Title	Nonverbal communication as rich interaction method with virtual agent in interactive virtual reality content
Sub Title	
Author	Stevanus, Kevin Kunze, Kai
Publisher	慶應義塾大学大学院メディアデザイン研究科
Publication year	2019
Jtitle	
JaLC DOI	
Abstract	
Notes	修士学位論文. 2019年度メディアデザイン学 第719号
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO40001001-00002019-0719

Master's Thesis
Academic Year 2019

Nonverbal Communication as Rich Interaction Method with Virtual Agent in Interactive Virtual Reality Content

Keio University
Graduate School of Media Design

Stevanus Kevin

A Master's Thesis
submitted to Keio University Graduate School of Media Design
in partial fulfillment of the requirements for the degree of
Master of Media Design

Stevanus Kevin

Master's Thesis Advisory Committee:

Professor Kai Kunze	(Main Research Supervisor)
Project Senior Assistant Professor Roshan Lalintha Peiris	(Sub Research Supervisor)

Master's Thesis Review Committee:

Professor Kai Kunze	(Chair)
Project Senior Assistant Professor Roshan Lalintha Peiris	(Co-Reviewer)
Professor Kazunori Sugiura	(Co-Reviewer)

Abstract of Master's Thesis of Academic Year 2019

Nonverbal Communication as Rich Interaction Method with Virtual Agent in Interactive Virtual Reality Content

Category: Design

Summary

Alongside recent rapid growth of VR in market, VR games has also considerably evolved. One of the emerging trend in VR game design is emphasizing element of interaction between user and virtual agent (VA). VR provides user with state-of-art immersive First Person Point of View, which immensely elevates immersion of interaction with VA compared to conventional screen based games. In current state however, these VR contents still a lot to be desired, especially in term of VA's behavior. Even though user provided with immersive visual, VAs on these contents are still only responding on button pressed by user instead of actively responding to user's actual action, potentially reducing user's immersion. In order to create a more immersive and engaging interaction, we propose concept of nonverbal communication as an input channel.

In this work, we introduced a real-world context aware VA that capable to recognizes and appropriately responds toward user's gaze and gesture, simulating real human behavior. Through our demo content, we showcased this approach to user and conducted studies on how this interaction models affects user's interaction experience with the VA. The result from two of our studies shows that all subjects had an overall better interaction experience with the VA compared to conventional VR interaction method.

Keywords:

Human-Computer Interaction, Virtual Reality, Virtual Agent

Keio University Graduate School of Media Design

Stevanus Kevin

Contents

Acknowledgements **vi**

1 Introduction **1**
1.1. Current State of Virtual Reality Content 1
1.2. Nonverbal Communication for Realistic User-VA Interaction . . . 2
1.2.1 Gaze 3
1.2.2 Gesture 4
1.3. Contribution 4
1.4. Structure 5
Notes 6

2 Related Works **7**
2.1. Eye Tracking as Input 7
2.2. Gesture as Input 9
2.3. Rich Interaction with Virtual Agent 10
2.4. Interaction with Virtual Agent in Virtual Reality 13
2.5. Summary 14
Notes 15

3 Concept Design **16**
3.1. Concept 16
3.2. Demo Content 17
3.3. The Teacher as Real World Context Aware Virtual Agent 20
3.3.1 Gaze 20
3.3.2 Gesture 22
Notes 27

4 Implementation **28**
4.1. System Architecture 28
4.1.1 System Overview 28
4.1.2 Hardware 28
4.1.3 Software 32
4.2. Prototypes 34
4.2.1 First Prototype (Virtual Gaze) 34
4.2.2 Second Prototype (Current Version) 34
Notes 35

5 Evaluation **37**
5.1. Pilot Study (First Prototype) 37
5.2. User Study (Second Prototype) 38
Notes 43

6 Conclusion **44**
6.1. Conclusion 44
6.2. Limitation 45
6.3. Extension 45
6.4. Future Works 46
Notes 46

References **47**

List of Figures

1.1 "The Inpatient" by Supermassive Games 2
1.2 "Summer Lesson" by Bandai Namco 2
1.3 VA initiating eye contact during conversation with player in "Summer Lesson" 4

2.1 Illustration of eye tracking process[1] 8
2.2 "L.A. Noire" by Rockstar Games 10
2.3 The Royal Corgi, a game of social gaze (Vidal et al. 2015) 12

3.1 Interaction with The Teacher in Demo Content 19
3.2 Pupil Labs HTC Vive Eye Tracking Add-on 23
3.3 Example of symbolic gestures (De Stefani et al. 2013) 25

4.1 Illustration of System Architecture and Data Flow 29
4.2 Gesture tracking and recognition by Kinect 31
4.3 Real-time eye tracking video feed by Pupil Capture 33
4.4 Virtual Gaze features 35
4.5 Current version setup, Kinect is facing user from front 36

5.1 Result of Virtual Gaze initial user test 38
5.2 Test subject experiencing our model on user test 39
5.3 Result of user test 41

List of Tables

3.1 List of implemented gaze based interaction concept 22

Acknowledgements

First and foremost, I want to express my deepest gratitude to my family, who has supported me unconditionally, both on my study and my life. Without their support and encouragement, I would not imagine myself continuing my study abroad in Japan, let alone staying for 3 years and finishing my graduate study in KMD.

I wish to thanks all of members of my real project, GEIST, especially Professor Kai Kunze who have guided and inspired me to be a better researcher during my study in KMD. Through his guidance and encouragement, I managed to broaden my horizon and pushed my limit, achieving things I never though I was capable of. I am also want to express my deepest gratitude to two of senior members of GEIST, Yun Suen Pai and Takuro Nakao, who have became a splendid mentors and good friends for me.

This whole work also would not be finished in a way I wished it to be without help from Assistant Professor Roshan Peiris, who provided me with his insight and knowledge to improve this work significantly.

Finally, I would also extend my appreciation towards all of fellow Master student of KMD 2017 September batch. Times I spend with these brilliant peoples has continuously droves me to be better. Their help and supports also largely contributes to smooth and rewarding finish of this work.

Chapter 1
Introduction

1.1. Current State of Virtual Reality Content

Extended Reality (XR) experiences, including Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR), is slowly but steadily growing as a digital media in various fields with entertainment, especially video games, undeniably being the most prominent form of implementation. In current video game industry, VR especially has become a well-established platform with several major developers released their own video game focused VR Head Mounted Device (HMD), such as Sony with Playstation VR[1], Oculus with Oculus Rift[2], and HTC with Vive[3]. Alongside continuous release of VR HMDs, amount of VR-supported game titles also steadily increasing with some high profile video game companies such as Bandai Namco Studio[4] and Capcom[5] releasing VR-supported version of their flagship IPs, indicating interest towards VR development from developers' side.

Due to it's strength in providing user with immersive virtual environment, first person Field of View is one of the most common discerning feature of VR games to simulate a realistic user experience. With this state-of-art immersive first person FOV experience as a focus, there are various emerging innovations in term of game design being enabled by utilizing VR technology. One of the interesting potential is immersive interaction with Virtual Agent (VA). Some of currently available VR games on market, such as "Summer Lesson" by Bandai Namco Games[6] and "The Inpatient" by Supermassive Games[7] are providing new kind of game genre by focusing on interaction between player and in-game character in first person FOV, simulating real-life like social interaction. While arguably similar type of interaction already exists in conventional game console, VR immersive first-person FOV experience provides new height of realism compared to conventional single screen based game or movie experience.

Figure 1.1 "The Inpatient" by Supermassive Games

Figure 1.2 "Summer Lesson" by Bandai Namco

At current stage however, previously mentioned games still leaves a lot to be desired, especially in term of agent's behavior. Being restricted by conventional input modality through conventional controller such as mouse - keyboard or gamepad, most of agent's action are simply waiting for player's input instead of being actively reacting towards user's action like how would real human would, which could results in reduction of immersion in term of interaction. For example, in "Summer Lesson", character "Allison" as an agent who players interacts with does not act differently whether player looking at her or not when she is talking to player, which seemingly unrealistic when compared to how normal people expected to react. This is completely understandable matter due to current limitation of commercially available popular VR HMD.

1.2. Nonverbal Communication for Realistic User-VA Interaction

Typical communication between humans consists of two aspect, verbal communication and nonverbal communication. Verbal communication commonly refers to words we use to communicate, whereas nonverbal communication refers to communication that is produced by means other than words (Argyle 1972). While both compliments each other, several previous studies reported that nonverbal behaviours, such as gestures, facial expressions, the way we use our voice, plays a more significant role during an interaction than its verbal counterpart (Mehrabian et al. 1971) (Wood 1972). Professor Mehrabian concluded that nonverbal aspect

of communication holds more importance when communicating matters of affection and attitude, stating verbal communication only holds 7% of importance, compared to 38% of vocal element and 55% of facial expression (Mehrabian et al. 1971). Knowing importance of nonverbal communication on human-to-human communication, nonverbal communication shows potential to help us achieve our goal of creating a immersive and engaging interaction model with VA.

However, nonverbal communication consists of various form of cues, and to implement all of it as HCI input method requires immense effort and complex setup. During the course of this work, we tackle this issue by incrementally implementing these elements of nonverbal communication. In current state of this work, we have implemented and tested two elements of nonverbal communication, in form of *gaze* and *gesture*.

1.2.1 Gaze

We started this project as gaze based input HCI study. While playing VR game titles we mentioned previously, one particular interesting similar point is how strong feeling of eye contact we felt initiated by the virtual agents, unlike any experience we had through conventional non-VR movies and games. As these games focused in interaction with VA, VAs in these games are designed to behave like how real human would during human-to-human interaction, including making eye contact with player while initiating conversation. Combined with VR's immersive First Person FOV, we believe this strong sensation of eye contact is a VR content's distinctive powerful tool to enhance user's immersion and feeling of co-presence with VA.

At current state of these games however, this "virtual eye contact" with VA is unfortunately only works one-way, as gaze initiated by user do not get recognized by the VA. From technical perspective, this is totally understandable as nearly all of the common consumer VR HMD do not have eye-tracking functionality installed, which makes it impossible to detect user's gaze. However, with recent surge of consumer level eye-tracking enabled VR HMD, we could see how this approach shows a lot of promises as default feature of VR contents in near future. Looking at lack of works addressing this topic, we consider use of user's gaze as an untapped potential to further enhance user's interaction with VA in VR content

Figure 1.3 VA initiating eye contact during conversation with player in "Summer Lesson"

and motivated us to start this work.

1.2.2 Gesture

Implementation of gesture as input module was a result of observation of test subjects behaviour during our user test. On our initial user test of gaze-based VR interaction model, two out of four of test subjects performed symbolic body gestures (waving hand, nodding head), which is also an element of nonverbal communication, even though they were not instructed to and informed that only their gaze will affect the VA. This occurrence lead us to assumption that body gestures is one of the more intuitive and natural nonverbal communication in interaction with VA in VR content. This finding also drove us to chose gesture as our second element of nonverbal communication to be implemented in our interaction model.

1.3. Contribution

This thesis mainly aims to explore potential implementation of elements of human-to-human nonverbal communication in order to enhance human-to-VA interaction

experience in VR environment. The outcome of this thesis is a interactive multimodular VR interactive content which demonstrates how our designed interaction model work, and comparison results with conventional VR human-to-VA interaction model.

To summarize it, following are contributions this work provides:

1. We propose user's nonverbal communication as a feasible interaction modality between user and virtual agent on VR platform.
2. Through our demo content, we present an example on how the concept of nonverbal communication could be implemented in VR content and how it positively impacts user's experience.

1.4. Structure

This thesis consists of 6 chapters, as following:

1. Introduction, which explains background and motivation behind this thesis.
2. Related Works, which shows preceding works and research related to aspects of this thesis.
3. Concept Design, which explains various concepts and decisions regarding our model's design process.
4. Implementation, which describes technically how our model was developed and set up.
5. Evaluation, which describes user tests we did in order to evaluate our model and it's result.
6. Conclusion, which states conclusion from this thesis and discussion regarding model's extensibility as well as future works.

Notes

1 Playstation VR : https://www.playstation.com/en-ae/explore/playstation-vr/

2 Oculus Rift : https://www.oculus.com/rift/

3 HTC Vive : https://www.vive.com/eu/

4 Bandai Namco Studio : https://www.bandainamcostudios.com/

5 Capcom : http://www.capcom.com/

6 Summer Lesson by Bandai Namco Games : https://summer-lesson.bn-ent.net/

7 The Inpatient by Supermassive Games : https://www.supermassivegames.com/games/the-inpatient

Chapter 2
Related Works

This chapter will describe previous related works, from academic studies as well as from industry or commercial products, which we deemed relevant to various elements of our work.

2.1. Eye Tracking as Input

In study of psychology, what a person is looking at is assumed to indicate thought "on top of the stack" of that person's cognitive processes (Just and Carpenter 1976). This hypothesis also means that provided a record of eye-movement and fixation, we could measures dynamic trace of where said person's attention is being directed in relation to a visual display.

Furthermore, Poole et. al (Poole and Ball 2005) in their work explained how this concept could be implemented in field of Human-Computer Interaction (HCI). For example, in a task scenario where participants are asked to search for an icon, longer-than-expected gaze on the icon before eventual selection would indicate that it lacks meaningfulness, and probably needs to be redesigned. On the same work, Poole also set on two main metrics used for eye tracking research, *fixations*, which are moments when the eyes are relatively stationary, taking in or "encoding" information, and *saccades*, which are quick eye movements occurring between fixations.

With rapid development of eye-tracking technology in recent years, eye-based interfaces has became more common, both in research field as well as in commercial application. In conventional desktop based scenario, various studies have already explored implementation of user's gaze for common practical HCI interaction such as target selection (Kumar et al. 2007) and text entry (Møllenbach et al. 2013). These studies concluded that user's gaze is reliable enough for each purposes,

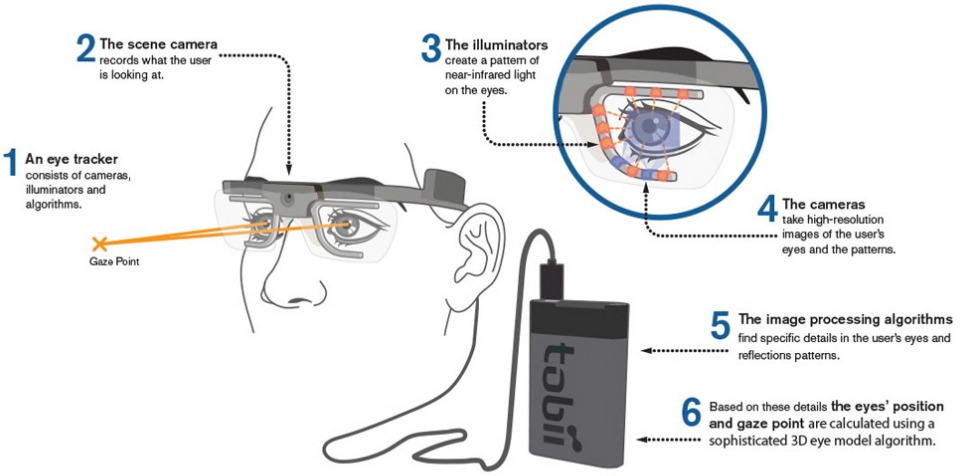

Figure 2.1 Illustration of eye tracking process[1]

showing promises of gaze as a potential feasible input modality in field of HCI. Typical recent eye tracking sensor for HCI purposes works by simultaneously emitting near-infrared light towards user eyes and capturing visual image of user's eyes, then based on those information, image-processing algorithm will calculate user's approximate gaze direction.

Eye tracking feature in VR environment on other hand, is still relatively new yet fast growing. Multiple VR HMD developers started to explore potential of eye tracking input in tandem with VR by developing VR HMD with built-in eye tracking functionality, such as Fove Inc. with FOVE [2] and HTC with Vive Pro Eye [3]. Eye tracking feature in VR HMD commonly applied by installing set of optical camera facing user eyes around each lenses to visually record user's eye direction and orientation. Utilizing this particular setup, multiple works used recording of user's eye for various purposes such as facial reenactment (Thies et al. 2018) and facial expression classification (Hickson et al. 2019).

2.2. Gesture as Input

Motion gestures (as opposed of pen gestures) was not commonly used for general computer or device operation (Ashbrook and Starner 2010). In video game industry however, body gestures have been implemented in several major video game console such as Nintendo Wii [4] and Microsoft Kinect for Xbox [5]. Since it's initial release for Xbox 360 on 2010, Kinect also has made available openly for desktop environment. A Kinect for Xbox One, the latest iteration of Kinect product line, features a RGB Camera and Infrared Depth-Finding Camera which capables of tracking up to 6 complete skeletons as well as position and orientation of 25 joints including thumb. Gesture recognition by Kinect is done by processing information of relative position and orientation of each bones and joint through machine learning process to match it's combination with prior trained gesture data. This capability of reliably tracking whole body movement plus it's off-the-shelf nature has made Kinect one of most popular sensor in gesture-based studies.

While actual body gesture consisted of movement of almost all of body parts, not all body parts are commonly explored and used for HCI purposes. Hand gesture input, as a smaller part of gesture-based input study, is one of the most explored direction in HCI study. One of the reason is hand being one of more dynamic and distinct part of the body for visual capture / camera based gesture detection approach. For example, Chai et al. (Chai et al. 2013) utilized visual information of user's hand shape and orientation as a natural input method to detect sign language and then translates it into text and speech. Agrawal et al. (Agrawal et al. 2013) brought this approach further by adding visual detection of user's head movement, specifically detecting user's head nod (vertical movement) and head shake (horizontal movement) to form a multimodal gesture based input system. Both of this works is taking a computer vision approach, by utilizing a camera to capture visual information of user's body part (hand and head) and applies algorithm to classify specific gestures. Both works also reported a high accuracy rate of gesture detection by their system, validating feasibility of gestures as input modality in HCI.

Figure 2.2 "L.A. Noire" by Rockstar Games

2.3. Rich Interaction with Virtual Agent

In field of Human-Robot Interaction (HRI), implementation of nonverbal cues, especially bodily gestures and gaze, are being studied and implemented in two directions. First direction is when the robot as agent reproduces a bodily cues in order to elicit more information or emotional response from the human side. Multiple works has explored this approach with generally positive results. Chidambaram came to a conclusion that a robot that interacts with nonverbal bodily cues have more effect on the compliance compared to a robot that only interacts with vocal cues alone (Chidambaram et al. 2012). In field of HCI, this approach also already relatively well explored. Eye gaze information inferred through avatar's head and eye movement found to improves quality of communication between multiple users that interacts through avatars (Garau et al. 2001). In video games, L.A. Noire[6] for example utilizes in-game character's realistic gaze and body gestures to challenge player to decide whether that character is lying or not.

Second direction is when the agent is capable of detecting nonverbal cues from the human side and response accordingly, which is also the direction we will apply for this work. Several study implemented human gaze as an interaction modality

with robot for various purposes such as conversation (Miyauchi et al. 2004) (Mutlu et al. 2009) and collaboration (Yoshikawa et al. 2006).

In scope of interaction with virtual character, one particular common use-case scenario of this approach is creation a conversational agent that capable of identifying and measuring user's attention by tracking user's gaze. Ishii et al. (Ishii et al. 2013) for example, developed a conversational virtual agent that capable to detect subject's disengaged state through their gaze. Through their experiment, they drew a conclusion that VA that probes subjects whenever they detect disengagement results in a improved subject's impression and engagement rating compared to VA that probes periodically. Furthermore, similar positive result also stated by Wang et al. (Wang and Gratch 2010), which developed a head gesture and gaze aware conversational VA. In a similar fashion, they compared subject's interaction experience against VA with 3 different behaviours, which are VA that behaves adapting toward subject's gaze and gesture (named "Rapport Agent"), VA that only continuously gazing at subject (named "Staring"), and VA that only gaze at subject direction occasionally (named "Ignoring"). The experiment's result align with Ishii's experiment result, stated that "participants experienced more rapport when the virtual representation of their conversation partner showed more attention, positivity and coordination: participants interacting with the Rapport Agent had greater subjective experience of rapport and exhibited more fluent speech when compared to an agent that only exhibited attention (Staring Agent) or an agent that exhibited none of the constituents of rapport (Ignoring Agent)". Both of these works successfully improved certain aspects of user's interaction with VA by implementing "real world context awareness" element to their VAs. Utilizing this context-awareness, VAs in both of these works were capable to enhance each subject's interaction experience by providing appropriate visual and/or audio feedback based on their nonverbal cues.

The Royal Corgi

Utilizing this concept in a more application oriented study, Vidal et al. (Vidal et al. 2015) developed "The Royal Corgi", a first-person social game experience that simulates concepts of social gaze inside the game. Vidal suggested concept of *social gaze interaction* as a way to augment user experience by making the

Figure 2.3 The Royal Corgi, a game of social gaze (Vidal et al. 2015)

computer react to the user's gaze in typical human-like reactions, with the aim to render interactions more immersive, natural, and make the user aware of the power of their own gaze. In "The Royal Corgi", player is being asked to initiate conversation with various Non-Playable Character (NPC) which each have unique personality and reacts differently towards player gaze. For example, whenever player talks to The Horse Instructor character which is an influential and proud character, player needs to be humble in front of her, and lower their eyes often while talking to her, otherwise she will take it as a lack of respect. On contrary, while talking with The Archivist character which has low self-esteem, the player is required to dominate him by keep staring at him in order to get his favour. In this case, user's gaze as an real-world context provides additional or alternative input to further customs VA's behavior to be more realistic and reactive towards user's condition. Through their study, they reported that test subjects describes social gaze interaction as "feels natural", "immersive", and "provides strong feelings of embodiment".

While this work is strongly resembles our goal, unlike our target platform of VR, this work was done in a conventional non-VR desktop PC environment. VR experience requires different hardware setup and application design compared

to conventional non-VR experience, which potentially could lead to completely different study result. Furthermore, they also described their system limitation of "our characters were not able to move their own eyes", which is an interesting point to be considered by us when we develop our own system.

2.4. Interaction with Virtual Agent in Virtual Reality

VR experience provides user with considerably more immersion compared to conventional desktop screen based experience. Fundamentally, commercially available VR HMD such as Oculus Rift[7] and HTC Vive is designed to completely covers user's vision and replaces it with immersive virtual view through lenses in front of user's both eye. Capitalizing this approach, majority of VR applications tend to adapts First-Person Field of View (FOV) to help evoke immersion and user's embodiment inside the content.

In field of HCI, VR itself is not completely new, dating back as far as 1968 (Sutherland 1968). However, as the technology was hardly accessible, it was not widely explored up until recent surge of commercially available VR HMDs. This last 5 years has seen vast growth in VR related study as well as it's implementation for various applications, including it's effect to human-VA interaction. One of the earliest study regarding human-VA we found was done by Rickel et al. (Rickel and Johnson 1998), who integrated an intelligent interactive VA, named STEVE, to assist user on doing procedural task in VR environment. STEVE is capable to understand basic questions from user such as "What should I do next?" and "Why?", then provides guide to user in form of speech, gaze, and gesture.

One common objective of implementing interaction with VA in VR enviroment we found in multiple study is to invoke or to control user's emotion. Bosse et al. (Bosse et al. 2018) in their study developed a "bad guy" VA in a VR environment with a goal to study how interaction with it could induce anxiety and stress to users. In this study, they created a VA they described as "intelligent virtual agents that take a negative or even aggressive stance towards the user". Combined with shock device to represent being hit in VR content as physical consequence in real environment, this experiment successfully increases user anxiety and stress. In

other study, Hartanto et al. (Hartanto et al. 2015) also take similar approach of using IVA, acting as a health support, as a way to control user's emotion. Taking into account user's real time physiological information (heart rate), the VA then verbally engages with user in a certain way to bring user's anxiety to a desired rate. They also reported to successfully raises and reduces user's rate of anxiety through this verbal interaction with IVA. Based on result of this two studies, we can draw a conclusion that similar to conventional desktop based HCI, interaction with VA in VR environment is a viable method to further enhance user's experience.

2.5. Summary

Based on literature review of related works we explained above, we gained positive indications regarding various aspects of our designed approach. First, multiple HCI studies explored use of user's gaze and gesture for various computer operation purposes and reported to performed with adequate accuracy rate. While not necessarily used in context of nonverbal communication, this indicates that both user's gaze and body gesture is indeed a feasible input method for HCI purposes, which supports our plan on using both as parts of our interaction model.

Second, multiple studies from both HCI and HRI precedently tried to incorporate element of nonverbal communication as additional interaction modality with VA, including gaze and gesture. Most of these works resulted in a positive note, either by improving user's engagement or increasing user's rapport towards the VA. One of the most significantly related work with our design in mind is "The Royal Corgi", which utilized user's gaze to simulates concept of social gaze in interaction between user and VA in a storytelling game. Test subject of this study reportedly expressed sensation of naturalness and immersion, showing promise of this kind of approach in increasing overall user's experience. However, most of these studies were conducted in a desktop computer based setup, which could resulted in different result compared to VR based setup experience we are trying to develop.

VR based human-VA interaction study itself while not completely non-existent, is relatively unexplored and most of the work we found was published during last 5 years. Through studies we found regarding human-VA interaction in VR, we

found that specifically designed VA is capable to induce or control certain feeling to user, which then being used for various purposes such as therapy and storytelling. This finding directly supports our approach of using specifically designed VA's behavior to enhance user's experience in VR content.

Notes

1 https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/how-do-tobii-eye-trackers-work/

2 https://www.getfove.com/

3 https://www.vive.com/eu/product/vive-pro-eye/

4 http://wii.com/

5 https://developer.microsoft.com/en-us/windows/kinect

6 L.A Noire by Rockstar Games : https://www.rockstargames.com/lanoire/

7 Oculus Rift : https://www.oculus.com/rift/

Chapter 3
Concept Design

This chapter will introduce our main design concept and design process, as well as various aspects behind our demo content, result artifact of this project which we use to showcase our model's features to users. Furthermore, for better understanding of each aspect of our model, we will also elaborate on decision behind those aspects, as well as various relevant elements such as our target user and hardware of choice.

3.1. Concept

From literature review process we summarized on chapter 2, we found 2 important finding which supports our approach of using nonverbal communication to improves user's interaction experience with VA in VR content. First, we found that nonverbal cues has been previously utilized as input modality for HCI purposes and performed with high rate of accuracy. This finding rationalized our decision of developing a nonverbal cues based input modality. However, nonverbal cues encompassed wide array of actions, therefore choosing the most effective and appropriate cues is imperative. In the beginning of this project, we started with only gaze as our modality as we considered gaze to be one of the most prominent implicit element of nonverbal communication. In context of HCI, user's gaze is also holds a lot of significance, as it indicates various such as user's Region of Interest (ROI), which lead us to prioritize gaze as our first nonverbal cues of choice. During pilot study of our first prototype, we then found another interesting occurrence where some of the subjects, in addition of using gaze interaction, were actually doing gesture to interact with the VA without being instructed. This lead us to incorporate bodily gestures, which is also part of nonverbal communication, as additional input in our current version of work besides gaze. By

applying multiple input modality, we are expecting even better improvement of user's experience.

Second, we also found that VA that specifically designed to reacted appropriately towards inputs from user are capable to induce certain feeling to user as well as increasing user's immersion towards the interaction. Based on this finding, we see a potential method of improving user's interaction by developing our VA as a "real-world context aware" agent, which acts as if it recognizes user's actions in real world. By responding appropriately towards user's nonverbal cues we mentioned above, we are expecting our VA to be able to provide a better interaction experience.

These 2 key points acts as main components of our model, hardware setup that capable to reliably recognize user's nonverbal cues (gaze and gesture) and VA that capable to reacts or gives feedback to user appropriately based on those cues.

3.2. Demo Content

In order to showcase how our interaction model works as well as to compares it with conventional VR interaction model, we created an interactive VR content. On deciding what kind of scenario would be appropriate and could effectively shows how each features works, there were 2 main consideration we set upon. Our first priority was looking for a specific real life scenario in which both gesture and social gaze action could be incorporated. Second, we also tried to decide on a scenario which our target user could generally relates to certain extent.

Based on these consideration, we created a story that revolves around interaction between a student and his/her teacher. Formal communication against someone with higher or respected position often involves sets of nonverbal manners, such as looking at said person eyes while he/she is talking to indicates respect and attention, compared to casual communication against someone with same social standing. Such scenario provides us with a lot of opportunity to utilize our gesture and gaze based input, acting as natural input method to simulate those manner. During the whole content, user will plays a role of a nameless high school student in a first-person point of view. We choose a role of user as a high school student as it is a generally common real life experience and relatable by most of

our target user.

Main objectives of this demo content for users is simply to interact with "The Teacher" with gaze and gesture. The whole story takes place in a closed classroom setting and approximately lasts for 3 minutes. Basic plot of the story is about how The Teacher call his student (player character) to a meeting on classroom to discuss about his/her bad grade, and what he/she he need to do in order to not fail a grade. The Teacher's dialogue and animation is designed in a way to encourage gestural response, such as waving and calling from a distance to attract user to wave back in response, or asking a yes or no question to attract affirmation/negation response.

As user's gaze direction is continuously tracked and heavily matters in our model, text based User Interface such as subtitle could potentially cause a distraction and be a liability to user's experience. To avoid that matter as well as to create a more immersive interface, we provided a voice over for each of The Teacher's dialogue. Additionally, to give user's reaction time to give a gestural response to The Teacher, we set a 10 second maximum pauses between each dialogues.

Addressing limitation of "The Royal Corgi" experience described on chapter 2, in where their characters cannot move their eyes, we added to our character functionality to dynamically move his eye depending on situations. For example, when talking to user, he will direct his eyes to user's direction, and when he is indicating certain object to user, he will move his eye direction towards that object. By doing so, we tried to give user a clear implication and indication on how they expected to react on certain condition. In similar fashion, we also equipped The Teacher with multiple animations to evoke gesture from users during the demo. For example, in a certain part of the story, he will waving his hand to user from distance, encouraging user to wave back.

Finally, based on all requirement and specification of the application part, we need to decide on what kind of device we will use to show this demo content to user. As our demo content is designed as a VR content, first and foremost VR HMD is most important hardware part to be considered. However, looking at the hardware requirement to execute all basic feature (except eye tracking and body gesture detection), there is no specific requirement that requires us to use special

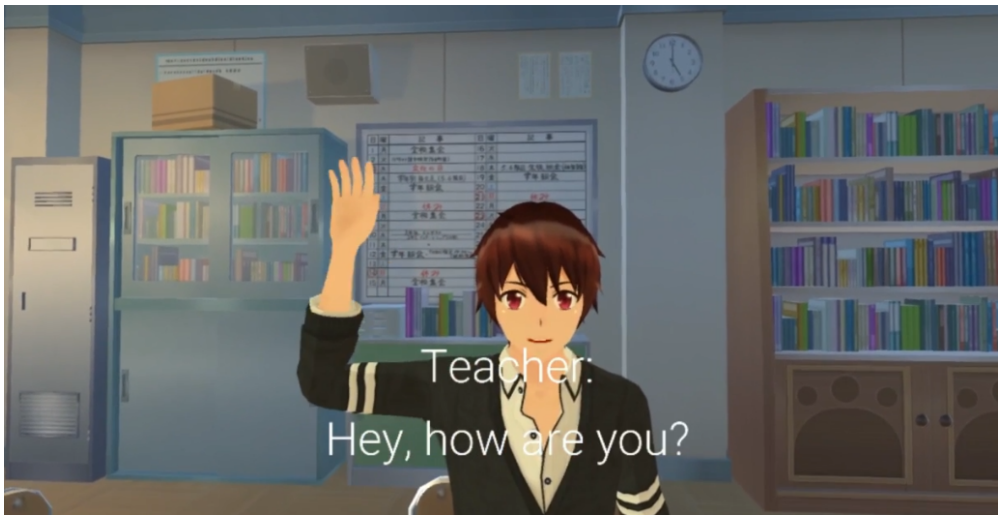

Figure 3.1 Interaction with The Teacher in Demo Content

VR HMD. Therefore, we decided to use HTC Vive[1], which is one of the high-end VR HMD that commercially available. First reason why we chose to use HTC Vive is because it has six degrees of freedom (6DoF) capability, which allows user to move around in VR environment by moving their actually moving their body. While our whole demo content is designed to be a seated experience, 6DoF could provide user with more immersion by showing user head movement and rotation (as opposed of only rotation with 3DoF) in VR environment. Second, being one of the most used VR HMD globally, HTC Vive is highly customisable and has a lot of dedicated 3rd party functional add-on, including eye tracking sensor we are using. Last, HTC Vive supports open source SteamVR SDK[2], which provides collection of libraries to support development of VR application.

Target User

Our demo content, which designed as an interactive VR content, was developed with two groups of target users in mind. The first and our main target user is active VR content consumer. We set this group of users as our main target because fundamentally our goal is to improve user experience in VR content. This group consists of generally young adult ranging between age of 16 to 34 years old [3] with decent amount of experience consuming VR content. With prior experience of

watching or playing VR content, users from this group are expected to gain most benefit by using our interaction model.

Second group of our target user is active video game consumers. Emphasizing focus on the interaction part, we consider video game as one of most prominent application for our interaction model. Similar with previous group, this group also mostly consists of young adult ranging between age of 18 to 35 years old [4] which regularly plays video game. This group of user is however considerably larger than previous group in number as video game in general is more accessible and commonly consumed by wider array of people.

3.3. The Teacher as Real World Context Aware Virtual Agent

As already described previously on chapter 2, VA that aware of and behaves accordingly towards user's nonverbal communication proved to improves overall user's interaction experience by giving appropriate visual and audio feedback. Based on this theory, we are also taking this approach of implementing real world context aware VA to improves user's experience during our demo content. In our demo content, "The Teacher" acts as a real world context aware virtual agent as he designed to be aware of real world context in form of user's gaze and body gestures.

3.3.1 Gaze

Our main objective of implementing gaze detection feature is to simulate element of social gaze on interaction with VA in VR environment. Therefore, first thing needed to be decided was what kind of social gaze concepts we want to implement on our demo content to showcase the feature to user. For this, we referred to Royal Corgi experience (Vidal et al. 2015) with some adjustment to fit story of our demo content. Summary of implemented concept with brief description could be seen on table 3.1. In total, there are 5 concepts of social gaze we chose to implement to the demo content, which are as following :

- *Seeking for interaction.* Looking at someone could be interpreted as desire

for interaction with the one who being looked at (Frischen et al. 2007). In our demo content, we implemented this concept by using it as a part trigger to advance the story when The Teacher calls user's from a far, symbolising The Teacher's acknowledgement of user's attention. This event triggered when user's gaze is directed at The Teacher direction when he call the user.

- *Cultural disrespect.* Certain cultures may interprets excessive direct eye contact as a sign of lack of respect (Frischen et al. 2007) (Argyle and Cook 1976). In our demo content, we implemented this concept as an optional event where The Teacher will act distressed whenever user stares at his face continuously for a while. This event triggered when user's gaze is directed at The Teacher's face direction continuously for 10 seconds.

- *Signs of intention.* While engaged in an interaction, people may also infer knowledge from monitoring the other person ' s gaze and predict what they are interested in or about to do (Castiello 2003). In our demo content, we implemented this concept as an optional event where The Teacher will ask user's whether he/she is on hurry whenever user stares at a clock on wall continuously for a while. This event triggered when user's gaze is directed at clock on wall direction continuously for 10 seconds.

- *Joint attention.* Whenever two people are engaged in a conversation, joint attention or shared attention could be occurred whenever one party keep looking towards certain another object and another party acknowledge this object as potential conversation topic by also looking at said object. In our demo content, we implemented this concept as a trigger to advance the story when The Teacher looking at direction of a book he told user to read, leading user's attention towards certain object in question. This event triggered when user's gaze is directed at the book direction direction, which also indicates user's acknowledgement of that book location.

- *Avoidance of interaction.* Gaze aversion, or avoiding to meet someone's eyes, prevents that person from initiating an interaction or could be interpreted as sign of unwillingness to be engaged in an interaction (Kleinke 1986). In our demo content, we implemented this as an optional event where The Teacher

Table 3.1 List of implemented gaze based interaction concept

Concept	Condition	Gaze Pattern	Teacher's Reaction	Potential use-case in video game
Seeking for interaction	Gaze at Teacher's face	Momentary	Acknowledges player's attention, greets	Automatically engage interaction with Non-Playable Character (NPC)
Cultural disrespect	Gaze at Teacher's face when it does not talking	Continuous	Distressed, confused	Implicit assesment of player's attitude and tendecy to customize storyline and gameplay
Signs of intention	Gaze at wall clock	Continuous	Ask question to player	Automatically triggers object related event, draw attention of NPC
Joint attention	Gaze at book on table when Teacher tells to	Momentary	Acknowledges player's attention, continue dialogue	Implicit acknowledgement of player's knowledge of certain game object
Avoidance of interaction	Do not gaze at Teacher's face at all when it talking	Continuous	Angry	Automatically disengage from interaction with NPC

will act angry and call out to user whenever user's gaze is continuously not directed toward The Teacher's face when he is talking. This event triggered when user's gaze is not directed at The Teacher's face direction at all for 15 seconds when he is talking to user, which could also indicates that user is not engaged or interested in conversation with The Teacher.

In order to enable all this concepts as a feature in demo content, we need a way to detect and translate user's gaze information into a usable data for demo content interaction purpose. As we are developing a VR application which requires use of HMD, normal desktop setup oriented eye tracking such as Tobii Eye Tracker 4C[5] could not be used as it cannot visually capture user's eye which is hidden under the HMD. To solve this, we are using Pupil Labs HTC Vive Eye Tracking Add-on [6], which designed to be installed inside VR HMD around each lenses, therefore allows it to capture visual feed of user's eyes even while wearing VR HMD.

3.3.2 Gesture

With our goal of providing a gesture-based input modality for interaction with VA, deciding on what kind of body gestures user's can intuitively relates was the first thing we need to settle. In order to decide what kind of gestures we will implement in our model, first we tried to classify gestures based on it's usage. Krauss et al. (Krauss et al. 1996) classifies gestures into 3 different classification, which are:

1. Adapters

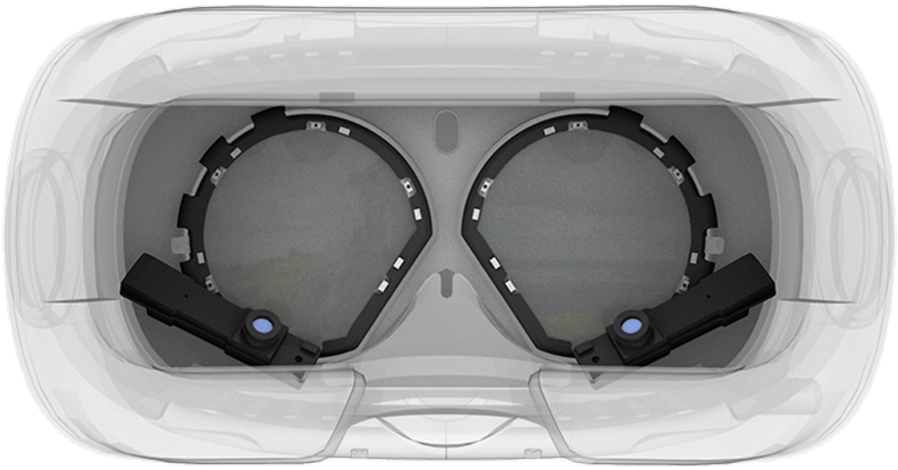

Figure 3.2 Pupil Labs HTC Vive Eye Tracking Add-on

Krauss defined adapters as *"adapters are not gestures as that term is usually understood. They are not perceived as communicatively intended, nor are they perceived to be meaningfully related to the speech they accompany, although they may serve as the basis for dispositional inferences (e.g., that the speaker is nervous, uncomfortable, bored, etc.)"*. Scratching, rubbing, tapping, and fidgeting are some example of adapters action.

2. Symbolic gestures

   Krauss defined symbolic gestures as *"hand configurations and movements with specific, conventionalized meanings"*. Symbolic gestures are used intentionally and serve a clear communicative function, as opposite of adapters. Most of the times symbolic gestures are being used in absence of speech, although it's also can be used accompanying speech to echoing a spoken word or substituting for something that was not said. Symbolic gestures includes commonly used gestures such as waving hand, O.K. sign, and thumbs-up.

3. Conversational gestures

   Krauss defined conversational gestures as set of gestures that fall between adapters symbolic gestures. He defines conversational gestures as *"move-*

ments that accompany speech, and seem related to the speech they accompany". Additionally, Krauss also states that conversational gestures, unlike symbolic gestures, always accompanies a speech.

Based on this classification and goal of this work in mind, symbolic gestures is the closest thing to what we aimed for. First reason being our system design of not using speech as input modality, which eliminates conversational gestures as potential candidate. Second, symbolic gestures fits our general goal of creating a nonverbal interaction modality, as it has clear and specific meanings behind it.

Next concern we needed to address was difference of gestures meanings across different cultures. Social bodily gestures and meaning behind it are known to varies between different cultures. For example, an "O.K." hand gesture, done by making a circle by touching point of index finger and thumb, is considered as a generally positive gesture in America, but considered to be a rude gesture in Brazil (Axtell 1999). In order to define a set of gestures we will implement in our demo content, we used a questionnaire to find out generally most common gestures for each of implication or messages we are will use in our demo content. We asked respondents regarding gestures they uses to indicates six message which we will potentially implement in our demo content, which are:

- Calling for someone's attention from a far
- Greetings / "Hi"
- Affirmative / "Yes"
- Negative / "No"
- Referring to myself
- Referring to another person

In total, 30 respondents with 14 different nationalities answered the questionnaire, ranging from Asian countries such as Japan, China, Philippine, Singapore, and Indonesia, to European countries such as Germany, Italy, France, and Russia. Results of each questions shows a single dominant gesture, with at least 60% of respondent chose said gesture. The result of each questions are as follows:

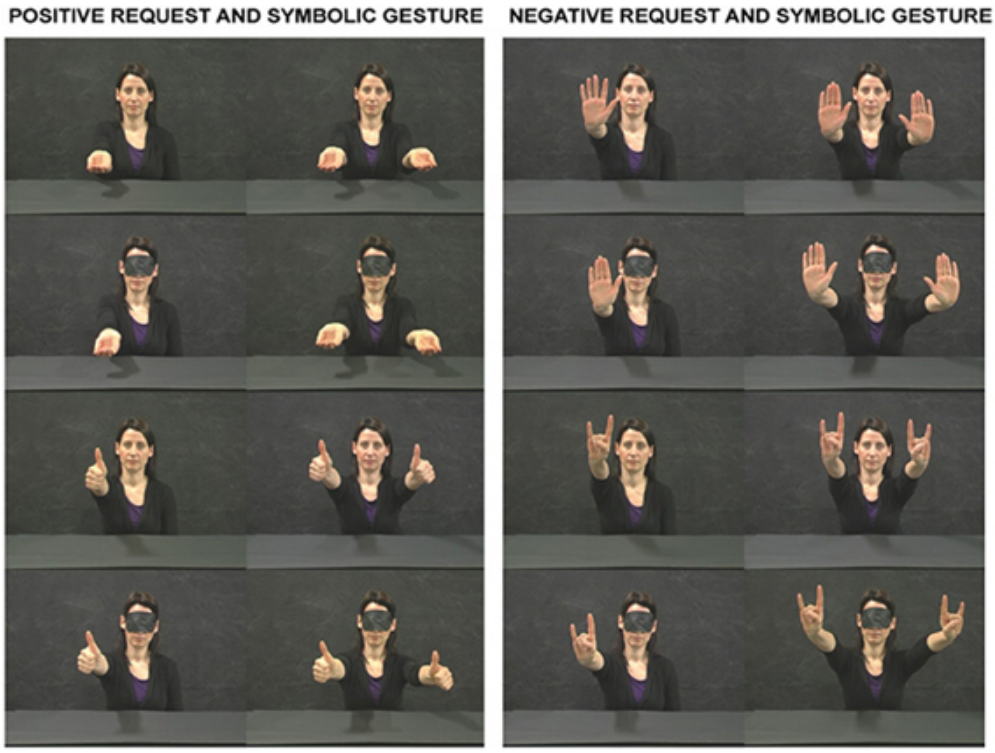

Figure 3.3 Example of symbolic gestures (De Stefani et al. 2013)

- Calling for someone's attention from a far : Wave / Raising Arm 97% (29/30)
- Greetings / "Hi" : Wave / Raising Arm 80% (24/30)
- Affirmative / "Yes" : Nod Head 60% (18/30)
- Negative / "No" : Shake Head Horizontally 73% (22/30)
- Referring to myself : Hand / Finger Pointing to Self 80% (24/30)
- Referring to another person : Hand / Finger Pointing to Other 77% (23/30)

Based on this results, it is safe to assume these most chosen gestures generally related with respective questioned meanings. These results will also be the base for gesture-based input module we will implement on demo content.

During designing process of story line for our demo content however, we decided to not using two out of these gestures, which are "Hand / Finger Pointing to Self" and "Hand / Finger Pointing to Other" as it do not fit the story. This decision left us to remaining three different gestures, which are "Wave / Raising Arm" (with 2 different implications), "Nod Head", and "Shake Head Horizontally". Based on this gestures, next thing needed to be decided was what kind of technologies we will use to reliably detect these gestures while using VR setup. "Nod Head" and "Shake Head Horizontally" which are head gestures do not require additional device, as most of VR HMD already has built-in position and orientation tracking functionality which theoretically could be used to detect head movement both vertically and horizontally. Therefore, for these two gestures, VR HMD alone is sufficient and additional sensor is not necessary. That left us with only one more gesture, which is "Wave / Raising Arm". To detect this gesture, initially two different sensors, Microsoft Kinect [7] and Leap Motion[8], were considered. During implementation test, we found that Leap Motion while being more efficient and portable compared to Kinect, has one major limitation, which is limited and unstable field of view as it is designed to be attached in front of the HMD. This limitation means in order for Leap Motion to detect user's hand gesture, it requires user's head to facing his/her gesturing hand to a certain degree, which potentially could restrict user's head and hand movement. Kinect on other hand,

while being less portable, is designed to be stationery and has wider field of view, which resulted in a more reliable "Wave / Raising Arm" gesture. Based on this consideration, we decided on using Kinect as our body gesture sensor device.

Notes

1 HTC Vive : https://www.vive.com/eu/

2 https://github.com/ValveSoftware/openvr

3 https://techjury.net/stats-about/virtual-reality/

4 https://www.statista.com/statistics/189582/age-of-us-video-game-players-since-2010/

5 Tobii Eye Tracker 4C : https://gaming.tobii.com/products/

6 Pupil Labs HTC Vive Eye Tracking Add-on : https://pupil-labs.com/blog/2016-08/htc-vive-eye-tracking-add-on/

7 Microsoft Kinect for Windows v2 : https://blogs.msdn.microsoft.com/kinectforwindows/2014/03/27/revealing-kinect-for-windows-v2-hardware/

8 Leap Motion : https://www.leapmotion.com/

Chapter 4
Implementation

In this chapter, we will describe how our system is built technically and explanation of two iterations we developed during the course of this project.

4.1. System Architecture

This section will describe each elements of current and latest version of our designed system.

4.1.1 System Overview

Following are the basic guidelines of how our system works:

- Playing interactive 3D VR content for user.
- Detects user gaze in real-time to be used as input to interact with VA on the content.
- Detects user head and hand gestures in real-time to be used as input to interact with VA on the content.
- VA on demo content gives visual (animation) and audio (dialogue) feedback accordingly based on user's gaze and gestures.

4.1.2 Hardware

Computer

VR content typically requires higher computer specification compared to normal 3D desktop application in order to smoothly renders image to both left and right

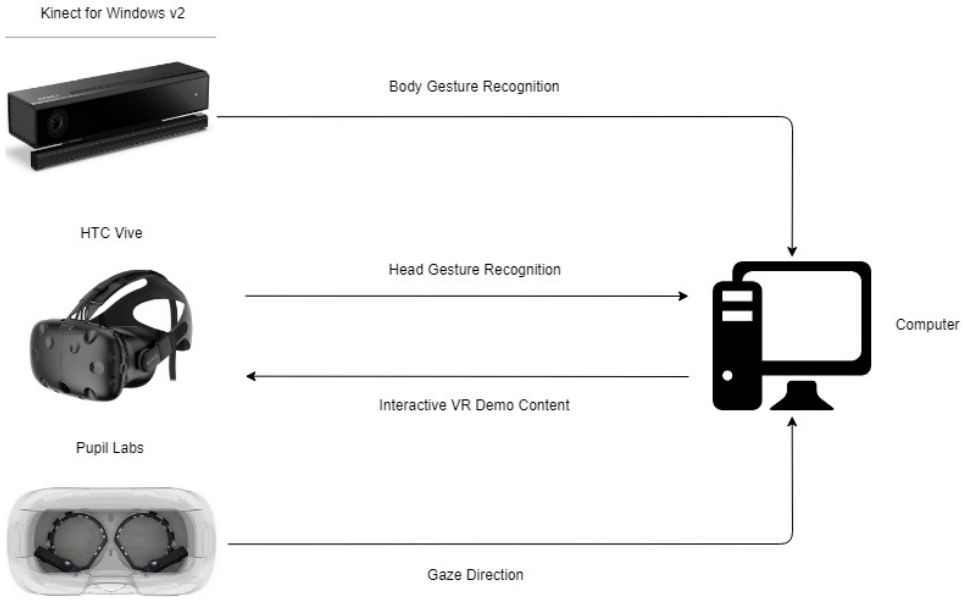

Figure 4.1 Illustration of System Architecture and Data Flow

eye screen simultaneously. Furthermore, our interaction model requires multiple separate processes (VR demo content by Unity, gesture detection by Kinect, eye tracking by Pupil Labs) to be ran simultaneously, which increase hardware demand to run whole system. Therefore, in order to run the whole system smoothly, a custom Alienware Area-51m high-end laptop by Dell, which capable of running whole system smoothly is used to showcase the demo content. Following is the specification of said machine :

- CPU : Intel Core i9-9990 (8 core, 16MB cache, up to 5GHz)
- GPU : NVIDIA GeForce RTX 2080 8GB GDDR6
- Memory : 32GB (16GB x 2) RAM DDR4-2400MHz
- Storage : 1TB SSHD (8GB SSD cache)
- Video output : HDMI 1.4, DisplayPort 1.2 or newer
- USB port : 3x USB 3.0
- Operating System : Windows 10 Home 64bit

HTC Vive

As mentioned on the previous chapter, there is no particular additional requirement for VR HMD except capabilities to track HMD movement to detect head gestures. For this model, HTC Vive which is one of commercialy available high-end VR HMD, is used as main HMD to show the VR content to users. Following are specifications of HTC Vive[1] :

- Screen : Dual AMOLED 3.6 '' diagonal
- Resolution : 1080 x 1200 pixels per eye (2160 x 1200 pixels combined)
- Refresh rate : 90 Hz
- Field of view : 110 degrees
- Safety features : Chaperone play area boundaries and front-facing camera
- Sensors : SteamVR Tracking, G-sensor, gyroscope, proximity
- Connections : HDMI, USB 2.0, stereo 3.5 mm headphone jack, Power, Bluetooth
- Input : Integrated microphone
- Eye Relief : Interpupillary distance and lens distance adjustment

Microsoft Kinect for Windows v2

To enable body gestures detection besides head gestures, Microsoft Kinect for Windows v2[2] is used. Microsoft Kinect is chosen because it's robust capabilities to real-time tracks whole body movement while also easily modified due to it's open source SDK. Following are Microsoft Kinect for Windows v2 specifications (Lachat et al. 2015):

- Infrared (IR) camera resolution : 512 × 424 pixels
- RGB camera resolution : 1920 × 1080 pixels
- Field of view : 70 × 60 degrees

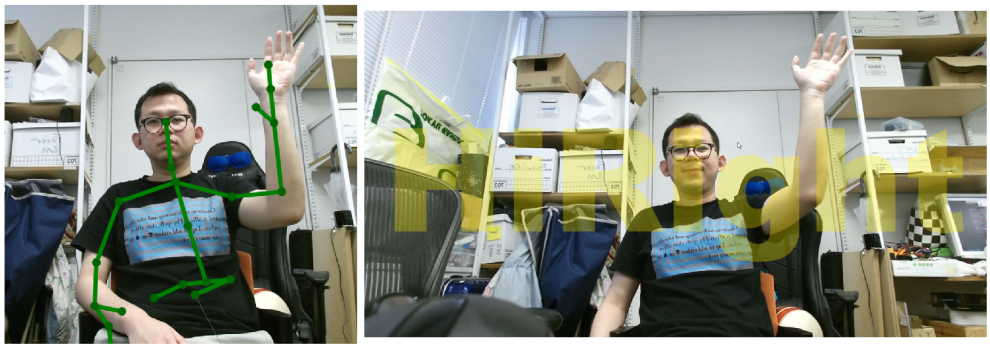

Figure 4.2 Gesture tracking and recognition by Kinect

- Framerate : 30 frames per second
- Operative measuring range : from 0.5 to 4.5 m
- Object pixel size (GSD) : between 1.4 mm (@ 0.5 m range) and 12 mm (@ 4.5 m range)

Pupil Labs

As already shortly described on chapter 3, for real-time eye tracking in VR setup, eye tracking sensor need to be installed inside the HMD to be able to capture visual image of user's eyes. In order to achieve that, we are using Pupil Labs HTC Vive Eye Tracking Add-on research grade eye tracking sensor which specifically built to fit each lenses of HTC Vive.

Following are specifications of Pupil Labs eye tracking sensor :

- Mono / Stereo : Both, depending on desired setup
- Tracking Frequency : 200Hz
- Field of View HTC Vive / Vive Pro: up to HMD limits.
- Gaze Accuracy : 1.0deg
- Gaze Precision : 0.08deg

- Camera Latency : 5.7ms
- Processing latency : 3-4ms on i5 CPU
- Resolution : 192x192
- Connection : USB 2.0
- Saturation : Interconnected bandwidth USB 2.0 60% saturation

4.1.3 Software

Unity

To build and run 3D VR interactive demo content, we are using Unity[3] game engine version 2018.2.11f. Unity is cross-platform game engine developed by Unity Technologies and one of the most used engine to develop VR content globally. Majority of demo content main component, including scenario flow and VA's behavior management is written completely by ourselves in C# language. Aside from that, we are also referring to multiple publicly available script for various purposes, which are as follows:

- hmd-eyes[4], which includes various scripts to support implementation of Pupil Labs in VR / AR application, provided by Pupil Labs.
- VR Gesture Recognized[5], which includes scripts to support detection of user's head nod (Vertical movement) and head shake (horizontal movement) through HMD movement.

Additionaly, we are also using some purchased assets from Unity Assetstore, which are:

- Assets_classroom by Argyle Co. Ltd [6], 3D classroom model to create a classroom environment in demo content.
- Taichi Character Pack by Game Asset Studio[7], 3D male humanoid model as a 3D visual representation of The Teacher as well as animation set to represent various expression.

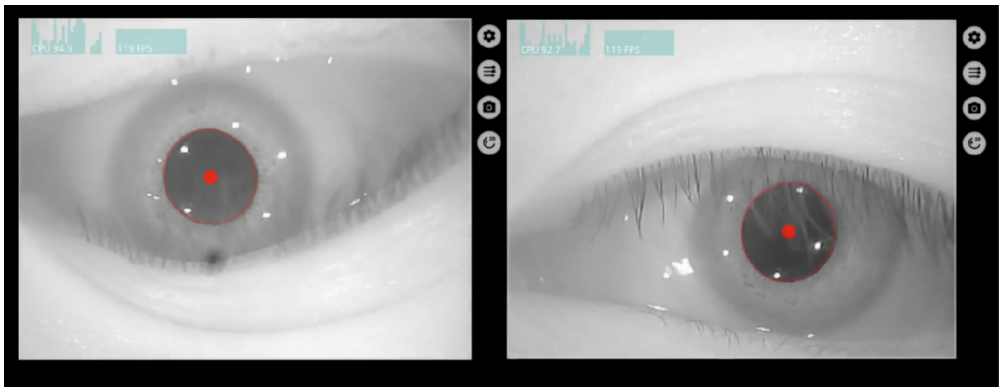

Figure 4.3 Real-time eye tracking video feed by Pupil Capture

GesturePak

GesturePak[8] is a gesture recognition library for Microsoft Kinect for Windows. GesturePak provides an interface for user to record body gesture through Kinect, which then could be saved as .xml file and be recognized also through Kinect in subsequent use. We are using GesturePak as it is allows a fast and easy creation and modification of custom gesture recognition database, which fits our requirement.

During the experience, GesturePak is running simultaneously with main application (demo content on Unity) and continuously streams any detected gestures to main application through User Datagram Protocol (UDP).

Pupil Capture

Pupil Capture[9] receives video and audio streams, detects your pupil, tracks your gaze, tracks markers in environment, streams data in real-time over the network, and records data in an open format. Pupil Capture is provided by Pupil Labs as a default interface for Pupil Labs eye tracking devices.

During the experience, Pupil capture is running simultaneously with main application (demo content on Unity) and continuously streams user's gaze information to main application. Streamed user's gaze information then translated into 2D vector data (x and y) on Unity application which indicates approximate user's gaze direction in VR environment.

IBM Watson Text to Speech

IBM Watson Text to Speech[10] is a cloud service that capable to convert written text into natural-sounding audio in a variety of languages and voice. We are utilizing IBM Watson Text to Speech service to generate all of The Teacher's dialogues audio.

4.2. Prototypes

4.2.1 First Prototype (Virtual Gaze)

First iteration of this work was titled Virtual Gaze, as it only incorporated gaze as input modality as opposed as gaze plus gesture in latest iteration. In this version of prototype, as gesture input concept has not implemented yet, Kinect was not used. Other difference regarding hardware was the VR HMD, with Oculus Rift DK2 being our HMD of choice compared to HTC Vive on current version, however there is no intended reason behind this change. This work was also published and presented at VRST 2018 as poster under title "Virtual Gaze : Exploring use of Gaze as Rich Interaction Method with Virtual Agent in Interactive Virtual Reality Content".

4.2.2 Second Prototype (Current Version)

Continuing from previous work, latest and current iteration of this work still focuses on building a nonverbal gesture based interaction with VA in VR content. As explained previously, compared to previous version this version incorporated user's gesture detection as additional input module by utilizing Microsoft Kinect for Windows v2 (for body gesture except head) and HTC Vive built-in position detection feature (for head gesture). Fundamentally, there is no significant difference in story of the demo content besides some added variation and modification of dialogues from previous version.

Figure 4.4 Virtual Gaze features

Notes

1 HTC Vive : https://www.vive.com/eu/

2 Microsoft Kinect for Windows v2 : https://blogs.msdn.microsoft.com/kinectforwindows/2014/03/27/revealing-kinect-for-windows-v2-hardware/

3 Unity : https://unity.com/

4 https://github.com/pupil-labs/hmd-eyes

5 https://github.com/korinVR/VRGestureRecognizer

6 https://assetstore.unity.com/packages/3d/environments/assets-classroom-98134

7 https://assetstore.unity.com/packages/3d/characters/taichi-character-pack-15667

8 https://github.com/carlfranklin/GesturePak2V1

9 https://github.com/pupil-labs/pupil/releases/tag/v1.12

10 https://www.ibm.com/watson/services/text-to-speech/

Figure 4.5 Current version setup, Kinect is facing user from front

Chapter 5
Evaluation

This chapter will describe in detail user test of two prototypes we developed, followed by it's result and insight we gain during these studies.

5.1. Pilot Study (First Prototype)

In order to evaluate how user reacts toward our model, we conducted an initial user study. In total, there were four participants (two male and two female) with average age of 25 years old. User test was done by asking each participant to try out the demo content with two different kinds of interaction model; gaze based model and HMD direction based mode. In gaze based model, application assumes participant's gaze as point of focus, while in HMD direction model application assumes center of HMD direction as the participant's point of focus, which is interaction model of some currently available VR video games, such as "The Inpatient"[1]. After each session, participants are asked to answer 9-point rating scale (from disagree to strongly agree) questionnaire regarding the demo experience aspects as designed by Bee et al. (Bee et al. 2010), which are Social Presence, Rapport, Engagement, Social Attraction, and Perception of Story.

Result

The results from the initial user test, as shown on figure 4.5, shows that Gaze based model outperforms HMD direction based model in all aspect of evaluation. Furthermore, to measure the significance between the two models, we ran a two-tailed t test for each aspects. Significant difference was found on aspect of rapport (t:3.31, p:0.02) and perception of story (t:2.9, p:0.03). Furthermore, all of the test subjects also expressed overall positive response towards the demo contents espe-

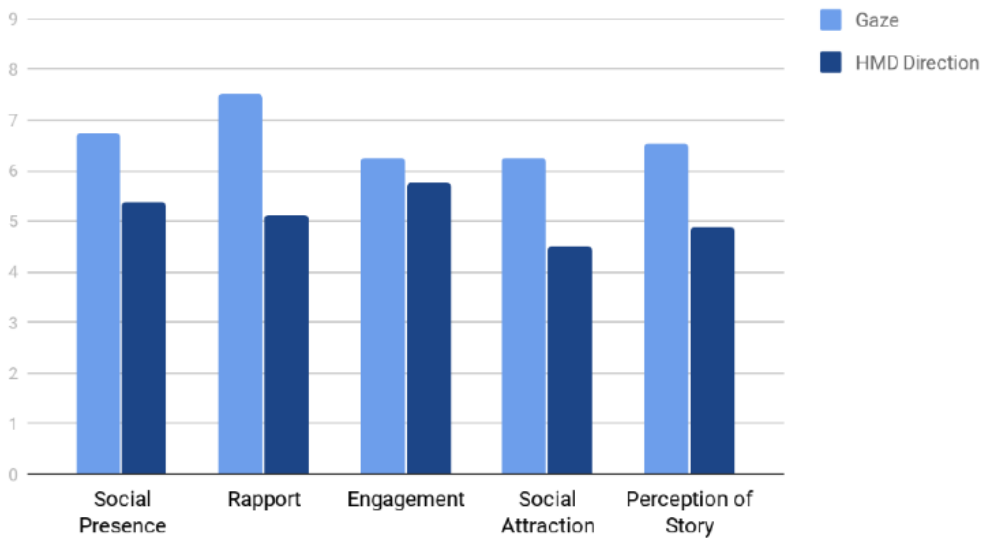

Figure 5.1 Result of Virtual Gaze initial user test

cially towards the gaze model, mentioning some positive evaluation such as "Eye tracking feature feels interesting and original", "New way to interact was fun", and "Better ease of use thanks to reduced head movement". One interesting insight we gain from the user test is how two of test participants actually performing body gestures during demo, particularly nodding head and waving hand, in respond to feedback from The Teacher, indicating how powerful interaction between player and agent in VR environment. While head gestures such as nodding could be considered normal and even used in some available VR games, hand gestures such as waving could be done more freely with our model due to a fully hands-free interaction, though it's relevancy might needs to be further proved. This finding is also our main motivation to add gesture input feature in our current prototype.

5.2. User Study (Second Prototype)

With goal of finding out how our model affects user experience compared to conventional VR content interaction model, we conducted a user study. As we incorporated additional variable compared to our previous model in form of gesture, additionally we also do additional tests by isolating our modality (gaze and ges-

Figure 5.2 Test subject experiencing our model on user test

ture) to study effect caused individually by each of these modality. This bring us to total of four models to be tested and compared, which are control model (no gaze and no gesture), gaze only model, gesture only model, and gaze + gesture model. In models that do not incorporate gesture modality (control model and gaze only model), users are given controller instead as a input device to choose options and progress story. In models that do not incorporate gaze modality (control model and gesture only model), instead of user's gaze, direction straight front of VR HMD is assumed by application as user's focus direction.

Similar to user test of previous version, we are referring to questionnaire by Bee et al. (Bee et al. 2010) to measures various experience aspects of interaction with VA, which are Social Presence (P), Rapport with the VA (R), Engagement (E), Social Attraction of the VA (A), and Perception of Story (S). In total, we asked 9 9-point rating scale (from disagree to strongly agree) questions to measure these 5 aspects, which are :

- ”I had the feeling that The Teacher was aware of me.” (P)
- ”I had the feeling of personal contact to The Teacher.” (P)
- ”I would have liked to continue the interaction with The Teacher.” (R)

- ”I had the feeling that The Teacher reacted on me.” (R)
- ”I enjoyed the meeting with The Teacher.” (E)
- ”I found it easy to interact with The Teacher.” (E)
- ”I had the feeling that The Teacher is concerned about me.” (A)
- ”The Teacher was sympathetic.” (A)
- ”I had no problems to empathize with the part of character I played.” (S)

In addition, after finishing all tests, we also conducted a verbal interview with test subjects to gather additional feedback regarding other aspects of interaction experience with VA, such as naturalness, ease of interaction, immersion, and enjoyment.

Total of 10 participants, consisted of 5 males and 5 females, with average age of 25.7 years old participated in this user test. All of the participant had prior experience of playing and/or watching VR contents. Each session in average requires 40 minutes. To reduce learning effect and ensure accuracy of measurement, order of testing was rotated between each subjects' session. Before the experiment begin, each subjects were asked to fill in consent form as well as briefed about basic information about the experiment. No information regarding story and setting were given before experiment in order to encourage subject to explore the content actively. As eye tracking calibration process is required for gaze models, calibration process was done before both gaze models and non-gaze models in order to hide the information from subject. Post-questionnaire was handed after each test

Result

Result of the post-questionnaire could be seen on figure 4.6. The result positively aligns with our prediction, where all of the nonverbal communication based models (gaze only, gesture only, and gaze + gesture model) is rated better in every measured aspect compared to control model. Our designed model is rated overall highest in term of Rapport with the Character (R) and Perception of the Story (S), and gesture only model is rated overall highest in term of Social Presence (P), Engagement (E), and Social Attraction of Character (A). Our assumption

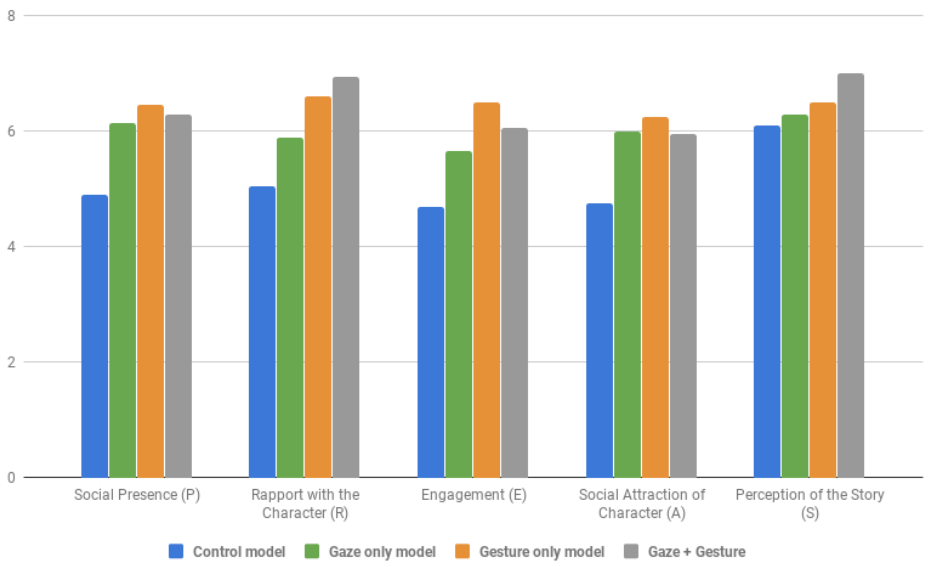

Figure 5.3 Result of user test

of reason behind this result is because gesture only model, which combines gesture based input with HMD direction based selection method, provides immersion in form of explicit gestural input while still retains accurate selection method of HMD direction (instead of less stable gaze direction). Based on this result, we could assume that simply adding more nonverbal interaction modality does not necessarily improves overall quality of the experience.

Gesture based input being the most prominent feature also being reflected in post-interview result, as majority of subjects expressed positive feedback regarding using gesture as input. For example, in term of naturalness, participant number 3 (P3),P4, P5, P8, P9, P10 expressed good experience, mentioning various things such as *"The tests without the controller feels more natural for the interaction between me and the teacher"* (P3), *"feels more natural and easier to operate"* (P4), and *"(gaze + gesture model) feels the best because the interaction feels natural, the teacher reacted to my actions like my head nod"* (P5). Additionally, P1 and P5 also described models with gesture modality as more "immersive" compared to those without. P1 in specific, provides interesting insight regarding two different input modality, citing *"Test without controller feels more immersive. However, test*

with controller feels more interactive yet less immersive.". On question regarding which part feels most interesting for them, P2 and P5 agreed on part where users are required to wave their hand when The Teacher is calling out to the from a far, which is a part we designed to introduce user to gesture based input.

While it did not rated best in any of the measurement points, gaze only model still rated better than control model and acquired an overall positive feedback from subjects during post-interview. When being prompted about which part of the demo was the most interesting for them, P3 and P9 chose part where users are required to look at a book The Teacher is looking (joint attention) which is a part where gaze based input is showcased. P3 described that part as *"(the book part) were the most interesting to me. It seems the teacher also acknowledges the moment I see the book"* while P9 answered "*The book part, I feel the teacher knows what I am looking."*. P7 answered *"The part where the teacher scolded me because I was not paying attention."*, which is the part where The Teacher scold user whenever they do not looking at his direction when he is talking (avoidance of interaction). Taking into account the fact that these subjects did not informed that their gaze influence the game, these positive feedback shows that implementation of user's gaze as additional input layer positively impacted user's experience.

Another interesting insight we found through this study is how subjects reacted towards The Teacher when they were not given a controller. As they were not being informed about availability of gesture input, not all subjects did a gesture right off the bat, though all of them eventually resorted to gesture. Besides doing gestures, during the experiment of models without controllers, 3 of the subjects (P1, P2, P4) tried to interact with The Teacher by talking to him, though our system do not support speech input. Based on this occurrence, implementation of affective speech recognition shows promise as future improvement for our model to further enhance user's interaction immersion.

Although in overall the experience was rated positively by the subjects and every subject successfully finished the experiment, there were also some problem reported by the subjects. Most of the problem reported were regarding system misinterpretation of user's gesture. P1 reported misinterpretation of his head nod gesture, mentioned *"when I nod my head to say yes, the teacher took it as no"*,

while P3 mentioned *"Some normal gestures seem to be slightly misinterpreted by the teacher"*. We expected this problem during our experiment because we expected subjects to move parts of their body actively during the experience, especially their head and hands, thus some wrong gesture reading by Kinect and HMD is prone to happen. Beside this problem, two of the test subjects (P4, P10) also expressed unnaturalness regarding The Teacher's 3D model, both in specific mentioned that The Teacher "looks too young to be a teacher". While this might irrelevant with scope of the experiment, this comment shows that VA's appearance and voice could be an important thing to be considered in studies regarding interaction with VA.

Notes

1 The Inpatient by Supermassive Games : https://www.supermassivegames.com/games/the-inpatient

Chapter 6
Conclusion

6.1. Conclusion

In this work, we developed a real-world context aware Intelligent Virtual Agent (IVA) that capable to recognizes and appropriately reacts toward user's nonverbal communication cues in VR environment, and through our demo content, we conducted a study on how this approach impacts user's interaction experience with that VA.

In the first prototype, we started this study by utilizing user's gaze, implementing eye tracking to enable a gaze-aware VA and concept of social gaze in interactive VR content. Pilot study of this prototype was overall a success, as all participant reported a better experience in every aspect of measurement, showing promises of our approach. Furthermore, from observation of this pilot study, we also gained insight of how some of the test subjects were doing social gestures during the study, which is also a part of nonverbal communication. This finding lead us to continue this study by developing the second prototype.

In the second and latest prototype, we introduced gesture recognition as additional modality with goal of further enhances user's interaction experience as well as to study how multiple nonverbal communication modality affects the experience. Result of our latest user test shows that while interaction model that incorporates nonverbal communication input modality indeed resulted in better experience, our multimodal system (gaze + gesture) do not resulted in better result compared to a single modality model (gesture only).

Looking back to the objective of this study, overall we could positively conclude that this study resulted in success. During both user study, all subjects managed to finish the interaction experience by resorting to gaze and gestures, either consciously or non-consciously. All subjects also expressed their preference

of using our interaction model compared to conventional VR interaction model. Likewise, through both of our prototypes and user tests, we proved that our approach of using nonverbal communication as input modality resulted in overall better interaction experience with VA in VR content compared to controller and HMD direction based conventional VR interaction model. Therefore, we could concluded that our interaction model is indeed feasible and positively impacts user's experience compared to conventional VR interaction model.

6.2. Limitation

One of the limitation we set on this work is position and posture of user while using our setup. As our body gesture detection sensor, Kinect, is set up stationarily in front of user, user is required to keep seated facing it and not turning around excessively during our demo content. While this does not caused a major problem during our study, this problem potentially could be further eliminated by employing a more sophisticated gesture tracking method such as OpenPose[1] in exchange with heavier processing load.

Another limitation on our demo content is how the story is fundamentally fixed and linear. While there is several dialogue and animation variation depending on user's input in certain part, most part of the demo's story line is fixed. For future work purpose, providing user with a more dynamic and less story-driven content could potentially resulted in better interaction experience measurement.

6.3. Extension

Based on our observation during course of the study and feedback from test subjects, there are multiple use case scenario from various field which potentially could gain benefit by adopting our approach as a framework. First, we believe that VR application with a goal to train it's user's communication skill (e.g. job interview training, public speaking training) is a strong application scenario for our framework, as some other work tried to tackle such application in desktop based setup. By tracking user's gaze and gesture, in application VA could dynamically evaluates and adapts it behaviour to match user condition.

Second, during process of literature study, we also found multiple works that explores use of IVA as a health consultant / therapist and resulted in a positive effect. In regard of this kind of application, we could see how our approach could be used as additional channel to measure various kinds of user's profiles (e.g. personality, anxiety) in higher accuracy, thus the VA could custom tailor the treatment based on this information.

Finally, being our main motivation for doing this project as described on chapter 1, video games is the most prominent target for this framework. We strongly believe that this approach holds potential in future commercial video games, as expressed by some of our test subject during user test. With recent trend of new built-in eye tracking high-end VR HMD such as HTC Vive Pro Eye[2] and FOVE[3], gaze based interaction such as one we introduced in this work is likely, if not obviously, will soon become a standard feature in future VR contents.

6.4. Future Works

As follow up of this study, there are multiple direction we would like to explore in future. First, instead of gesture which is arguably an explicit action, we are interested in incorporating another implicit action beside gaze, especially user's facial expression. By completely using implicit input modality, we are expecting a more intuitive control over the VA's behavior.

Second, based on our observation of what subjects did during our user test as described on chapter 4, while it is not an element of "nonverbal" communication, we are interested in incorporating affective speech recognition as one of input channel in similar study. Effect of speech based input in conjunction with what we have studied so far about nonverbal communication is interesting potential direction we would like to explore in the future.

Notes

1 OpenPose : https://github.com/CMU-Perceptual-Computing-Lab/openpose

2 HTC Vive Pro Eye : https://enterprise.vive.com/ca/product/vive-pro-eye/

3 FOVE : https://www.getfove.com/

References

Agrawal, A., R. Raj, and S. Porwal (2013) "Vision-based multimodal human-computer interaction using hand and head gestures," in *2013 IEEE Conference on Information Communication Technologies*, pp. 1288–1292, April.

Argyle, Michael (1972) "Non-verbal communication in human social interaction.."

Argyle, Michael and Mark Cook (1976) "Gaze and mutual gaze.."

Ashbrook, Daniel and Thad Starner (2010) "MAGIC: A Motion Gesture Design Tool," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pp. 2159–2168, New York, NY, USA: ACM.

Axtell, R.E. (1999) *Do's and Taboos of Humor Around the World: Stories and Tips from Business and Life*: Wiley.

Bee, Nikolaus, Johannes Wagner, Elisabeth André, Thurid Vogt, Fred Charles, David Pizzi, and MO Cavazza (2010) "Gaze behavior during interaction with a virtual character in interactive storytelling,": IFAAMAS.

Bosse, Tibor, Tilo Hartmann, Romy A.M. Blankendaal, Nienke Dokter, Marco Otte, and Linford Goedschalk (2018) "Virtually Bad: A Study on Virtual Agents That Physically Threaten Human Beings," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pp. 1258–1266, Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Castiello, Umberto (2003) "Understanding other people's actions: intention and attention.," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 29, No. 2, p. 416.

Chai, Xiujuan, Guang Li, Yushun Lin, Zhihao Xu, Yili Tang, Xilin Chen, and Ming Zhou (2013) "Sign language recognition and translation with kinect," in *IEEE Conf. on AFGR*, Vol. 655.

Chidambaram, Vijay, Yueh-Hsuan Chiang, and Bilge Mutlu (2012) "Designing Persuasive Robots: How Robots Might Persuade People Using Vocal and Nonverbal Cues," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '12, pp. 293–300, New York, NY, USA: ACM.

De Stefani, Elisa, Alessandro Innocenti, Claudio Secchi, Veronica Papa, and Maurizio Gentilucci (2013) "Type of gesture, valence, and gaze modulate the influence of gestures on observer's behaviors," *Frontiers in Human Neuroscience*, Vol. 7, p. 542.

Frischen, Alexandra, Andrew Bayliss, and Steven Tipper (2007) "Gaze Cueing of Attention: Visual Attention, Social Cognition, and Individual Differences," *Psychological bulletin*, Vol. 133, pp. 694–724.

Garau, Maia, Mel Slater, Simon Bee, and Martina Angela Sasse (2001) "The Impact of Eye Gaze on Communication Using Humanoid Avatars," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pp. 309–316, New York, NY, USA: ACM.

Hartanto, Dwi, Willem-Paul Brinkman, Isabel L Kampmann, Nexhmedin Morina, Paul GM Emmelkamp, and Mark A Neerincx (2015) "Home-based virtual reality exposure therapy with virtual health agent support," in *International Symposium on Pervasive Computing Paradigms for Mental Health*, pp. 85–98, Springer.

Hickson, S., N. Dufour, A. Sud, V. Kwatra, and I. Essa (2019) "Eyemotion: Classifying Facial Expressions in VR Using Eye-Tracking Cameras," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1626–1635, Jan.

Ishii, Ryo, Yukiko I. Nakano, and Toyoaki Nishida (2013) "Gaze Awareness in Conversational Agents: Estimating a User's Conversational Engagement

from Eye Gaze," *ACM Trans. Interact. Intell. Syst.*, Vol. 3, No. 2, pp. 11:1–11:25.

Just, Marcel Adam and Patricia A. Carpenter (1976) "Eye fixations and cognitive processes," *Cognitive Psychology*, Vol. 8, pp. 441–480.

Kleinke, Chris L (1986) "Gaze and eye contact: a research review.," *Psychological bulletin*, Vol. 100, No. 1, p. 78.

Krauss, Robert M, Yihsiu Chen, and Purnima Chawla (1996) "Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us?" in *Advances in experimental social psychology*, Vol. 28: Elsevier, pp. 389–450.

Kumar, Manu, Andreas Paepcke, Terry Winograd, and Terry Winograd (2007) "EyePoint: practical pointing and selection using gaze and keyboard," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 421–430, ACM.

Lachat, Elise, Hélène Macher, Tania Landes, and Pierre Grussenmeyer (2015) "Assessment and Calibration of a RGB-D Camera (Kinect v2 Sensor) Towards a Potential Use for Close-Range 3D Modeling," *Remote Sensing*, Vol. 7, pp. 13070–13097.

Mehrabian, Albert et al. (1971) *Silent messages*, Vol. 8: Wadsworth Belmont, CA.

Miyauchi, Dai, Arihiro Sakurai, Akio Nakamura, and Yoshinori Kuno (2004) "Active Eye Contact for Human-robot Communication," in *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, pp. 1099–1102, New York, NY, USA: ACM.

Møllenbach, Emilie, John Paulin Hansen, and Martin Lillholm (2013) "Eye Movements in Gaze Interaction," *Journal of E M D R Practice and Research*, Vol. 6, No. 2, pp. 1–15.

Mutlu, Bilge, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita (2009) "Footing in Human-robot Conversations: How Robots

Might Shape Participant Roles Using Gaze Cues," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, HRI '09, pp. 61–68, New York, NY, USA: ACM.

Poole, Alex and Linden J. Ball (2005) "Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future," in *Prospects" , Chapter in C. Ghaoui (Ed.): Encyclopedia of Human-Computer Interaction. Pennsylvania: Idea Group, Inc.*

Rickel, Jeff and W. Lewis Johnson (1998) "STEVE (Video Session): A Pedagogical Agent for Virtual Reality," in *Proceedings of the Second International Conference on Autonomous Agents*, AGENTS '98, pp. 332–333, New York, NY, USA: ACM.

Sutherland, Ivan E. (1968) "A Head-mounted Three Dimensional Display," in *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*, AFIPS '68 (Fall, part I), pp. 757–764, New York, NY, USA: ACM.

Thies, Justus, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nie(2018) "FaceVR: Real-Time Gaze-Aware Facial Reenactment in Virtual Reality," *ACM Trans. Graph.*, Vol. 37, No. 2, pp. 25:1–25:15.

Vidal, Melodie, Remi Bismuth, Andreas Bulling, and Hans Gellersen (2015) "The royal corgi: Exploring social gaze interaction for immersive gameplay," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 115–124, ACM.

Wang, Ning and Jonathan Gratch (2010) "Don'T Just Stare at Me!," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pp. 1241–1250, New York, NY, USA: ACM.

Wood, Marion M. (1972) "Book Reviews : KINESICS AND CONTEXT: ESSAYS ON BODY MOTION COMMUNICATION. Ray L. Birdwhistell. Philadelphia, University of Pennsylvania Press, 1970," *The ABCA Journal of Business Communication*, Vol. 9, No. 3, pp. 68–69.

Yoshikawa, Yuichiro, Kazuhiko Shinozawa, Hiroshi Ishiguro, Norihiro Hagita, and Takanori Miyamoto (2006) "Responsive Robot Gaze to Interaction Partner.," 08.