *2009 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE. 2009, pp. 50–57.

[104] Thomas Maran et al. "Visual attention in real-world conversation: Gaze patterns are modulated by communication and group size". In: *Applied Psychology* (2020).

[105] Joseph Edward McGrath. *Groups: Interaction and performance*. Vol. 14. Prentice-Hall Englewood Cliffs, NJ, 1984.

[106] David McNeill. *Gesture and thought*. University of Chicago press, 2008.

[107] Albert Mehrabian et al. *Silent messages*. Vol. 8. 152. Wadsworth Belmont, CA, 1971.

[108] Dushyant Mehta et al. "XNect: Real-time Multi-Person 3D Motion Capture with a Single RGB Camera". In: vol. 39. 4. 2020. DOI: 10.1145/3386569.3392410. URL: http://gvv.mpi-inf.mpg.de/projects/XNect/.

[109] Ross E Mitchell et al. "How many people use ASL in the United States? Why estimates need updating". In: *Sign Language Studies* 6.3 (2006), pp. 306–335.

[110] Liliane Momeni et al. "Watch, Read and Lookup: Learning to Spot Signs from Multiple Supervisors". In: *Asian Conference on Computer Vision*. 2020.

[111] Bradford W Mott and James C Lester. "U-director: a decision-theoretic narrative planning architecture for storytelling environments". In: *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. ACM. 2006, pp. 977–984.

[112] R Charles Murray and Kurt VanLehn. "DT Tutor: A Decision-TheoreticDynamic Approach for Optimal Selection of Tutorial Actions". In: *International Conference on Intelligent Tutoring Systems*. Springer. 2000, pp. 153–162.

[113] Carol Neidle and Christian Vogler. "A new web interface to facilitate access to corpora: Development of the ASLLRP data access interface (DAI)". In: *Proc. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC*. Vol. 3. Citeseer. 2012.

[114] David T Nguyen and John Canny. "More than face-to-face: empathy effects of video framing". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2009, pp. 423–432.

[115] David T Nguyen and John Canny. "Multiview: improving trust in group video conferencing through spatial faithfulness". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2007, pp. 1465–1474.

[116] Martin A Nowak and David C Krakauer. "The evolution of language". In: *Proceedings of the National Academy of Sciences* 96.14 (1999), pp. 8028–8033.

[117] Brid O'Conaill, Steve Whittaker, and Sylvia Wilbur. "Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication". In: *Human-computer interaction* 8.4 (1993), pp. 389–428.

[118] Claire O'Malley et al. "Comparison of face-to-face and video-mediated interaction". In: *Interacting with computers* 8.2 (1996), pp. 177–192.

[119] Charles A O'Reilly and Karlene H Roberts. "Relationships among components of credibility and communication behaviors in work units." In: *Journal of Applied Psychology* 61.1 (1976), p. 99.

[120] Sergio Orts-Escolano et al. "Holoportation: Virtual 3d teleportation in real-time". In: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM. 2016, pp. 741–754.

[121] Carol Padden. "The deaf community and the culture of deaf people". In: *Readings for diversity and social justice: An anthology on racism, antisemitism, sexism, heterosexism, ableism, and classism* (2000), pp. 343–351.

[122] Ye Pan and Anthony Steed. "A comparison of avatar-, video-, and robot-mediated interaction on users' trust in expertise". In: *Frontiers in Robotics and AI* 3 (2016), p. 12.

[123] J Pearl. "Probabilistic Reasoning in Intelligent Systems". In: (1988).

[124] Barbara Pease and Allan Pease. *The definitive book of body language: The hidden meaning behind people's gestures and expressions*. Bantam, 2008.

[125] Tomislav Pejsa et al. “Room2room: Enabling life-size telepresence in a projected augmented reality environment”. In: *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 2016, pp. 1716–1725.

[126] Suwannee Phitakwinai, Sansanee Auephanwiriyakul, and Nipon Theera-Umpon. “Thai sign language translation using fuzzy c-means and scale invariant feature transform”. In: *International Conference on Computational Science and Its Applications*. Springer. 2008, pp. 1107–1119.

[127] Jean Piaget and Bärbel Inhelder. *The psychology of the child*. Vol. 5001. Basic books, 1969.

[128] Helmut Prendinger and Mitsuru Ishizuka. “The empathic companion: A character-based interface that addresses users’affective states”. In: *Applied Artificial Intelligence* 19.3-4 (2005), pp. 267–285.

[129] Iasonas Kokkinos Rıza Alp Güler Natalia Neverova. “DensePose: Dense Human Pose Estimation In The Wild”. In: 2018.

[130] Rod D Roscoe and Danielle S McNamara. “Writing pal: Feasibility of an intelligent writing strategy tutor in the high school classroom.” In: *Journal of Educational Psychology* 105.4 (2013), p. 1010.

[131] Daniel Roth et al. “Technologies for social augmentations in user-embodied virtual reality”. In: *25th ACM Symposium on Virtual Reality Software and Technology*. 2019, pp. 1–12.

[132] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[133] L Russell et al. “Estimated marginal means, aka least-squares means”. In: *The American Statistician* 34 (2018), pp. 216–221.

[134] Adam Schembri et al. “Building the British sign language corpus”. In: *Language Documentation & Conservation* 7 (2013), pp. 136–154.

[135] Ralph Schroeder. "Comparing avatar and video representations". In: *Reinventing Ourselves: Contemporary Concepts of Identity in Virtual Worlds*. Springer, 2011, pp. 235–251.

[136] Shayle R Searle, F Michael Speed, and George A Milliken. "Population marginal means in the linear model: an alternative to least squares means". In: *The American Statistician* 34.4 (1980), pp. 216–221.

[137] Chris Segrin. "The effects of nonverbal behavior on outcomes of compliance gaining attempts". In: *Communication Studies* 44.3-4 (1993), pp. 169–187.

[138] Leslie D Setlock, Pablo-Alejandro Quinones, and Susan R Fussell. "Does culture interact with media richness? The effects of audio vs. video conferencing on Chinese and American dyads". In: *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. IEEE. 2007, pp. 13–13.

[139] Erin Shaw et al. "Building a case for agent-assisted learning as a catalyst for curriculum reform in medical education". In: *Proceedings of the International Conference on Artificial Intelligence in Education*. 1999, pp. 509–516.

[140] Lei Shi et al. "Adaptive Spectral Graph Convolutional Networks for Skeleton-Based Action Recognition". In: *CoRR* abs/1805.07694 (2018). arXiv: 1805.07694. URL: http://arxiv.org/abs/1805.07694.

[141] Tomas Simon et al. "Hand Keypoint Detection in Single Images using Multiview Bootstrapping". In: *CVPR*. 2017.

[142] Ozge Mercanoglu Sincan et al. "ChaLearn LAP Large Scale Signer Independent Isolated Sign Language Recognition Challenge: Design, Results and Future Research". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2021, pp. 3472–3481.

[143] Mel Slater et al. “Small-group behavior in a virtual and real environment: A comparative study”. In: *Presence: Teleoperators & Virtual Environments* 9.1 (2000), pp. 37–51.

[144] Harrison Jesse Smith and Michael Neff. “Communication behavior in embodied virtual reality”. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. 2018, pp. 1–12.

[145] William Steptoe et al. “Eye-tracking for avatar eye-gaze and interactional analysis in immersive collaborative virtual environments”. In: *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM. 2008, pp. 197–200.

[146] William Steptoe et al. “Lie tracking: social presence, truth and deception in avatar-mediated telecommunication”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2010, pp. 1039–1048.

[147] William C Stokoe, Dorothy C Casterline, and Carl G Croneberg. *A dictionary of American Sign Language on linguistic principles*. Linstok Press, 1976.

[148] Susan G Straus. “Getting a clue: The effects of communication media and information distribution on participation and performance in computer-mediated and face-to-face groups”. In: *Small group research* 27.1 (1996), pp. 115–142.

[149] Susan G Straus and Joseph E McGrath. “Does the medium matter? The interaction of task type and technology on group performance and member reactions.” In: *Journal of applied psychology* 79.1 (1994), p. 87.

[150] Andrea Tartaro and Justine Cassell. “Authorable virtual peers for autism spectrum disorders”. In: *Proceedings of the Combined workshop on Language-Enabled Educational Technology and Development and Evaluation for Robust Spoken Dialogue Systems at the 17th European Conference on Artificial Intellegence*. 2006.

[151] Du Tran et al. “A closer look at spatiotemporal convolutions for action recognition”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018, pp. 6450–6459.

[152] Carol Bloomquist Traxler. “The Stanford Achievement Test: National norming and performance standards for deaf and hard-of-hearing students”. In: *Journal of deaf studies and deaf education* 5.4 (2000), pp. 337–348.

[153] Michael Twyman. “Using pictorial language: A discussion of the dimensions of the problem”. In: *Designing usable texts*. Elsevier, 1985, pp. 245–312.

[154] K. Vanlehn et al. “The Andes physics tutoring system: Lessons learned”. In: *International Journal of Artificial Intelligence in Education* 15.3 (2005), pp. 147–204.

[155] Gül Varol et al. “Read and Attend: Temporal Localisation in Sign Language Videos”. In: *CVPR*. 2021.

[156] Manuel Vazquez-Enriquez et al. “Isolated Sign Language Recognition With Multi-Scale Spatial-Temporal Graph Convolutional Networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 3462–3471.

[157] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: `https://www.stats.ox.ac.uk/pub/MASS4/`.

[158] Peter H Waxer. “Nonverbal cues for anxiety: An examination of emotional leakage.” In: *Journal of abnormal psychology* 86.3 (1977), p. 306.

[159] Steve Whittaker. “Theories and Methods in Mediated Communication: Steve Whittaker”. In: *Handbook of discourse processes*. Routledge, 2003, pp. 246–289.

[160] Ronnie Wilbur and Avinash C Kak. “Purdue RVL-SLLL American sign language database”. In: (2006).

[161] Nelson Wong and Carl Gutwin. "Support for deictic pointing in CVEs: still fragmented after all these years'". In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 2014, pp. 1377–1387.

[162] Yuxin Wu et al. *Detectron2*. `https://github.com/facebookresearch/detectron2`. 2019.

[163] Sijie Yan, Yuanjun Xiong, and Dahua Lin. "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition". In: *CoRR* abs/1801.07455 (2018). arXiv: `1801.07455`. URL: `http://arxiv.org/abs/1801.07455`.

[164] Boram Yoon et al. "The Effect of Avatar Appearance on Social Presence in an Augmented Reality Remote Collaboration". In: *IEEE VR*. ACM. 2019.

[165] Morteza Zahedi et al. "Combination of tangent distance and an image distortion model for appearance-based sign language recognition". In: *Joint Pattern Recognition Symposium*. Springer. 2005, pp. 401–408.

[166] Morteza Zahedi et al. "Continuous sign language recognition-approaches from speech recognition and available data resources". In: *Second Workshop on the Representation and Processing of Sign Languages: Lexicographic Matters and Didactic Scenarios*. 2006, pp. 21–24.

[167] Da Zhang et al. "S3D: single shot multi-span detector via fully 3D convolutional networks". In: *arXiv preprint arXiv:1807.08069* (2018).

[168] Yuxiao Zhou et al. "Monocular Real-time Hand Shape and Motion Capture using Multi-modal Data". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 0–0.