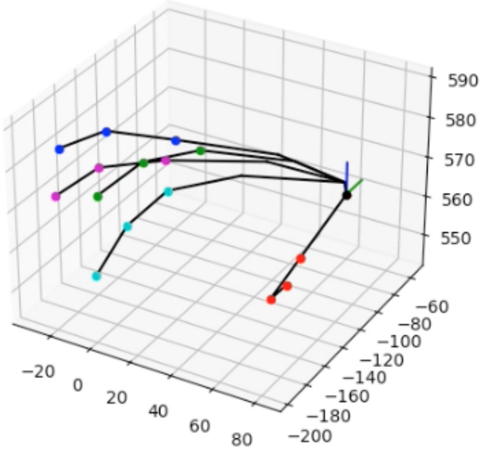
DEPARTMENT OF COMPUTER
SCIENCE & ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
MADRAS
CHENNAI - 600036

Enabling Realistic and Quantified User Interfaces in Virtual Reality



A Thesis

Submitted by

JOSEPH HOSANNA RAJ I.

For the award of the degree

Of

DOCTOR OF PHILOSOPHY

January 2023

THESIS CERTIFICATE

This is to undertake that the Thesis titled **ENABLING REALISTIC AND QUANTIFIED USER INTERFACES IN VIRTUAL REALITY**, submitted by me to the Indian Institute of Technology Madras, for the award of Ph.D., is a bona fide record of the research work done by me under the supervision of Prof. B. Ravindran and Prof. M Manivannan. The contents of this Thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Place: Chennai 600-036
Date: January 23, 2023

Joseph Hosanna Raj I.
Research Scholar

Prof. B. Ravindran
Research Guide

Prof. M. Manivannan
Research Co-Guide

LIST OF PUBLICATIONS

I. REFEREED JOURNALS BASED ON THE THESIS

1. **Isaac, J. H. R.**, Manivannan, M., & Ravindran, B. (2022). Single Shot Corrective CNN for Anatomically Correct 3D Hand Pose Estimation. *Frontiers in Artificial Intelligence* 5:759255. DOI: 10.3389/frai.2022.759255
2. **Isaac, J. H. R.**, Manivannan, M., & Ravindran, B. (2021). Corrective Filter Based on Kinematics of Human Hand for Pose Estimation. *Frontiers in Virtual Reality*, 2, 92. https://doi.org/10.3389/frvir.2021.663618
3. **Isaac, J. H. R.**, Vasudevan, M. K., & Manivannan, M. (2020). Effect of Visual Awareness of the Real Hand on User Performance in Partially Immersive Virtual Environments: Presence of Virtual Kinesthetic Conflict. International Journal of Human-Computer Interaction, 36(16), 1540-1550. https://doi.org/10.1080/10447318.2020.1768667
4. Vasudevan, M. K., **Isaac, J. H. R.** (equal contribution), Sadanand, V., & Manivannan, M. (2020). Novel virtual reality based training system for fine motor skills: Towards developing a robotic surgery training system. International Journal of Medical Robotics and Computer Assisted Surgery, 16(6), 1-14. https://doi.org/10.1002/rcs.2173

II. PRESENTATIONS IN CONFERENCES

1. **Isaac, J. H. R.**, Krishnadas, A., Damodaran, N., & Manivannan, M., Effect of control movement scale on visual haptic interactions. *EuroHaptics 2018*, pp.150-162, (2018)

III. PUBLICATIONS IN CONFERENCE PROCEEDINGS

1. **Isaac, J. H. R.**, Krishnadas, A., Damodaran, N., & Manivannan, M. (2018). Effect of control movement scale on visual haptic interactions. In D. Prattichizzo, H. Shinoda, H. Z. Tan, E. Ruffaldi, & A. Frisoli (Eds.), Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 10893 LNCS, pp. 150-162). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-93445-7_14

ACKNOWLEDGEMENTS

First and foremost, I would like to give thanks to the many people who contributed to give the thesis its present shape. It is a pleasant moment to express my gratitude to the people behind my life for guiding, motivating and inspiring me.

I would like to express my deep and sincere gratitude to my thesis advisor Prof. B. Ravindran, Head of Robert Bosch Center for Data Science and Artificial Intelligence (RBC-DSAI) and Professor in the Department of Computer Science and Engineering. Its was a great opportunity to pursue research under his guidance and tutelage. I have learned a lot and still learning from my advisor, and gained much needed knowledge and perspectives in the field of Machine Learning.

I would like to extend my sincere thanks to my thesis co-advisor Prof. M. Manivannan, Founder of Merkel Haptic Systems, Head of the Touch Lab, Department of Applied Mechanics (Biomedical Engineering). Under his guidance, I learned various interdisciplinary fields, namely Haptics and Virtual Reality, and extremely grateful to him. He not only helped me in the research ideas but also motivated me to push my limits with his wisdom.

It gives me great pleasure to thank Prof. Susan Elias, Professor & Deputy Director, Centre for Advanced Data Science, Vellore Institute of Technology, Chennai. Her guidance was very helpful and it was due to her mentorship that led me to the opportunity to pursue my Ph. D. in IITM. I would also like to thank Dr. Venkatraman Sadanand, Pediatric Neurosurgeon in Loma Linda, California, United States for his guidance during several projects and his great sense of humour.

I owe my sense of gratitude to the former and current HODs Prof. P. Sreenivasa Kumar, Prof. Krishna Moorthy Sivalingam, and Prof. C. Chandra Sekhar. I would also like to thank my DC members Prof. P. Sreenivasa Kumar (chair), Dr. Arun Kumar Thittai, Dr. Kaushik Mitra, Dr. Arun Rajkumar for their support in reviewing my thesis and for their feedback and suggestions during my seminar and DC meetings.

I would like to give special thanks to my senior, best friend and alumni of the Touch lab, Dr. Ravali Gourishetti for giving me great motivation and backing me in every step of my time in IITM. I would also like to thank my friend in the Computer Science department and alumnus Dr. Vikram Singh for being with me and helping in many machine learning related aspects. I would also like to thank Mr. V. Prabhu who helped in the fabrication process for many of my experiments and helped in many hardware-related aspects with his extensive knowledge.

My thanks go to all my friends in the Touch Lab who helped me in my research, namely Rahul Kumar Ray, Balaji Pujari, Debadutta Subhudhi, Dr. Madhan Kumar, Varun Durai, Aayush Maloo, Sujay Kotwale, Arul Praveen, Bose Karthik Yathati, Payal Patel and Tarun Kumar Yellapu. I would like to thank all my friends in the RISE Lab who helped me in my research, namely Tarun Kumar, Dr. Pathanjali SLPSK and Priyesh Vijayan. They made my time in IITM pleasant and joyful.

Finally, I would like to thank my family for their limitless support and love for me at all times. I also express my thanks to my elder brother, Dr. Ebenezer RHP Isaac for his helpful tips and support during my research.

Last but not least, I would like to thank the Lord for giving me the strength, knowledge, ability and opportunity to undertake this research study and to persevere and complete it. Without His blessings, this achievement would not have been possible.

- Joseph Hosanna Raj I.

ABSTRACT

KEYWORDS: Human-Computer Interaction; 3D Hand Pose estimation; Hand Biomechanics; Anatomical Filter; Convolutional Neural Network; Fitts's Law; Scaled Interaction; Microscopic Selection Task.

Interaction is one of the three pillars (other pillars are Immersion and Presence) of Virtual Reality (VR) which facilitates the feeling of the other two to the user. A user can interact with the 3D virtual world through actions such as pointing, clicking, rotating the head, and so on involving the coordination between human haptic and visual systems. Currently, with the advent of feature-rich virtual reality hardware such as the HTC Vive and the Oculus Touch, there are many ways for the user to interact with the virtual environment. For example, the user can move objects in VR by holding and moving the controller in real space and interact by pressing the various buttons on the controller. Pointing (or selection) and reaching a target in VR constitute two of the fundamental interactions using a virtual cursor.

This thesis contains novel methods to enhance the user experience in virtual reality by proposing solutions to two distinct problems. We will briefly describe these two problems in detail. The first part of the thesis addresses problems related to 3D hand pose estimation. Depth-based 3D hand trackers are expected to estimate highly accurate poses of the human hand given the image. One of the critical problems in tracking the hand pose is the generation of realistic predictions. This thesis proposes a novel "anatomical filter" that accepts a hand pose from a hand tracker and generates the closest possible pose within the real human hand's anatomical bounds. The filter works by calculating the 26-DoF vector representing the joint angles and correcting those angles based on the real human hand's biomechanical limitations. The proposed filter can be plugged into any hand tracker to enhance its performance. The filter has been tested on two state-of-the-art 3D hand trackers. The empirical observations show that our proposed filter improves the hand pose's anatomical correctness and allows a smooth trade-off with pose error. The filter achieves the lowest prediction error when used with state-of-the-art trackers at 10% correction.

The second part of the thesis implements the anatomical filter as a Convolutional Neural Network (CNN). In this work, we present the Single Shot Corrective CNN (SSC-CNN) framework to tackle the problem at the architecture level. In contrast to previous works which uses post-facto pose filters, SSC-CNN predicts the hand pose that implicitly conforms to the human hand's biomechanical bounds and rules in a single forward pass. The model was trained and tested on the HANDS2017 and MSRA datasets. Experiments show that our proposed model shows comparable accuracy to the state-of-the-art models. However, the previous methods have high anatomical errors whereas our model is free from such errors. Experiments also show that the ground truth provided in the datasets used also suffer from anatomical errors and an Anatomical Error Free (AEF) version of the datasets namely AEF-HANDS2017 and AEF-MSRA was created.

The third part of the thesis involves scaled motions in Human-Computer Interactions. Although the human hand is a complex system which can perform multiple actions, when the kinaesthetic actions are scaled in a system, the applications are limitless. In this thesis, we examine the effect of control movement scale on user's kinaesthetic actions. We use Fitts's Law for quantifying the user's performance on different scales and to verify if higher control movement scale, in general, can be better than natural movements in tasks which require extended accuracy. The experiment consists of a Wacom™ tablet as an input device connected to a system. The tablet provides means for scaling the kinaesthetic input movement of a user. The experiment is a modified version of the classical multi-directional tapping task. It was performed on 16 healthy participants with ages between 20 to 48 years. The Fitts's regressions were visualised and the Z-scores were computed. It was found that the performance of the participants increases with the scale and has an optimum scale at 1:3.3 before reducing rapidly.

The fourth part of the thesis improves the 2D quantification of user performance measure to 3D interfaces. Considering 3D interactions in Virtual-Reality (VR), it is critical to study how visual awareness of real hands influences users' scaled interaction performance in different VR environments. We used Fitts's law to analyze user performance with five different Control-Display (CD) ratios (1:1 to 1:5). Fifteen participants performed a 3D selection task in three different setups: Head-Mounted Display (HMD), and two variations of the Active One-walled 3D Projection (AOP), with and without visual awareness of the real hand (AOP-A and AOP-B, respectively). The results show that the throughput of AOP-B is significantly higher than that of the AOP-A and HMD (p = 0.00001 and 0.0002, respectively) which suggests the existence of a conflict between the kinesthetic and visual real-hand movements, which we term as Virtual Kinesthetic Conflict (VKC). To reduce VKC during scaled movements, tasks should be designed such that the visual awareness of the real hand is avoided.

The last part of the thesis extends the Fitts's law as a means for training fine-motor skills such as Microscopic Selection Task (MST) for robot-assisted surgery using Virtual Reality with objective quantification of performance. We also introduce Vibrotactile Feedback (VTFB) to study its impact on training performance. We use a VR-based environment to perform MST with varying degrees of difficulties. Using a well-known Human-Computer Interaction paradigm and incorporating VTFB, we quantify the performance: speed, precision, and accuracy. MST with VTFB showed statistically significant improvement in performance metrics leading to faster completion of MST with higher precision and accuracy compared to that without VTFB. The addition of VTFB to VR-based training for robot-assisted surgeries may improve performance outcomes in real robotic surgery. VTFB, along with proposed performance metrics, can be used in training curricula for robot-assisted surgeries.

Using the several approaches mentioned above, we conclude that: 1) Using biomechanical bounds and rules when predicting hand poses enhances the user's visual experience, and 2) using quantified interactions in VR can be useful for maximizing the user performance with such interactions and also useful for training medical procedures, such as MST.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

2D	2-Dimensional
3D	3-Dimensional
3DJE	3 Dimensional Joint Error
AE	Anatomically Error
AEF	Anatomically Error Free
AR	Augmented Reality
AOP	Active One-walled 3D Projection
CD	Control-Display
CNN	Convolutional Neural Network
DoF	Degrees of Freedom
GPU	Graphics Processing Unit
HCI	Human-Computer Interface
HMD	Head Mounted Display
ICVL	Imperial Computer Vision and Learning Lab
IITM	Indian Institute of Technology Madras
MSRA	Microsoft Research Asia
NYU	New York University
SSC-CNN	Single Shot Corrective Convolutional Neural Network
VR	Virtual Reality

CHAPTER 1

INTRODUCTION

Interaction is one of the three pillars (other pillars are Immersion and Presence) of Virtual Reality (VR) (Mütterlein, 2018) which facilitates the feeling of the other two to the user. A user can interact with the 3D virtual world through actions such as pointing, clicking, rotating the head, and so on involving the coordination between human haptic and visual systems. Currently, with the advent of feature-rich virtual reality hardware such as the Oculus Quest, HTC Vive, and Oculus Touch, the user can interact via hand tracking or controllers to interact with the virtual environment (Egger *et al.*, 2017; Borrego *et al.*, 2018). For example, the user can move objects in VR by holding and moving the controller in real space and interact by pressing the various buttons on the controller or use their hands and manipulate the objects as if they are present in real life. These basic interactions, namely pointing (or selection) and reaching a target in VR constitute two of the fundamental interactions using a virtual cursor.

1.1 Realistic 3D Hand Tracking

3D hand pose estimation is the task of predicting the pose of the hand in 3D space provided the depth (or 2D) image of the hand. It is used in many fields such as human-computer interactions (Naik *et al.*, 2006; Yeo *et al.*, 2015; Lyubanenko *et al.*, 2017), gesture recognition (Fang *et al.*, 2007), Virtual Reality (VR), and Augmented Reality (AR) (Lee *et al.*, 2019; Ferche *et al.*, 2016; Cameron *et al.*, 2011; Lee *et al.*, 2015). Earlier, tracking methods used markers such as coloured gloves by Wang and Popović (2009) so that they can be tracked using cameras and other such sensors. With the advent of deep learning in computer vision, commercial systems such as Oculus™ and LeapMotion™ are shifting from marker-based tracking methods to purely vision-based hand tracking. The shift prevents the need to wear cumbersome equipment, which affects the user experience. However, marker-less pose estimation is a challenging task as there are several factors such as the complexity of the hand poses, background noise, and occlusions. Model-based tracking (de La Gorce *et al.*, 2008; Hamer *et al.*, 2009; Oikonomidis *et al.*, 2010; Stenger *et al.*, 2001) creates a 3D model of the hand and aligns it according to the visual data provided. Tagliasacchi *et al.* (2015) made a fast 3D model-based tracking using gradient-based optimization to track the hand position and pose. Sridhar *et al.* (2015) made a robust detection-guided optimization strategy to track the complicated poses of the human hand at speeds of about 50 frames per second (fps) without a Graphical Processing Unit (GPU).

A key problem overlooked by several state-of-the-art models is the realism of the output hand pose. Current state-of-the-art models focus on the accuracy of the model rather than the overall anatomical correctness of the model and reported low errors in benchmark tests such as the ICVL (Tang *et al.*, 2016), NYU (Tompson *et al.*, 2014), MSRA (Sun *et al.*, 2015), BigHand2.2M (Yuan *et al.*, 2017) and HANDS2017. It is possible to train a model to match almost all hand joints when tested; however, when the error is a finger bent in the opposite direction, it can affect the user experience. The error can also affect the human system, leading to false information and mismatch in the motor cortex and the visual system as pointed out by Pelphrey *et al.* (2005). Another issue is the validity of the ground truth of the benchmark tests as it is uncertain if the ground truth hand poses are truly accurate representations of the real human hand with no anatomical errors.

1.2 Scaled Human-Computer Interactions

A scaled motion of the virtual cursor is used in most applications in order to achieve more immersion (Biocca and Delaney, 1995; Slater and Wilbur, 1997) and allows the user to interact with more objects in the virtual world (Wilson *et al.*, 2018; Xie *et al.*, 2010; Steinicke *et al.*, 2008*b*,*a*; Jaekl *et al.*, 2005). The concept of cursor movement scaling in human-computer interaction is otherwise called as Control-Display (CD) ratio. It is different from redirected touching (in works done by Azmandian *et al.* (2016)) due to the fact that the scaling is constant and is dependent on the kinaesthetic motion of the user only. According to MacKenzie and Riddersma (1994), it is the ratio between the physical movement of the input device in the real world and the corresponding movement of the cursor in the virtual world. Changing the CD ratio allows the virtual cursor to make movements which can be larger or smaller than the movement of the real hand. It also increases the effective workspace of the environments which is a difficult task in terms of hardware. When changing the CD ratio, the effective performance of the user must be known since it is crucial for the optimized usage of the HCI device for the task that the device is meant to do. In order to objectively quantify the interaction performance, a HCI law called as Fitts's law is used. The Fitts's Law (Fitts, 1954) states that there is a relation between the time it takes for a task to be completed and the difficulty of the task. This relation is linear, and the modified Fitts's Law by MacKenzie and Riddersma (1994) is used in this thesis, which is developed in analogy with Shannon's information theory (Shannon, 1948).

One of the problems observed in current HCI devices is the lack of quantified interaction options with regards to 3D interfaces. Fitts's Law in particular is originally used for 2D based interactions, namely using pen and paper. MacKenzie then updated this model to accommodate HCI devices such as the keyboard, mouse, trackball, etc. However such well established quantifications is not present for currently trending 3D interfaces such as the Oculus Rift and the HTC Vive. This problem shall be addressed in this thesis.

1.3 Motivation

One of the critical problems in hand tracking is the realism of the output. This problem of hand pose realism has been studied in a partial aspect as "highly accurate tracking" in earlier work such as those by Sharp *et al.* (2015) as increasing the tracker's accuracy and reducing the poses' overall position-based error. This error can disrupt the immersiveness of the individual during the virtual experience. Moreover, from a human perspective (Pelphrey *et al.*, 2005), the error can affect the internal human system leading to false information and mismatch in the motor cortex and the visual system.

Regarding HCI devices, there are well-established studies and laws which provides optimum settings and characteristics for maximum user performance in certain tasks. However, there is a lack of such established performance characteristics and optimum settings for 3D HCI devices such as the HTC Vive or 3D projectors. These are the primary motivations for pursuing research on the topic of **Realistic and Quantified Human-Computer Interaction**.

1.4 Hypothesis

In light of the problems discussion in the previous section, we frame the following hypotheses with respect to the two problem statements:

(a) Whether improving the realism of hand poses predicted by the state-of-the-art trackers using biomechanical aspects improves the accuracy as compared to the poses without implementing such concepts.

(b) Can a hand pose estimator guarantee zero anatomical error while maintaining low deviation from the ground truth pose.

(c) Can higher control movement scale, in general, can be better than natural kinaesthetic movements (where control movement scale = 1:1) in tasks which require extended accuracy using 2D interfaces.

(d) Can higher control movement scale, in general, can be better than natural kinaesthetic movements (where control movement scale = 1:1) in tasks which require extended accuracy using 3D interfaces.

(e) Can HCI based quantification methods be used as a training tool for important tasks such as computer based surgery.

1.5 Aim and Scope

This thesis aims to develop a 3D hand tracker that will utilize biomechanical constraints as a closed form equation. This thesis also analyses existing 3D interfaces to provide an optimum setting using Fitts's Law. The scope of this thesis is limited to solve two distinct problems in Human-Computer Interaction (HCI), namely realism of 3D hand tracking and quantification of 2D and 3D HCI.

1.6 Research Objectives

The following research objectives were set based on the problems discussed in the previous sections:

(a) Realistic 3D Hand Tracking

- Design a biomechanical filter that can correct the hand pose such that it guarantees zero anatomical error while maintaining low deviation from the ground truth pose.
- Design a neural network that incorporates the biomechanical filter in the architecture level such that it increases the speed of computation and efficiency of the network.

(b) Scaled Human-Computer Interactions

- Analyse the effect of changing the control display ratio of 2D interfaces and quantify the scaled interactions of the user using those 2D interfaces.
- Upgrade the quantification method to analyse and quantify 3D interfaces.
- Utilise the 3D quantification method in an applied environment such as robotic surgery.

1.7 Contributions of the Thesis

(a) We propose a filter that functions on the human hand's biomechanical principles and kinematics. This filter's novelty is the use of bounds and rules derived from the human hand's biomechanical aspects to produce a more realistic rendering of the hand pose. The filter can be plugged into any hand tracker and enhance its performance.

(b) We integrate the filter defined in the previous contribution inside a CNN architecture called Single Shot Corrective CNN (SSC-CNN) and build a hand pose estimator that guarantees zero anatomical error while maintaining low deviation from the ground truth pose. We also show that these anatomical rules and bounds were not maintained when creating the HANDS2017 and the MSRA hand datasets, and an Anatomical Error Free (AEF) version of the datasets called AEF-HANDS2017 and AEF-MSRA was created.

(c) We examine the effect of control movement scale on user's kinaesthetic actions. We use the Fitts' Law for quantifying the user's performance on different scales and to show that at an optimum control movement scale, users perform better than natural movements.

(d) Using the Fitts's Law from the previous contribution, we examine how the effect of visual awareness of the real hand influence the performance of interaction with the virtual objects using a virtual cursor in a Virtual Environment (VE). To study the effect, we have compared the user performance of same 3D tasks in two different VEs, namely the Active One-walled 3D Projection (AOP), and the Head Mounted Display (HMD).

(e) Using the Fitts's Law from the previous contribution, we create a VR based training system that allows for the objective evaluation and learning of psychomotor skills for robotic surgery, specifically fine motor skills. We include vibrotactile feedback (VTFB) in our training system to improve the performance of fine motor skills and augment the experience of human-computer interaction. This study proposes a performance index and compares the performance with and without VTFB.

1.8 Organisation of the Thesis

The thesis starts with Chapter 2 detailing a filter that will correct the hand pose from state-of-the-art hand trackers which in turn increases the realisticity of the output poses. Chapter 3 integrates the filter into the architecture of the hand tracker itself which increases efficiency of computations and provides a better hand pose. Codes for Chapter 2 and 3 are available online[1]. Chapter 4 describes the scaled HCI interactions in 2D and explains the use of Fitts's Law with respect to 2D interfaces with an experiment using the Wacom graphics tablet. Chapter 5 extends the Fitts's Law to 3D interfaces and utilizes the law to quantify two distinct interfaces, namely the 3D projector and the Head Mounted Display. Chapter 6 further extends the use of Fitts's Law as a training interface for surgical tasks. Finally, Chapter 7 is the conclusion of the current work along with future scopes for 3D hand tracking as well as scaled interactions.

[1] `https://github.com/RBC-DSAI-IITM/SSCCNN`

CHAPTER 2

ANATOMICAL FILTER FOR HAND POSE ESTIMATION

2.1 Introduction

One of the critical problems in hand tracking is the realism of the output. This problem of hand pose realism has been studied in a partial aspect as “highly accurate tracking" in earlier work as increasing the tracker’s accuracy and reducing the poses’ overall position-based error. Many studies overlooked this problem by focusing solely on the accuracy of the hand tracking models. Such models have low errors in benchmark tests such as the NYU (Tompson *et al.*, 2014), ICVL (Tang *et al.*, 2016), HANDS2017 and BigHand2.2M (Yuan *et al.*, 2017). However, high accuracy does not always translate to realistic hand output. Such an example is a hypothetical case of a hand pose that matches all joint positions of the actual hand pose except one joint, which is at an anatomically implausible angle from the previous joint (such as a finger bent backward). This error can disrupt the immersiveness of the individual during the game or simulation. Moreover, from a human perspective (Pelphrey *et al.*, 2005), the error can affect the internal human system leading to false information and mismatch in the motor cortex and the visual system. Other solutions to this problem include inverse kinematics-based solutions such as Wang and Popović (2009) and using kinematic priors such as Thayananthan *et al.* (2003). However, these solutions are tailor-made for their hand trackers and not built for generic use. Another limitation on using kinematic priors is that such models need to be trained using data collected through several methods such as reference models (Sun *et al.*, 2015) or sensors (Yuan *et al.*, 2017). Hence any anatomic errors present in the data is then propagated in the kinematic model as well. Hence, this problem is the focus and motivation of our work.

This chapter proposes a filter that functions on the human hand’s biomechanical principles and kinematics. This filter’s novelty is the use of bounds and rules derived from the human hand's biomechanical aspects to produce a more realistic rendering of the hand pose. The hand is an articulated body with joints and corresponding bounds (Gustus *et al.*, 2012), and the filter is created using these rules and bounds. The input is the pose of the human hand in the form of joint locations and angles from the hand tracker and outputs the closest possible hand as per the real human hand's bounds. The filter can be plugged into any hand tracker and enhance its performance. Later in Section 2.5.2, we show that the proposed filter improves the realism of the hand poses predicted by the state-of-the-art trackers as compared to the poses without using the filter. We also elaborate on the filter rules and bounds in Section 2.3.

2.2 Related Work

In this section, we discuss a few state-of-the-art methods for 3D hand tracking. Joo *et al.* (2014) proposed a real-time hand tracker using the Depth Adaptive Mean Shift algorithm, a variant of the classic computer vision method known as CAM - Shift by Bradski (1998). It tracks the hand in real-time, however, only in two dimensions due to the limitations of traditional computer vision techniques. Other similar 2D-based trackers include works done by El Sibai *et al.* (2017) and Held *et al.* (2016). Taylor *et al.* (2016) proposed an efficient and fast 3D hand tracker algorithm that utilizes only the CPU to track the hand using iterative methods. This method's drawback is that the hand is treated as a smooth body, and the joints and bones are not distinguished in the model, frequently resulting in anatomically implausible hand structures when tracking.

Recent state-of-the-art models utilize deep learning to achieve highly accurate 3D trackers with low errors in the order of millimeters. Deep learning provides new perspectives to computer vision problems with 3D Convolutional Neural Networks (CNNs) (Ge *et al.*, 2017; Simon *et al.*, 2019) and other such models. There are many survey works and literature available in the field of hand tracking concerning appearance and model-based hand trackers using depth images such as those done by Deng *et al.* (2018); Sagayam and Hemanth (2017); Li *et al.* (2019); Dang *et al.* (2019). Model-based tracking (de La Gorce *et al.*, 2008; Hamer *et al.*, 2009; Oikonomidis *et al.*, 2010; Stenger *et al.*, 2001) creates a 3D model of the hand and aligns it according to the visual data provided. Tagliasacchi *et al.* (2015) made a fast 3D model-based tracking using gradient-based optimization to track the hand position and pose. The drawback of this method is that a wristband must be worn on the hand to be tracked, and the model does not incorporate the angular velocity bounds of the human hand. Although the angle bounds are incorporated in the model, during certain conditions, the hand pose derived from the algorithm results in hand poses, which are impossible for a natural hand. Other models still suffer from heavy computational requirements, such as 3D CNNs, which require voxelization (Ge *et al.*, 2017) of the image for pose estimation. Works done by Sharp *et al.* (2015); Wan *et al.* (2018); Malik *et al.* (2018*b*); Wan *et al.* (2018); Xiong *et al.* (2019); Kha Gia Quach *et al.* (2016) proposed fast 3D hand trackers with high accuracy, but at the expense of heavy computational algorithms and can track only a single hand. Works such as Deng *et al.* (2017); Misra and Laskar (2017); Roy *et al.* (2017) utilize deep learning for hand tracking but in 2D.

Focusing on realism and multi-hand interaction, Mueller *et al.* (2019) proposed a model that uses a single depth camera to track hands while they move and interact with each other. It can also take the fingers' collision with the other hand into account to a certain degree. It was trained using available and synthetically created data as well. This method's drawback is that it is computationally expensive and cannot predict poses when the hand moves very fast. There are also discrepancies in some interactions when the calibration is imperfect.

To the best of our knowledge, none of the existing hand tracking approaches have explicitly corrected the predicted pose by using a filter based on the biomechanics principle as is being proposed in this Chapter. The main contributions of this Chapter are:

(a) A filter based on the human hand's biomechanics, ensuring that the output of the hand tracker conforms to the rules of true human hand kinematics and enhances the immersiveness of the end application.

(b) An approach of adding a modular filter that can be easily plugged into an existing hand tracker with little or no modifications.

(c) Increased hand pose realism of the output of the tracker to which the filter is attached to.

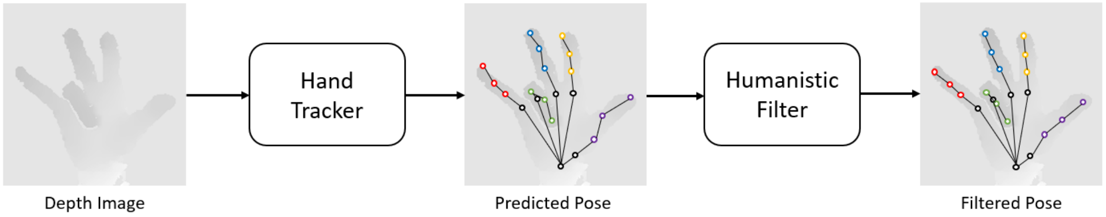


Fig. 2.1: Overview of the anatomical filter. The depth image first passes through the hand tracker, and then a pose is retrieved (the unaltered pose). This pose then passes through the anatomical filter, and then the filtered pose is given as the output.

2.3 Anatomical filter

The anatomical filter takes the pose from the tracker as input and then adjusts the individual joint angles according to their biomechanical limits. The overview of the filter is shown in Figure 2.1. Section 2.3.1 describes the construction and working of the anatomical filter. Section 2.3.2 describes the anatomical bounds and rules used to create the anatomical filter.

2.3.1 Filter construction

The filter utilizes the anatomical bounds and corrects the hand pose according to those bounds. The first step is to calculate the joint angles since most hand trackers' output is the joint's location in 3D space and not the joint's angle of rotation. The angle of the joints is computed separately using 3D transformations such that each joint chain is aligned on the XY plane. Then, the Euler angles are calculated using the vectors computed from each pair of joints.

The second step is to calculate the deviation of each joint from its limit. Considering the current joint angle of a particular joint as $\theta_{\mathrm{c}} = [\theta_x, \theta_y, \theta_z]$, where θ_x, θ_y, and θ_z are the individual angles to each axis, the anatomical error of the particular joint is derived in equation 2.1.

$$\varepsilon_{\mathrm{d}}^{\theta} = \begin{cases} \theta_{\mathrm{d}} - \theta_{\mathrm{upper}} & \text{if } \theta_{\mathrm{d}} > \theta_{\mathrm{upper}} \\ \theta_{\mathrm{lower}} - \theta_{\mathrm{d}} & \text{if } \theta_{\mathrm{d}} < \theta_{\mathrm{lower}} \\ 0 & \text{otherwise} \end{cases} \quad \text{where } d = x, y, z \tag{2.1}$$

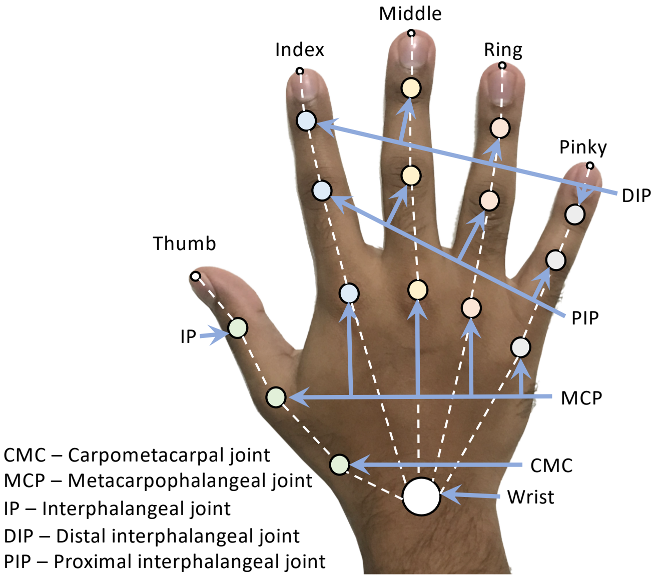

Middle
Index
Ring
Pinky
DIP
Thumb
PIP
IP
MCP
CMC
Wrist
CMC – Carpometacarpal joint
MCP – Metacarpophalangeal joint
IP – Interphalangeal joint
DIP – Distal interphalangeal joint
PIP – Proximal interphalangeal joint

Fig. 2.2: Structure of the human hand.

The third step is to correct the joint's angle using the error derived from equation 2.1. In order to control the amount of correction made, we introduce a parameter α and is shown in equation 2.2. The intuition is to vary the strength of the filter and study the effect of the filter's strength on the overall hand pose accuracy.

$$\theta_{\mathrm{d(new)}} = \begin{cases} \theta_{\mathrm{d}} - \alpha * \varepsilon_{\mathrm{d}}^{\theta} & \text{if } \theta_{\mathrm{d}} > \theta_{\mathrm{upper}} \\ \theta_{\mathrm{d}} + \alpha * \varepsilon_{\mathrm{d}}^{\theta} & \text{if } \theta_{\mathrm{d}} < \theta_{\mathrm{lower}} \\ \theta_{\mathrm{d}} & \text{otherwise} \end{cases} \tag{2.2}$$

where $d = x, y, z$ and $\alpha \in [0, 1]$. If $\alpha = 0$, then there is no correction and the resultant angle is the original angle. If $\alpha = 1$, then the angle is 100$\%$ corrected based on the hand's biomechanical rules. For example, if a chain of joints of a finger has an error of 30° and is corrected by a factor $\alpha = 0.5$, then 15° will be corrected in the chain of joints. This in turn enables us to see the apparent shift of the joints from the ground truth pose and hence its effect on the overall 3D joint error from the ground truth.

2.3.2 Biomechanics of the hand

In the human hand, there are 27 bones with 36 articulations and 39 active muscles (Ross and Lamperti, 2006), as shown in Figure 2.2. According to Kehr and Graftiaux (2017), the lower arm's distal area consists of the distal radio-ulnar joint, the thumb and finger carpometacarpal (CMC) joints, palm, and fingers. These muscles map up to 19 degrees of freedom with complex functions such as grasping and object manipulation. The key joints for the movements of the hand are:

(a) Metacarpophalangeal (MCP) joint

(b) Distal interphalangeal (DIP) joint

(c) Proximal interphalangeal (PIP) joint

(d) Carpometacarpal (CMC) joint

The wrist is simplified to six degrees of freedom (DoF), consisting of 3 DoFs for movement and 3 DoFs for rotation across the three axes. The thumb's CMC joint is integrated into the wrist and is an important joint since it enables a wide range of hand movements by performing the thumb's opposition. According to Chim (2017), the CMC joint has 3 DoFs: 45° abduction and 0° adduction, 20° flexion and 45° extension, and 10° of rotation.

There are five MCP joints in which the first MCP joint is connected to the thumb's CMC joint. The remaining four MCP joints are attached to the wrist of the hand. The MCP joint of the thumb is a 2 DoF joint that provides flexion 80° and extension 0°, abduction 12° and adduction 7°. The remaining MCP joints are also 2 DoF joints and provide flexion 90° and extension 40°, as well as abduction 15° and adduction 15°. Clear illustrations and details regarding these bounds can be found in works done by Hochschild (2015) and Ross and Lamperti (2006).

There are two types of interphalangeal (IP) joints: the distal and proximal (DIP and PIP) joints. The thumb only has a single IP joint, while the other fingers have both DIP and PIP joints. The PIP joints provide flexion 130° and extension 0°. The DIP joints, including the thumb IP joint, provide flexion 90° and extension 30°.

These rules and bounds are shown in Table 2.1 and incorporated into the construction of the filter. When the filter activates, each joint of the hand-pose is compared with these rules and then corrected to output a hand-pose that conforms to the hand's biomechanics.

Table 2.1: Angular bounds for each joint of the hand derived from the works of Hochschild (2015) and Ross and Lamperti (2006)

Joint	**Maximum angle**	**Minimum angle**
CMC abduction and adduction	45°	0°
CMC extension and flexion	45°	-20°
Thumb MCP flexion and extension	80°	0°
Thumb MCP abduction and adduction	12°	-7°
Thumb IP flexion and extension	90°	-30°
Index, middle, ring and pinky MCP flexion and extension	90°	-40°
Index, middle, ring and pinky MCP abduction and adduction	15°	-15°
Index, middle, ring and pinky PIP flexion and extension	130°	0°
Index, middle, ring and pinky DIP flexion and extension	90°	-30°

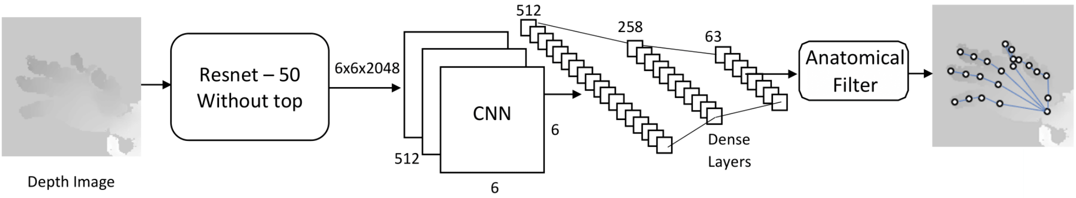


Fig. 2.3: Architecture of the model used for the hand tracker with the anatomical filter.

2.4 Baseline hand-tracking model

To compare the state-of-the-art trackers with the anatomical filter, we made a simple hand tracker to serve as a baseline model. The baseline model is trained with the filter attached to compare with the other state-of-the-art models that were not trained with such filters.

2.4.1 Architecture

We created our hand tracker using the ResNet-50 (He *et al.*, 2016) as a backbone with transfer learning (Torrey and Shavlik, 2010) to utilize the powerful model for 3D hand pose detection. The architecture is shown in Figure 2.3, and the process diagram is shown in Figure 2.1. Since the ResNet originally performs classification using a softmax layer, we use the model without the top classification layer, which results in an output of size 6 × 6 × 2048. After pre-processing, the input image size is 176 × 176 × 1, which is then replicated for the three channels as the input to the backbone model should be a 3-channel image. The output features from the backbone model are then compressed by passing it through a single convolutional layer of size 512 × 6 × 6. The resultant features are flattened (to size 512 × 1) and then pass through two fully connected dense layers of sizes 258 and 63, respectively. The first dense layer uses a ReLU activation function, whereas the last layer uses a linear activation function. This output is filtered using our anatomical filter, and then the estimated pose is retrieved. The code was built using Keras and used the Adam optimizer (Kingma and Ba, 2014) with the learning rate set to 0.00035. The model trained on the entire training data with 20% of the data for validation until there was no improvement in validation error for five epochs.

2.4.2 Dataset used

The dataset used for the evaluation is the HANDS2017 (Yuan *et al.*, 2017), which consists of more than 900,000 images for training and 99 video segments of depth images for testing pose estimators. The images consist of various poses that are complex and challenging to estimate the correct pose. Our model is first used without any filter to evaluate it on the dataset, and then the anatomical filter is used to correct the hand pose. Then the whole system is re-evaluated with a grid search to incorporate all possible α values. To use the filter on the current state-of-the-art A2J model (Xiong *et al.*, 2019) and V2V-Posenet (Moon *et al.*, 2018), the “frames" subset of the HANDS2017 dataset is used, which contains 295510 independent hand images that cover a wide variety of challenging hand poses.

2.5 Results and analysis

The focus of this Chapter is on improving the realism of the predicted hand poses. We designed the following experiments to demonstrate that our proposed method can work with any pose prediction model.

(a) We study the effect of the filter on the output of various state-of-the-art trackers. We chose a simple baseline model, the A2J model, and the V2V-Posenet model as the trackers. We show that the outputs are more realistic when corrected by the anatomical filter.

(b) We quantify the anatomical error and show how the filter reduces this error with various configurations.

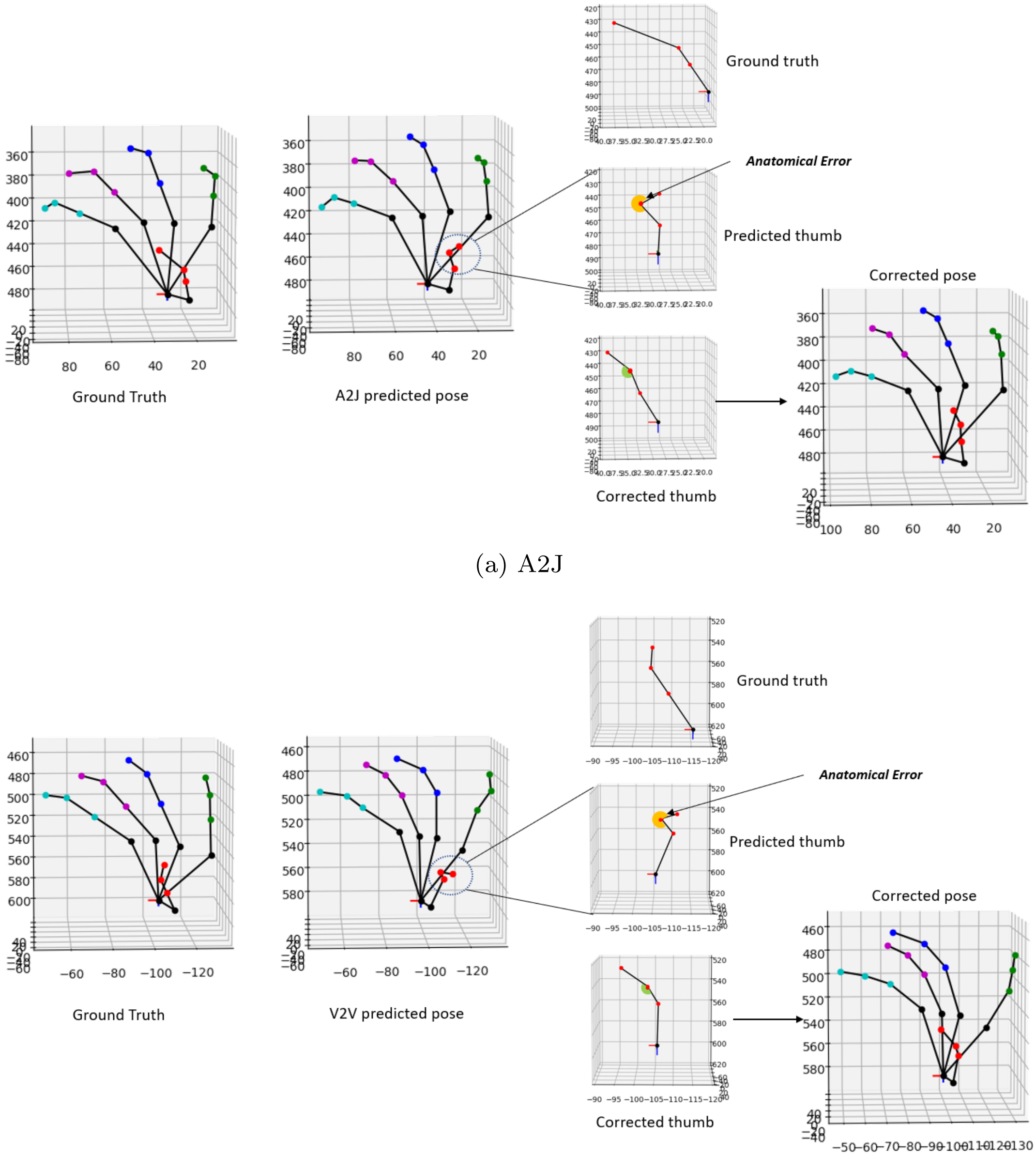
(c) We study the effect of α on the baseline model using the filter.

(d) We show the best-case and worst-case scenarios of the filter correction.

(e) We test the error of the state-of-the-art models using the filter with various configurations.

2.5.1 Filter function on the state-of-the-art trackers

To understand the filter's function, Figure 2.4 shows the working of the filter for a single frame of the dataset. Figure 2.4a shows the A2J model prediction of a simple pose in the dataset and our filter's correction of the pose. The figure shows that the thumb is bent in an anatomically implausible manner, shown in detail (selected by a dotted circle). The highlighted angle in yellow is known as the anatomical error (shown in Figure 2.4a), and the anatomical filter corrects this error. The corrected angle is shown in green, and the process is repeated for all joints. The resulting pose is shown in Figure 2.4a as the corrected pose. A similar scenario is shown in Figure 2.4b for the V2V-Posenet model. These discrepancies in the poses disrupt the user experience if used in an immersive application such as gaming or simulation-based training programs. Our filter corrects these errors at the minor expense of overall 3D error, resulting in a smoother application experience.



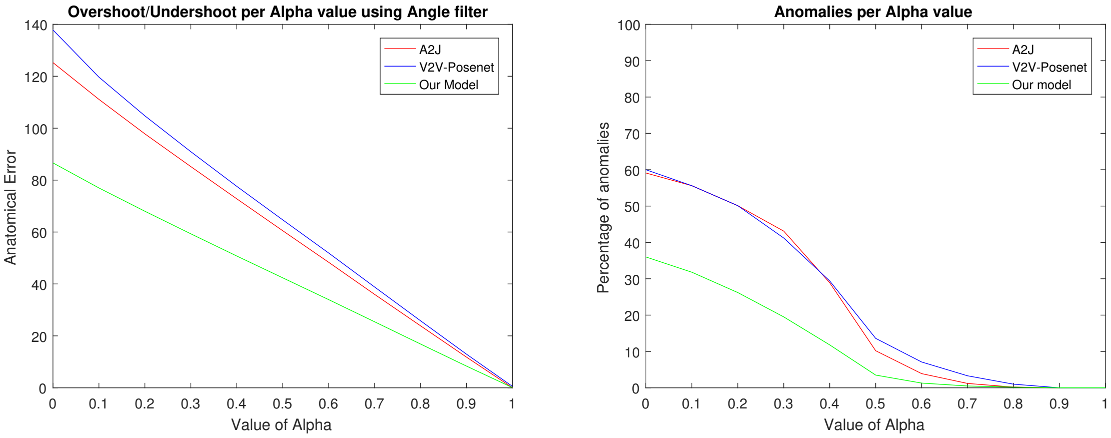
(a) A2J

(b) V2V-Posenet

Fig. 2.4: The type of corrections performed by the anatomical filter. The dotted circles indicate an anomaly in the joint. In the zoomed graph, the yellow semi-circle denotes the joint which has an error, and the green semi-circle denotes the corrected joint.

Table 2.2: Percentage of poses with anatomical anomalies at the specified ranges, comparing the baseline model with the state-of-the-art models with no correction ($\alpha = 0$). The test was performed on a subset of 20000 test images of the HANDS2017 dataset.

Model	Percentage of poses with anatomical anomalies		
	0°-50°	51°-100°	>100°
Baseline model	35.6%	28.1%	36.3%
A2J	23.3%	17.4%	59.3%
V2V	20.8%	19%	60.2%
After Anatomical filter (Any model)	0%	0%	0%



(a)

(b)

Fig. 2.5: Graphical visualization of the anatomical errors of two state-of-the-art models, namely A2J (Xiong *et al.*, 2019) and V2V-Posenet (Moon *et al.*, 2018) compared to our model using the angle filter attached to the end of the model for every value of α. The x-axis corresponds to the value of α used for the filter. The y-axis in Figure 2.5a corresponds to the model's anatomical error, which is the mean joint degree that overshoots or undershoots the anatomical bounds of the corresponding joint of the hand. In Figure 2.5b, the y-axis corresponds to the percentage of frames in which the anatomical error exceeded 100 degrees.

2.5.2 Anatomical anomaly test

To quantify the direct factor relating to the anatomical structure-based realism of the human hand pose, we derive a quantity that we refer to as the **anatomical error** and shown in Equation 2.3. This error is derived for the three models and shown in Figure 2.5, which is the mean joint degree that overshoots or undershoots the anatomical bounds of the corresponding joint of the hand. The higher the error, the more “unreal" the given hand pose is according to the hand’s anatomical structure. The error is high for both the A2J and V2V-Posenet models, which reduces smoothly as α increases. This reduction is because α directly controls these errors in the filter. Figure 2.5b shows the percentage of frames in which the hand pose has an anatomical error above 100 degrees. The quantified results for these tests are shown in Table 2.2. We infer from the graph and table that our model predicts more realistic poses with lower anatomical errors with a small trade-off with 3D Joint Position Error.

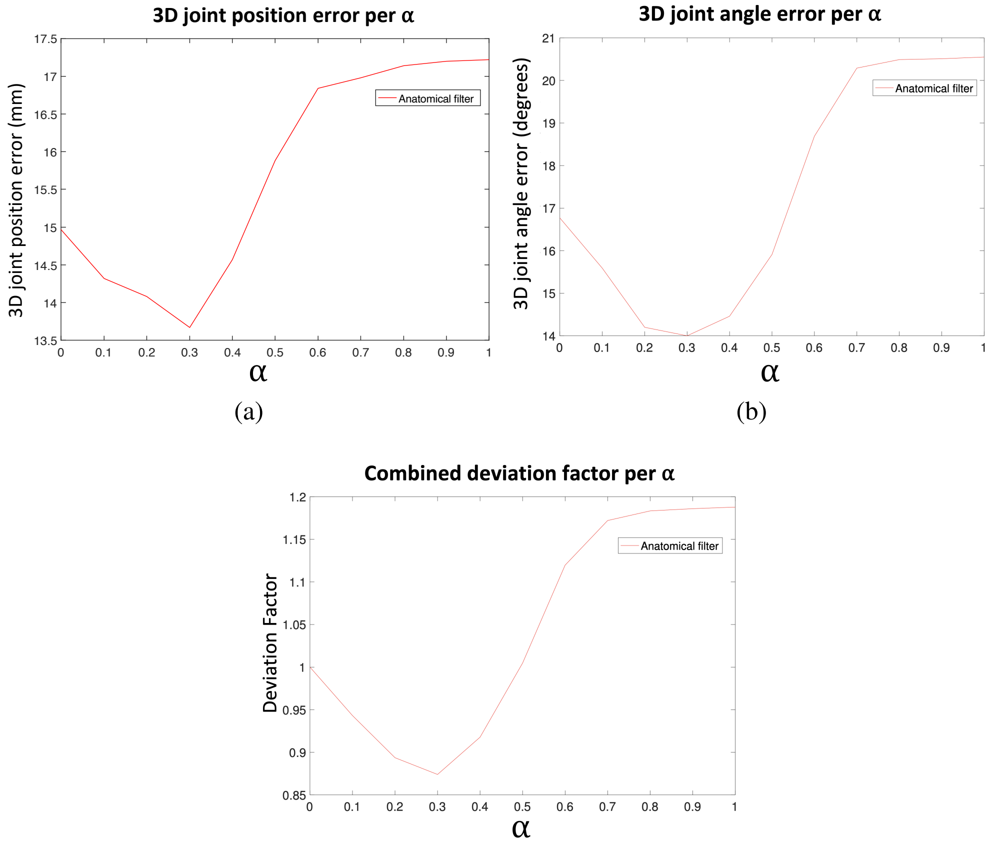
$$\text{Anatomical Error} = \begin{cases} \theta - \theta_{\text{upper}} & \text{if } \theta > \theta_{\text{upper}} \\ \theta_{\text{lower}} - \theta & \text{if } \theta < \theta_{\text{lower}} \end{cases} \tag{2.3}$$

2.5.3 Effect of α on our model using the anatomical filter

The mean **3D joint position error** is usually computed for 3D hand tracking models, which is computed by calculating the individual 21 joint distances from the estimated model to the ground truth pose and deriving the mean of that sum. The mean is then computed for each video segment. To measure the hand pose’s error, we introduce a metric known as **3D joint angle error** and is given in equation 2.4.

$$\text{3D joint angle error} = \frac{1}{26}\sum_{i=1}^{26} |\theta^i_{Predicted} - \theta^i_{Actual}| \tag{2.4}$$

The 3D joint angle error is similar to the position error; however, this error measures the difference between the 26-DoF vector derived from the joint locations as per Section 2.3.2. Together, these two errors represent the **3D joint pose error**. First, the 3D joint position and angle errors of our model are calculated for different α values. A graphical representation of the results is shown in Figure 2.6. The x-axis is the α



(a)

(b)

(c)

Fig. 2.6: Graphical visualization of the results computed for every value of α used in the filter on our custom model. The x-axis corresponds to the value of α used for the filter. In Figure 2.6a, the y-axis corresponds to the mean 3D joint position error of the model, which is the mean distance of each joint of the estimated pose to the joint of the corresponding ground truth pose. In Figure 2.6b, the y-axis corresponds to the 3D joint angle error of the model, which is the mean error between the 27 DoF vector of the estimated pose to the corresponding vector of the ground truth pose. Finally, in Figure 2.6c, the y-axis corresponds to the deviation factor, which is the value the error (both joint position and angle errors together) deviates from the point where the filter was not used (unfiltered error).

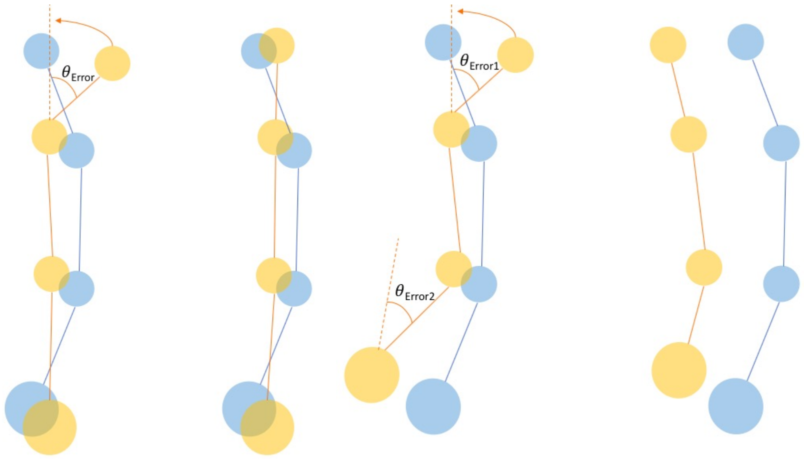
Table 2.3: 3D Joint Errors (3DJE) and 3D Angle Errors (3DAE) derived from the HANDS2017 dataset with all the models.

Model	Filter Used	Lowest 3DJE (mm)	AE at given α	Lowest 3DAE (°)	3DJE with $\alpha = 1$	AE at $\alpha = 1$
Ours	Unfiltered	14.97	88 ($\alpha = 0$)	16.32°	-	0
	Anatomical filter	13.67	61 ($\alpha = 0.3$)	14.13°	17.24	0
A2J	Unfiltered	8.57	125 ($\alpha = 0$)	9.57°	-	0
	Anatomical filter	**8.53**	112 ($\alpha = 0.08$)	**9.56**°	9.62	0
V2V	Unfiltered	9.95	137 ($\alpha = 0$)	12.2°	-	0
	Anatomical filter	**9.94**	121 ($\alpha = 0.075$)	**12.18**°	11.21	0

set for the filter as per equation 2.2. The y-axis represents a different measure for each sub-figure in Figure 2.6. In Figure 2.6a, the y-axis corresponds to the mean 3D joint position error. In Figure 2.6b, the y-axis corresponds to the mean degree error of the model. Finally, in Figure 2.6c, the y-axis corresponds to the deviation factor, which is the value the error deviates from the point where the filter was not used (unfiltered error). Since there are two error metrics computed, each error's deviation is computed separately and then combined using the arithmetic mean. This method is possible since the deviation factor has no unit. For example, a deviation factor of one means that the error did not change from the unfiltered model, and the filter is of no use. However, if the deviation factor is lower than one, then the new model performs better than the unfiltered model and vice versa if the factor is above one. Figure 2.6c shows that the deviation factor is lowest at $\alpha = 0.3$. Hence the model shows the best results when the filter is set at 30% strength. Beyond that value, the deviation factor steadily increases to a point beyond one. This decrease is shown quantitatively in Table 2.3, where the error of the filter is lower than that of the other configurations when $\alpha = 0.3$.

2.5.4 Best-case and worst-case scenarios

When the filter corrects the hand's pose based on the hand biomechanics, inevitably, the hand pose drifts from the original pose. This drift can either make the pose closer to the ground truth or defer from it. The former is the best-case scenario, while the latter is the worst-case scenario. The scenarios are shown in Figure 2.7. The yellow



(a) Best-case (b) Worst-case

Fig. 2.7: Simple 2D illustration for the best-case and worst-case corrections performed by the anatomical filter. The blue circles indicate the joint locations of a single index finger from the ground truth. The yellow circles indicate the position of the estimated joints from the hand tracker.

dots correspond to the predicted joints' position, and the blue dots correspond to the ground truth joints' position. The yellow dots must be as close to the corresponding blue dots as possible, ideally overlapping them. The first case is the positive scenario where one joint error occurred in the pose. When the anatomical filter corrected this pose, the error was reduced. The second case is the non-ideal scenario where the error resides in the bottom joint. When this error is corrected, the secondary joints above the corrected joint all shift their positions, hence drifting from the ground truth. The final correction shifts the distance even more, increasing the total error. This shift results in a hand pose that conforms to the rules. However, the overall pose after correction is now further from the ground truth than the uncorrected poses.

2.5.5 Effect of α on state-of-the-art models using the anatomical filter

In order to study the effect of the filter on the overall 3D joint position error, the filter was tested on the current state-of-the-art A2J model (Xiong *et al.*, 2019) and V2V-Posenet (Moon *et al.*, 2018) using the “frames" subset of the HANDS2017 dataset. Figure 2.8 shows the results of the test using various configurations of the angle filter described in equation 2.2. The position errors at $\alpha = 0$ are the reported errors of

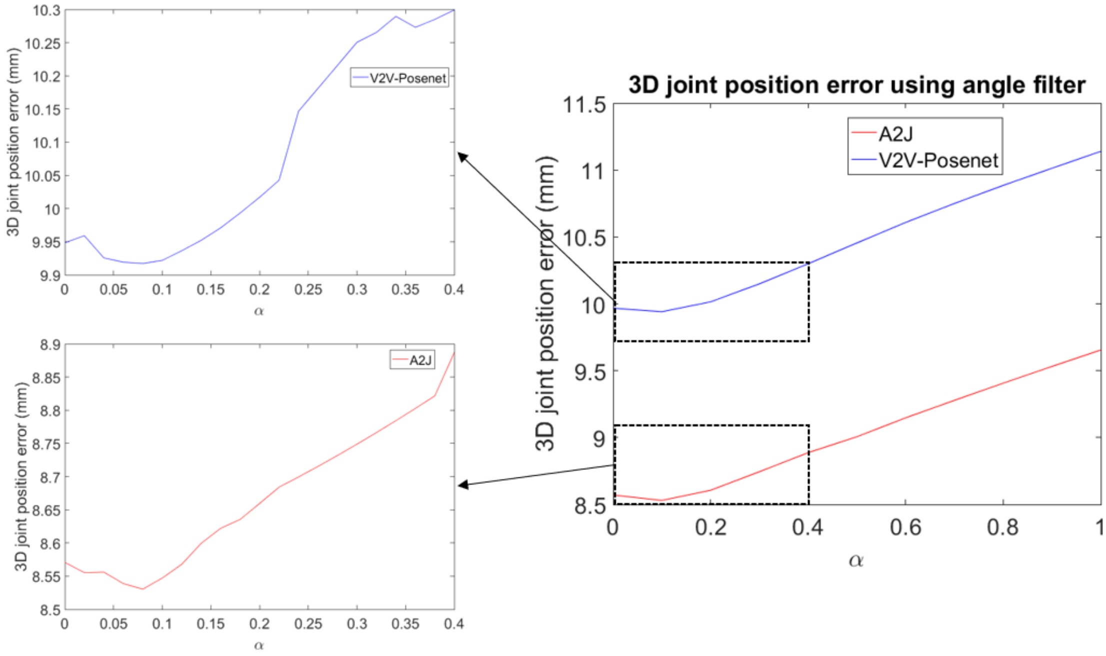


Fig. 2.8: Graphical visualization of the results of two state-of-the-art models, namely A2J (Xiong *et al.*, 2019) and V2V-Posenet (Moon *et al.*, 2018) using the angle filter attached to the end of the model for every value of α. The x-axis corresponds to the value of α used for the filter. The y-axis is the mean 3D joint position error of the model, which is the mean distance of each joint of the estimated pose to the corresponding ground truth pose. Since the improvement is minor, a zoomed version of the selected regions is also shown for the respective models.

8.570 mm and 9.95 mm, respectively, as reported by Xiong *et al.* (2019) and Moon *et al.* (2018). When increasing the filter's strength, the error slightly reduces (8.530 mm and 9.94 mm) and then increases monotonically beyond that value. To visualize the minor changes that occur when α ranges from 0 to 0.4, a smaller test was also performed with alpha ranging from 0 to 0.4 with a step size of 0.02. This test is done for both the A2J model and the V2V-Posenet model, and the individual graphs are also shown in Figure 2.8. From the figure, we derive that at $\alpha = 0.08$, the filter improves the A2J model and $\alpha = 0.075$ for V2V Posenet since the error reduces at the filter strength, seen from both the main graph and the zoomed graphs. The 3D joint error at $\alpha = 0.1$ is 17.24 for the baseline model and 9.62 for the A2J model with $\alpha = 0.08$ and 11.21 for the V2V Posenet model with $\alpha = 0.075$. This shows that the simple baseline model has comparable performance to the state-of-the-art models in terms of anatomical correctness, and using the filter in the model improves the overall performance of the model significantly.

2.6 Summary, limitations and future work

This chapter proposed the anatomical filter, which functions on the human hand's biomechanical principles. The filter is modular and can be easily plugged into existing hand trackers with little or no modifications. The results showed that the filter does improve the current state-of-the-art trackers when $\alpha = 0.1$, and it was also shown that the state-of-the-art trackers have high errors in terms of anatomical rules and bounds.

The filter's computational requirements are high since the angles and bounds are calculated and compared for each joint in the hand. This process increases the time taken to estimate output for each input frame and runs at lower speeds when running real-time tracking. Our future work is to optimize the filter to compute angles and bounds in fewer functions and reduce the time taken to estimate the filtered pose. Optimized methods such as inverse kinematics-based modeling done by Aristidou (2018) can effectively correct the joints in real-time. Future works also include utilizing the law of mobility as per works of Manivannan *et al.* (2009), which states that the two-point discrimination improves from proximal to distal body parts. Hence, the filter's strength can be changed from the hand’s proximal parts towards the hands’ distal part. Other future works include enhanced optimizations such as implementing the filter function into the model architecture instead of attaching the filter at the end of the model. The baseline model used in this chapter highlights the importance of using anatomical rules during training and can improve the model’s accuracy, not only in anatomical correctness but also in pose error. Using the filter inside the model may also reduce training and testing time and reduce excessive computations.

CHAPTER 3

INTEGRATION OF ANATOMICAL FILTER TO A CNN - SINGLE SHOT CORRECTIVE CNN

3.1 Introduction

Hand pose estimation in 3D is the task in which the input is a 3D (or 2D) image of the hand, and the output is the predicted pose of the hand in 3D space. It is used in many fields such as human-computer interactions (Naik *et al.*, 2006; Yeo *et al.*, 2015; Lyubanenko *et al.*, 2017), gesture recognition (Fang *et al.*, 2007), Virtual Reality (VR), and Augmented Reality (AR) (Lee *et al.*, 2019; Ferche *et al.*, 2016; Cameron *et al.*, 2011; Lee *et al.*, 2015). With the advent of deep learning in computer vision, commercial systems such as Oculus™ and LeapMotion™ are shifting from marker-based tracking methods to purely vision-based hand tracking. The shift obliviates the need to wear cumbersome equipment, which affects the user experience. However, marker-less pose estimation is a challenging task as there are several factors such as the complexity of the hand poses, background noise, and occlusions.

A key problem overlooked by several state-of-the-art models is the realism of the output hand pose. Current state-of-the-art models focus on the accuracy as per the closeness to the ground truth pose of the model rather than the overall anatomical correctness of the model and report low errors in benchmark tests such as the ICVL (Tang *et al.*, 2016), NYU (Tompson *et al.*, 2014), MSRA (Sun *et al.*, 2015), BigHand2.2M (Yuan *et al.*, 2017) and HANDS2017 (Yuan *et al.*, 2017). It is possible to train a model to match almost all hand joints when tested; however, when the error is caused by a finger bent in the opposite direction (as shown in Figure 3.1), it can have a negative effect on the user experience. The error can also negatively affect the human system, leading to false information and mismatch in the motor cortex and the visual system (Pelphrey *et al.*, 2005).

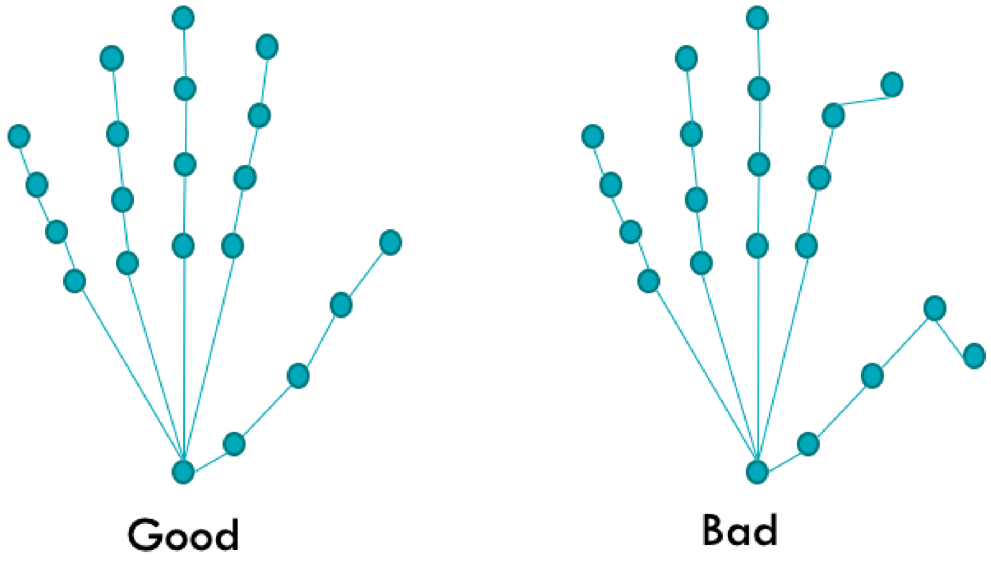


Fig. 3.1: Example of an anatomically incorrect pose. Although most of the joints match the original pose, since two joints are in abnormal angles, the whole pose is considered implausible.

Hence, in this Chapter, we focus on improving the realism of the predicted hand pose and the validity of the dataset. The main metric used for comparison in this Chapter is the anatomical error of the hand pose, which is computed by using the joint angles measured for each joint in the hand pose after prediction. These joint angles were compared with the true biomechanical bounds of the hand (discussed in Section 2.3.2). The absolute error between the true bound and predicted joint angle was then calculated for every joint and added together. This value is denoted as the **anatomical error**, and its unit is in degrees. The mean anatomical error of the joints was also reported, and this process was repeated for every hand pose prediction. During the experiments, we observe that the ground truth of the dataset itself contains many anatomical errors in many instances. We address this issue by proposing a new corrected ground truth that conforms to the anatomical bounds of a true human hand.

Earlier approaches to incorporate anatomical information usually take the form of a hand pose filter applied post-facto after the prediction of the pose to correct for anatomical errors (Aristidou, 2018; Chen Chen *et al.*, 2013; Tompson *et al.*, 2014). Post-processing often leads to significant computational overhead. We present a novel approach that we call the Single Shot Corrective CNN (SSC-CNN) that provides a highly accurate hand pose estimation with no anatomical errors by applying corrective functions in the forward pass of the neural network using three separate networks. The term "Single Shot" implies that the model will process the hand-pose and ensure the anatomical correctness in a single forward pass of the network. This ensures that the initial prediction from the network is free of anatomical errors and prevents the need for any correction using a post-processing function.

3.2 Related Works

This section discusses hand pose estimation methods that use deep learning algorithms and hand pose estimators with biomechanics-related features such as anatomical bounds.

3.2.1 Pose Estimation with Deep Learning

Hand pose estimation using deep learning algorithms can be classified into discriminative and model-based methods. The former category directly regresses the joint locations of the hand using deep networks such as CNNs (Chen *et al.*, 2019; Cai *et al.*, 2019; Simon *et al.*, 2017; Xiong *et al.*, 2019; Poier *et al.*, 2019; Rad *et al.*, 2018; Moon *et al.*, 2018; Guo *et al.*, 2017; Ge *et al.*, 2017; Malik *et al.*, 2018*a*). The latter category abstracts a model of the human hand and fits the model with minimum error (such as the mean distance between ground truth and predicted hand pose joints) on the input data (Malik *et al.*, 2018*b*; Vollmer *et al.*, 1999; Ge *et al.*, 2018*b*; Oberweger and Lepetit, 2017; Taylor *et al.*, 2016). Directly regressing the joint locations achieves high accuracy poses but suffers from issues such as the hand's structural properties. Works done by Li and Lee (2019) and Xiong *et al.* (2019) used cost functions taking only the joint locations of the hands into account and no structural properties of the hand. Moon *et al.* (2018) proposed the V2V Posenet, which converts the 2D depth image into a 3D voxelized grid and then predicts the joint positions of the hand. The cost function of the V2V algorithm used the joint locations alone for training and did not consider biomechanical constraints such as the joint angles.

3.2.2 Pose Estimation with Biomechanical Constraints

Biomechanical constraints are well studied in earlier works to enable anatomically correct hand poses using structural limits of the hands (Wan *et al.*, 2019; Dibra *et al.*, 2017; Tompson *et al.*, 2014; Melax *et al.*, 2013; Sridhar *et al.*, 2013; Xu and Cheng, 2013; Cobos *et al.*, 2008; Ryf and Weymann, 1995; Aristidou, 2018; Spurr *et al.*, 2020; Chen Chen *et al.*, 2013; Poier *et al.*, 2015). Some works, such as those done by Cai *et al.* (2019) used refinement models to adjust the poses with limits and rules. However, most of these works (Aristidou, 2018; Chen Chen *et al.*, 2013; Cobos *et al.*, 2008; Melax *et al.*, 2013; Ryf and Weymann, 1995; Sridhar *et al.*, 2013; Tompson *et al.*, 2014; Xu and Cheng, 2013; Li *et al.*, 2021) apply the rules and bounds after estimating the pose of the hand using post-processing methods such as inverse kinematics and bound penalization. Recent works used biomechanical constraints for hand pose estimation using 2D images in the neural network's cost function to penalize the joints. Malik *et al.* (2018*b*) incorporated structural properties of the hand such as the finger lengths and inter-finger joint distances to provide an accurate estimation of the hand pose. The drawback of this method is that the joints' angles are not considered for estimating the pose. Hence the resulting hand pose can still output a pose in which the joint angles can exceed the human joint bounds. Works such as those by Zhou *et al.* (2017); Sun *et al.* (2017) successfully implemented bone length-based constraints on human pose estimation but only on the whole body and not for the intricate parts of the hand such as finger length constraints. The model designed by Spurr *et al.* (2020) achieved better accuracy when tested on 2D datasets; however, the model was weakly supervised, and bound constraints were soft. Hence there are poses where the joint angles exceed the anatomical bounds. Li *et al.* (2021) used a model-based iterative approach by first applying the PoseNet (Choi *et al.*, 2020) and then computing the motion parameters. The drawback of this approach is that it depends on the PoseNet for recovering the primary joint positions and fails to operate if PoseNet fails to predict the pose. Moreover, the resulting search space of the earlier networks still includes implausible hand poses as these models only rely on the training dataset to learn the kinematic rules. We encoded the biomechanical rules as a closed-form expression that does not require any form of training. SSC-CNN's search space is hence much smaller than the aforementioned

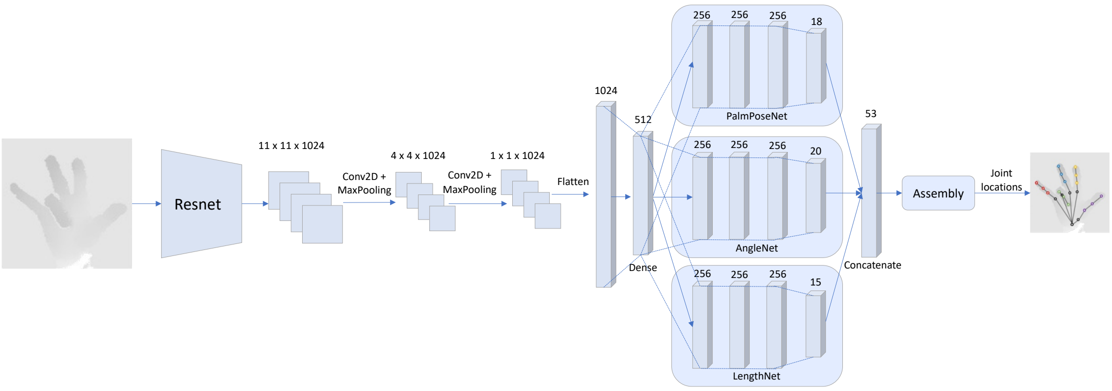


Fig. 3.2: Framework Architecture of the SSC-CNN. The proposed architecture uses part of the Resnet50 model for feature extraction. The features then pass through two sets of convolutional layers and max-pooling layers and then flatten to a common dense layer. This layer is then fed as input to three sub-nets: (1) The PalmPoseNet, which outputs an 18-dimensional vector corresponding to the 3D positions of the palm joints (root joint, MCPs and CMC), (2) the AngleNet, which outputs a 20-dimensional vector corresponding to the joint angles of the hand and (3) the LengthNet which outputs an 18-dimensional vector corresponding to the length of each finger segment of the hand. The features are then concatenated and sent as input to the assembly, which then provides the joint locations as output.

models. In our approach, the hand joint locations and their respective angles are predicted, and the bounds were implicitly applied to the model such that the joint angle always lies between them. Also, as pointed out in Section 3.5, many datasets themselves are not free from anatomical errors due to errors during annotation, and hence learning kinematic structures based on the dataset alone might lead to absorbing those errors into our model. To the best of our knowledge, our work is the first to propose incorporating anatomical constraints implicitly into the neural architecture.

3.3 Proposed Framework

In this chapter, we present a framework that provides hand poses that conform to the hand's biomechanical rules and bounds as explained in section 2.3.2. This goal is achieved by applying the rules implicitly to the forward pass of the neural network. The code for this model is publicly available[1] and the overall architecture is shown in Figure 3.2.

[1] https://github.com/RBC-DSAI-IITM/SSCCNN

3.3.1 SSC-CNN Architecture

The Resnet50 model (He *et al.*, 2016) is used as a backbone for the SSC-CNN (architecture shown in Figure 3.2). The layers up to "**conv4_block6_out**" are used (after 6 block computations of the Resnet50), and the weights were transferred from the model trained on the ImageNet dataset (Deng *et al.*, 2009). Using the transferred weights achieved better results as compared to random initialization of all the weights in the Resnet layers. An input image of size $176 \times 176 \times 3$ is provided to the pre-trained Resnet50 model. The output of this layer is then fed to a convolutional layer (1024 filters of size 3×3 with ReLU activation) and then a max-pooling layer (2×2). The size of the features at this time is $4 \times 4 \times 1024$, which is sent through another convolutional layer and max-pooling with the same configuration as before and then flattened to a 1024-dimensional vector. The compressed set of features is passed to a single dense layer of size 512 using ReLU activation, which is called the *common dense layer*. This is then sent to three individual networks for regressing the hand's various characteristics, which then predicts the pose of the hand using an assembler. The three individual networks are called: (1) PalmPoseNet, (2) AngleNet, and (3) LengthNet.

PalmPoseNet

The PalmPoseNet predicts the joint locations of the root joint, the CMC joint of the thumb, and 4 MCP joints (the thumb MCP is excluded as the root joint is the CMC joint). These joints do not have any strong biomechanical bounds and are dependent on the user's palm-size and structure. Hence to make the model robust, these points are directly regressed by the PalmPoseNet. The 512 features from the common dense layer are taken as input to three dense layers, which have 256 nodes, each using the sigmoid activation function. The features then pass to a final dense layer with 18 nodes which also uses a sigmoid activation function, and these 18 points correspond to the 3D location of the six joints.

AngleNet

The AngleNet provides the angle of each joint of the fingers. As there are five fingers, including the thumb, and each finger has four angles associated with it (as explained in Section 2.3.2), there are a total of 20 angles that are regressed by the AngleNet. The 512 features from the common dense layer are taken as input to three dense layers, which have 256 nodes, each using the sigmoid activation function. The features then pass to a final dense layer with 20 nodes that uses a sigmoid activation function. These 20 features are then used in the composition of the hand pose as described in Section 3.3.2.

LengthNet

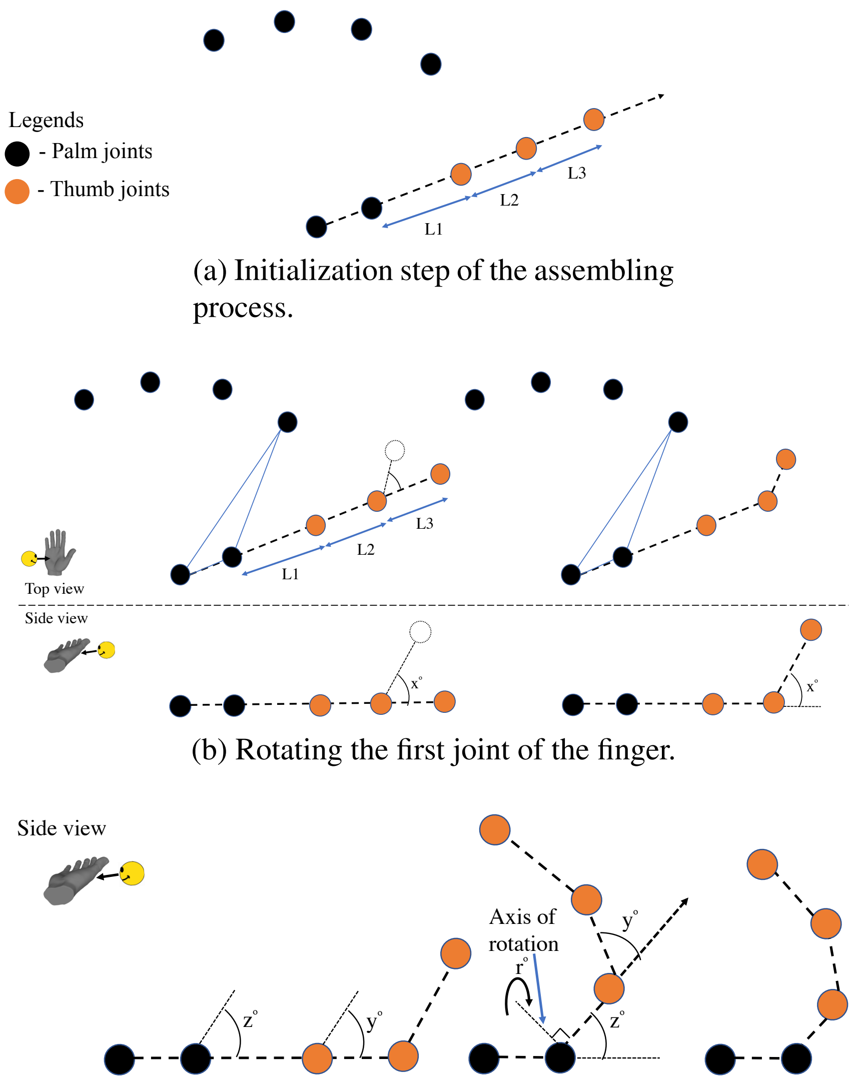
The LengthNet provides the length of the individual segments of the fingers of the hand, such as the length of the part between the thumb CMC to the thumb MCP and the thumb MCP to thumb IP. The 512 features from the common dense layer are taken as input to three dense layers, which have 256 nodes, each using the sigmoid activation function. The features then pass to a dense layer with 15 nodes that uses a sigmoid activation function. These 15 features are relative values to calculate the segments' lengths which are then used in the composition of the hand pose as described in Section 3.3.2.

3.3.2 Assembly of the Pose

The assembly is a non-trainable portion of the architecture responsible for constructing the resulting hand pose based on the values from the previous individual networks. A sample process flow of the assembly for one finger (the thumb) is shown in Figure 3.3, and this process repeats for each finger.

The first step is to take the 3D positions from the PalmPoseNet and use these points as the reference for the fingers. Taking the thumb as an example sequence, the next step is to use the root joint and the CMC joint as line points and extrapolate the line beyond the CMC joint for placing the thumb joints as shown in Figure 3.3a. The length of each segment between the joints is taken from the LengthNet.

The LengthNet output vector is from a sigmoid function and ranges from 0 to 1. These values are multiplied with a hyper-parameter (γ) which is the longest possible



(a) Initialization step of the assembling process.

(b) Rotating the first joint of the finger.

(c) Rotation of next two joint with final rotation using the axis of rotation as shown.

Fig. 3.3: Overall sample process of the assembly to create a thumb of the hand pose. The initialization step shown in Figure 3.3a uses the 3D positions of the joints directly regressed from the PalmPoseNet and then uses the root joint with the CMC joint to extrapolate a 3D line. The three joints are then placed on this line, and the length of each segment between the joints is taken from the LengthNet. The tip joint is first rotated using the axis, which passes the adjacent joint and is parallel to the plane created by the root joint, CMC joint, and the index MCP joint. The second and third joints rotate similar to the previous joint, as shown in Figure 3.3b using their axes of rotation. The last rotation uses an axis perpendicular to the line that passes the CMC and the current third joint position.

length of a finger segment. Works done by Sunil (2004) and chan Jee and Yun (2016) include studies where the individual parts of the hand are measured. Using this data, we set $\gamma = 80$ mm which is the maximum length of an average finger segment as per these studies. Using a sigmoid-based output for the individual lengths provides finer control for the model during training.

After extrapolating the thumb joints, the next step is to rotate the joints to their corresponding angles, as shown in Figure 3.3b. The values from AngleNet are used for setting the angles of rotation. These values also range from 0 to 1 as the sigmoid activation function is used. Each value is then multiplied according to the biomechanical range of the joint. This ensures that the range of the angle does not overshoot or undershoot the range of the joint and is given in equation 3.1.

$$\theta^i = (A^i * (\theta^i_{\text{Upper}} - \theta^i_{\text{Lower}})) + \theta^i_{\text{Lower}} \tag{3.1}$$

where θ^i is the i-th joint angle, A is a vector from the AngleNet, θ^i_{Lower} is the lower bound of θ^i and θ^i_{Upper} is the upper bound of θ^i. For example, the thumb IP ranges from $-30°$ (considering extension as negative) to $80°$ (flexion as positive) and if $A^i = 0.2$ then $\theta^i = (0.2 * (80 - (-30)) + (-30) = -8$. This value lies in the range $[-30, 80]$.

To rotate the joint by an angle, a reference plane is required. The root joint is always used as one point of the reference plane, while two adjacent MCP joints will be used as the other two points for each finger. The plane used for the thumb rotation is formed by the root joint, CMC joint, and the index MCP joint. Similarly, for the index finger, the index MCP and the middle MCP is used. For the last finger, i.e., the pinky, the pinky MCP and the ring MCP are used, and the rotation signs are inverted.

The finger's tip is the first joint to be rotated, as shown in Figure 3.3b. To rotate the joint, an axis of rotation must be calculated. This axis is created using a vector from the adjacent joint, which lies on the reference plane and is perpendicular to the line from the first joint to the adjacent joint. After the first joint rotation, the next joint in the chain is rotated using an axis vector constructed in a similar fashion to the first joint and originating from the next adjacent joint. The second joint rotation is also applied to the first joint using the same axis of rotation. The chain then continues for the third joint using the axis originating from the next adjacent joint in the line where the first and second joint also rotates. After the three rotations are performed, the last rotation takes place with an axis originating from the last joint (CMC in the case of the thumb and MCP for other fingers) and is projected perpendicular to the line from the current joint to the previous joint. The three joints are then rotated around this axis as shown in Figure 3.3c. This whole process is repeated for each finger, resulting in the overall pose of the hand.

3.3.3 Loss Function of SSC-CNN

As the assembly module is non-trainable, the loss function is calculated using the 53-dimensional vector after the concatenation phase. The assembly process is invertible, and hence the joint locations of the ground truth are converted to the target 53-dimensional vector, and the loss function is calculated and given in equation 3.2.

$$L = \frac{1}{6}\sum_{i=1}^{6}\left\|\hat{P}_i - P_{\mathrm{GT}_i}\right\|^2 + \frac{1}{15}\sum_{i=1}^{15}\left|\hat{L}_i - L_{\mathrm{GT}_i}\right| + \frac{1}{20}\sum_{i=1}^{20}\left|\hat{A}_i - A_{\mathrm{GT}_i}\right| \tag{3.2}$$

where $\hat{P}$ is the predicted vector of joint locations from the PalmPoseNet, $\hat{L}$ is the predicted vector of lengths derived from the LengthNet, and $\hat{A}$ is the predicted vector of angles derived from the AngleNet. $P_{\mathrm{GT}}, L_{\mathrm{GT}}, A_{\mathrm{GT}}$ are the ground truth vectors which are derived by using the reverse assembly process. The loss L consists of three parts, (1) the mean Euclidean distance between the ground truth and predicted PalmPoseNet joint locations, (2) the mean absolute difference between the ground truth and predicted lengths, and (3) the mean absolute difference between the ground truth and predicted angles. Hence the gradients are computed on the pre-final output that comes before the assembly phase and not on the assembled pose (joint locations) of the hand.

3.3.4 Dataset used

The proposed framework was tested on two popular datasets, namely the MSRA (Sun *et al.*, 2015) and HANDS2017 (Yuan *et al.*, 2017) datasets. These datasets were used as they use the true joint locations such as the MCP joint and CMC joint locations compared to the edge centers used by the NYU (Tompson *et al.*, 2014) dataset. The MSRA dataset comprises about 70000 images, and HANDS2017 has more than 900000 train images and 250000 test images. The SSC-CNN was trained on these dataset's training sets and tested their respective test sets using the same architecture. During training, no anatomical corrections were made on the datasets' ground truth to maintain consistency with other state-of-the-art models during comparison.

3.4 Experiments Performed

As this framework focuses primarily on the anatomical correctness of the pose instead of the supposed accuracy as reported by other papers, we performed comparative tests of anatomical correctness on other state-of-the-art models and datasets along with the accuracy metrics.

3.4.1 Error of model after external correction

To study the change in the accuracy of the model when correcting the anatomical error of the model, a corrector module was designed based on our earlier work (Isaac *et al.*, 2021) so that it can take the hand poses of the current state-of-the-art models as input and correct the anatomical errors of the model. This module is plugged into each test model and used to correct the anatomical error, and the construction along with the details are explained in Section 2.3.1.

3.4.2 Ground truth validation

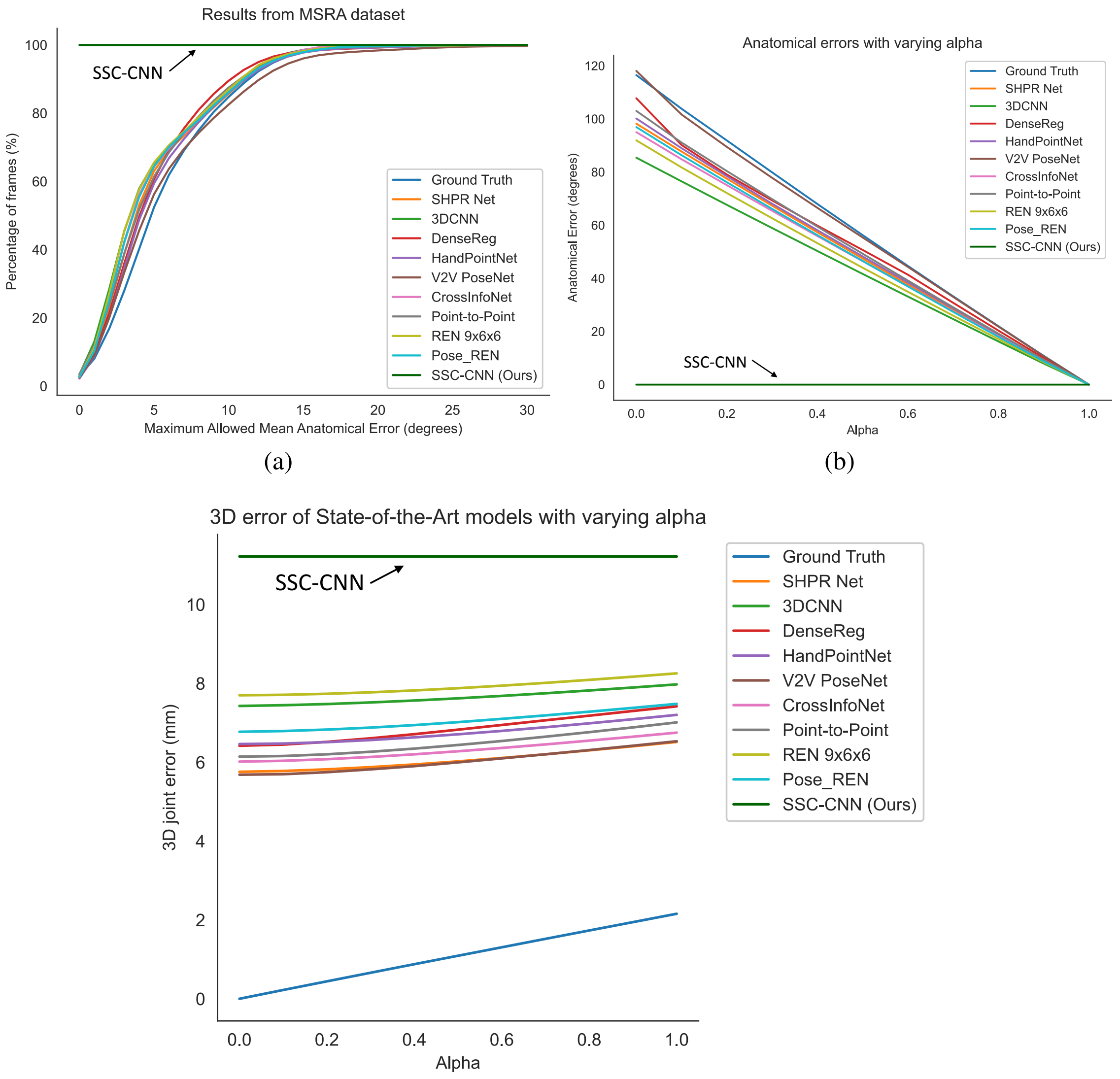
To validate the anatomical correctness of the ground truth, the anatomical error of the ground truth labels is calculated, and the ground truth is also compared with itself after external correction using the corrector module described in Section 2.3.1.

Table 3.1: 3D Joint Errors (3DJE) and Anatomical Errors (AE) derived from 40000 images from the HANDS2017 dataset with all the models. The first column contains the name of the model, the second column has the errors of the model with no external correction, the third column has the Anatomical Error (AE) of the model without correction and the fourth column has the errors of the model after correction.

Model	3DJE with no correction (mm)	AE with no correction (°)	3DJE with correction (mm)
SSC-CNN (Ours)	9.48	**0**	**9.48**
A2J Xiong *et al.* (2019)	**8.65**	125	9.73
V2V Posenet Moon *et al.* (2018)	10.42	135	11.27
Ground truth	0	131	2.38

Table 3.2: 3D Joint Errors (3DJE) and Anatomical Errors (AE) derived from 20000 images from the MSRA dataset with all the models. The first column contains the name of the model, the second column has the errors of the model with no external correction, the third column has the Anatomical Error (AE) of the model without correction, the fourth column has the errors of the model after correction and the last column has the errors of the model when compared to a correction version of the ground truth.

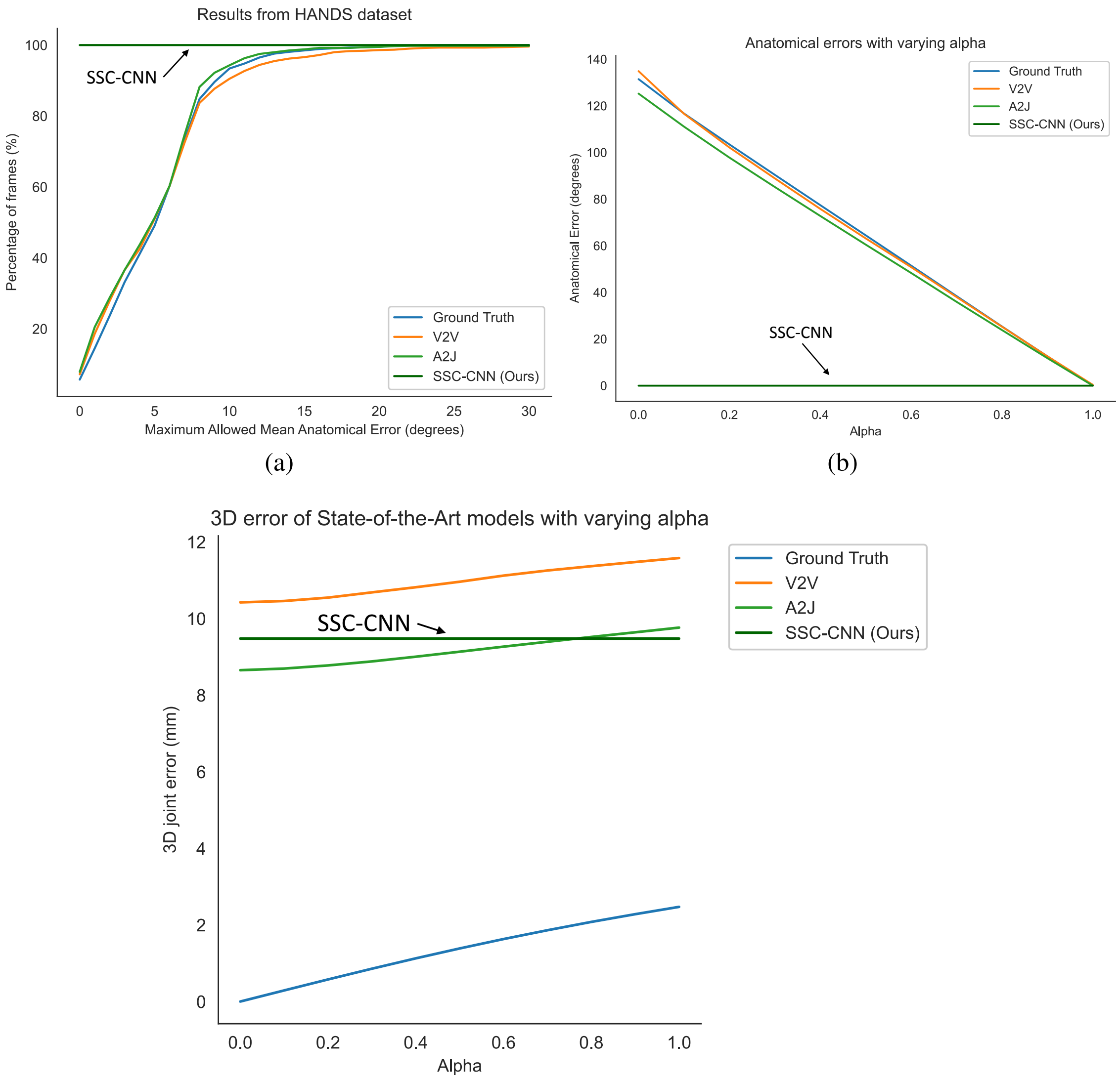
Model	3DJE with no correction (mm)	AE with no correction (°)	3DJE with correction (mm)	3DJE with correction compared to corrected ground truth (mm)
SSC-CNN (Ours)	11.42	**0**	11.42	11.32
SHPR Net Chen *et al.* (2018)	7.86	98	8.56	7.92
3DCNN Ge *et al.* (2017)	9.48	85	10.05	9.55
DenseReg Wan *et al.* (2018)	7.73	107	**8.37**	7.80
HandPointNet Ge *et al.* (2018*a*)	8.31	100	9.14	8.55
V2V Posenet Moon *et al.* (2018)	**7.59**	118	15.37	14.84
CrossInfoNet Du *et al.* (2019)	7.96	103	8.41	**7.75**
Point-to-Point Ge *et al.* (2018*b*)	7.71	95	8.51	7.91
REN 9x6x6 Wang *et al.* (2018)	9.79	91	10.20	9.66
Pose REN Chen *et al.* (2020)	8.65	96	9.19	8.59
Ground truth	0	116	1.89	0



(a) (b)

(c)

Fig. 3.4: Comparison of the anatomical errors and the 3D joint errors of various state-of-the-art models along with our proposed model using the MSRA hand dataset. The ground truth is also shown for comparison as it has high anatomical errors. The error can be due to the noises during the recording of the labels.



(a)

(b)

(c)

Fig. 3.5: Comparison of the anatomical errors and the 3D joint errors of various state-of-the-art models along with our proposed model using the HANDS hand dataset. The ground truth is also shown for comparison as it has high anatomical errors. This can be due to the noises during the recording of the labels.

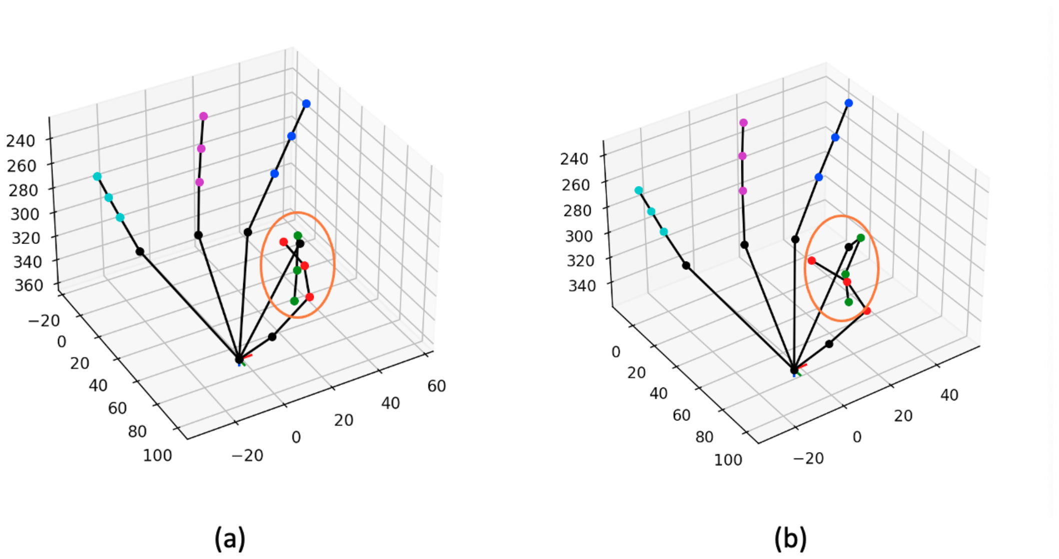


Fig. 3.6: Qualitative comparison between the pose from the MSRA dataset using (a) SSC-CNN and (b) the same pose from Wang *et al.* (2018). The circle shows the part which is anatomically wrong in (b) while its correctly shown in (a).

3.5 Experiment Results and Discussion

Figure 3.4 and Figure 3.5 shows the comparison of the models using the MSRA hand gesture (Sun *et al.*, 2015) and HANDS2017 (Yuan *et al.*, 2017) datasets respectively. A qualitative comparison is shown in Figure 3.6 between a pose from SSC-CNN and another state-of-the-art model. For the first graph of the two sets (Figure 3.4a and Figure 3.5a), the x-axis shows the maximum allowed mean anatomical error (calculated as mean per joint per hand), and the y-axis denotes the percentage of frames of the dataset, which is up to the specified mean anatomical error. For context, the steeper the curve is in the graph, the better the model in terms of anatomical correctness. Our model has no anatomical errors and hence the steepest line in both datasets. The second part of the set (Figure 3.4b and Figure 3.5b) shows the total anatomical error (calculated as mean per hand and not per joint to show the difference) of the model per hand frame using the correction module set at each value of α at steps of 0.1. The third graph (Figure 3.4c and Figure 3.5c) represents the 3D joint error which is the mean Euclidean distance from the predicted joint to the ground truth joint. The ground truth used in the test is not anatomically corrected and is the original ground truth. The lowermost line seen in both graphs is the dataset's ground truth compared with itself after anatomical

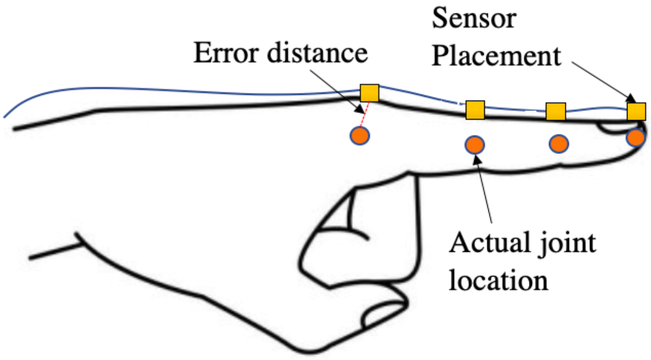


Fig. 3.7: Illustration to show the error in the true position of the hand when measuring using a sensor placed on top of the finger. There is a small gap since the sensor placement is superficial, and the true position of joints that lie inside the hand will have large errors.

correction. As seen in the graphs, the ground truth itself has high anatomical errors, and a likely cause of this anatomical discrepancy is the method used in creating the datasets.

As shown in the works of Oberweger *et al.* (2016), the ground-truth curated in the HANDS and MSRA datasets are not exact representations of the real hand poses. The MSRA dataset uses a combination of the author's hand pose estimator as a reference with manual editing, which is tedious and prone to human errors, as seen in Table 3.3. The HANDS2017 dataset was recorded using the Ascension Trakstar™ [2] which is reported to have an accuracy of ±1.4 mm and is attached on top of the finger during recording. As shown in Figure 3.7, if the sensor is placed on top of the finger during the recording of poses, the joint's actual position will be at an offset from the recorded position of the hand. Hence the ground truth may not always be the actual position of the hand for many frames. With anatomically incorrect models, the error to the ground truth (non-corrected) can tend to 0. However, our algorithm emphasizes anatomic correctness over the closeness to the ground truth. Hence this resulted in a relatively higher 3D joint error of 9.48 mm using the HANDS2017 dataset and 11.42 mm using the MSRA dataset as compared to the state-of-the-art models. However, our model shows comparable results when using the correction module as these models have very high anatomical errors, and correcting these errors increases the 3D joint location error. To help the community for future hand tracking related works, we also provide our cor-

[2] https://tracklab.com.au/products/brands/ndi/ascension-trakstar/, Accessed January 2022

Table 3.3: Anatomical Errors (AE) of the HANDS2017 dataset and MSRA dataset ground truth and the 3DJE of the corrected ground truth to the non-corrected version.

Dataset	AE with no correction (°)	AE with correction (°)	3DJE after correction (mm)
HANDS2017	131	0	2.38
MSRA	116	0	1.89

rector module publicly available to correct the ground truth of the HANDS2017 and MSRA datasets to create an Anatomical Error-Free (AEF) version of those datasets.

To study the effect of the sub-networks, ablation studies were performed by removing the subnetworks and regressing all the joints of the hand directly. The resultant hand poses did not conform to the biomechanical rules and had joints rotated by abnormal angles as well as abnormally long finger segments at times. This behavior shows that the subnetworks ensure that the hand pose conforms to the angle bounds and proper finger lengths.

Table 3.1 and Table 3.2 contain the 3D error of the hand pose estimators before and after the application of the corrector module. The anatomical error before using the module is also shown in the tables. As seen in the table, our model predicts poses with no anatomical errors and has the same 3DJE as these bounds are implicitly coded in the model's architecture, and the resulting hand poses always conform to these bounds. In contrast, other models have large anatomical errors and deviate after correction.

3.6 Conclusion, Limitations and Future Works

We proposed a novel framework called the SSC-CNN for 3D hand pose estimation with biomechanical constraints. The network has biomechanical rules and bounds encoded in the architecture level such that the resulting hand poses always lie inside the biomechanical bounds and rules of the human hand, and no post-processing is required to correct the poses. Our framework was compared to several state-of-the-art models with two datasets. Experiments have shown that the SSC-CNN has comparable results but with no anatomical errors, whereas the state-of-the-art models have very high anatomical errors. The ground truth of the datasets also has anatomical errors, and anatomically error-free versions were created.

Our framework has a limitation in which the training phase requires data pre-processing to derive the joint angles as these angles were not available in the datasets used. Another limitation is that our hand pose estimator does not consider the velocity of the joint movements when correcting them. The angular velocity of the joints also has biomechanical constraints, and these will be incorporated in future works for the model. Although the model is highly robust for varying palm sizes, extreme cases like estimating the hand poses of children may result in inaccurate poses as the dataset used for training does not cover young children's hands and can be investigated in a future work.

Future works also include using synthetic datasets such as the MANO hands (Romero *et al.*, 2017) so that the ground truth will be assured of the hands' true location along with children's hand poses. Using these synthetic datasets, we can also compare the spectrum of poses covered by the currently available datasets and hence cover a broader spectrum of poses for training. Analyzing the history of the hands' motion using methods such as recurrent neural networks (Yoo *et al.*, 2020) instead of processing only one instance of the hand can avoid erratic motions during self-occlusions and will be investigated in another study for adding the feature to the SSC-CNN. The history can include the velocity and acceleration of the joint motions, which also have biomechanical bounds and further enhance the pose realism during hand motion tracking.

CHAPTER 4

QUANTIFICATION OF 2D HCI

4.1 Introduction

Although the human hand is a complex system that can perform multiple actions, when the kinaesthetic actions are scaled in a system, the applications are limitless. This concept of scaling the user's actions can be explained by the Control-Display (CD) ratio in a visual-haptic interaction. It is the ratio of the kinaesthetic movement made using the input device in the real world to the visual movement made by the object or cursor that the device controls. When the display size is fixed, this ratio will then become the control movement scale. In this Chapter, the term haptics refers to kinaesthetic sensing alone, and therefore the utilized haptic devices are used as kinaesthetic input devices, and the output feedback is using visual display alone. Based on this assumption, kinaesthetic input devices can be classified broadly into two categories: scaled and non-scaled input devices.

Scaled input devices translate the user's kinaesthetic input into a larger visual output in the system. The CD ratio in such input devices is generally other than 1:1 (for example, a computer mouse (Smith *et al.*, 1999)). When the user moves the mouse (assuming small bursts of constant acceleration), the cursor moves a larger distance on the screen. Although this is a scaled movement, users find it adjustable, even sometimes more comfortable than a regular action. This example shows that scaled movement already exists in current Human-Computer Interaction (HCI).

Non-scaled input devices translate the user's kinaesthetic input to a similar output in the system. The CD ratio of this type of input device is 1:1. Examples of this kind of movement include single-finger gestures performed on touch screens, and smartphones (Sears and Jacko, 2007). The kinaesthetic movement is not scaled since the distance moved with the finger is the same as that on the visual screen.

4.1.1 Motivation

From our observations, kinaesthetic movement with a larger control movement scale for some users is easier and more efficient for them than regular one-to-one visual feedback. An example is an experienced user of a computer mouse performing the operations with relative ease. Sometimes they seem to be more accurate in moving to the target icon. Our motivation is to leverage this fact and verify if higher control movement scale, in general, can be better than natural kinaesthetic movements (where control movement scale = 1:1) in tasks that require extended accuracy (such as a telesurgery where the doctor's movement of the surgical tools can be scaled and mapped to the movement of the robot in the patient's body). Such scaled movements can also help to increase the effective workspace of haptic devices, such as the OmniPhantom® where the workspace is restricted to a small region. The well-known Fitts' Law (Fitts, 1954) is used to verify this hypothesis.

4.2 Related Work And Literature

There are two related concepts in the literature about the scale effects: control gain and CD ratio. According to Accot and Zhai (2001), if the display size is kept constant, then both the concepts will point towards the control movement scale.

Hess (1973) performed one of the earliest experiments related to human performance in scaled conditions. His study showed that human performance is an inverted U-shaped function of control gain or CD ratio, where an optimum value will exist in the medium range of the function and decreases when away from this range. This was also confirmed by works of Zhai *et al.* (1996), Boff and Lincoln (1988), and Boff *et al.* (1986). Works of Langolf *et al.* (1976) also suggested that the human performance is non-linear when the experiments are performed with different task conditions.

Arnaut and Greenstein (1990) conducted two experiments on a touch tablet and a trackball in which the display output magnitude, control input magnitude (movement scale), display target width, Fitts' Index of difficult and control target width were varied. They showed that increasing the movement scale increased the gross movement time but decreased the fine adjustment time. For a touch tablet, the total completion time was a U-shaped function. However, when using a trackball, they showed that the greater movement scale increased the total completion time monotonically. Overall they concluded that gain and Fitts' Index of difficulty must be combined to be a more useful predictor.

Jellinek and Card (1990) performed experiments with the computer mouse and analyzed the user performance against the control gain. Their result was also an inverted U-shaped performance-gain function. However, they argued against it, saying that the performance loss was due to the loss of relative measurement resolution at high control gain. This is also called a quantization effect.

Casiez *et al.* (2008) corrected the quantization effect by using high-resolution displays to show the output of the system. They performed extensive studies based on the CD ratio and its effect on user performance. According to them, common operating systems (OS) use the Pointer acceleration (PA) as the default behavior for the computer mouse. It dynamically manipulates CD gain between the visual display and the kinaesthetic input device as a function of the velocity of the movement. This means that the CD gain is high when the velocity of the kinaesthetic device is high and vice versa. They assumed the fact that when the target (such as an icon) is far away from the current position, then the cursor must move a large distance in a short time. Conversely, when the target is nearby, the movement must slow down to make the finer adjustments. Constant gain (CG) is the simpler method for manipulating CD gain via a constant multiplier regardless of device movement characteristics. They encourage to use of pointer acceleration rather than constant gain as a base technique for comparing new pointing technique performances. This, however, was contradicted by Arnaut and Greenstein (1986) where he showed that the control gain in graphics tablets reduces the performance, and having a gain of 1.0 (no gain) is best for the absolute mode of cursor control.

In light of the above points, the best parameter is the control movement scale since the results derived from using the control gain and CD ratio contradict each other in the literature. In this Chapter, we keep the display size constant so that the parameter in effect is the control movement scale.

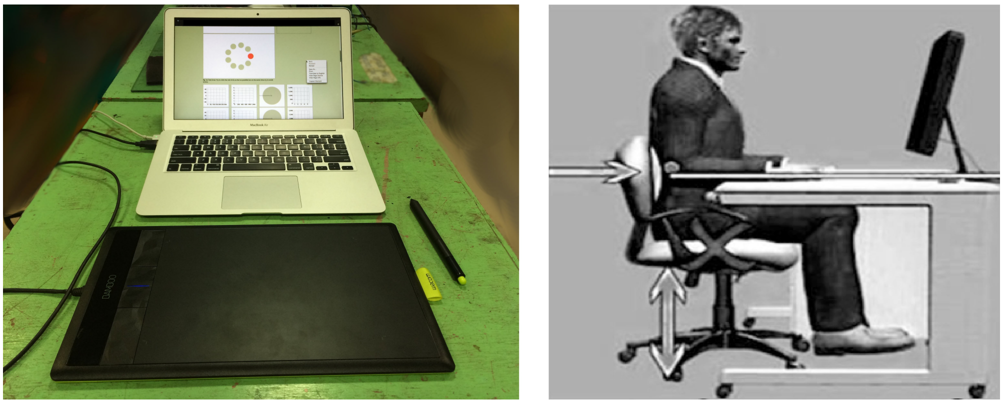


Fig. 4.1: Experimental setup for the 2D Fitts's Law based experiment. The left side figure is the apparatus used for conducting the experiment. It consists of a Wacom™ Graphics tablet which is connected to a system. The right side figure is from Raghu Prasad *et al.* (2013) and is the standard sitting posture that is maintained for the experiment. It ensures proper comfort during the experiment.

4.3 Experimental Apparatus And Procedure

4.3.1 Apparatus Set-up

The hypothesis is that kinaesthetic movement with a larger control movement scale for users is easier and more efficient for them than regular one-to-one visual feedback. This experiment is to test that hypothesis and verify if higher control movement scale, in general, can be better than natural kinaesthetic movements with the help of Fitts' Law. It involves a standard multi-directional tapping task. The input device is a Wacom™ Graphics tablet which converts the user's kinaesthetic movement into a cursor movement on the screen. (Figure 4.1). The Wacom tablet was selected because of the availability of settings for scaling the kinaesthetic input movement of the user. The sitting posture of the participant is the same as the standards followed by Raghu Prasad *et al.* (2013). The participant is positioned such that their arms gently fall on the table without much pressure or stress on their wrists, elbows, or shoulders. The screen is positioned at a constant distance from the eyes of the participants. The feet are on the ground with a comfortable bend in the knee at around 90 degrees approximately. These ensure that the participant is in a comfortable position so that no physical attributes disturb the performance rates derived from the experiment.

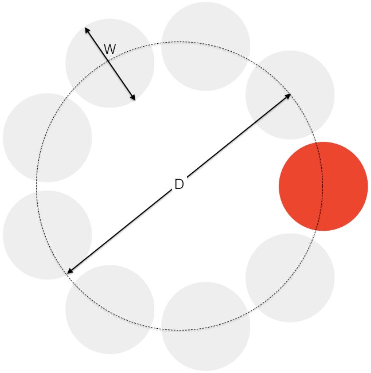


Fig. 4.2: Multi-directional tapping experiment. W is the diameter of the small circles while D is the diameter of the larger circle made by the small circles.

4.3.2 Participants' Specifics

Sixteen healthy participants performed this experiment aged between 20 to 48 years, consisting of eleven male participants and five female participants. All of them were screened for any difficulty in performing kinaesthetic actions and sensing relative positions of the body and limbs (proprioception). No participant was known to have any such difficulty. All of the participants were right-handed. They followed the procedure as per specifications, and their sitting posture was as shown in Figure 4.1.

4.3.3 Task to Perform

The task described in this Chapter is a modified version of the multi-directional tapping experiment, which was outlined in the ISO 9241-9 standard (2002). The experiment composes of a ring of 9 circles, as shown in Figure 4.2. W is the diameter of the small circles, while D is the diameter of the larger circle made by the small circles. The goal is to click on the highlighted circle (which will be in red) in the shortest time possible. After each click, another circle will be highlighted, and this process will continue. The pattern of clicking will follow a star-shaped manner similar to the original tapping experiment. Hence the user must traverse the whole range of the outer circle every time, ensuring the constant D with W. After 9 clicks (completing a full star), the circles will be arranged with a different D and W. The user will perform this task for 5 different settings of D and W. This concludes a test for a particular scale setting in the tablet. This entire test is performed for four different scale settings (1:2, 1:2.4, 1:3.3, and 1:4.9) of the tablet. This constitutes a full experiment on one participant.

4.3.4 Variables Measured

Following the conventions of the original Fitts' Law (Fitts, 1954), the index of difficulty (I_d) is defined as

$$I_d = -\, log_2 \frac{W_s}{2A} \text{ bits/response} \tag{4.1}$$

where A is the average amplitude of the particular class of movements and W_s is the tolerance range (or target width) in inches. The use of $2A$ rather than A ensures that the Index will be greater than zero for all practical situations.

Another parameter is the binary Index of performance (I_p), which expresses the results as a performance rate. In the original version, Fitts declared that for a task in which A and W_s are fixed for a series of movements, $I_{p(original)}$ is defined as

$$I_{p(original)} = -\frac{1}{t}\, log_2 \frac{W_s}{2A} \text{ bits/s} \tag{4.2}$$

where t is the average time in seconds per movement. Many modifications to the original Fitts' law are found in the literature to obtain better regression fits for the model. One of the well-known amendments is by MacKenzie (1992) in which he modified the I_d as explained in equation 4.3. The modified Index is denoted as I'_d for clarity in this Chapter which is always positive and provides a better fit with observations. This new Index is in the model used in equation 4.4, which is developed in analogy with Shannon's information theory (Shannon, 1948).

$$I'_d = log_2\, (\frac{2A}{W_s} + 1) \tag{4.3}$$

$$t = a + bI'_d \tag{4.4}$$

$$I_p = 1/b \tag{4.5}$$

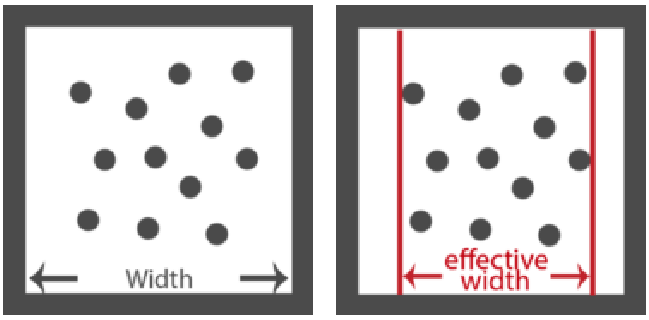


Fig. 4.3: Illustration for effective width. In this image, the two boxes are targets for testing purposes. The dots that are inside the boxes represent the end-points made by the user during the experiment. The left side image shows the actual width, while the right side image shows the effective width W_e of the test. This can only be measured after the full test is performed and cannot be measured during the test. Image from Jude *et al.* (2016)

Where t the is movement time (MT) in seconds, and I'_d is in bits. The reciprocal of b is the Index of performance (I_p), and its unit is bits per second. The intercept a is not of importance in this context. I_p is found to carry the informational aspect of system performance and is also termed as throughput (Zhai, 2004). It is a measure of human performance, and this modified version is used as an indicator of hand-eye coordination in this study.

From the experiment performed as per Section 4.3.3, the following values are calculated :

Movement Time (MT):

When a circle gets highlighted, the time it takes for the participant to look at the target and then click on it gets recorded. The average duration for a participant for each effective Index of difficulty is noted. This duration is the MT and is noted in seconds.

Effective Index of Difficulty Per Test (ID_e):

Jude *et al.* (2016) compared different methods of reporting and visualizing Fitts' regressions. They have shown that the visualization is clearer when the regression is based on means per user and ID_e which is the Effective Index of difficulty. It is a modification of the regular Index, which uses the effective width W_e (Figure 4.3) rather than the width defined in the conventional Fitts' Law. This experiment was successful as the fit was almost perfect with the Pearson's r coefficient close to 1. In this Chapter, equation 4.3 is modified and used to measure the Effective Index of difficulty of the test. Instead of using the amplitude $2A$, we use the effective diameter of the larger circle (D_e). The target width W is now the effective width W_e. The definitions of D_e and W_e are the same as per Jude *et al.* (2016). The unit of I_d is bits. Even though the same D and W are provided for all participants, their effective Index of difficulty will change because D_e and W_e will differ.

$$ID_e = log_2 \left(\frac{D_e}{W_e} + 1 \right) \tag{4.6}$$

Index of Performance (I_p)

There were tests performed on different settings of D and W for each participant. The values of ID_e from these calculations will be rounded to the first decimal and then noted. This will result in 9 MT values for each ID_e. The mean MT is then calculated and noted. After plotting these points, a line is fitted, and the inverse of the slope is noted as the I_p for that particular scale the tablet was set on. The unit is bits per sec.

4.3.5 Analysis and Visualization of Data

Our experiment uses the visualization methods of Jude *et al.* (2016) for showing the various results and observations. This enables us to provide strong deductions since the variance of the raw data was unequal and can be seen in the box plot (see Figure 4.4). From the figure, it is evident that the variances are not equal when the experiments are conducted on higher scales. Hence this form of visualization cannot be used to deduce any conclusion. The data requires further processing to arrive at a solid conclusion, for which we adopt a different form of visualization from Jude *et al.* (2016). This results in multiple MTs for each ID_e. From this, the mean MT is calculated for each ID_e. When the I_p is calculated for this set of data, it results in less variance as compared to individual results that can be seen in Figure 4.6.

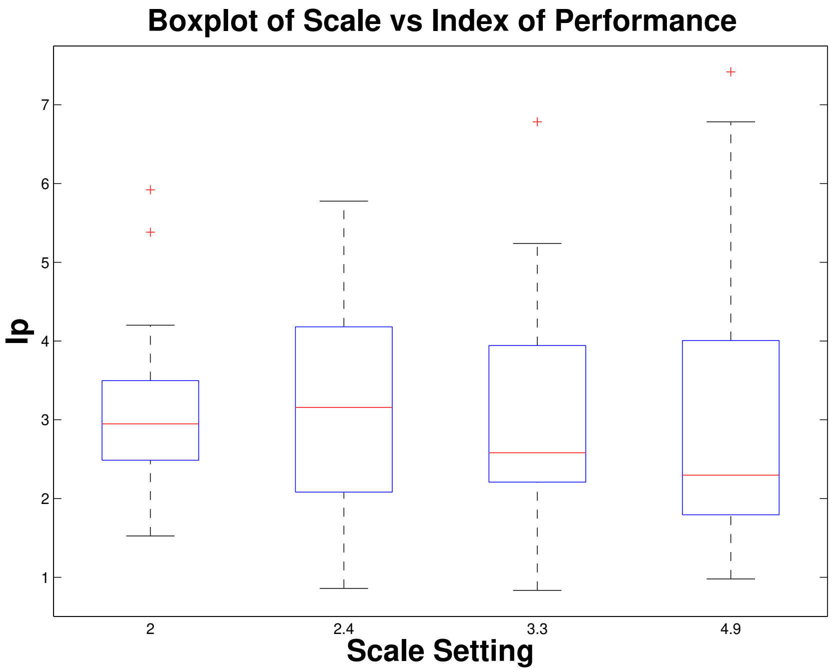


Fig. 4.4: The box-plot of the raw I_p values taken from all subjects individually for each scale setting. Observations cannot be made from this data as it has very high variance at the higher scales.

4.4 Results And Discussion

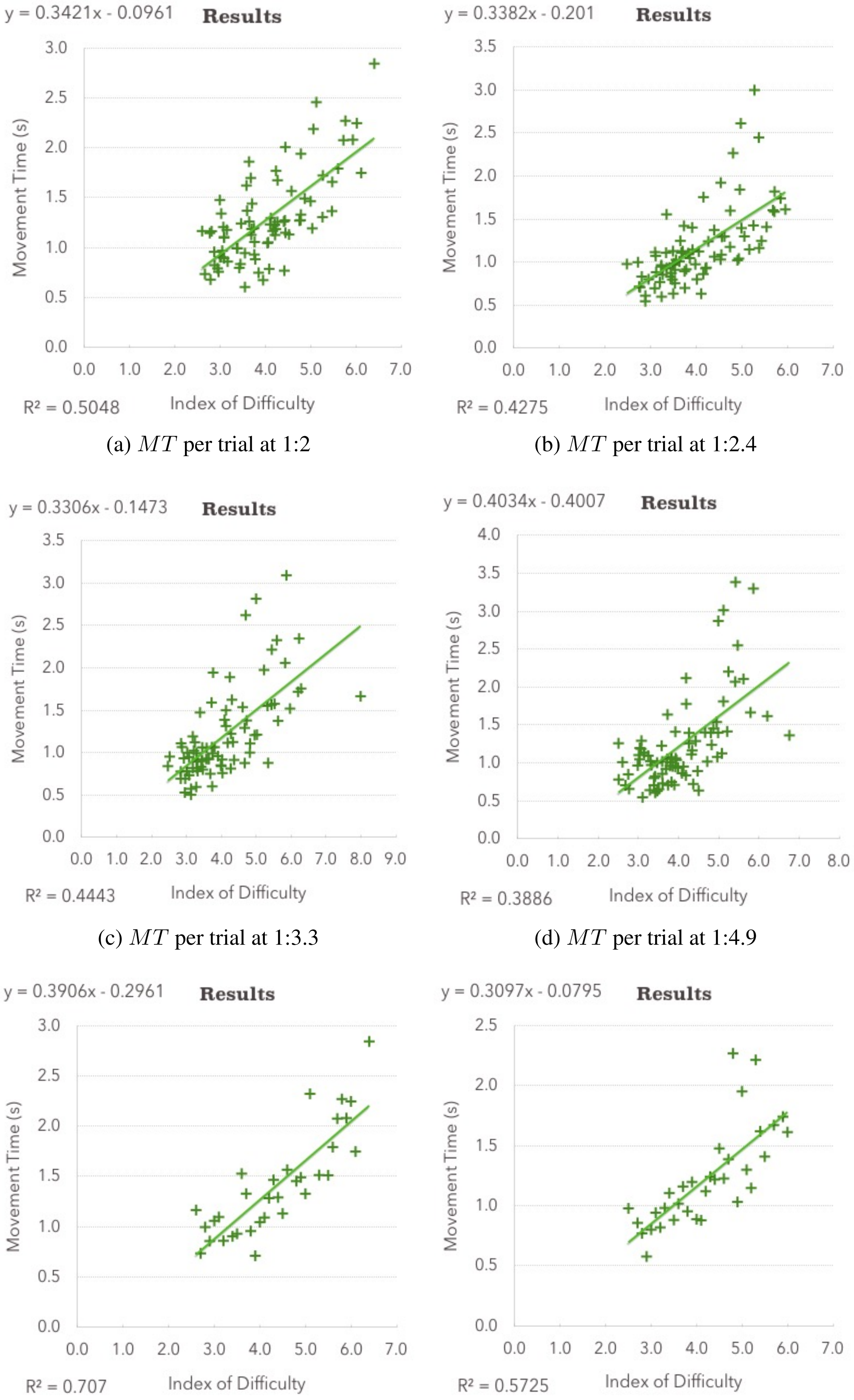
The MT and ID_e for different scales of all participants are consolidated and shown in Figure 4.6. From this data, the following observations can be noted:

4.4.1 Observations

(a) The black line in Figure 4.6c shows that I_p at scale 1:2.4 increased by 1.15% from the I_p at 1:2. Similarly the I_p at scale 1:3.3 increased by 3.48% and the I_p at scale 1:4.9 decreased by 15.2% (refer Table 4.1). Hence this figure shows that more participants performed better along the range 1:2 to 1:3.3.

(b) The red line in Figure 4.6d shows that I_p at scale 1:2.4 increased by 26.12% from the I_p at 1:2. Similarly the I_p at scale 1:3.3 increased by 29.25% and the I_p at scale 1:4.9 increased by 1.69% (refer Table 4.2). Hence these figures show that more participants performed better along the range 1:2.4 to 1:3.3.

(c) From Table 4.1 and Table 4.2 it is clear that the I_p from means per ID_e is statistically better since the fit of the line is more accurate (Pearson correlation coefficient). Although the standard error is relatively lower in the case of I_p calculated from MT per trial, the difference is significantly more for the Pearson's coefficient. Hence Figure 4.6d has more significance than Figure 4.6c.



(a) MT per trial at 1:2

(b) MT per trial at 1:2.4

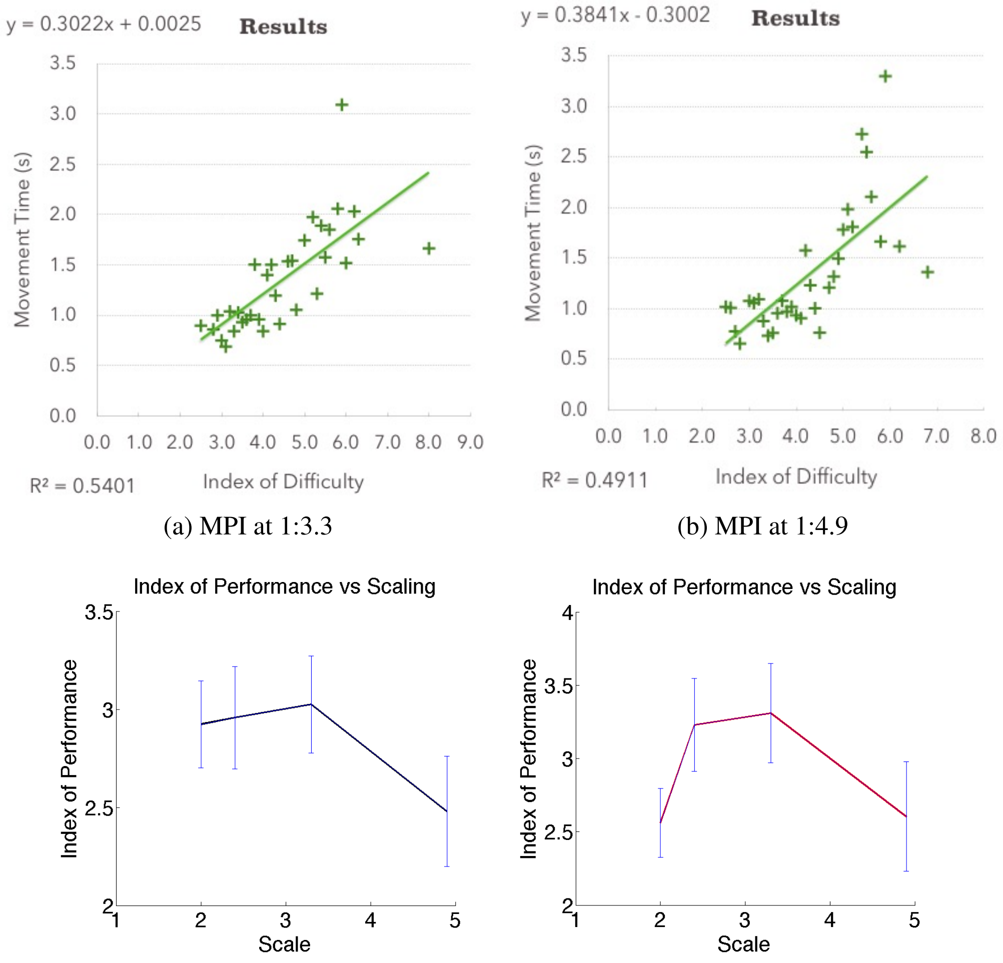
(c) MT per trial at 1:3.3

(d) MT per trial at 1:4.9

(e) MPI at 1:2

(f) MPI at 1:2.4

Fig. 4.5: Resulting graphs and different visualizations derived from the experiment conducted on 16 participants.



(a) MPI at 1:3.3

(b) MPI at 1:4.9

(c) I_p from MT per trial

(d) I_p from Means per ID_e

Fig. 4.6: Resulting graphs and different visualizations derived from the experiment conducted on 16 participants. Figure 4.6c is adopted from Jude *et al.* (2016) and shows the I_p values calculated from the slopes of the top row along with the error bars. The X-axis here is the scale of the task with respect to 1 unit (as in 1:x). Similarly, Figure 4.6d is also adopted from Jude *et al.* (2016) shows the I_p values calculated from the middle row along with the error bars, and its axes are the same as Figure 4.6d.

(d) Figures 4.6c and 4.6d show that I_p increases at first then takes a steep drop at higher scales.

(e) The Z-scores were computed for the slopes of the lines in Figure 4.5 and 4.6 (shown in Table 4.3 and Table 4.4 respectively). The blue cells indicate the combination for which the null hypothesis ($\mu_r = \mu_c$) is rejected for the alternate hypothesis being $\mu_r < \mu_c$ where μ_r is the slope at the corresponding scale setting of the row, and μ_c is similar for the column.

(f) When the slope for one scale setting is significantly smaller than the other slope, it implies that the I_p for that setting is significantly larger (since it's the inverse).

(g) Table 4.3 shows that the Z-score for the pairwise test between scales 1:3.3 and 1:2 is -1.83. Hence the null hypothesis cannot be rejected in favor of the alternate hypothesis at $\alpha = 0.05$ ($z_{0.05} = 1.96$) but can be rejected at $\alpha = 0.07$ ($z_{0.07} = 1.81$) which is very close. This difference becomes more prominent in Table 4.4.

(h) Table 4.4 shows that the means for the I_p at scales 1:2.4 and 1:3.3 is significantly larger than the other two scales with the level of significance at 0.05 (refer Table 4.4)

4.4.2 Discussion

Each individual participant is unique in their performance range. The Index of Performance scores significantly differed across scale levels with a significance of p<.05 [F(15, 48) = 5.49, p = $2.91 * e^{-}6$]. Some participants use input devices frequently, while others use input devices on a need-to-use basis. In our work, the participants who use input devices frequently include those who do activities such as gaming or large applications such as video editing and graphics design. In order to derive more results, further processing was performed as detailed in Section 4.3.5.

Looking at Figure 4.6d, a larger number of participants perform better along the range 1:2.4 to 1:3.3, which is also supported by the pairwise Z-scores. This setting corresponds to the larger I_p at the scale of 1:3.3. This fact was also proved by Accot and Zhai (2001) in their experiments with varying scales . According to them, the best scale for a graphics tablet is 2-3, which corresponds to the optimum ranges in this experiment.

As shown in Figures 4.6c and 4.6d, I_p increases significantly when the scale increases and lowers at the larger scale. This means that at the higher levels of scales, human performance decreases. The current results show that there is an optimum scale

Table 4.1: I_p from MT per trial

Scale	**Index of Performance I_p**	**% Change from the I_p at scale 1:2**	**Standard Error of the slope** $\sigma(1/I_p)$	**Pearson's** r^2
1:2	2.9231	+ 0%	0.03811	0.5048
1:2.4	2.9568	+ 1.15%	0.04431	0.4275
1:3.3	3.0248	+ 3.48%	0.04109	0.4443
1:4.9	2.4789	- 15.2%	0.05693	0.3886

Table 4.2: I_p from Means per ID_e

Scale	**Index of Performance I_p**	**% Change from the I_p at scale 1:2**	**Standard Error of the slope** $\sigma(1/I_p)$	**Pearson's** r^2
1:2	2.5602	+ 0%	0.04516	0.707
1:2.4	3.2289	+ 26.12%	0.04807	0.5725
1:3.3	3.3091	+ 29.25%	0.05008	0.5401
1:4.9	2.6035	+ 1.69%	0.07023	0.4911

Table 4.3: Z-scores calculated from the Means per trial (α = 0.05)

Scale	**1:2**	**1:2.4**	**1:3.3**	**1:4.9**
1:2		0.62	1.837	***-8.002***
1:2.4	-0.62		1.12	***-8.089***
1:3.3	-1.837	-1.12		***-9.27***
1:4.9	8.002	8.089	9.27	

Table 4.4: Z-scores calculated from the Means per ID_e (α = 0.05)

Scale	**1:2**	**1:2.4**	**1:3.3**	**1:4.9**
1:2		6.93	7.416	0.439
1:2.4	***-6.93***		-0.611	***-4.945***
1:3.3	***-7.416***	0.611		***-5.374***
1:4.9	0.439	4.945	5.374	

between the scale 1:2.4 and 1:3.3 that has a better I_p than that of the rest. The visualizations of Jude *et al.* (2016) makes the result clear by enhancing the plots. This enhancement increases the Pearson's *r* coefficient to higher values and hence provides a clearer result even though the variation between participants is high. Since the number of points is high and Fitts' Law is valid for the multi-tapping task, averaging the points between the I_d for a better fit is possible and hence done to increase the fit of the lines.

As shown in Figure 4.4, the values of the indices vary a lot at higher scales. This means that the values that are derived from the higher scales are slightly unstable. This was rectified when the different visualizations were used.

When the user moves the graphics tablet pen and touches the various points on the board, they experience a force, albeit constant, on contact. This can be an exceptional case of haptic feedback, considering that the user is feeling a force when hitting the pen on the board.

4.5 Conclusion And Future Work

In this chapter, we have examined the effect of the control movement scale on the user's kinaesthetic actions. We have used the Fitts' Law for quantifying the user's performance on different scales and have shown that at an optimum control movement scale, users perform better than natural movements. The Fitts' Law also proved to show its validity in scaled conditions with adequate r^2 values in all the scales. The Fitts' regressions were visualized, and it was found that the performance of the participants increases significantly when the scale increases and has an optimum range as well. Earlier studies involved using Fitts' Law with other older devices contradicted each other based on either control gain and control display ratio. We have kept control gain constant and varied only control display ratio for the experiment and the results echoed with studies involving control display ratio (Zhai *et al.*, 1996; Boff and Lincoln, 1988; Boff *et al.*, 1986).

The experiment discussed in this chapter (the modified multi-directional tapping task) was successful in deriving critical data about the characteristics of the participants, and further details such as a common trend in the group and classifications within groups can be acquired when more participants are present.

For future works, we plan to perform the experiment on more participants and on higher scales for more derivations. We also intend to experiment with different haptic input devices such as an OmniPhantom® and surgical devices which require precision, such as the Da Vinci® surgical system. We plan to derive results based on the different devices to get their characteristic curves. In this case, the experiment may also be extended to incorporate 3D movement in the system with varying haptic feedback. The scale may also be less than one for some systems, and this factor needs investigation.

The present work assumed a typical computer input device position for the experiment. If this position changes, the scales may vary with respect to the relative functional reach of the user. This can be further studied in a future experiment.

CHAPTER 5

QUANTIFICATION OF 3D HCI

5.1 Introduction

Interaction is one of the three pillars (other pillars are Immersion and Presence) of Virtual Reality (VR) (Mütterlein, 2018) which facilitates the feeling of the other two to the user. A user can interact with the 3D virtual world through actions such as pointing, clicking, rotating the head, and so on, involving the coordination between human haptic and visual systems. Currently, with the advent of feature-rich virtual reality hardware such as the HTC Vive and the Oculus Touch, there are many ways for the user to interact with the virtual environment (Egger *et al.*, 2017; Borrego *et al.*, 2018). For example, the user can move objects in VR by holding and moving the controller in real space and interact by pressing the various buttons on the controller. Pointing (or selection) and reaching a target in VR constitute two of the fundamental interactions using a virtual cursor.

A scaled motion of the virtual cursor is used in most applications in order to achieve more immersion (Biocca and Delaney, 1995; Slater and Wilbur, 1997) and allows the user to interact with more objects in the virtual world (Wilson *et al.*, 2018; Xie *et al.*, 2010; Steinicke *et al.*, 2008*b*,*a*; Jaekl *et al.*, 2005). The concept of cursor movement scaling in human-computer interaction is otherwise called as Control-Display (CD) ratio and is explained in Section 4.1. It is different from redirected touching (in works such as Azmandian *et al.* (2016)) due to the fact that the scaling is constant and is dependent on the kinaesthetic motion of the user only.

Fitts's Law (Fitts, 1954) states that there is a linear relation between the time it takes for a task to be completed and the difficulty of the task. The modified Fitts's Law (MacKenzie and Riddersma, 1994) is used in this chapter, which is developed in analogy with Shannon's information theory (Shannon, 1948).

Fitts's Law is used in any earlier studies as shown in Table 5.1 with the correspond-

Table 5.1: Prior works using Fitts's Law. The first column contains the author name and year of the study. The second column denotes the application of the particular study. The third column shows the variant of the law which is used for the study.

Author and Year	**Application**	**Usage of Law**
Arnaut and Greenstein (1986)	2D touch tablet	Variable gain factor
Jellinek and Card (1990)	Powermice and user performance	Reciprocal tapping task
Johnsgard (1994)	VR Glove and mouse	Variable gain factor
Chun *et al.* (2004)	Visio-haptic workstations	Tapping task
Blanch *et al.* (2004)	Computer mouse	Semantic Pointing
Dominjon *et al.* (2006)	Force - feedback interface	Mass perception
Casiez *et al.* (2008)	Performance study in pointing tasks	2D pointing
Teather and Stuerzlinger (2011)	Fish tank VR	3D pointing tasks
Fu *et al.* (2011)	Visio-haptic co-location	Point-to-point reaching
Zeng *et al.* (2012)	3D space	3D hand gestures
Cha and Myung (2013)	3D target arrangement	3D pointing tasks
Teather *et al.* (2014)	Fish tank VR	3D point-selection
Pfeiffer and Stuerzlinger (2015)	Vibratory & EMS feedback	Hand pointing task
Shen and Zhou (2017)	VR design system	Drawing tasks
Holmes *et al.* (2017)	Stroke rehabilitation	4 variants, different tasks
Hansen *et al.* (2018)	Gaze based interactions	Click and dwell interaction
Qian and Teather (2017)	Eye-tracking HMD	Eye-based selection
Isaac *et al.* (2018)	Visio-haptic interactions	5 scales, tapping task

ing author and the variant of the Fitts's Law used in the study. An early study from Hess (1973) showed that the performance of an average human would increase to a point and then reduce (inverted U-shaped function) when increasing the control gain or CD ratio. Works of many other authors also found this inverted U-shaped function using the Fitts's Law (Zhai *et al.*, 1996; Boff and Lincoln, 1988; Boff *et al.*, 1986; MacKenzie, 1992; Blanch *et al.*, 2004; Langolf *et al.*, 1976; Arnaut and Greenstein, 1990). Raghu Prasad *et al.* (2013) conducted experiments for finding the minimum movement time (MT) in a 2D task which indicates human performance. The task type is based on force variation, and it uses the individual's index finger. Fitts's Law was modified for a force-based task in a virtual environment. It does not consider the position of a force applied or the limb movements. Their study proved that Fitts's Law could work under these constraints. They also proved that Fitts's Law describes the relationship between the index of difficulty and the time it takes to do one task in the context of virtual environment-related force-based tasks. Other studies with Fitts's Law include works by Johnsgard (1994) in which they compared the performance of a virtual glove to that of the computer mouse. Using Fitts's Law, they verified that the mouse was superior to the VR glove at the time.

When the Law is extended to 3D, there are other factors, such as the depth perception of the user, which affect the Law. According to Balakrishnan (2004), Fitts's Law may not hold in all conditions when used in VR environments. Teather and Stuerzlinger (2011) shown that the Law holds well when used in 3D pointing tasks with a stereoscopic CRT monitor. Their study has shown that the depth perception of the target affected the performance of the participants. This was also shown by works of Chun *et al.* (2004) and works of Pfeiffer and Stuerzlinger (2015). With respect to 3D coordinated hand movements, Zeng *et al.* (2012) performed experiments using the Kinect, and Coelho and Verbeek (2014) performed experiments using the Leap Motion controller. Both experiments proved that Fitts's Law holds well in the given constraints. Hansen *et al.* (2018) and Qian and Teather (2017) implemented Fitts's Law in VR for gaze-based tasks, and both works show that the gaze-based interactions have a lesser throughput as compared to other modalities such as mouse and controllers. Fu *et al.* (2011) used Fitts's Law to validate visual-haptic co-location based movements with varying rotations of the virtual environment. His study showed that the Law holds well irrespective of the rotation of the experiment in the azimuth angle.

A loss of performance reported by Jellinek and Card (1990) occurred due to the loss of resolution of movement at larger scales and resulted in an inverted U-shaped performance-gain function similar to other authors. The performance increases up to a certain scale, and at larger scales, the performance of the user reduces because the cursor movement is not smooth but jagged. This phenomenon is termed the quantization effect.

A solution to the quantization effect, as reported from Casiez *et al.* (2008) is found by increasing the resolution of the display that shows a smooth motion of the cursor at larger scales. He debated that the CD gain should be variable and not constant. This concept is used in the computer mouses of conventional operating systems. The distance traveled by the mouse is a function of the acceleration of the mouse. Hence the mouse moves the cursor at a larger pace when moved fast and vice versa. Hence the cursor can move slower when needed by the user during high precision tasks. However, studies by Arnaut and Greenstein (1986) show that variable CD gain in graphics tablets does not increase the performance of the user. He showed that a CD ratio of 1:1 is optimum for efficient control of the cursor. However, this was contradicted by Isaac *et al.* (2018), where they have shown that the performance was an inverted U-shaped function of the CD ratio when using a graphics tablet. He used a Wacom tablet in which the display was seen on a monitor and not on the tablet. Hence users perform better when the CD ratio is set to a particular range than a natural 1:1 movement when the users' hand is not visible to them during the interaction. It was also noticed during the experiment that the users' hand was not visible to them during the interaction. Hence the users' hand visibility could play a vital role in the interaction performance, which was not studied before.

5.1.1 Objectives

The objective of this Chapter is to examine how the effect of visual awareness of the real hand influences the performance of interaction with the virtual objects using a virtual cursor in a VE. To study the effect, we have compared the user performance of the same 3D tasks in two different VEs, namely the Active One-walled 3D Projection (AOP) and the Head Mounted Display (HMD). Fitts's Law is used to objectively quantify the user performance in both the VEs. There is no visual awareness of the real hand in the case of HMD, however, the visual awareness of the real hand is predominant in AOP.

In this Chapter, we have found that the visual awareness of the real hand decreases the user performance in virtual environments. This decrease is due to a conflict which we refer to as Virtual Kinaesthetic Conflict (VKC), and to avoid the conflict, we also provide guidelines in this Chapter.

5.2 Methodology

We have designed an experiment to study the influence of the visual awareness of the real hand in AOP and HMD. When using the AOP, the user can see both the cursor and their hand during the experiment. However, when using the HMD, the user can not see their real hand (the physical hand). They can only see the cursor that is in the VE.

5.2.1 Virtual Environment Setup - Active One-walled 3D Projection (AOP)

The first environment for the experiment is the Active One-walled 3D Projection, which is made by using a Benq W1050 projector [1]. The schematic of the experimental setup using AOP is shown in Figure 5.1. The participant stands wearing a pair of active 3D glasses in an open space of about 2m×2m wide. The participant holds a Vive controller with their dominant hand and is made to face a projection screen that is 2.5m away from the participant's starting position (to prevent the participant from hitting the screen during the experiment). A 3D projector is positioned above the participant, which projects the interaction task on the screen for visual feedback. The projector is fixed such that the participant does not occlude the projection, which avoids any shadow on the projection.

Two variations of the AOP are designed in order to ascertain the effect of visual hand awareness on user performance in the experiment, namely AOP type A and type B, as shown in Figure 5.2. AOP type A (AOP-A) is the first variant in which the participant

[1] `https://www.benq.com/en/projector/cinehome-home-cinema/w1050.html` (accessed April 30, 2019)

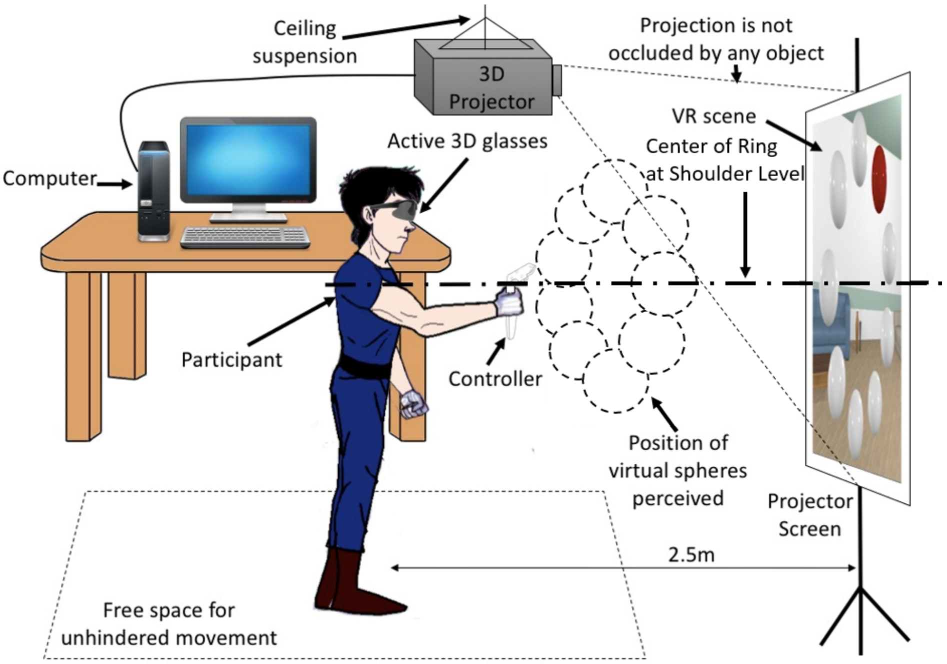

Ceiling suspension
3D Projector
Projection is not occluded by any object
VR scene
Center of Ring at Shoulder Level
Active 3D glasses
Computer
Participant
Controller
Position of virtual spheres perceived
Projector Screen
2.5m
Free space for unhindered movement

Fig. 5.1: A schematic diagram of the Active one-walled 3D projection. The participant stands wearing a pair of active 3D glasses in an open space of about 2m×2m wide. The participant holds a Vive controller with their dominant hand and is made to face a projection screen that is 2.5m away from the participant's starting position. A 3D projector is positioned above the participant, which projects the interaction task on the screen.

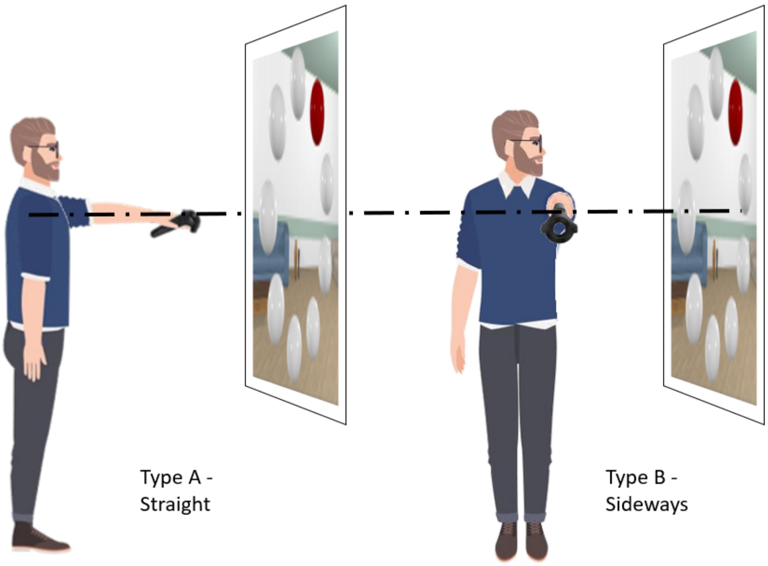


Fig. 5.2: Two variations of the experiment performed with the AOP. AOP type A (AOP-A) is the first variant in which the participant faces the screen with their body facing forward towards the screen as well and moves their hand in front of them. In the second variant, the participant turns their body 90° clockwise and performs the task while facing the screen but with the rest of the body facing away from the screen.

faces the screen with their body facing forward towards the screen as well and moves their hand in front of them; hence they see their real hand as well as the virtual cursor. In the second variant, AOP type B (AOP-B), the participant turns their body 90° to the right and performs the experiment while facing the screen but with the rest of the body facing away from the screen. This ensures that the participant does not see their real hand while performing the experiment.

5.2.2 Virtual Environment Setup - Head Mounted Display (HMD)

The schematic of the experimental setup using the HMD is shown in Figure 5.3. The experiment was conducted in a room with an ambient temperature set at 24°C. The participant wears the Vive HMD[2] and stands in a clear space of dimensions 2m×2m. There are no obstructions to the movement of the participant's hands and legs. The participant then holds a Vive controller using their dominant hand and faces a ring of spheres in the VE. When starting the experiment, the participant holds their hand straight in front of them (similar to the avatar in Figure 5.3) and presses a button on the controller. This action sets the center of the ring of spheres to be at the same level as the current position of the controller. This calibration ensures that the center of the vertical ring is calibrated to be the same height as the participant's shoulder level. The participant can then move forward or backward to position the ring at arm's length. The latency between the movement of the Vive controller and the movement of the cursor in the VR environment was measured and is in the range of 10-20 ms. Hence the effect of latency on the participant can be neglected since the latency of the system is well below the allowable limits as reported by Gourishetti *et al.* (2018) for visual-haptic feedback.

Software Specifications

The virtual environment used for both VEs is created using the Unity[3] 3D game engine along with the SteamVR SDK[4] and Windows 10 operating system. The assets used in the virtual environment are all open source.

[2] `https://www.vive.com/eu/`
[3] `https://unity.com/` (accessed April 30, 2019)
[4] `https://www.steamvr.com/en/` (accessed April 30, 2019)

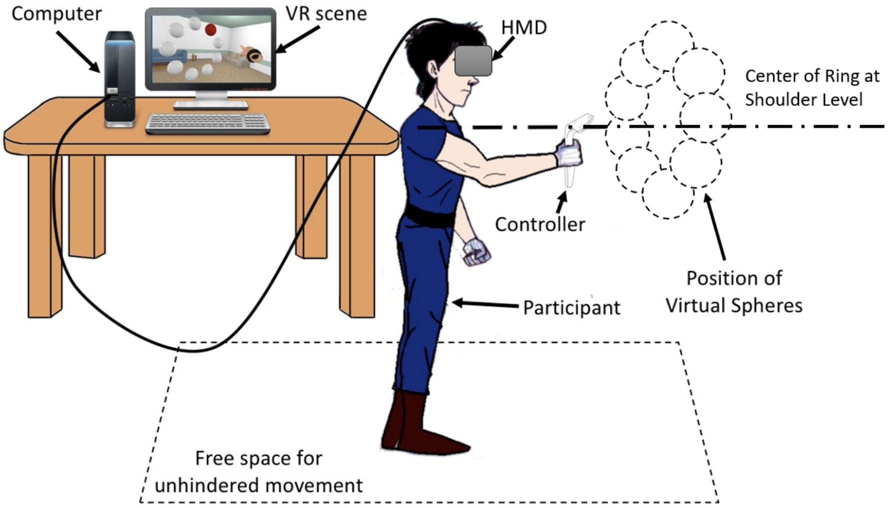


Fig. 5.3: A schematic diagram of the HTC Vive. The participant stands wearing the HMD in an open space of about 2m×2m wide. The HMD is connected to the computer using a long cable. The participant holds a Vive controller with their dominant hand and is made to face an interaction task in the VE.

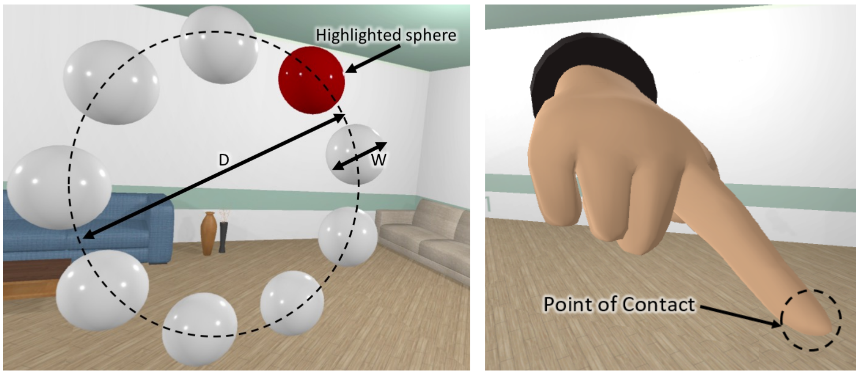
5.2.3 Participants' specifics

The experiment involves 15 participants, with ages ranging from 20 to 30 years. We got informed consent from all participants by filling out a consent form before the experiment. They were all naive for the experiment, and they were given a general introduction to the VE using a trial experiment. All participants had little or no gaming experience since experienced gamers may cause a bias in the experiment. They all could see stereo imagery, were right-handed, and had normal or corrected vision. The height of the participants did not matter since the test is calibrated to the height of each participant.

5.2.4 Experimental Procedure and Task

The task performed in all the environments is a modified version of the multi-tapping experiment outlined in the ISO 9241-9 standard (2000)[5]. The test is similar to that performed by Isaac *et al.* (2018), and it is modified for a 3D environment; hence, the circles are now spheres.

[5] Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Requirements for non-keyboard input devices, International Organization for Standardization, 2000



(a) 3D view

(b) Cursor used

Fig. 5.4: 3D view of the modified multi-tapping experiment outlined in the ISO 9241-9 standard. As shown in (a), nine spheres of diameter W each are arranged in a large ring of diameter D. One sphere is then highlighted in red which indicates the current target of the participant. The participant uses the cursor shown in (b) to select the sphere as fast as possible. The tip of the index finger is used as the point of selection of the cursor.

Table 5.2: Set of (D,W) pairs used for the experiment. These pairs are selected such that they cover a good range of ID values for the experiment. The table is sorted by descending order of ID.

Set ϕ	**D (m)**	**W (m)**	**ID (bits)**
Pair 1	0.35	0.05	3
Pair 2	0.3	0.1	2
Pair 3	0.15	0.05	2
Pair 4	0.25	0.1	1.80
Pair 5	0.3	0.15	1.58
Pair 6	0.2	0.15	1.22
Pair 7	0.25	0.2	1.16
Pair 8	0.2	0.2	1
Pair 9	0.1	0.1	1
Pair 10	0.15	0.2	0.81

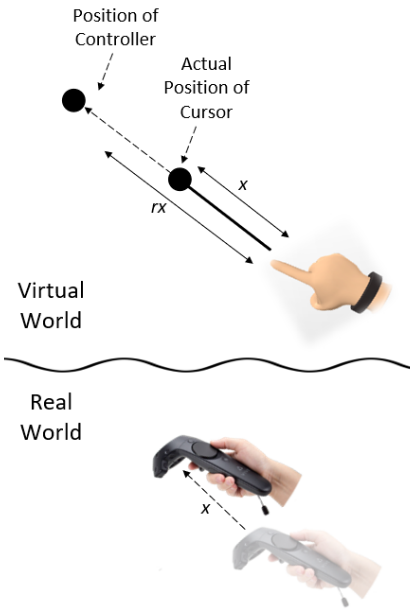


Fig. 5.5: Illustration explaining the concept of scale or CD ratio. When the Vive controller moves a distance x in the real world, the cursor moves by a distance rx. The value of r ranges from 1 to 5 as in scale 1:r

Task to perform

The experiment consists of a virtual environment with a ring of 9 spheres, as shown in Figure 5.4. The plane of the ring is vertical and perpendicular to the floor of the virtual world. The task is to select one of the 9 spheres, which is highlighted in red in the virtual space, by clicking on it using the Vive controller as fast as possible and traverse to the next highlighted sphere. They must repeat the process until there are no more spheres available to click on. They are encouraged to click as close to the center of the sphere as possible.

The distance from the floor to the center of the ring is adjustable and is calibrated to be the shoulder-height of the participant when beginning the test. The diameter of the ring is D while the diameter of the spheres in the ring is W. The values D and W of the spheres are taken from a predefined set (denoted as ϕ) of pairs shown in Table 5.2. The cursor is a generic hand model which is pointing with its index finger. Whenever the tip of the index finger collides with the highlighted sphere, the controller provides vibrotactile feedback.

Changing the CD ratio

The scale in this Chapter refers to the ratio between the distance traveled by the participant's Vive controller to the distance traveled by the virtual cursor, as shown in Figure 5.5. When the scale is set to 1:1, the displacement of the cursor is the same as the displacement of the HTC Vive controller. When the scale is set to a value higher than 1:1, the cursor moves more than the actual displacement of the Vive controller with respect to the center of the user (central chest region) taken as the origin. The origin is chosen such that the participant can return to a starting point rather than drifting through the environment, especially in AOP. Each trial is repeated for five different scales (1:1, 1:2, 1:3, 1:4, and 1:5).

Experiment Protocol

Step 1: The participant is instructed to read the informed consent and fill the questionnaire before and after the experiment.

Step 2: The participant is made to stand at the designated location and hold the Vive controller.

Step 3: **If using AOP-A:** The participant is made to wear the active 3D glasses and made to stand 2.5m in front of the 3D projector screen. The participant's head and torso face the screen.

If using AOP-B: The participant is made to wear the active 3D glasses and made to stand 2.5m in front of the 3D projector screen. The participant's torso is at an angle of 90° away from the screen, as shown in Figure 5.2 with their head facing the screen.

If using HMD: The participant is made to wear the HMD and made to stand such that there is no obstruction to their movement by the cables or any other objects.

Step 4: The task in Section 5.2.4 is then explained to the participant in detail.

Step 5: A trial experiment is provided to the participant to get acquainted with the VR environment and the experiment.

Step 6: After conducting the trial experiment, the set ϕ is shuffled to provide a random order of (D, W) pairs to the participant. This shuffled set is denoted as ϕ'. A random CD ratio is then set (from 1:1, 1:2, 1:3, 1:4, 1:5).

Step 7: A pair is taken from ϕ', and the participant is instructed to perform the task.

Step 8: After finishing a pair, the next pair is provided from ϕ' and repeated until all pairs in ϕ' are completed.

Step 9: After the experiment, a break is given for 5 minutes, and then steps 7 and 8 are repeated with another random CD ratio (from 1:1, 1:2, 1:3, 1:4, 1:5).

Step 10: Step 9 is repeated for all five CD ratios.

Step 11: Breaks are provided intermittently if the participant demands.

This protocol is then repeated for AOP-A, AOP-B, and HMD. We ensured that adaptation (bias) did not occur to the participant by making them perform the experiments for each VE on different days, providing at least 5 minutes between trials for a particular VE. The CD ratio per trial is given at random to the participant in order to prevent a learning bias in the participant. It is ensured that all participants perform the experiment in all VEs (AOP-A, AOP-B, and HMD introduced in different orders to each participant) using all CD ratios (1:1, 1:2, 1:3, 1:4, 1:5).

5.2.5 Parameters Measured

In order to quantify the performance of the participant during the task in the VE, Fitts's Law is used as shown in equation 5.1.

$$t = a + b\,\mathrm{ID} \tag{5.1}$$

where t the is movement time (MT) in seconds and ID is the index of difficulty in bits which is shown in equation 5.2. The intercept a is not of importance in this context since the slope b is the only variable required.

$$ID = log_2\left(\frac{D}{W} + 1\right) \tag{5.2}$$

where D is the distance between targets and W is the target width. In our experiment, this corresponds to the ring diameter and sphere diameter, respectively. The Throughput (also known as the index of performance, I_p) is shown in equation 5.3.

$$I_p = 1/b \tag{5.3}$$

I_p is a measure of human performance in bits per second, and this modified version is used as an indicator of hand-eye coordination in this study.

In order to calculate the Throughput, the following parameters are recorded in the experiment:

Movement Time (MT):

When the participant clicks on the first red sphere, a timer starts and then records the elapsed time when the participant clicks on the next red circle. This recording repeats eight times (the first click of the trial is excluded), hence providing eight time-stamps for completing one ring. From these values, the mean time is calculated and is denoted as the Movement Time (MT), and its unit is in seconds.

Effective Index of Difficulty Per Test (ID_e):

The concept of effective width (W_e) and effective diameter (D_e) is similar to the terms stated by Jude *et al.* (2016) and Isaac *et al.* (2018). The equation 5.4 is the modified form of the Fitts's Law used in this experiment and follows the updated ISO 9241-411 (Secretary, 2012). Instead of using the diameter D, the effective diameter of the larger circle (D_e) is used, which is the average distance traversed by the participant when they move the cursor from one sphere to another. The target width W is now the effective width W_e, which is the mean of the distance of the cursor to the center of the sphere in which it is clicked. D_e and W_e are computed per trial of the participant. The unit of ID_e is bits. Even though the same D and W are provided for all participants, their ID_e change because D_e and W_e differ for each participant based on their performance. The set of D and W values used is shown in Table 5.2.

$$ID_e = log_2 \left(\frac{D_e}{W_e} + 1 \right) \tag{5.4}$$

Throughput (I_p):

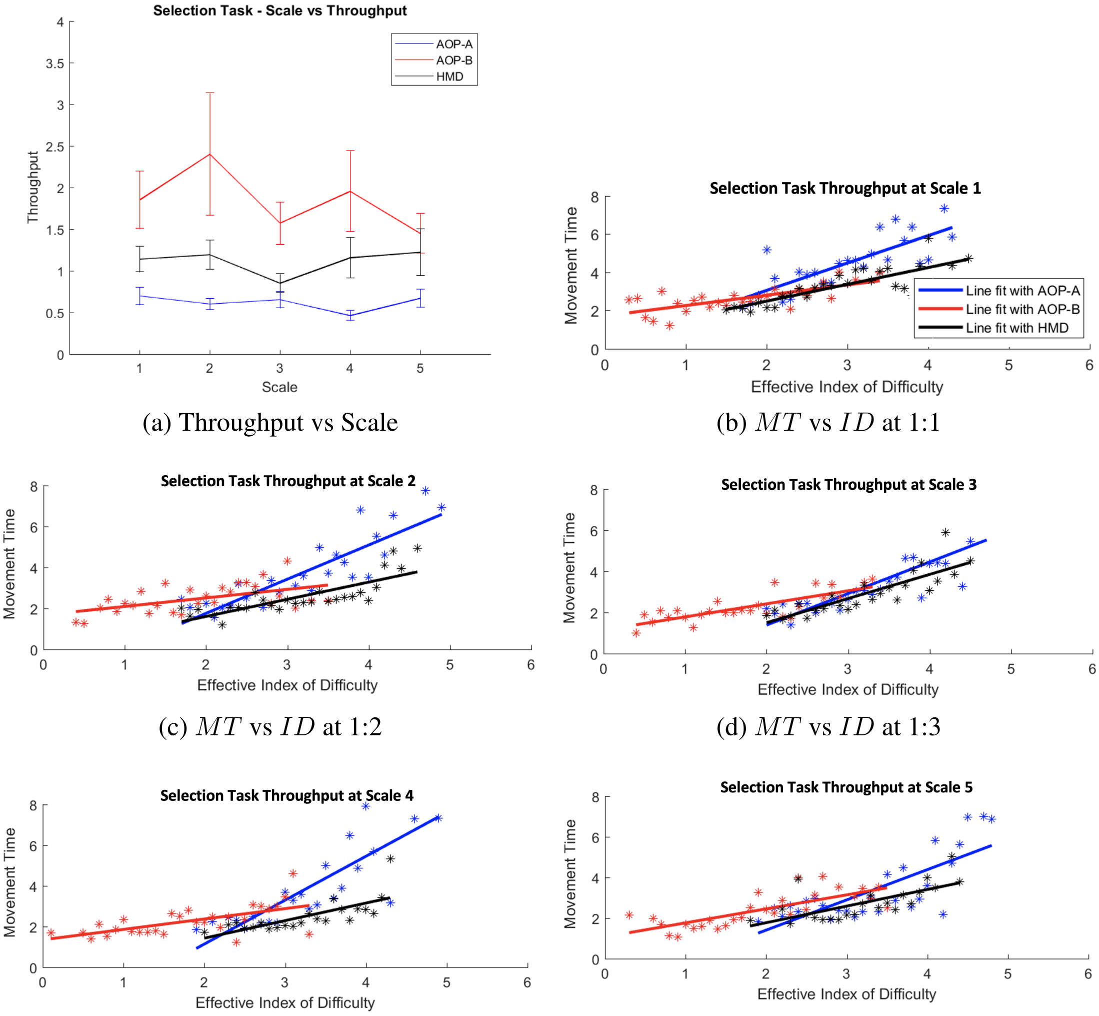
Throughput refers to the number of targets selected per unit time. The participant performs the test with ten different sets of D and W. These trials yield ten different values of D_e and W_e, which are then used to calculate ID$_e$ and fit a line on ID$_e$ vs. MT. The inverse of the slope of the line is the effective Throughput (I_p) of the participant. The linear fit for ID (which uses D and W) is also shown in the results for comparison.

5.2.6 Analysis and Visualization of Data

The data collected from the tests are visualized by using the methods from Jude *et al.* (2016). The raw data from the subjects are collated together and then grouped per environment, per CD ratio. That data set is then grouped per ID_e rounded to one decimal point. This rounding procedure is to cluster the points so that they can be used to calculate the average MT for all those points. Figure 5.6b to Figure 5.6f shows the points after the clustering and the calculation of the mean MT per unique ID_e value. A line is then fit on these points using the least-squares method, and the slope of the line is computed. The inverse of that slope corresponds to the Throughput achieved collectively by the subjects in the virtual environment and the particular CD ratio. This process is then repeated for each scale and each VE. The collated set of throughputs is shown in Figure 5.6a as three trend lines that correspond to the three VR scenarios performed. The black points and the black line correspond to the data collected from the subjects using the HMD. Similarly, the blue and red data correspond to AOP-A and AOP-B, respectively.

5.3 Results and Discussion

Based on the analysis of the data from the experiment, as shown in Figure 5.6b to Figure 5.6f, we observe that the relation between ID_e and MT is linear and has a positive slope. The range of ID values used in the experiment is the same for all three experiments and ranges from 0.81 bits to 3 bits, as shown in Table 5.2. These values correspond to different ID_e values as reported by Jude *et al.* (2016). The range of ID_e largely depends on the users' overall performance during the experiment. Higher ID_e values imply that the participant performs a more difficult task than the given task hence they are more precise in their movements and select the spheres closer to the center. This reduces W_e, which, in turn, increases ID_e. However, if the participant selects the sphere closer to the edge of the sphere, the ID_e reduce. This is seen in the subplots of Figure 5.6b to Figure 5.6f in which the range of ID_e is low in AOP-B but



(a) Throughput vs Scale

(b) MT vs ID at 1:1

(c) MT vs ID at 1:2

(d) MT vs ID at 1:3

(e) MT vs ID at 1:4

(f) MT vs ID at 1:5

Fig. 5.6: Consolidated Results which shows the MT versus ID obtained from the experiment, which shows one plot for each scale setting (1:1, 1:2, 1:3, 1:4, 1:5) used in the experiment. The data collected from the tests are visualized by using the methods from Jude *et al.* (2016). The raw data from the subjects are collated together and then grouped per environment and per CD ratio. That data set is then grouped per ID_e rounded to one decimal point. This rounding procedure is to cluster the points so that they can get be used to calculate the average MT for all those points. In each plot, the black asterisks correspond to the experiment conducted with HMD, the blue corresponds to the experiment conducted with AOP-A, and the red corresponds to AOP-B.

Parameter	AOP-A & AOP-B	AOP-A & HMD	AOP-B & HMD
t-stat	-7.192	-6.310	4.099
P($T <= t$) two-tail	$9.3132 * 10^{-5}$	0.0002	0.0034

Table 5.3: Paired t-Tests of the data

is higher in HMD and AOP-A. This could imply that participants in AOP-B perform effectively easier tasks by selecting the spheres close to the edges. In other words, the improvement could be due to the fact that the real hand is not visible in AOP-B.

Significance Tests

The data in Figure 5.6 shows that the performance of the subjects is significantly different across the three VEs. The ANOVA test shows that the three throughput sets (set of five values for each VE) are significantly different with $\alpha = 0.05$ [F(2,12) = 34.06, p = 1.128 * $e^{-}5$]. Paired t-tests were also performed between the sets, and the results are shown in Table 5.3, which shows that each set of throughputs (from AOP-A, AOP-B, and HMD) are significantly different from each other, and the performance using AOP-B is significantly larger than both HMD and AOP-A. Table 5.3 also shows that the performance using the HMD is significantly larger than AOP-A. Table 5.4, Table 5.5, and Table 5.6 are the Z scores calculated to examine the significance with respect to the throughputs of each scale for AOP-A, AOP-B, and HMD, respectively. The blue color squares indicate that the Throughput using the CD ratio in the row is significantly lower than that in the corresponding column. Table 5.6 shows that there is no particular range where the Throughput is significantly high. However, there is a significant performance drop at the CD ratio of 1:3. Table 5.4 shows that the CD ratio 1:1, 1:3, and 1:5 have a reduced throughput; however, the difference is minor. Table 5.5 shows that the throughputs at CD ratios 1:2 and 1:4 are significantly high.

Virtual kinaesthetic Conflict

The main finding of this study is that the Throughput of AOP-A is significantly lower than that of the HMD. However, the Throughput of AOP-B is significantly higher than

Scale	1:1	1:2	1:3	1:4	1:5
1:1		3.894	1.559	9.448	0.882
1:2	***-3.894***		***-2.194***	7.599	***-2.877***
1:3	-1.559	2.194		8.052	-0.659
1:4	***-9.448***	***-7.599***	***-8.052***		***-8.502***
1:5	-0.882	2.877	0.659	8.502	

Table 5.4: Paired Z test for AOP-A with $\alpha = 0.05$. The blue color squares indicate that the throughput using the CD ratio in the row is significantly lower than the CD ratio in the corresponding column. For example, the performance of CD ratio 1:2 is significantly lower than 1:1 (Z = -3.894).

Scale	1:1	1:2	1:3	1:4	1:5
1:1		***-3.563***	3.431	-0.919	5.1
1:2	3.563		5.607	2.665	6.503
1:3	***-3.431***	***-5.607***		***-3.687***	1.83
1:4	0.919	***-2.665***	3.687		4.965
1:5	***-5.1***	***-6.503***	***-1.83***	***-4.965***	

Table 5.5: Paired Z test for AOP-B with $\alpha = 0.05$. The blue color squares indicate that the throughput using the CD ratio in the row is significantly lower than the CD ratio in the corresponding column. For example, the performance of CD ratio 1:1 is significantly lower than 1:2 (Z = -3.563).

Scale	**1:1**	**1:2**	**1:3**	**1:4**	**1:5**
1:1		-1.225	7.789	-0.325	-1.265
1:2	1.225		8.345	0.555	-0.428
1:3	***-7.789***	***-8.345***		***-5.309***	***-5.799***
1:4	0.325	-0.555	5.309		-0.805
1:5	1.265	0.428	5.799	0.805	

Table 5.6: Paired Z test for HMD with $\alpha = 0.05$. The blue color squares indicate that the throughput using the CD ratio in the row is significantly lower than the CD ratio in the corresponding column. For example, the performance of CD ratio 1:3 is significantly lower than 1:1 (Z = -7.789).

both HMD and AOP-A. This shows that the user performance significantly increases when the user does not have visual awareness of their hands. Conversely, if the visual awareness of the hand is present, the participant experiences a conflict between the movement of the virtual cursor and the movement of the real hand at higher CD ratios. This visual conflict requires some time to get adjusted and hence reflected in the performance of the participant.

We define this conflict as a virtual kinaesthetic conflict (VKC). It is a phenomenon in which the user perceives a difference in the positions of their hand (kinesthesia) and the virtual cursor (visual feedback). The conflict mainly occurs when the virtual movements are scaled with respect to the users' hand movements. This concept of cursor movement scaling in human-computer interaction is otherwise called as Control-Display (CD) ratio. According to MacKenzie and Riddersma (1994), it is the ratio between the physical movement of the input device in the real world and the corresponding movement of the cursor in the virtual world. Changing the CD ratio allows the virtual cursor to make movements that can be larger or smaller than the movement of the real hand. In a VR system such as the HMD, modifying the CD ratio of the virtual cursor has little influence on the user's kinaesthetic perception. In this case, the user is not visually aware of the real hand, and the user can adjust his/her kinaesthetic sense according to the movement of the virtual cursor. In the case of VR systems like CAVE or 3D projection, the user is visually aware of both the real hand and the virtual cursor. In such a scenario, there is a conflict between the visual and kinaesthetic information (Lambrey *et al.*, 2002), which is VKC.

A related issue is well discussed in the Augmented Reality literature as the registration problem (Azuma, 1997). Registration problems also exist in Virtual Environments but are not as severe as it is in AR. VKC explains factors more complicated than registration problems. Several experimental studies (Jones *et al.*, 2010; McGuire and Sabes, 2009; Sarlegna and Sainburg, 2007; Sober and Sabes, 2005; Pouget *et al.*, 2002) have shown that even for simple kinaesthetic tasks, such as reaching for an object with a hidden hand, the brain constructs a visual representation of the movement. This imaginary visual representation usually matches that of the virtual cursor in VR. However, when the representation does not match with that of the virtual cursor in a scaled movement, the brain detects a conflict which is the VKC.

Relation between CD Ratio and VKC

Human motor behavior is essentially a trade-off between speed and accuracy. This property is well modeled and described by Fitts' Law, assuming the human body has a limited capacity to transmit information while executing the motor task. This trade-off or limited capacity leads to optimum CD ratios, as shown in previous literature on Fitts' Law without considering the VKC. VKC occurs only when the CD ratio is other than unity (lesser than or greater than unity). The degree of VKC depends on the CD ratio and other factors pertaining to the virtual environment. According to our hypothesis, the VKC occurs when the user is visually aware of both the real hand and the virtual cursor. In this kind of interaction, the user's perception adjustment to the virtual cursor is difficult, and it affects the user's performance in the interaction with the virtual world.

The existence of an optimum CD ratio with VKC can be explained in terms of control theory. The sensory and motor control loop of the human movement as a system may resonate more in this optimum CD ratio (McMahon, 1984). This means that the user's sensory-motor system may adapt more to the particular CD ratio, which is considered the optimum CD ratio. More precisely, the visual appearance of the real hand may be considered as a noise added to the sensory-motor control loop, which may degrade

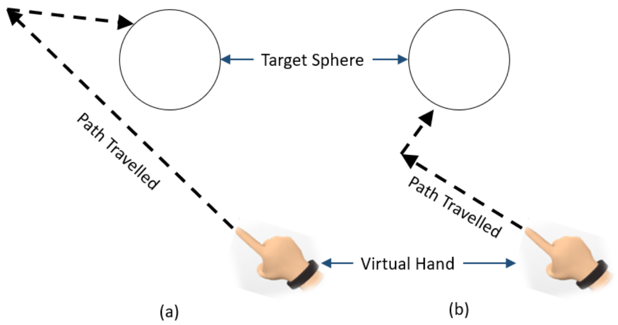


Fig. 5.7: The overshooting and undershooting phenomenon. The left side (a) demonstrates the overshooting, whereas the right side (b) demonstrates undershooting.

the overall user's interaction performance.

The actual human motor circuit has inherent time delays, and values have been reported from about 30 ms for a spinal reflex and up to $200 - 300$ ms for a visually guided response. VKC may cause further time delays, which may be of the same order as that of the inherent time delays. From our observations, we show that varying the CD ratios affect user performance when the real hand is visible, essentially due to the time delays added by VKC. This is because there is a visual conflict during the time the user is trying to correct the movement of the cursor while also seeing their real hand. This mismatch may cause a time delay in the movements, resulting in a further loss of performance.

It can also be understood in the form of overshooting and undershooting the target, as shown in Figure 5.7. The overshooting is shown in Figure 5.7a and is the phenomenon that occurs when the CD ratio is too high, and the user misses the target by moving the cursor beyond the target initially, then corrects the movement by coming back to the target. The second phenomenon is undershooting, as shown in the second in Figure 5.7b, where the cursor moves at an inadequate distance initially, and then the user corrects the movement by moving again to the target. When the CD ratio is lesser than the optimum CD ratio, undershooting often occurs, hence reducing performance. Conversely, when the CD ratio is more than the optimum CD ratio, overshooting occurs more often, which again reduces performance. Hence at the optimum CD ratio, the level of undershooting and overshooting could be at a minimum. From our observations from the experiments, the VKC may have caused a time delay in the movements

(McMahon, 1984) which resulted in undershooting and overshooting, and in turn, the loss of performance.

Guidelines To Avoid VKC

In order to reduce the effects of VKC, we propose a set of guidelines to follow when scaled motions are involved in the virtual environment to improve the 3D interaction performance.

- For the HMD VE, the display mounted on the head should be fully opaque, and the user should not be able to see their real hand movements in the virtual world.
- For the Active One-walled 3D Projection, the user must perform tasks in which visual awareness of the real hand is not present. This is due to the presence of VKC when visual awareness of the real hand is present. The participant should also be positioned such that they face straight at the display with no parallax, and their shadow does not fall on the display. Their arm and hand movements should not be in their field of view.
- For partially immersive environments (such as AOP-A) where the real hands ought to be visible to the participants, scaled movements should be avoided.
- For partially immersive environments (such as AOP-A) where the real hands ought to be visible to the participants and scaled movements are necessary, then the scales that cause minimum VKC should be used.

5.4 Conclusion and Future Work

In this chapter, we examined how the effect of visual awareness of the real hand influences the performance of interaction with the virtual objects using a virtual cursor. Earlier studies such as those by Zhai *et al.* (1996); Boff and Lincoln (1988); Boff *et al.* (1986) shown that users perform better when the CD ratio is set to a particular range than a natural 1:1 movement when the users' hand is not visible to them during the interaction. It was also noticed during the experiment that the users' hand was not visible to them during the interaction. The main focus of this experiment was the effect of user hand visibility on the performance of scaled interactions.

We validated the effect by comparing the performance of the same task in two different VEs, namely the Active One-walled 3D Projection (AOP) and Head Mounted Display (HMD). Two variations of the AOP were performed, namely AOP type A

(AOP-A), where the participant faces the screen with their torso facing forward towards the screen, and AOP type B (AOP-B), where the participant faces the screen with their torso rotated 90° clockwise from the screen. Fitts's Law was used as a tool to quantify the performance in the two VEs, and we have shown that the performance in AOP-A is significantly lower than that of the HMD, but the performance in AOP-B is significantly higher than both HMD and AOP-A. We observed that the CD ratio of 1:2 is optimum in AOP-B, and there is no optimum CD ratio for HMD and AOP-A. This performance drop in AOP-A from AOP-B was due to the conflict between the virtual hand and the real hand. This conflict we referred to as the Virtual Kinaesthetic Conflict or VKC. Guidelines for avoiding the VKC have been provided.

Future works include investigating the presence of VKC in other VEs such as the CAVE (CAve Virtual Environment) and monocular head-based displays. While we used the selection task in the current study, other tasks can also be investigated to find the optimum performance and CD ratio, such as reaching tasks and manipulation tasks. We can also study the effect of CD ratios lesser than unity on these tasks for different VEs.

The current work studied the VKC effect of scaled movements in translation. Any general movement may be considered as a combination of both translatory and rotatory movements. Therefore, in our future study, it is critical to study the VKC effect of scaled movements in rotation.

The cursor used for the experiment is a generic hand model. This was selected due to its analogy with real human hands pointing and selecting an object. Further studies can be done to see the effects of using non-human objects as cursors, such as 3D arrow marks.

CHAPTER 6

QUANTIFICATION OF 3D MICROSCOPIC SELECTION TASK

6.1 Introduction

Surgical robots are increasingly being used in almost every multi-specialty hospital. Currently, they have multiple applications in pelvic, abdominal, select-chest, and neurosurgical procedures (Patel, 2005; Benway *et al.*, 2009; Gutt *et al.*, 2004; Kiaii *et al.*, 2000; Gharagozloo *et al.*, 2008; Rizun *et al.*, 2004). The technology of surgical robots today has advanced sufficiently to provide a high degree of freedom in the robots' arm movements, control, and stereoscopic vision. The improved user interface of these technologies has led to better on-the-job performance and faster learning curves in the use of surgical robots such as da Vinci® (Ballantyne and Moll, 2003), ROSA® (Gonzalez-Martinez *et al.*, 2014; Lefranc and Peltier, 2016) and Senhance™ (Alletti *et al.*, 2018). However, some of these robots lack the nuances of haptic (touch sensation) feedback that would allow the surgeon to feel tissue through the robot. Surgeons need surgical robots that result in decreased complications, better outcomes, and shortened operating room times. While the future development of surgical robots will undoubtedly involve haptics, it would be useful to have surgical training systems that simulate such an environment. We develop such a training system for surgeons who use surgical robots with haptics such as Senhance™ (Alletti *et al.*, 2018). We also demonstrate that our training system renders faster and improved task performance when haptics is used.

Task performance skills such as speed, accuracy, sensitivity, and precision are referred to as psychomotor skills. They play a vital role in surgical performance. Surgical robotic systems typically seek to help surgeons perform surgeries with higher accuracy, faster responses to intraoperative complications, and increased dexterity. Robotic surgical training systems should, therefore, have scenarios, challenges, evaluations, and quantification of these skills. Quantification would then help gauge and improve the surgeon's learning curves. Our work achieves this by developing a model that applies Human-Computer-Interface (HCI) laws (Dubey *et al.*, 2014) to robotic surgical training. HCI laws allow the evaluation and quantification of skills necessary for the performance of fine motor tasks involving interaction between humans and machines. Using HCI laws, we develop a quantitative index of performance. We are then able to evaluate training outcomes using this Index of performance which can be incorporated into current robotic surgery training curricula. In the future, we plan to incorporate a module to train for contextual intraoperative complications.

Modern curricula for robot-specific surgical training through simulation are being developed and improved to revolutionize robotic surgeries (Kassite *et al.*, 2019). These simulations are aimed at developing self-driven, and mentor-free skills using fully-immersive 3D virtual environments, which is often referred to as Virtual Reality (VR) Lee and Lee (2018). However, neither quantifiable psychomotor skills nor haptic feedback is present in these training simulations. We show that incorporating both of these features improves learning curves and prepares surgeons for the use of surgical robots with haptic feedback.

6.1.1 Objective Quantification of Psychomotor Skills in Training

Performance during robotic surgery training is currently evaluated subjectively through cognitive, technical, and non-technical skills. Technical skills (TS) include psychomotor skills, user perception, visuospatial orientation, and adaptation (Collins *et al.*, 2018). Subjective evaluations can sometimes have observer biases. Hence, we need training systems that will challenge, evaluate and train surgeons in all these aspects objectively. In the operating room, TS is currently assessed using three methods: (i) subjectively by Global Evaluative Assessment of Robotic Skills (GEARS) which involves scoring videos of the actual surgery on a numerical scale called Likert scale (Volpe *et al.*, 2015), (ii) objectively by evaluating intraoperative videos using assessment tools such as Prostatectomy Assessment and Competency Evaluation (PACE) (Hussein *et al.*, 2017), and (iii) automated performance metrics (APMs) (Hung *et al.*, 2018).

Each of these methods has a minor drawback in evaluating and improving a surgeon's performance. GEARS using the Likert scale is easy to use but is subjective and can therefore have inter-observer variability. The PACE method provides useful information on the entirety of individual steps in the procedure but does not provide a breakdown of performance skills within each step. Both GEARS and PACE are tools used after the actual surgery is completed rather than in a pre-surgical training system. On the other hand, APM measures the robot's performance for specific tasks undertaken by the operator surgeon rather than the skills of the surgeon. Thus, pre-surgical assessment and training of surgeons in technical skills are still needed (Strauss *et al.*, 2006). The reason technical skills are important from a surgeon's perspective is that they reduce operating time, reduce patient exposure to operative risks, improve the handling of intraoperative complications, and reduce post-operative morbidities (Cheng *et al.*, 2018). Our training system permits a quantifiable evaluation of the technical skills of the surgeon before performing actual surgeries.

Approaching a target with increased speed, accuracy, precision and manipulating the target with improved dexterity are essential psychomotor skills demanded of the surgeon. This task becomes much more challenging when the target is in the microscopic (millimeters) scale, where the surgeon has to zoom in and zoom out the visual magnification while operating. Most surgical licensing loard exams, therefore, evaluate critical thinking during intraoperative complications and assess the speed and precision subjectively, without quantification of skill levels (Levi, 2016; Billiar *et al.*, 2009). A meta-analysis by Cheng *et al.* (2018) has shown that the post-operative complications are increased by 14% for every 30 minutes increase in the operating time. Therefore, it has become essential for a surgeon to get trained in psychomotor skills such as speed, accuracy, precision, and dexterity in a surgical training curriculum.

Our training system allows for the objective evaluation and learning of psychomotor skills. Improvement in psychomotor skills of the trainees is achieved using Selection tasks to choose a specific object with distinct characteristics (color, shape, etc.) among a collection of objects in a 3D virtual environment. This object to be selected is referred to as a target.

One of the unique requirements for robot-specific surgical training is the incorporation of variation in the levels of difficulty, which are defined by various parameters.

One among them is the movement scale which refers to the ratio of the movement of the input device (surgeon's console) to the movement of the robotic arm. Other parameters include distance to target and size of the target. These parameters are used to enhance the capability of the user to control and manipulate small objects. When the objects are small, the task becomes more challenging. Such selection tasks are termed Microscopic Selection Tasks (MST). Training for MST is essential, because movement scales and visual magnifications (Chapuis and Dragicevic, 2011) are incorporated into current surgical robots (Palep, 2009). We thus provide challenging scenarios to the user in performing and training for MST.

6.1.2 Importance of Haptic Feedback in Training

Haptics is the science of the sense of touch. In robotic surgery, the incorporation of haptics is an important advancement as it allows the robot operator to feel tissues being handled by the robot. When the surgeon is able to feel tissue characteristics through a robot, it is called haptic feedback. Psychomotor skills training with haptic feedback has shown improvements in the learning curve for surgical simulations in virtual reality (Prasad *et al.*, 2016; Basdogan *et al.*, 2004). One of the unique challenges in robotic surgery is the lack of haptic feedback, which includes tactile and proprioceptive sensations (Srinivasan, 1995). Tactile sensation involves a sense of pressure and vibration, whereas proprioceptive sensation involves the sense of position, movement, and forces. Most surgical robots today do not have the capability to sense and transmit such information to the surgeon. In a surgical robot console, a surgeon relies entirely on visual cues from 3D cameras. This lack of haptic feedback makes the surgical task difficult (Abiri *et al.*, 2019), especially for novice surgeons. Therefore, tactile feedback could assist surgeons using robotic consoles (Abiri *et al.*, 2019; Bethea *et al.*, 2004; Pacchierotti *et al.*, 2015). Moreover, to deal with intraoperative complications, handling of tissue, and suturing, haptic feedback plays a vital role. Intraoperative bleeding during surgeries is a critical complication that is dealt with by accurately and precisely identifying the source of the bleed and gently stopping the bleed (Billiar *et al.*, 2009). Studies by Ebrahimi *et al.* (2016) and Kontarinis and Howe (1995) reported that Vibro-Tactile Feedback (VTFB) enhances the performance of manipulation tasks in virtual environments by reducing reaction times. However, these studies did not quantify the psychomotor skills required for robotic surgery. Koehn and Kuchenbecker (2015) reported that the VTFB is preferred by both surgeons and non-surgeons in simulated robotic surgery. Although this work involves robotic surgery, psychomotor skills were not quantified in their study. More recent studies have shown that VTFB is a vital sensory adjunct to surgeons operating a surgical robot (Okamura, 2004, 2009; Westebring-van der Putten *et al.*, 2008; Van der Meijden and Schijven, 2009; Wedmid *et al.*, 2011). Thus, the addition of VTFB during training can improve the learning curve, especially for psychomotor skills. We include haptic feedback in our training system to augment the experience of human-computer interaction and study the performance index with and without VTFB.

6.1.3 Aim and Contribution

Our goals are: (1) To objectively quantify the psychomotor performance of a Microscopic Selection Task (MST) in a fully-immersive 3D virtual environment that can be used to train surgeons in the use of surgical robots. (2) To find an optimum movement scale by measuring the psychomotor performance in MST at each scale. (3) To verify if the performance of the MST improves with VTFB.

Our contributions to this work:

(a) We determine a surgeon-specific optimal movement scale.

(b) We introduce VTFB as a tool for improvement of performance in MST.

(c) We introduce a method for quantifying psychomotor skills in MST by adapting existing HCI laws. Our work can be used as a tool to quantify and improve psychomotor performance during the training of surgeons in a robotic surgical curriculum (Collins *et al.*, 2018).

6.2 Law of Human-Computer Interaction (HCI) for MST

A surgical robot's console is a Human-Machine-Interface (HMI). Interaction with such interfaces is referred to as Human-Computer-Interaction (HCI). With the advent of HMI and HCI tools (Wania *et al.*, 2006; Ha, 2014), there are many ways to objectively quantify psychomotor performance using any input device such as the aforementioned surgi-

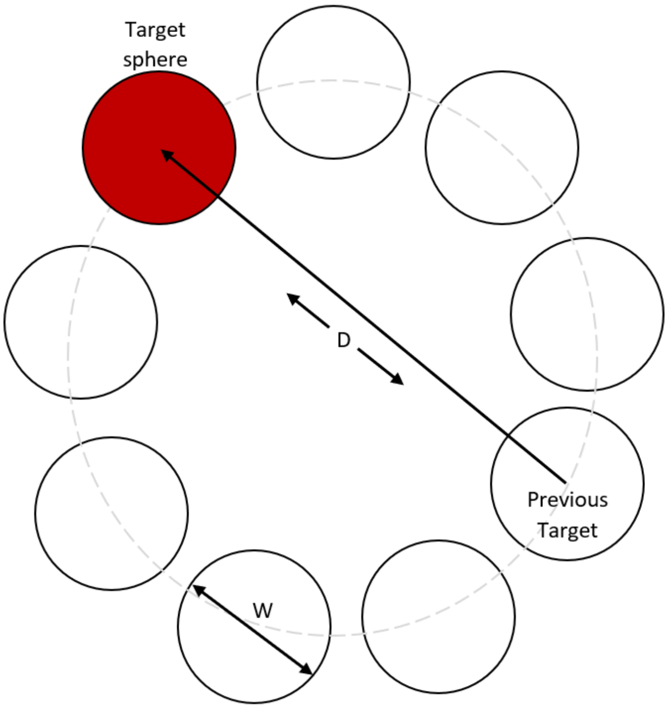


Fig. 6.1: A typical Fitts's multi-tapping task (MacKenzie and Isokoski, 2008) which depicts nine spheres, each with diameter W, arranged in a circle of diameter D. D and W are varied to change the difficulty level of the task.

cal robot's console. One such method is Fitts's law which is a widely accepted powerful tool for modeling human movement. Although Fitts's law was introduced in 1954 originally (Fitts, 1954), the first application of it to HCI was by Card *et al.* (1978) in 1978 for comparison of different input devices. Fitts's law is a linear relation between task completion time and the difficulty of the task. The difficulty of the task is described by two parameters as shown in Figure 6.1. The first parameter, D, is the distance the cursor (which is driven by the user's hand) has to travel to the target (red ball). The second parameter is W which is the width of each ball. Each level of difficulty that the user encounters is defined by the different values of D and W. Mathematically, the performance of an individual person for a given task is described by a linear equation shown in Equation 6.1

$$t = a + b\,\mathrm{ID} \tag{6.1}$$

where 't' is the Movement Time (MT) to reach the target in seconds, ID (Index of Difficulty) is a measure of the difficulty of a given task, a refers to the minimum time required to complete the easiest task ($ID = 0$) and b describes how MT changes as ID changes. Both a and b depend on the choice of the input device, which is the surgical robotic console. The measure of the Index of difficulty is computed by the ratio D/W. Thus, as the distance D decreases or the width W increases, the task becomes progressively easier as the surgeon's hand has to traverse shorter distances to reach a larger target. Similarly, as D increases or W decreases, the task becomes progressively harder. Mathematically, the values of D/W can, therefore, range from 0 to infinity. In order to scale this down, the next step in constructing the measure of the Index of Difficulty is to use the logarithm of D/W. Although this makes the scale more manageable, we still have to deal with the possibility of log(0). Hence we measure the Index of Difficulty by calculating the log of ($D/W + 1$). For this purpose, we use a log base 2 scale so that when $D = W$, the Index of Difficulty becomes 1. Hence, we define the Index of difficulty according to the equation 6.2 below given by Soukoreff and MacKenzie (2004).

$$\text{ID} = log_2 \left(\frac{D}{W} + 1\right) \tag{6.2}$$

A measure of performance of a user in carrying out the selection of targets is termed as throughput (MacKenzie and Isokoski, 2008). It refers to the number of targets selected per unit time. In the context of robotic surgery, throughput indicates how quickly the surgeon selects the target during robotic surgery training.

High throughput is demanded in robot-assisted surgeries, especially Urological and Neurological surgeries, which involve a precise selection of microscopic tissues quickly. The movement scale settings are already existing in the da Vinci® surgical robot and are given as Normal (1:1/2), Fine (1:1/3), and Ultra-fine (1:1/5) scales. However, quantification of performance in these scales has not been reported yet.

6.2.1 Literature review on 3D Fitts's MST

The concept of scaling in visual and motor domains are widely studied (Coutrix and Masclet, 2015; Browning and Teather, 2014; Chapuis and Dragicevic, 2008). In Chapter 4, we studied various movement scales for 2D tapping tasks using Fitts's law. The results have shown that the performance was an inverted U-shaped function of the movement scale. The same tapping tasks can be extended to 3D Virtual Reality (VR), where the user perceives depth information. Current literature has extended Fitts's law application from 2D to 3D virtual environments where depth influences performance (Balakrishnan, 2004; Teather *et al.*, 2014; Pfeiffer and Stuerzlinger, 2015). Each of these papers mentions methods by which Fitts's law in the 2D virtual world can now be applied to a 3D virtual environment. However, Balakrishnan (2004) suggests that Fitts's law may not always hold in all VR conditions. Other studies suggest that Fitts's law fails in selection tasks where the target size is of few pixels (Chapuis and Dragicevic, 2008, 2011). The problems in the acquisition of small targets are well established in the HCI research literature (Chapuis and Dragicevic, 2011). Teather *et al.* (2014) have shown that Fitts's law holds when the 3D pointing task is performed with the stereoscopic monitor. They also observed that depth perception influences the performance of the participants. This result was confirmed later by Chun *et al.* (2004) and Pfeiffer and Stuerzlinger (2015).

Fitts's law holds well in 3D coordinated hand movements (Coelho and Verbeek, 2014; Zeng *et al.*, 2012) in which hand tracking is done by IR tracking devices such as Kinect™ and Leap Motion™. The law is also used for the gaze-based tasks (Hansen *et al.*, 2018; Qian and Teather, 2017), which have shown that the throughput is lesser compared to that of using a mouse or hand-held controllers.

Our previous work (Isaac *et al.*, 2018) found an optimum movement scale in a 2D environment by considering four different movement scales in the macro range (1:2, 1:2 4, 1:3.3, 1:4.9). The following are the differences in the present work compared to our previous work: First, the environment used in the present study is 3D, which involves depth effect, and it is an entirely immersive VR space, whereas the previous study is a desktop screen which is not fully-immersive. Second, the movement scales considered in the present work are less than one (micro-scale), whereas the previous work involved macro scales. When the movement scale is less than 1, there is a diminution of output movement for any specific input movement. Third, the present study's comparison of throughput with and without tactile feedback was not part of the earlier study. Fourth, unlike the earlier work, which used conventional Fitts's law, the current work uses modified Fitts's law which is defined in the next section. Finally, the previous

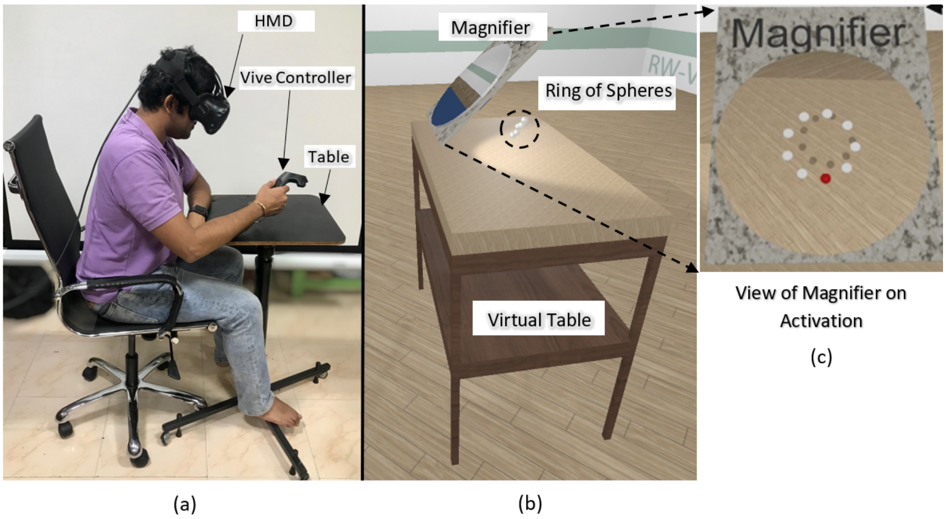


Fig. 6.2: The setup used to perform the experiment. The subject is made to sit in a chair comfortably, and their arms are resting on the table in front, as shown in Figure 6.2a. Figure 6.2b is the view of the experiment in the virtual environment, while Figure 1b shows the real-world perspective of the experiment. The subject is made to sit comfortably on a chair with their arms resting on the table to prevent fatigue during the experiment. The virtual dummy is shown to represent the position of the subject during the experiment and is not present during the actual experiment. A magnifier is shown on the left and is positioned in front of the subject that offers an FoV of 50°. The ring of circles is illuminated by a small light that causes a shadow that is used for depth cues.

work uses throughput as the only performance measure of the task, whereas the current work emphasizes quantification of performance in an MST through speed, dexterity, and precision along with throughput.

6.3 Materials and Methods

Our objective is to propose a training system and demonstrate its effectiveness in improving the performance of virtual psychomotor tasks. In this section, we show the construction and working of the training system. We also explain the parameters used to objectively quantify and analyze the performance of the user. The experiment is performed in a 3D virtual environment designed in such a way that the subject can perform MST, similar to operating a generic surgical robot console.

6.3.1 Apparatus Setup and Specifications

The experiment was conducted in a room with the temperature set at 24°C (this is considered a comfortable temperature given the geographical and cultural context of the venue where the experiment was conducted), where the subject is made to sit in a chair comfortably and their arms resting on the table in front, as shown in Figure 6.2a.

The experiment was conducted using the HTC Vive™ which consists of a head-mounted display (HMD) with two base stations for tracking, positioned at opposite ends of the room. The Organic Light Emitting Diode (OLED) display embedded in the HMD provides a refresh rate of 90 Hz and an FoV of 110 degrees, making sure that the user is completely immersed in the VR environment. The user holds a Vive controller, as shown in Figure 6.2a, for interacting with the virtual environment. They are used to track the position of the user's hands in real-time with sub-millimeter level accuracy and map them into the virtual environment space. The Vive™ controller also provides vibrotactile feedback (VTFB) by means of an embedded linear resonant actuator (LRA). It is an electromagnetic device that can produce vibrations at 235 Hz. The latency between the movement of the Vive™ controller and the movement of the cursor in the VR environment was measured and is in the range of 10-20 ms. The effect of latency on the subject can be neglected since the latency of the system is well below the allowable limits as per Gourishetti *et al.* (2018) for visual-haptic feedback. The virtual environment shown in Figure 6.2b is created using the Unity 3D game engine (Helgason, 2004) along with the SteamVR SDK (Valve, 2003).

6.3.2 Subject Selection Criteria for the Study

We conducted the experiment with fifteen subjects with a mean age of 25.3 years ± 4.7 years.

Inclusion criteria: Healthy subjects in the age range from 21 - 30 years. None of the subjects should have any prior knowledge of our hypothesis, experimental environment, or experience in VR.

Exclusion criteria: Subjects with any neurological motor or sensory disorders. Presence of any visual deficits despite corrected vision.

The experimental protocol stated in Section 6.3.3 below was explained to all the subjects clearly. Any questions they had were answered to their satisfaction, and they were given a trial to familiarize themselves with the equipment before the start of the actual experiment. Subjects received no remuneration, and there were no fees to participate in this study.

6.3.3 Experimental Procedure

The experimental task performed by the subjects is a modified version of the ISO 9241-9 standard (2002) multi-tapping experiment ISO (2000).

Virtual Environment Setup

The VR environment consists of a virtual room in which the subject is made to sit in front of a table, as shown in Figure 6.2a. Every subject carries out the entire study with a fixed level of zoom in order to be able to see microscopic objects. In a totally immersed 3D VR environment, zooming in to see small objects can create a vertiginous effect on the user. In the real world, such an effect can be nullified by the user by taking his view off the field and looking at a standard 1:1 magnified world. However, in our experimental protocol, since the subject is using an HMD, the only way to look away from the zoomed VR world is by dismounting the HMD. This creates a noticeable and laborious interruption to the experiment. We, therefore, devised a visual magnifier in the VR environment that allows the subject to zoom in on the target. During any vertiginous episode, the subject can look away from the magnifier, even in the virtual environment, to get back to a 1:1 VR world without zoom. This avoids the necessity for the HMD to be dismounted. The experiment can then continue smoothly when the subject returns to look through the magnifier. Thus, we have placed in the VR environment a virtual visual magnifier. This offers a 50° field of view.

Through the virtual visual magnifier, the subject can see nine virtual spheres of diameter W arranged in a large ring of diameter D on a plane inclined 50° to the table, as shown in Figure 6.2b. The spheres are all colored white, and one of the spheres is highlighted in red, which is the target sphere. The subject holds the Vive™ controller using their dominant hand to control a virtual cursor, which mimics a typical surgical

Table 6.1: Set of (D,W) pairs used for the experiment. These pairs are selected such that they cover a good range of ID values for the experiment. The table is sorted by descending order of ID.

Set	D (mm)	W (mm)	ID (bits)
Pair 1	5	8	0.70
Pair 2	8	6	1.22
Pair 3	10	4	1.81
Pair 4	15	6	1.81
Pair 5	20	4	2.58
Pair 6	15	2	3.08
Pair 7	10	1	3.46
Pair 8	20	1	4.39

tool. The ring of spheres is illuminated by the magnifier such that a shadow is formed behind it. The shadow serves as a depth cue for the subject during the task. Another depth cue is provided by the stereoscopic rendering of the magnifier (the view is rendered for each eye separately, thereby mimicking a real magnifier). The origin of the nine spheres is the lowermost sphere in the ring. This origin is chosen such that the subject can rest their hands on the table during the experiment and avoid fatigue during the task.

The task to Perform - Microscopic Selection Task

The task of the subject is to move the cursor to the target sphere and select it as fast as possible by clicking the Vive controller. The subject is encouraged to click as close to the center of the target sphere as possible. Once the target sphere is selected by the cursor, another sphere diagonally opposite to the current sphere becomes the target and the process repeats. This repetition occurs once for each sphere (totaling nine repetitions), and then the D and W change to a new set. The set of D and W values is predefined initially, as shown in Table 6.1. Each pair of (D, W) is given to each subject exactly once from Table 1 in random order until all the pairs are exhausted. This reduces the learning bias that may occur during the experiment. The (D, W) pairs used in this experiment consists of smaller ranges as compared to that from Table 5.2 as the task in question is the Microscopic Selection Task. The first four pairs (Pairs 1 to

4) corresponds to small tasks such as ENT procedures involving small bones in the ear. The next 4 pairs (Pairs 5 to 8) are tougher tasks which can be compared to micro tasks such as brain aneurysm surgery (Linfante and Wakhloo, 2007).

There are two different variations of our experiment; one is with VTFB every time the cursor collides with the target to be selected, and another is without VTFB in which the subject purely relies on the visual cues only. The subjects perform the experiment in both variations.

Movement Scale

The movement scale in this work refers to the ratio between the distance traveled by the subject's Vive controller (x) to the distance traveled by the virtual cursor (x/r), as shown in Figure 6.3. When the Vive controller moves a distance x in the real world, the cursor moves by a distance $\frac{x}{r}$. When $r = x$, the movement of the cursor is the same as the movement of the Vive controller. When $r < x$, the movement of the cursor is lesser than the movement of the Vive controller. The experiment involves five different scales (1:1, 1:$\frac{1}{2}$, 1:$\frac{1}{3}$, 1:$\frac{1}{4}$, 1:$\frac{1}{5}$).

Protocol of the experiment

The 15 subjects were divided into two groups, as follows: Group A with 7 subjects and Group B with 8 subjects.

Step 1: Instruct the subject to read the informed consent and fill out an initial questionnaire before and a feedback questionnaire after the experiment.

Step 2: Request the subject to sit comfortably on a chair and to hold the Vive controller while resting their arms on the desk. After sitting, help them wear the HMD.

Step 3: Explain to the subject that their task is to look through the virtual magnifier, to select the target sphere in the virtual space as close to the center of the sphere as possible by clicking on the Vive controller, and then to traverse as fast as possible to the next red target sphere.

Step 4: Inform the subject to look away from the magnifier within the VR environment in the event of any vertiginous episode resulting from the zooming effect of the magnifier.

Step 5: Provide a preliminary trial task to the subject to get acquainted with the VR environment and the experiment.

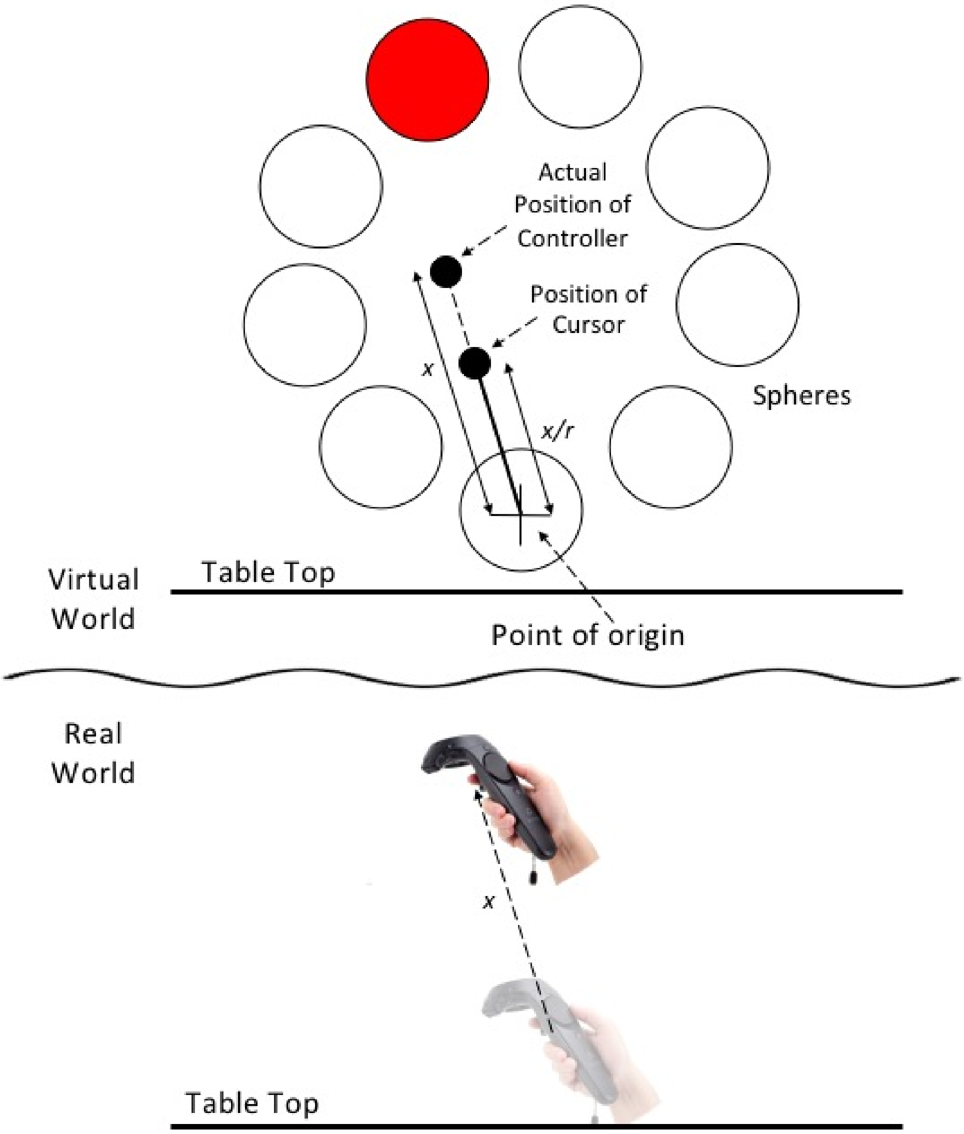


Fig. 6.3: Illustration explaining the concept of scale. When the Vive controller moves a distance x in the real world, the cursor moves by a distance $\frac{x}{r}$. The value of r ranges from 1 to 5 as in scale 1:$\frac{1}{r}$

Step 6: Begin the actual experiment. At the end of each set, provide a break of 2 minutes.

Step 7: Provide the tasks with VTFB first and then without VTFB for group A, whereas for group B, provide the tasks in the reverse order.

Step 8: Instruct the subject clearly that they can stop the experiment at any point in time if they feel any discomfort in performing the tasks.

If the experiment is stopped for any reason during a particular (D, W) pair trial, then that trial is repeated, and data from the previous incomplete (D, W) pair trial is discarded.

6.3.4 Quantification of the Microscopic Selection Task

Task Parameters

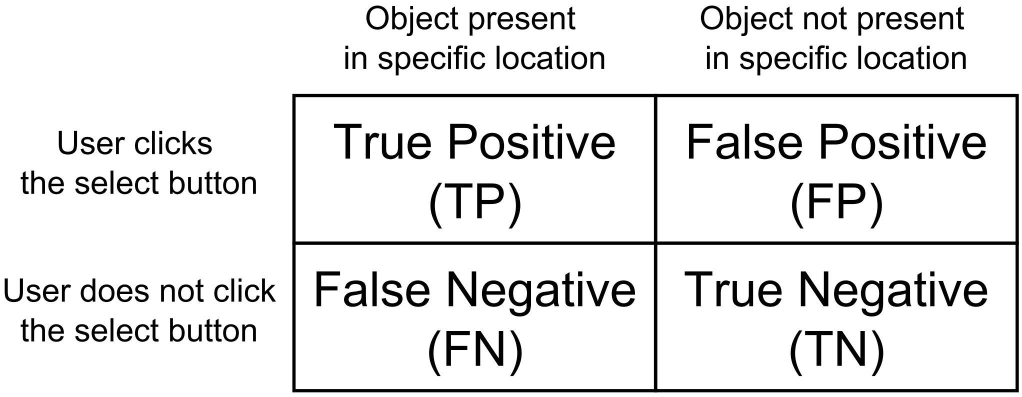
Several parameters are collected from the experiment performed on all the subjects. These parameters are then used to find the performance of the subject using Fitts's Law.

Movement Time (MT)

Movement Time is defined as the average time taken for the subject to select the nine spheres for a particular (D, W) pair. For each (D, W) pair, the timer gets initiated from the instant the subject selects the first target sphere until the next target sphere. Consequently, 8 values get recorded for each trial. The average of these values indicates Movement Time (MT) in seconds for that particular (D, W) pair.

Effective Index of Difficulty (ID_e)

In certain surgical tasks, surgeons must select small tissues more accurately, sometimes at the cost of speed during events such as debridement, cauterization, resection, and suturing. Clicking at the center of the target sphere emphasizes the need to select target tissues accurately in a VR environment. Based on these requirements, the D and W given in Equation 6.2 are modified into effective diameter (D_e) and effective width (W_e) (Jude *et al.*, 2016; Isaac *et al.*, 2018). W_e represents the average distance between the point where the cursor is clicked and the center of the sphere, D_e represents the average



	Object present in specific location	Object not present in specific location
User clicks the select button	True Positive (TP)	False Positive (FP)
User does not click the select button	False Negative (FN)	True Negative (TN)

Fig. 6.4: 2 x 2 Confusion Matrix for MST

distance from one sphere to another traversed by the subject with the cursor. Hence Equation 6.2 is modified to give the effective index of difficulty ID_e as in Equation 6.3

$$\text{ID}_e = log_2\left(\frac{D_e}{W_e} + 1\right) \tag{6.3}$$

ID_e is unique for each subject and depends on their performance during the experiment. Performing tasks with low W_e and high D_e results in a better throughput. Hence the subject is encouraged to click as close to the center of the target as possible to reduce (W_e) and in the shortest time possible.

Throughput

Each subject performs a trial with 8 different ID values, as shown in Table 6.1. After completion of all the trials for a given subject, these ID values translate into 8 different ID_e values with their respective MTs. These are then used to plot the relation between ID_e and MT by linear regression of data obtained for MT vs. ID_e. The effective throughput (I_p) is calculated as the inverse of the slope of the linear relation between ID_e and MT. This parameter is the main factor in quantifying the performance of the subject in the experiment.

Sensitivity and Positive Predictive Value

We define True Positive Rate (TPR) as sensitivity which is defined below, and we define Precision as Positive Predictive Value (PPV) which is also defined below. Since our objective is to compare the performance with VTFB and without VTFB, we calculate the relative change in TPR and Precision (Powers, 2011; Ozenne *et al.*, 2015; Simon and Boring III, 1990; Schechter, 1998). In order to calculate the sensitivity and PPV, the number of True Positives (TP), the number of False Positives (FP), and the number of False Negatives (FN) are recorded, as shown in Figure 6.4. A TP is when the subject correctly clicks inside the target sphere during a trial. An FP is when the subject clicks outside the target sphere. An FN is when the subject enters the target but fails to click.

From these values, the sensitivity and PPV are calculated for each movement scale as given in equations 6.4 and 6.5.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{6.4}$$

$$\text{PPV} = \frac{TP}{TP + FP} \tag{6.5}$$

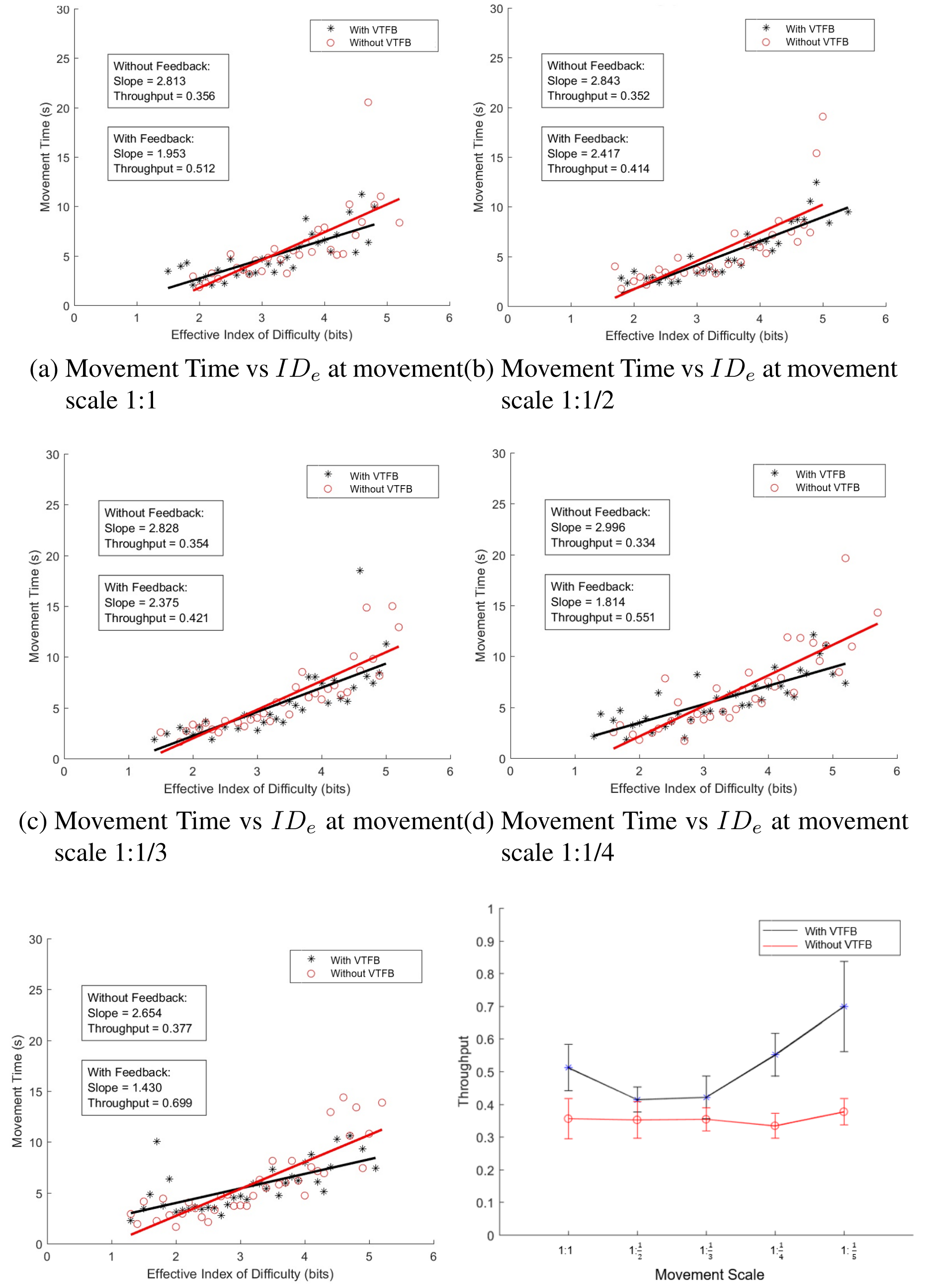
Using the sensitivity of the tasks with VTFB and the tasks without VTFB, the relative change in sensitivity and PPV are calculated as follows

$$\text{Relative change in Sensitivity} \quad = \quad \frac{\text{Sensitivity}_{\text{with VTFB}} - \text{Sensitivity}_{\text{without VTFB}}}{\text{Sensitivity}_{\text{without VTFB}}} \tag{6.6}$$

$$\text{Relative change in PPV} \quad = \quad \frac{\text{PPV}_{\text{with VTFB}} - \text{PPV}_{\text{without VTFB}}}{\text{PPV}_{\text{without VTFB}}} \tag{6.7}$$

Data Analysis

According to Jude *et al.* (2016), using a means-per-user and mean-of-means method instead of the method by Soukoreff and MacKenzie (2004) improves the goodness-of-fit (R^2) and Pearson's *r* coefficient. This is due to the fact that the variance of the MT increases when the difficulty of the task increases; hence fitting a line to these points could have errors. Therefore, in our work, we implement the mean-of-users and then the mean-of-means approach, which reduces these errors.



(a) Movement Time vs ID_e at movement scale 1:1

(b) Movement Time vs ID_e at movement scale 1:1/2

(c) Movement Time vs ID_e at movement scale 1:1/3

(d) Movement Time vs ID_e at movement scale 1:1/4

(e) Movement Time vs ID_e at movement scale 1:1/5

(f) Throughput vs. movement scale with and without VTFB

Fig. 6.5: Consolidated Results showing MT versus ID obtained from the experiment which shows one plot for each scale setting (1:1, 1:$\frac{1}{2}$, 1:$\frac{1}{3}$, 1:$\frac{1}{4}$, 1:$\frac{1}{5}$) used in the experiment. In each plot, the black asterisks correspond to the experiment conducted with tactile feedback, and the red circles correspond to the experiment conducted without tactile feedback. Similarly, the black line and the red line correspond to the linear fit using the black points and the red points, respectively. The inverse of the slope of these fits is the throughput (subplot f) of the participants in the specified scale setting. We can observe that the slope is positive in all the cases (MT increases with ID), leading to the validity of Fitts's law. The corresponding error bars represent the standard deviation of each scale.

6.4 Results and Discussion

Our main objective is to quantify the psychomotor performance of a surgeon, training to carry out microscopic selection tasks (MST). This is conducted in a fully-immersive 3D virtual environment where the trainee undergoes tasks with varying levels of difficulty. We have designed our experiments to find a user-specific optimum movement scale by measuring throughput in MST for each scale (1:1, 1:$\frac{1}{2}$, 1:$\frac{1}{3}$, 1:$\frac{1}{4}$, 1:$\frac{1}{5}$). Moreover, the throughput, sensitivity, and PPV improvements for the tasks with and without VTFB are also studied.

6.4.1 Fitts's Task

The movement Time (MT) data collected from our experiments (with and without VTFB) are consolidated and shown in Figure 6.5, where each plot corresponds to a scale (1:1, 1:$\frac{1}{2}$, 1:$\frac{1}{3}$, 1:$\frac{1}{4}$, 1:$\frac{1}{5}$), which shows that the MT increases as ID_e increases. In each scale setting, 15 subjects performed the experiment twice (with and without VTFB). This results in two sets of points (ID_e , MT), one with VTFB and another one without VTFB. In order to calculate the throughput, a linear regression model is used. The ID_e is rounded to one decimal point, resulting in multiple MTs for each ID_e. The mean of these MTs is calculated per unique ID_e and plotted in Figure 6.5. Any point with an MT higher than the 20s (chosen as it is 2-sigma away from the mean MT) was considered an outlier and was removed from the plot. The remaining points were then used for the linear fit. The linear fit is evident in Figure 6.5 for both the tasks with and without VTFB. For both linear fits, Pearson's *r* coefficient is above 0.8, and the goodness-of-fit is above 0.6.

The ID$_e$, in our experiment, ranges from 1 bit to 5.5 bits, where the latter can be considered a most difficult task. In all the cases, we have a visual magnifier with fixed magnification through which the entire ring of spheres can be visualized.

6.4.2 Significance Tests

Group A with 7 subjects performed tasks without VTFB first and then with VTFB, and group B with 8 subjects performed the same tasks in the reverse order. The movement

Table 6.2: Pairwise Z-score test for throughput with VTFB. $\alpha = 0.05$ and $Z_\alpha = 1.645$.

Scale	**1:1/1**	**1:1/2**	**1:1/3**	**1:1/4**	**1:1/5**
1:1/1		***6.776***	***5.324***	***-2.342***	***-6.983***
1:1/2	***-6.776***		-0.536	***-10.607***	***-11.582***
1:1/3	***-5.324***	0.536		***-8.232***	***-10.594***
1:1/4	***2.342***	***10.607***	***8.232***		***-5.697***
1:1/5	***6.983***	***11.582***	***10.594***	***5.697***	

Table 6.3: Pairwise Z-score test for throughput without VTFB. $\alpha = 0.05$ and $Z_\alpha = 1.645$.

Scale	**1:1/1**	**1:1/2**	**1:1/3**	**1:1/4**	**1:1/5**
1:1/1		0.25	0.15	***1.683***	***-1.643***
1:1/2	-0.25		-0.157	1.472	***-2.048***
1:1/3	-0.15	0.157		***2.203***	***-2.558***
1:1/4	***-1.683***	-1.472	***-2.203***		***-4.613***
1:1/5	***1.643***	***2.048***	***2.558***	***4.613***	

Time (MT) data from these two groups were analyzed to check whether there is a component of learning that can bias the throughput results when VTFB is introduced after the user has already performed the MST without VTFB. The Paired t-test analysis has shown that the comparison of throughput between these two groups is not significant ($t(4) = -2.193, p = 0.151$, for α = 0.05), which leads to the inference that the order of introducing VTFB does not influence the performance. Therefore, all the data are combined together for the quantification of psychomotor performance.

Pairwise Z-score tests were performed between the throughput of movement scales combining both groups (A and B) with and without VTFB. The results of the tests are shown in Table 6.2 and Table 6.3. In what follows, we refer to a cell in the table by its corresponding row and column. The green cell indicates that the throughput of the movement scale in that cell's row is significantly (p < 0.05) greater than the corresponding throughput of the movement scale in its column. Conversely, the blue cell indicates that the throughput of the movement scale in the cell's row is significantly (p < 0.05) lesser than the throughput of the movement scale in its corresponding column.

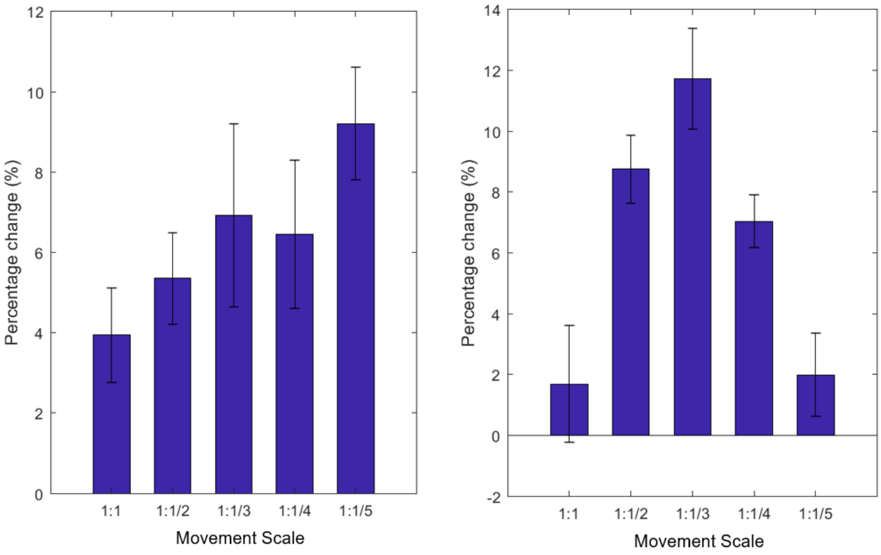
According to Table 6.2, at our significance level (α=0.05), there is an increase in throughput when the scale is set to 1:$\frac{1}{4}$ and 1:$\frac{1}{5}$ with VTFB. This shows that finer movements in the virtual world increase the overall throughput of the subject when performing the MST. However, there is also a performance drop at scales 1:$\frac{1}{2}$ and 1:$\frac{1}{3}$. This could be due to the trade-off between ID$_e$ and MT. When performing tasks at a smaller scale, movement Time increases, which reduces the throughput. When reducing the scale even further, the subject clicks the target more towards the center, which increases ID$_e$. This increase is more than the increase in MT, which results in a lower slope, improving the throughput. Hence the scales 1:$\frac{1}{4}$ and 1:$\frac{1}{5}$ yield higher throughput in terms of precise movements. Natural 1:1 movements have higher throughput in terms of short MTs.

The interpretation of each cell in Table 6.3 is similar to that of Table 6.2. In Table 6.3, the throughput is high when the scale is 1:$\frac{1}{5}$ without VTFB compared to other movement scales. However, the overall magnitude of the throughput across all scales is significantly lower than that with VTFB. This is also seen in a paired t-test, where there is a significant difference in throughput with VTFB (μ = 0.52, σ = 0.12) and without VTFB (μ = 0.35, σ = 0.01); t(8) = 3.14, p = 0.013. This proves that using VTFB is preferable when performing MST. On the whole, the results suggest that the throughput is higher for the tasks with VTFB leading to the need for VTFB in surgical robots and training simulations for robotic surgery.

6.4.3 Role of Tactile Feedback

It is evident from Figure 6.5f that throughput is improving as the movement scale decreases for tasks with VTFB. In our study, with the range of five movement scales used, it can be seen that movement scales less than $\frac{1}{3}$ result in significant ($p < 0.05$) improvement in throughput. However, for the tasks without VTFB, there is no improvement in throughput as the movement scale decreases. In the current literature, the relation between the throughput and the movement scale is an inverted U-shaped curve for macro scales. However, our results do not exhibit this relationship with or without VTFB. The reason could be that the experiment involves micro scales.

The sensitivity and PPV are calculated from the TP, FP, and FN data collected during the experiment. Relative change (percentage) of the sensitivity and PPV with VTFB with respect to that without VTFB are calculated and shown in Figure 6.6a and Figure 6.6b, respectively. When we consider Figure 6.6a, there is a significant ($p < 0.05$)



(a) Relative change in sensitivity

(b) Relative change in PPV

Fig. 6.6: The relative change of sensitivity and PPV for each scale. Each bar in the plots is the relative change of the sensitivity and PPV of the tasks with and without VTFB. The corresponding error bars represent the standard deviation of each scale.

improvement in the relative sensitivity for the cases with and without VTFB. It is observed that the sensitivity is improving for the case with VTFB, especially on scales 1:1/3 and 1:1/5. Moreover, the relative PPV in Figure 6.6b shows a significant ($p < 0.05$) improvement for cases with and without VTFB, especially on a scale of 1:1/3. The overall effect is that the use of VTFB improves the performance of the task as the movement scale decreases.

6.4.4 Robotic Surgical Skills Training Simulation

Surgeons are being trained to use surgical robots. This is now part of surgical training curricula in the last few years (Collins *et al.*, 2018; Satava *et al.*, 2019). Virtual Reality simulations can be considered a better training tool as the skills learning method can be self-driven, mentor free, and it can use modifiable levels of difficulties and challenges (Lee and Lee, 2018; Mazur *et al.*, 2018; Goldenberg *et al.*, 2018). There are three different skills considered in training for the use of surgical robots: technical, non-technical, and cognitive skills.

Training surgeons to use surgical robots and monitoring performance prior to actual patient use is essential to the efficient utilization of both surgeons' skills and robots' capabilities. It is this training, learning, and monitoring using HCI laws that provide the possibility for the surgeon to perform better and for the robots to be fully utilized. Training systems such as the one proposed in this work seek to fulfill these requirements.

6.5 Conclusions

The objective of this study is to quantify the psychomotor performance of microscopic selection tasks (MST) in a fully immersive 3D virtual environment that can be used in training with surgical robots. We have adapted Fitts's law which has been used extensively in the literature on HCI to quantify psychomotor performance. We conclude from our throughput curve that there is no optimum movement scale less than 1:1, and there is a significant improvement in the throughput of MST with vibrotactile feedback (VTFB). From our experiment, we infer that the implementation of VTFB in real and training scenarios for surgical robots has a significant impact on performance, as mentioned in our results. Furthermore, the quantification method described here can be implemented in psychomotor skills assessment for robotic surgery and training curricula. The experiment can be tested in surgical robots such as da Vinci® , ZEUS® , and ROSA® in the future to quantify psychomotor skills along with the incorporation of VTFB in those robots. The goal to improve both the sensitivity and PPV along with speed for target manipulation must form an essential part of an adequate training module in robotic surgery training curricula.

The current study can be extended in many ways: 1) The haptic feedback considered here is a high-frequency vibration. In addition, we can incorporate proprioceptive feedback into the simulation. 2) It involves only one-handed tasks; it can be extended to two-handed Fitts's tasks mimicking the surgical robotic console. 3) It considers a fixed magnification for a virtual visual magnifier, and this can be extended to variable magnification levels for studying the performance of MST under different magnification levels. 4) One of the inclusion criteria for subjects in this study is that they are naive to the performance of MST. Future work can involve surgical trainees and experienced surgeons. 5) Finally, the communication delay between the real and virtual environments is not considered in the current study. Communication delay is usually present when real surgical robots use haptics. In our future work, we can simulate the delay due to the tactile or proprioceptive feedback and study the resulting psychomotor performance.

The microscopic selection task described here, along with the quantification by Fitts's law, can be a psychomotor performance assessment tool. By using the task and analysis described here, mentor-free skills learning in robotic surgery through a quantifiable VR simulation can be achieved.

CHAPTER 7

CONCLUSIONS

The aim of this thesis was to develop a 3D hand tracker that will utilize biomechanical constraints as a closed-form equation. This thesis also analyzed existing 3D interfaces to provide an optimum setting using Fitts's Law.

7.1 Objectives Achieved

The objectives of the research were achieved as follows:

7.1.1 Objective 1 - Realistic 3D Hand Tracking

Biomechanical filter

The first part of the objective was to design a biomechanical filter that can correct the hand pose such that it guarantees zero anatomical error while maintaining low deviation from the ground truth pose (Chapter 2). The filter is modular and can be easily plugged into existing hand trackers with little or no modifications. The results showed that the filter does improve the current state-of-the-art trackers when used in 10% strength, and it was also shown that the state-of-the-art trackers have high errors in terms of anatomical rules and bounds.

Integrate the biomechanical filter to a CNN

The second part of the objective was to design a neural network that incorporates the biomechanical filter at the architecture level such that it increases the speed of computation and efficiency of the network (Chapter 3). We proposed a novel framework called the SSC-CNN for 3D hand pose estimation with biomechanical constraints. The network has biomechanical rules and bounds encoded in the architecture level such that the resulting hand poses always lie inside the biomechanical bounds and rules of the human hand, and no post-processing is required to correct the poses. Our framework was compared to several state-of-the-art models with two datasets. Experiments have shown that the SSC-CNN has comparable results but with no anatomical errors, whereas the state-of-the-art models have very high anatomical errors. The ground truth of the datasets also has anatomical errors, and anatomically error-free versions were created.

7.1.2 Objective 2 - Scaled Human-Computer Interactions

Quantification of 2D HCI

The first part of this objective was to analyze the effect of changing the control display ratio of 2D interfaces and quantify the scaled interactions of the user using those 2D interfaces (Chapter 4). We have used the Fitts' Law for quantifying the user's performance on different scales and have shown that at an optimum control movement scale, users perform better than natural movements. The Fitts' Law also proved to show its validity in scaled conditions with adequate r^2 values in all the scales. The Fitts' regressions were visualized, and it was found that the performance of the participants increases significantly when the scale increases and has an optimum range as well. The experiment discussed in Chapter 3 (the modified multi-directional tapping task) was successful in deriving critical data about the characteristics of the participants, and further details such as a common trend in the group and classifications within groups can be acquired when more participants are present.

Quantification of 3D HCI

The second part of this objective was to upgrade the quantification method to analyze and quantify 3D interfaces (Chapter 5). We validated the effect by comparing the performance of the same task in two different VEs, namely the Active One-walled 3D Projection (AOP) and Head Mounted Display (HMD). Two variations of the AOP were performed, namely AOP type A (AOP-A), where the participant faces the screen with their torso facing forward towards the screen, and AOP type B (AOP-B), where the participant faces the screen with their torso rotated 90° clockwise from the screen. Fitts's Law was used as a tool to quantify the performance in the two VEs, and we have shown that the performance in AOP-A is significantly lower than that of the HMD, but the performance in AOP-B is significantly higher than both HMD and AOP-A. We observed that the CD ratio of 1:2 is optimum in AOP-B, and there is no optimum CD ratio for HMD and AOP-A. This performance drop in AOP-A from AOP-B was due to the conflict between the virtual hand and the real hand. This conflict we referred to as the Virtual Kinaesthetic Conflict or VKC. Guidelines for avoiding the VKC have been provided.

Quantification of 3D microscopic selection task

The third part of this objective was to utilize the 3D quantification method in an applied environment such as robotic surgery (Chapter 6). We have adapted Fitts's Law which has been used extensively in the literature on HCI to quantify psychomotor performance. We conclude from our Throughput curve that there is no optimum Movement Scale less than 1:1, and there is a significant improvement in the Throughput of MST with vibrotactile feedback (VTFB). From our experiment, we infer that the implementation of VTFB in real and training scenarios for surgical robots has a significant impact on performance, as mentioned in our results. Furthermore, the quantification method described here can be implemented in psychomotor skills assessment for robotic surgery and training curricula. The experiment can be tested in surgical robots such as da Vinci® , ZEUS® , and ROSA® in the future to quantify psychomotor skills along with the incorporation of VTFB in those robots. The goal to improve both the Sensitivity and PPV along with speed for target manipulation must form an essential part of an adequate training module in robotic surgery training curricula.

In light of the objectives discussed, the answers to the hypotheses stated before are as follows:

(a) Improving the realism of hand poses predicted by the state-of-the-art trackers using biomechanical aspects improves the accuracy as compared to the poses without implementing such concepts.

(b) A hand pose estimator can, in fact, guarantee zero anatomical error while maintaining low deviation from the ground truth pose.

(c) Higher control movement scale, in general, can be better than natural kinaesthetic movements (where control movement scale = 1:1) in tasks that require extended accuracy using 2D interfaces.

(d) Higher control movement scale, in general, can be better than natural kinaesthetic movements (where control movement scale = 1:1) in tasks that require extended accuracy using 3D interfaces.

(e) HCI-based quantification methods can be used as a training tool for important tasks such as computer-based surgery.

7.2 Limitations

- The biomechanical filter's computational requirements are high since the angles and bounds are calculated and compared for each joint in the hand. This process increases the time taken to estimate output for each input frame and runs at lower speeds when running real-time tracking.

- The SSC-CNN has a limitation in which the training phase requires data preprocessing to derive the joint angles as these angles were not available in the datasets used. Another limitation is that our hand pose estimator does not take the velocity of the joint movements into consideration when correcting them. The angular velocity of the joints also has biomechanical constraints, and these will be incorporated in future works for the model. Although the model is highly robust for varying palm sizes, extreme cases like estimating the hand poses of children may result in inaccurate poses as the dataset used for training does not cover young children's hands and can be investigated in a future work.

- Regarding the quantification of 2D interfaces, the work assumed a typical computer input device position for the experiment. If this position changes, the scales may vary with respect to the relative functional reach of the user.

- Regarding the quantification of 3D interfaces, the cursor used for the experiment is a generic hand model. This was selected due to its analogy with real human hands pointing and selecting an object.

7.3 Recommendations for Future Work

(a) For the biomechanical filter, our future work is to optimize the filter to compute angles and bounds in fewer functions and reduce the time taken to estimate the filtered pose. Optimized methods such as inverse kinematics-based modeling (Aristidou, 2018) methods can effectively correct the joints in real-time. Future works also include utilizing the Law of mobility as per Manivannan *et al.* (2009), which states that the two-point discrimination improves from proximal to distal body parts. Hence, the filter's strength can be changed from the hand's proximal parts towards the hands' distal part. Other future works include enhanced optimizations such as implementing the filter function into the model architecture itself instead of attaching the filter at the end of the model. The baseline model used in Chapter 2 highlights the importance of using anatomical rules during training and can improve the model's accuracy, not only in anatomical correctness but also in pose error.

(b) Future works for SSC-CNN include using synthetic datasets such as the MANO hands (Romero *et al.*, 2017) so that the ground truth will be assured of the hands' true location along with children's hand poses. Using these synthetic datasets, we can also compare the spectrum of poses covered by the currently available datasets and hence cover a broader spectrum of poses for training. Analyzing the history of the hands' motion using methods such as recurrent neural networks (Yoo *et al.*, 2020) instead of processing only one instance of the hand can avoid erratic motions during self-occlusions and will be investigated in another study for adding the feature to the SSC-CNN. The history can include the velocity and acceleration of the joint motions, which also have biomechanical bounds and further enhance the pose realism during hand motion tracking.

(c) Regarding future works on the quantification of 2D interfaces, we can perform the experiment on more participants and on higher scales for more derivations. We also intend to experiment with different haptic input devices such as an OmniPhantom and surgical devices which require precision, such as the Da Vinci® surgical system. We plan to derive results based on the different devices to get their characteristic curves. In this case, the experiment may also be extended to incorporate 3D movement in the system with varying haptic feedback. The scale may also be less than one for some systems, and this factor needs investigation.

(d) Future works regarding the quantification of 3D interfaces include investigating the presence of VKC in other VEs such as the CAVE (CAve Virtual Environment) and monocular head-based displays. While we used the selection task in the current study, other tasks can also be investigated to find the optimum performance and CD ratio, such as reaching tasks and manipulation tasks. We can also study the effect of CD ratios lesser than unity on these tasks for different VEs.

(e) The current study for the quantification of the 3D microscopic selection task can be extended in many ways: 1) The haptic feedback considered here is a high-frequency vibration. In addition, we can incorporate proprioceptive feedback into the simulation. 2) It involves only one-handed tasks; it can be extended to two-handed Fitts's tasks mimicking the surgical robotic console. 3) It considers a fixed magnification for a virtual visual magnifier, and this can be extended to variable magnification levels for studying the performance of MST under different magnification levels. 4) One of the inclusion criteria for subjects in this study is that they are naive to the performance of MST. Future work can involve surgical trainees and experienced surgeons. 5) Finally, the communication delay between the real and virtual environments is not considered in the current study. Communication delay is usually present when real surgical robots use haptics. In our future work, we can simulate the delay due to the tactile or proprioceptive feedback and study the resulting psychomotor performance.

CHAPTER 8

Pilot study for comparing realistic hand poses

8.1 Introduction

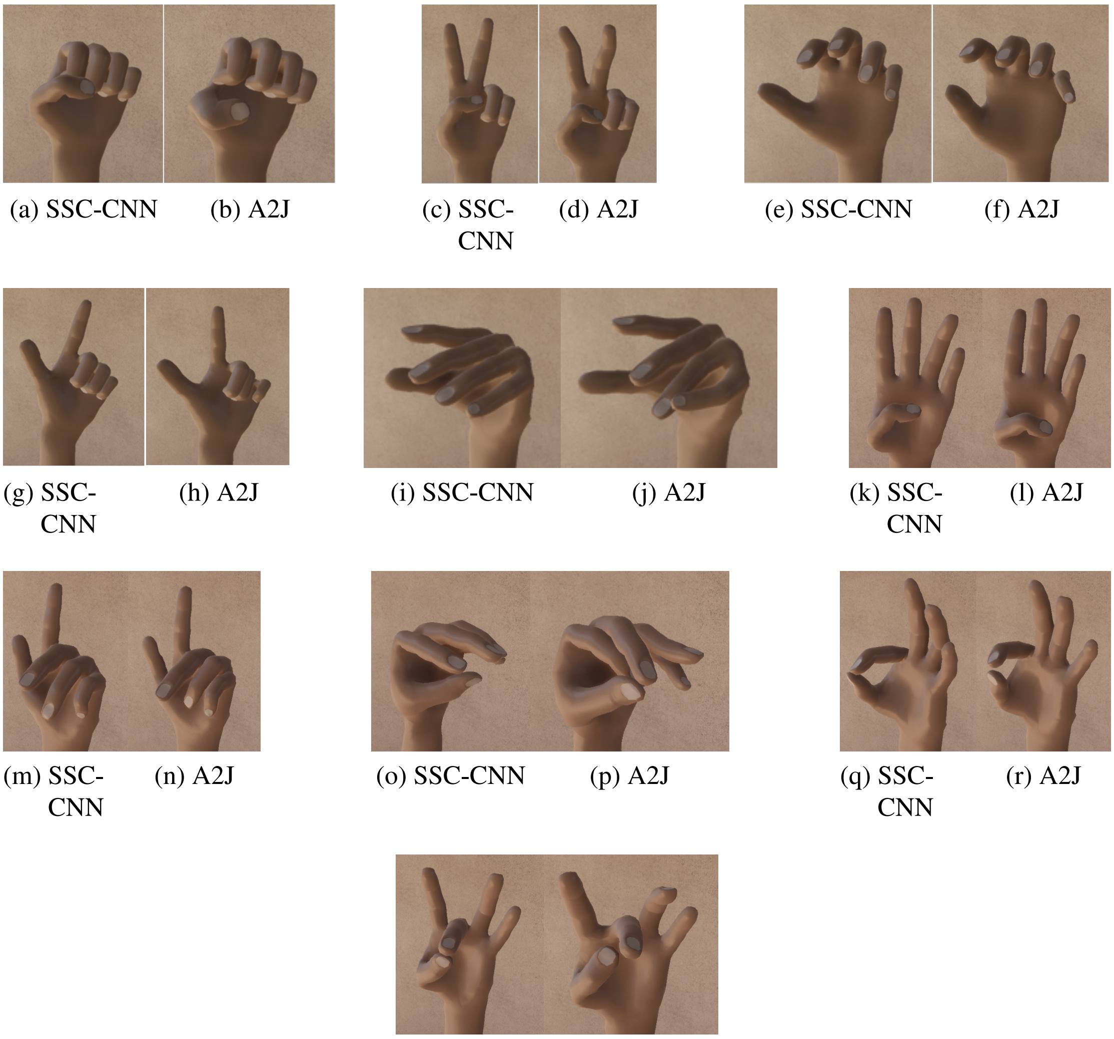
In order to study the perceived realism of the hand poses rendered by the SSC-CNN hand tracker, we compared the hand poses predicted by the SSC-CNN to the same pose predicted by the A2J hand tracker.

8.2 Details of the pilot study

We have conducted a pilot study on 15 participants and shown each of them the results of 10 hand frames by the SSC-CNN and the same frames by the A2J hand tracker. The pose predicted from the models are shown to the participants but order of presentation of the two poses from the two models was randomized (sometimes the pose from A2J is shown first with the pose from SSC-CNN shown second and vice versa). The question we ask the participant is which hand pose is more natural and realistic (first or second). The reply of each participant is then noted down.

8.3 Results of pilot test

From the results, all 15 pointed out poses from the SSC-CNN as the more natural pose as compared to that from the A2J tracker. When questioned for details, some participants stated that the fingers look awkward in few of the poses that were from the A2J tracker. The pose are shown in Figure 8.1



(a) SSC-CNN (b) A2J (c) SSC-CNN (d) A2J (e) SSC-CNN (f) A2J

(g) SSC-CNN (h) A2J (i) SSC-CNN (j) A2J (k) SSC-CNN (l) A2J

(m) SSC-CNN (n) A2J (o) SSC-CNN (p) A2J (q) SSC-CNN (r) A2J

(s) SSC-CNN (t) A2J

Fig. 8.1: The rendered poses for pilot test of realism.

REFERENCES

1. **Abiri, A.**, **Y.-Y. Juo**, **A. Tao**, **S. J. Askari**, **J. Pensa**, **J. W. Bisley**, **E. P. Dutson**, and **W. S. Grundfest** (2019). Artificial Palpation in Robotic Surgery using Haptic Feedback. *Surgical endoscopy*, **33**(4), 1252–1259.

2. **Accot, J.** and **S. Zhai** (2001). Scale Effects in Steering Law Tasks. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01. ACM, New York, NY, USA. ISBN 1-58113-327-8, `doi:10.1145/365024.365027`.

3. **Alletti, S. G.**, **C. Rossitto**, **S. Cianci**, **E. Perrone**, **S. Pizzacalla**, **G. Monterossi**, **G. Vizzielli**, **S. Gidaro**, and **G. Scambia** (2018). The Senhance™ Surgical Robotic System ("Senhance") for Total Hysterectomy in Obese Patients: A Pilot Study. *Journal of robotic surgery*, **12**(2), 229–234.

4. **Aristidou, A.** (2018). Hand Tracking With Physiological Constraints. *The Visual Computer*, **34**(2), 213–228.

5. **Arnaut, L. Y.** and **J. S. Greenstein** (1986). Optimizing the Touch Tablet: The Effects of Control-Display Gain and Method of Cursor Control. *Human Factors*, **28**(6), 717–726. ISSN 00187208, `doi:10.1177/001872088602800609`.

6. **Arnaut, L. Y.** and **J. S. Greenstein** (1990). Is Display/control Gain a Useful Metric for Optimizing an Interface? *Human Factors*, **32**, 651–663. ISSN 00187208.

7. **Azmandian, M.**, **M. Hancock**, **H. Benko**, **E. Ofek**, and **A. D. Wilson** (2016). Haptic Retargeting: Dynamic Repurposing of Passive Haptics for Enhanced Virtual Reality Experiences. *In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16. ACM, New York, NY, USA. ISBN 978-1-4503-3362-7, `doi:10.1145/2858036.2858226`.

8. **Azuma, R. T.** (1997). A Survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments*, **6**(4), 355–385, `doi:10.1162/pres.1997.6.4.355`.

9. **Balakrishnan, R.** (2004). "Beating" Fitts' Law: Virtual Enhancements for Pointing Facilitation. *International Journal of Human Computer Studies*, **61**(6), 857–874. ISSN 10715819, `doi:10.1016/j.ijhcs.2004.09.002`.

10. **Ballantyne, G. H.** and **F. Moll** (2003). The Da Vinci Telerobotic Surgical System: The Virtual Operative Field and Telepresence Surgery. *Surgical Clinics*, **83**(6), 1293–1304.

11. **Basdogan, C.**, **S. De**, **J. Kim**, **M. Muniyandi**, **H. Kim**, and **M. A. Srinivasan** (2004). Haptics in Minimally Invasive Surgical Simulation and Training. *IEEE computer graphics and applications*, **24**(2), 56–64.

12. **Benway, B. M.**, **S. B. Bhayani**, **C. G. Rogers**, **L. M. Dulabon**, **M. N. Patel**, **M. Lipkin**, **A. J. Wang**, and **M. D. Stifelman** (2009). Robot Assisted Partial Nephrectomy Versus Laparoscopic Partial Nephrectomy for Renal Tumors: A Multi-Institutional Analysis of Perioperative Outcomes. *The Journal of urology*, **182**(3), 866–873.

13. **Bethea, B. T.**, **A. M. Okamura**, **M. Kitagawa**, **T. P. Fitton**, **S. M. Cattaneo**, **V. L. Gott**, **W. A. Baumgartner**, and **D. D. Yuh** (2004). Application of Haptic Feedback to Robotic Surgery. *Journal of Laparoendoscopic & Advanced Surgical Techniques*, **14**(3), 191–195.

14. **Billiar, T.**, **D. Andersen**, **J. Hunter**, **F. Brunicardi**, **D. Dunn**, **R. E. Pollock**, and **J. Matthews** (2009). *Schwartz's principles of surgery*. McGraw-Hill Professional.

15. **Biocca, F.** and **B. Delaney** (1995). *Immersive Virtual Reality Technology*, 57–124. L. Erlbaum Associates Inc., USA. ISBN 0805815503. doi:10.5555/207922.207926.

16. **Blanch, R.**, **Y. Guiard**, and **M. Beaudouin-Lafon** (2004). Semantic Pointing: Improving Target Acquisition With Control-Display Ratio Adaptation. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04. ACM, New York, NY, USA. ISBN 1-58113-702-8, doi:10.1145/985692.985758.

17. **Boff, K. R.**, **L. Kaufman**, and **J. P. Thomas** (1986). *Handbook of Perception and Human Performance, Vol. 2: Cognitive Processes and Performance.*. Wiley New York. ISBN 0-471-82956-0.

18. **Boff, K. R.** and **J. E. Lincoln** (1988). Engineering Data Compendium. Human Perception and Performance, Volume 3. *Wright Patterson Air Force Base Ohio*, **72**, 863.

19. **Borrego, A.**, **J. Latorre**, **M. Alcaniz**, and **R. Llorens** (2018). Comparison of Oculus Rift and HTC Vive: Feasibility for Virtual Reality-Based Exploration, Navigation, Exergaming, and Rehabilitation. *Games for health journal*.

20. **Bradski, G. R.** (1998). Computer Vision Face Tracking for Use in a Perceptual User Interface.

21. **Browning, G.** and **R. J. Teather** (2014). Screen Scaling: Effects of Screen Scale on Moving Target Selection. *In CHI'14 Extended Abstracts on Human Factors in Computing Systems*. ACM.

22. **Cai, Y.**, **L. Ge**, **J. Liu**, **J. Cai**, **T.-J. Cham**, **J. Yuan**, and **N. M. Thalmann** (2019). Exploiting Spatial-Temporal Relationships for 3D Pose Estimation via Graph Convolutional Networks. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*.

23. **Cameron, C. R.**, **L. W. DiValentin**, **R. Manaktala**, **A. C. McElhaney**, **C. H. Nostrand**, **O. J. Quinlan**, **L. N. Sharpe**, **A. C. Slagle**, **C. D. Wood**, **Y. Y. Zheng**, *et al.* (2011). Hand Tracking and Visualization in a Virtual Reality Simulation. *In 2011 IEEE systems and information engineering design symposium*. IEEE.

24. **Card, S. K.**, **W. K. English**, and **B. J. Burr** (1978). Evaluation of Mouse, Rate-Controlled Isometric Joystick, Step Keys, and Text Keys for Text Selection on a CRT. *Ergonomics*, **21**(8), 601–613.

25. **Casiez, G.**, **D. Vogel**, **R. Balakrishnan**, and **A. Cockburn** (2008). The Impact of Control-Display Gain on User Performance in Pointing Tasks. *Human-Computer Interaction*, **23**(3), 215–250, `doi:10.1080/07370020802278163`.

26. **Cha, Y.** and **R. Myung** (2013). Extended Fitts' Law for 3D Pointing Tasks Using 3D Target Arrangements. *International Journal of Industrial Ergonomics*, **43**(4), 350–355. ISSN 01698141, `doi:10.1016/j.ergon.2013.05.005`.

27. **chan Jee, S.** and **M. H. Yun** (2016). An Anthropometric Survey of Korean Hand and Hand Shape Types. *International Journal of Industrial Ergonomics*, **53**, 10–18. ISSN 0169-8141, `doi:https://doi.org/10.1016/j.ergon.2015.10.004`.

28. **Chapuis, O.** and **P. Dragicevic** (2008). Small Targets: Why Are They So Difficult to Acquire. *Laboratoire de Recherche en Informatique, Tech. Rep.*

29. **Chapuis, O.** and **P. Dragicevic** (2011). Effects of Motor Scale, Visual Scale, and Quantization on Small Target Acquisition Difficulty. *ACM Transactions on Computer-Human Interaction (TOCHI)*, **18**(3), 13.

30. **Chen, X.**, **G. Wang**, **H. Guo**, and **C. Zhang** (2020). Pose Guided Structured Region Ensemble Network for Cascaded Hand Pose Estimation. *Neurocomputing*, **395**, 138–149. ISSN 0925-2312, `doi:https://doi.org/10.1016/j.neucom.2018.06.097`.

31. **Chen, X.**, **G. Wang**, **C. Zhang**, **T.-K. Kim**, and **X. Ji** (2018). SHPR-NET: Deep Semantic Hand Pose Regression From Point Clouds. *IEEE Access*, **6**, 43425–43439.

32. **Chen, Y.**, **Z. Tu**, **L. Ge**, **D. Zhang**, **R. Chen**, and **J. Yuan** (2019). SO-HANDNET: Self-Organizing Network for 3D Hand Pose Estimation With Semi-Supervised Learning. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*.

33. **Chen Chen, F.**, **S. Appendino**, **A. Battezzato**, **A. Favetto**, **M. Mousavi**, and **F. Pescarmona** (2013). Constraint Study for a Hand Exoskeleton: Human Hand Kinematics and Dynamics. *Journal of Robotics*, **2013**, 910961. ISSN 1687-9600, `doi:10.1155/2013/910961`.

34. **Cheng, H.**, **J. W. Clymer**, **B. P.-H. Chen**, **B. Sadeghirad**, **N. C. Ferko**, **C. G. Cameron**, and **P. Hinoul** (2018). Prolonged Operative Duration Is Associated With Complications: A Systematic Review and Meta-Analysis. *journal of surgical research*, **229**, 134–144.

35. **Chim, H.** (2017). Hand and Wrist Anatomy and Biomechanics: A Comprehensive Guide. *Plastic and reconstructive surgery*, **140**(4), 865.

36. **Choi, H.**, **G. Moon**, and **K. M. Lee** (2020). Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery From a 2d Human Pose. *In European Conference on Computer Vision*. Springer.

37. **Chun, K.**, **B. Verplank**, **F. Barbagli**, and **K. Salisbury** (2004). Evaluating Haptics and 3D Stereo Displays Using Fitts' Law. *In Proceedings. Second International Conference on Creating, Connecting and Collaborating through Computing*. doi:10.1109/HAVE.2004.1391881.

38. **Cobos, S.**, **M. Ferre**, **M. A. Sanchéz-Urán**, **J. Ortego**, and **C. Peña** (2008). Efficient Human Hand Kinematics for Manipulation Tasks. *In 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*. ISBN 9781424420582. ISSN 02766574, doi:10.1109/IROS.2008.4651053.

39. **Coelho, J.** and **F. Verbeek** (2014). Pointing Task Evaluation of Leap Motion Controller in 3D Virtual Environment. *Creating the Difference, Proceedings of the Chi Sparks 2014 Conference*.

40. **Collins, J. W.**, **P. Dell'Oglio**, **A. J. Hung**, and **N. R. Brook** (2018). The Importance of Technical and Non-Technical Skills in Robotic Surgery Training. *European urology focus*.

41. **Coutrix, C.** and **C. Masclet** (2015). Shape-Change for Zoomable Tuis: Opportunities and Limits of a Resizable Slider. *In IFIP Conference on Human-Computer Interaction*. Springer.

42. **Dang, Q.**, **J. Yin**, **B. Wang**, and **W. Zheng** (2019). Deep Learning Based 2D Human Pose Estimation: A Survey. *Tsinghua Science and Technology*, **24**(6), 663–676. ISSN 18787606, doi:10.26599/TST.2018.9010100.

43. **de La Gorce, M.**, **N. Paragios**, and **D. J. Fleet** (2008). Model-Based Hand Tracking With Texture, Shading and Self-Occlusions. *In 2008 IEEE Conference on Computer Vision and Pattern Recognition*.

44. **Deng, J.**, **W. Dong**, **R. Socher**, **L.-J. Li**, **K. Li**, and **L. Fei-Fei** (2009). Imagenet: A Large-Scale Hierarchical Image Database. *In 2009 IEEE conference on computer vision and pattern recognition*. Ieee.

45. **Deng, X.**, **Y. Zhang**, **S. Yang**, **P. Tan**, **L. Chang**, **Y. Yuan**, and **H. Wang** (2017). Joint Hand Detection and Rotation Estimation Using CNN. *IEEE Transactions on Image Processing*, **27**(4), 1888–1900. ISSN 10577149, doi:10.1109/TIP.2017.2779600.

46. **Deng, X.**, **Y. Zhang**, **S. Yang**, **P. Tan**, **L. Chang**, **Y. Yuan**, and **H. Wang** (2018). Joint Hand Detection and Rotation Estimation Using CNN. *IEEE Transactions on Image Processing*, **27**(4), 1888–1900. ISSN 1941-0042, doi:10.1109/TIP.2017.2779600.

47. **Dibra, E.**, **T. Wolf**, **C. Oztireli**, and **M. Gross** (2017). How to Refine 3D Hand Pose Estimation From Unlabelled Depth Data? *In 2017 International Conference on 3D Vision (3DV)*. IEEE.

48. **Dominjon, L.**, **S. Richir**, **J. Burkhardt**, **P. Richard**, and **A. Lecuyer** (2006). Influence of Color/Display Ratio on the Perception of Mass of Manipulated Objects in

Virtual Environments. *In Proceedings of the 2005 IEEE Conference 2005 on Virtual Reality*, VR '05. IEEE Computer Society, Washington, DC, USA. ISBN 0-7803-8929-8, doi:10.1109/vr.2005.49.

49. **Du, K.**, **X. Lin**, **Y. Sun**, and **X. Ma** (2019). Crossinfonet: Multi-Task Information Sharing Based Hand Pose Estimation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

50. **Dubey, A. K.**, **K. Gulabani**, **A. Guwalani**, and **R. Rathi** (2014). Pragmatic Study on HCI Laws. *International Journal of Science, Engineering and Technology Research (IJSETR)*, **3**(12).

51. **Ebrahimi, E.**, **S. V. Babu**, **C. C. Pagano**, and **S. Jörg** (2016). An Empirical Evaluation of Visuo-Haptic Feedback on Physical Reaching Behaviors During 3D Interaction in Real and Immersive Virtual Environments. *ACM Trans. Appl. Percept.*, **13**(4), 19:1–19:21. ISSN 1544-3558, doi:10.1145/2947617.

52. **Egger, J.**, **M. Gall**, **J. Wallner**, **P. Boechat**, **A. Hann**, **X. Li**, **X. Chen**, and **D. Schmalstieg** (2017). HTC Vive Mevislab Integration via OpenVR for Medical Applications. *PloS one*, **12**(3), e0173972.

53. **El Sibai, R.**, **C. A. Jaoude**, and **J. Demerjian** (2017). A New Robust Approach for Real-Time Hand Detection and Gesture Recognition. *2017 International Conference on Computer and Applications (ICCA)*, 18–25, doi:10.1109/COMAPP.2017.8079780.

54. **Fang, Y.**, **K. Wang**, **J. Cheng**, and **H. Lu** (2007). A Real-Time Hand Gesture Recognition Method. *In 2007 IEEE International Conference on Multimedia and Expo*. IEEE.

55. **Ferche, O.**, **A. Moldoveanu**, and **F. Moldoveanu** (2016). Evaluating Lightweight Optical Hand Tracking for Virtual Reality Rehabilitation. *Romanian Journal of Human-Computer Interaction*, **9**(2), 85.

56. **Fitts, P. M.** (1954). The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement. *Journal of Experimental Psychology*, **47**(6), 381–391. ISSN 0022-1015, doi:10.1037/h0055392.

57. **Fu, M. J.**, **A. D. Hershberger**, **K. Sano**, and **M. Cenk Çavuşoglu** (2011). Effect of Visuo-Haptic Co-Location on 3D Fitts' Task Performance. *IEEE International Conference on Intelligent Robots and Systems*, 3460–3467. ISSN 2153-0858, doi:10.1109/IROS.2011.6048296.

58. **Ge, L.**, **Y. Cai**, **J. Weng**, and **J. Yuan** (2018*a*). Hand Pointnet: 3D Hand Pose Estimation Using Point Sets. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

59. **Ge, L.**, **H. Liang**, **J. Yuan**, and **D. Thalmann** (2017). 3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation From Single Depth Images. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

60. **Ge, L.**, **Z. Ren**, and **J. Yuan** (2018*b*). Point-to-Point Regression Pointnet for 3D Hand Pose Estimation. *In Proceedings of the European Conference on Computer Vision (ECCV).*

61. **Gharagozloo, F.**, **M. Margolis**, and **B. Tempesta** (2008). Robot-Assisted Thoracoscopic Lobectomy for Early-Stage Lung Cancer. *The Annals of thoracic surgery*, **85**(6), 1880–1886.

62. **Goldenberg, M. G.**, **J. Y. Lee**, **J. C. Kwong**, **T. P. Grantcharov**, and **A. Costello** (2018). Implementing Assessments of Robot-Assisted Technical Skill In urological education: a systematic review and synthesis of the validity evidence. *BJU international*, **122**(3), 501–519.

63. **Gonzalez-Martinez, J.**, **S. Vadera**, **J. Mullin**, **R. Enatsu**, **A. V. Alexopoulos**, **R. Patwardhan**, **W. Bingaman**, and **I. Najm** (2014). Robot-Assisted Stereotactic Laser Ablation in Medically Intractable Epilepsy: Operative Technique. *Operative Neurosurgery*, **10**(2), 167–173.

64. **Gourishetti, R.**, **J. H. R. Isaac**, and **M. Manivannan** (2018). Passive Probing Perception: Effect of Latency in Visual-Haptic Feedback. *In* **D. Prattichizzo**, **H. Shinoda**, **H. Z. Tan**, **E. Ruffaldi**, and **A. Frisoli** (eds.), *Haptics: Science, Technology, and Applications*. Springer International Publishing, Cham. ISBN 978-3-319-93445-7.

65. **Guo, H.**, **G. Wang**, **X. Chen**, **C. Zhang**, **F. Qiao**, and **H. Yang** (2017). Region Ensemble Network: Improving Convolutional Network for Hand Pose Estimation. *In 2017 IEEE International Conference on Image Processing (ICIP).* IEEE.

66. **Gustus, A.**, **G. Stillfried**, **J. Visser**, **H. Jörntell**, and **P. van der Smagt** (2012). Human Hand Modelling: Kinematics, Dynamics, Applications. *Biological cybernetics*, **106**(11-12), 741–755.

67. **Gutt, C. N.**, **T. Oniu**, **A. Mehrabi**, **A. Kashfi**, **P. Schemmer**, and **M. W. Büchler** (2004). Robot-Assisted Abdominal Surgery. *British journal of surgery*, **91**(11), 1390–1397.

68. **Ha, J. S.** (2014). a Human-Machine Interface Evaluation Method Based on Balancing Principles. *Procedia Engineering*, **69**, 13–19.

69. **Hamer, H.**, **K. Schindler**, **E. Koller-Meier**, and **L. V. Gool** (2009). Tracking a Hand Manipulating an Object. *In 2009 IEEE 12th International Conference on Computer Vision.*

70. **Hansen, J. P.**, **V. Rajanna**, **I. S. MacKenzie**, and **P. Bækgaard** (2018). A Fitts' Law Study of Click and Dwell Interaction by Gaze, Head and Mouse with a Head-mounted Display. *In Proceedings of the Workshop on Communication by Gaze Interaction*, COGAIN '18. ACM, New York, NY, USA. ISBN 978-1-4503-5790-6, doi:10.1145/3206343.3206344.

71. **He, K.**, **X. Zhang**, **S. Ren**, and **J. Sun** (2016). Deep Residual Learning for Image Recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition.*

72. **Held, D.**, **S. Thrun**, and **S. Savarese** (2016). Learning to track at 100 FPS with deep regression networks. *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9905 LNCS. ISBN 9783319464473. ISSN 16113349, `doi:10.1007/978-3-319-46448-0_45`.

73. **Helgason, D.** (2004). Unity Game Engine. `https://unity.com/`.

74. **Hess, R. A.** (1973). Nonadjectival rating scales in human response experiments. *Human Factors*, **15**(3), 275–280, `doi:10.1177/001872087301500311`.

75. **Hochschild, J.** (2015). *Functional Anatomy for Physical Therapists*. Thieme. ISBN 9783131768711.

76. **Holmes, D.**, **D. Charles**, **P. Morrow**, **S. McClean**, and **S. McDonough** (2017). Leap Motion Controller and Oculus Rift Virtual Reality Headset for Upper Arm Stroke Rehabilitation. *Virtual Reality: Recent Advances in Virtual Rehabilitation System Design*, 20–22.

77. **Hung, A. J.**, **J. Chen**, and **I. S. Gill** (2018). Automated Performance Metrics and Machine Learning Algorithms to Measure Surgeon Performance and Anticipate Clinical Outcomes in Robotic Surgery. *JAMA surgery*, **153**(8), 770–771.

78. **Hussein, A. A.**, **K. R. Ghani**, **J. Peabody**, **R. Sarle**, **R. Abaza**, **D. Eun**, **J. Hu**, **M. Fumo**, **B. Lane**, **J. S. Montgomery**, *et al.* (2017). Development and Validation of an Objective Scoring Tool for Robot-Assisted Radical Prostatectomy: Prostatectomy Assessment and Competency Evaluation. *The Journal of urology*, **197**(5), 1237–1244.

79. **Isaac, J. H. R.**, **A. Krishnadas**, **N. Damodaran**, and **M. Manivannan** (2018). Effect of Control Movement Scale on Visual Haptic Interactions. *In* **D. Prattichizzo**, **H. Shinoda**, **H. Z. Tan**, **E. Ruffaldi**, and **A. Frisoli** (eds.), *Haptics: Science, Technology, and Applications*. Springer International Publishing, Cham. ISBN 978-3-319-93445-7.

80. **Isaac, J. H. R.**, **M. Manivannan**, and **B. Ravindran** (2021). Corrective Filter Based on Kinematics of Human Hand for Pose Estimation. *Frontiers in Virtual Reality*, **2**, 92. ISSN 2673-4192, `doi:10.3389/frvir.2021.663618`.

81. **ISO** (2000). Dis 9241-9 ergonomic requirements for office work with visual display terminals (vdts)-part 9: Requirements for non-keyboard input devices. *International Standard, International Organization for Standardization*.

82. **Jaekl, P. M.**, **M. R. Jenkin**, and **L. R. Harris** (2005). Perceiving a Stable World During Active Rotational and Translational Head Movements. *Experimental Brain Research*, **163**(3), 388–399. ISSN 1432-1106, `doi:10.1007/s00221-004-2191-8`.

83. **Jellinek, H. D.** and **S. K. Card** (1990). Powermice and User Performance. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, number April in CHI '90. ACM, New York, NY, USA. ISBN 0-201-50932-6, `doi:10.1145/97243.97276`.

84. **Johnsgard, T.** (1994). Fitts' Law With a Virtual Reality Glove and a Mouse: Effects of Gain. *Graphics Interface*, 8–8. ISSN 07135424.

85. **Jones, S. A. H.**, **E. K. Cressman**, and **D. Y. P. Henriques** (2010). Proprioceptive Localization of the Left and Right Hands. *In Experimental Brain Research*. ISBN 0022100920798. ISSN 00144819, `doi:10.1007/s00221-009-2079-8`.

86. **Joo, S. I.**, **S. H. Weon**, and **H. I. Choi** (2014). Real-Time Depth-Based Hand Detection and Tracking. *The Scientific World Journal*, **2014**. ISSN 1537744X, `doi:10.1155/2014/284827`.

87. **Jude, A.**, **D. Guinness**, and **G. M. Poor** (2016). Reporting and Visualizing Fitts's Law: Dataset, Tools and Methodologies. *In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16. ACM, New York, NY, USA. ISBN 978-1-4503-4082-3, `doi:10.1145/2851581.2892364`.

88. **Kassite, I.**, **T. Bejan-Angoulvant**, **H. Lardy**, and **A. Binet** (2019). A Systematic Review of the Learning Curve in Robotic Surgery: Range and Heterogeneity. *Surgical endoscopy*, **33**(2), 353–365.

89. **Kehr, P.** and **A. G. Graftiaux** (2017). Hand and Wrist Anatomy and Biomechanics: A Comprehensive Guide. *European Journal of Orthopaedic Surgery & Traumatology*, **27**(7), 1029–1029. ISSN 1432-1068, `doi:10.1007/s00590-017-1991-z`.

90. **Kha Gia Quach**, **Chi Nhan Duong**, **Khoa Luu**, and **T. D. Bui** (2016). Depth-Based 3D Hand Pose Tracking. *In 2016 23rd International Conference on Pattern Recognition (ICPR)*.

91. **Kiaii, B.**, **W. D. Boyd**, **R. Rayman**, **W. Dobkowski**, **S. Ganapathy**, **G. Jablonsky**, and **R. Novick** (2000). Robot-Assisted Computer Enhanced Closed-Chest Coronary Surgery: Preliminary Experience Using a Harmonic Scalpel® and Zeus™. *In Heart Surgery Forum*, volume 3. FORUM MULTIMEDIA PUBLISHING.

92. **Kingma, D. P.** and **J. Ba** (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

93. **Koehn, J. K.** and **K. J. Kuchenbecker** (2015). Surgeons and Non-Surgeons Prefer Haptic Feedback of Instrument Vibrations During Robotic Surgery. *Surgical endoscopy*, **29**(10), 2970–2983.

94. **Kontarinis, D. A.** and **R. D. Howe** (1995). Tactile Display of Vibratory Information in Teleoperation and Virtual Environments. *Presence: Teleoperators & Virtual Environments*, **4**(4), 387–402.

95. **Lambrey, S.**, **I. Viaud-Delmon**, and **A. Berthoz** (2002). Influence of a Sensorimotor Conflict on the Memorization of a Path Traveled in Virtual Reality. *Cognitive Brain Research*, **14**(1), 177–186. ISSN 0926-6410, `doi:https://doi.org/10.1016/S0926-6410(02)00072-1`. Multisensory Proceedings.

96. **Langolf, G. D.**, **D. B. Chaffin**, and **J. A. Foulke** (1976). An Investigation of Fitts' Law Using a Wide Range of Movement Amplitudes. *Journal of Motor Behavior*, **8**(2), 113–128. ISSN 19401027, `doi:10.1080/00222895.1976.10735061`.

97. **Lee, G. I.** and **M. R. Lee** (2018). Can a virtual reality surgical simulation training provide a self-driven and mentor-free skills learning? investigation of the practical influence of the performance metrics from the virtual reality robotic surgery simulator on the skill learning and associated cognitive workloads. *Surgical endoscopy*, **32**(1), 62–72.

98. **Lee, J.**, **M. Sinclair**, **M. Gonzalez-Franco**, **E. Ofek**, and **C. Holz** (2019). Torc: A Virtual Reality Controller for In-Hand High-Dexterity Finger Interaction. *In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

99. **Lee, P.-W.**, **H.-Y. Wang**, **Y.-C. Tung**, **J.-W. Lin**, and **A. Valstar** (2015). Transection: Hand-Based Interaction for Playing a Game Within a Virtual Reality Game. *In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*.

100. **Lefranc, M.** and **J. Peltier** (2016). Evaluation of the Rosa™ Spine Robot for Minimally Invasive Surgical Procedures. *Expert review of medical devices*, **13**(10), 899–906.

101. **Levi, A. D.** (2016). *Goodman's Neurosurgery Oral Board Review*. Oxford University Press, Oxford, UK. ISBN 9780190636937. URL `https://oxfordmedicine.com/view/10.1093/med/9780190636937.001.0001/med-9780190636937`.

102. **Li, G.**, **Z. Wu**, **Y. Liu**, **H. Zhang**, **Y. Nie**, and **A. Mao** (2021). 3D Hand Reconstruction From a Single Image Based on Biomechanical Constraints. *The Visual Computer*. ISSN 1432-2315, `doi:10.1007/s00371-021-02250-y`.

103. **Li, R.**, **Z. Liu**, and **J. Tan** (2019). A Survey on 3D Hand Pose Estimation: Cameras, Methods, and Datasets. *Pattern Recognition*, **93**, 251–272. ISSN 0031-3203, `doi:https://doi.org/10.1016/j.patcog.2019.04.026`.

104. **Li, S.** and **D. Lee** (2019). Point-to-pose voting based hand pose estimation using residual permutation equivariant layer. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **2019-June**, 11919–11928. ISSN 10636919, `doi:10.1109/CVPR.2019.01220`.

105. **Linfante, I.** and **A. K. Wakhloo** (2007). Brain aneurysms and arteriovenous malformations: advancements and emerging treatments in endovascular embolization. *Stroke*, **38**(4), 1411–1417.

106. **Lyubanenko, V.**, **T. Kuronen**, **T. Eerola**, **L. Lensu**, **H. Kälviäinen**, and **J. Häkkinen** (2017). Multi-camera finger tracking and 3D trajectory reconstruction for HCI studies. *In International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer.

107. **MacKenzie, I. S.** (1992). Fitts' law as a research and design tool in human-computer interaction. *Human-computer interaction*, **7**(1), 91–139. ISSN 0737-0024, `doi:10.1207/s15327051hci0701_3`.

108. **MacKenzie, I. S.** and **P. Isokoski** (2008). Fitts' Throughput and the Speed-Accuracy Tradeoff. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

109. **MacKenzie, I. S.** and **S. Riddersma** (1994). Effects of Output Display and Control-Display Gain on Human Performance in Interactive Systems. *Behaviour and Information Technology*, **13**(5), 328–337. ISSN 13623001, `doi:10.1080/01449299408914613`.

110. **Malik, J.**, **A. Elhayek**, **S. Ahmed**, **F. Shafait**, **M. I. Malik**, and **D. Stricker** (2018*a*). 3DAIRSIG: A Framework for Enabling In-Air Signatures Using a Multi-Modal Depth Sensor. *Sensors*, **18**(11), 3872.

111. **Malik, J.**, **A. Elhayek**, and **D. Stricker** (2018*b*). *Structure-Aware 3D Hand Pose Regression from a Single Depth Image*, volume 2. Springer International Publishing. ISBN 978-3-030-01789-7, `doi:10.1007/978-3-030-01790-3`.

112. **Manivannan, M.**, **R. Periyasamy**, and **V. Narayanamurthy** (2009). Vibration Perception Threshold and the Law of Mobility in Diabetic Mellitus Patients. *Primary Care Diabetes*, **3**(1), 17–21. ISSN 1751-9918, `doi:https://doi.org/10.1016/j.pcd.2008.10.006`.

113. **Mazur, T.**, **T. R. Mansour**, **L. Mugge**, and **A. Medhkour** (2018). Virtual Reality–Based Simulators for Cranial Tumor Surgery: A Systematic Review. *World neurosurgery*, **110**, 414–422.

114. **McGuire, L. M. M.** and **P. N. Sabes** (2009). Sensory Transformations and the Use of Multiple Reference Frames for Reach Planning. *Nature Neuroscience*. ISSN 10976256, `doi:10.1038/nn.2357`.

115. **McMahon, T. A.** (1984). *Muscles, reflexes, and locomotion*. Princeton University Press.

116. **Melax, S.**, **L. Keselman**, and **S. Orsten** (2013). Dynamics Based 3D Skeletal Hand Tracking. *In Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*.

117. **Misra, S.** and **R. H. Laskar** (2017). Multi-Factor Analysis of Texture and Color-Texture Features for Robust Hand Detection in Non-Ideal Conditions. *In Proc. of the 2017 IEEE Region 10 Conference (TENCON)*. ISBN 9781509011346.

118. **Moon, G.**, **J. Yong Chang**, and **K. Mu Lee** (2018). V2V-Posenet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation From a Single Depth Map. *In Proceedings of the IEEE conference on computer vision and pattern Recognition*.

119. **Mueller, F.**, **M. Davis**, **F. Bernard**, **O. Sotnychenko**, **M. Verschoor**, **M. A. Otaduy**, **D. Casas**, and **C. Theobalt** (2019). Real-Time Pose and Shape Reconstruction of Two Interacting Hands With a Single Depth Camera. *ACM Transactions on Graphics (TOG)*, **38**(4).

120. **Mütterlein, J.** (2018). The Three Pillars of Virtual Reality? Investigating the Roles of Immersion, Presence, and Interactivity. *In Proceedings of the 51st Hawaii International Conference on System Sciences*. doi:10.24251/hicss.2018.174.

121. **Naik, G. R.**, **D. K. Kumar**, **V. P. Singh**, and **M. Palaniswami** (2006). Hand Gestures for HCI Using Ica of Emg. *In ACM International Conference Proceeding Series*, volume 237.

122. **Oberweger, M.** and **V. Lepetit** (2017). Deepprior++: Improving fast and accurate 3D hand pose estimation. *In Proceedings of the IEEE international conference on computer vision Workshops*.

123. **Oberweger, M.**, **G. Riegler**, **P. Wohlhart**, and **V. Lepetit** (2016). Efficiently creating 3D training data for fine hand pose estimation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*.

124. **Oikonomidis, I.**, **N. Kyriazis**, and **A. A. Argyros** (2010). Markerless and Efficient 26-Dof Hand Pose Recovery. *In Asian Conference on Computer Vision*. Springer.

125. **Okamura, A. M.** (2004). Methods for Haptic Feedback in Teleoperated Robot-Assisted Surgery. *Industrial Robot: An International Journal*, **31**(6), 499–508.

126. **Okamura, A. M.** (2009). Haptic Feedback in Robot-Assisted Minimally Invasive Surgery. *Current opinion in urology*, **19**(1), 102.

127. **Ozenne, B.**, **F. Subtil**, and **D. Maucort-Boulch** (2015). The Precision–recall Curve Overcame the Optimism of the Receiver Operating Characteristic Curve in Rare Diseases. *Journal of Clinical Epidemiology*, **68**(8), 855–859. ISSN 0895-4356, doi:https://doi.org/10.1016/j.jclinepi.2015.02.010.

128. **Pacchierotti, C.**, **D. Prattichizzo**, and **K. J. Kuchenbecker** (2015). Cutaneous Feedback of Fingertip Deformation and Vibration for Palpation in Robotic Surgery. *IEEE Transactions on Biomedical Engineering*, **63**(2), 278–287.

129. **Palep, J. H.** (2009). Robotic Assisted Minimally Invasive Surgery. *Journal of Minimal Access Surgery*, **5**(1), 1.

130. **Patel, V.** (2005). Robotic-Assisted Laparoscopic Dismembered Pyeloplasty. *Urology*, **66**(1), 45–49.

131. **Pelphrey, K. A.**, **J. P. Morris**, **C. R. Michelich**, **T. Allison**, and **G. McCarthy** (2005). Functional Anatomy of Biological Motion Perception in Posterior Temporal Cortex: An Fmri Study of Eye, Mouth and Hand Movements. *Cerebral cortex*, **15**(12), 1866–1876.

132. **Pfeiffer, M.** and **W. Stuerzlinger** (2015). 3D Virtual Hand Pointing With EMS and Vibration Feedback. *In 2015 IEEE Symposium on 3D User Interfaces (3DUI)*. `doi:10.1109/3DUI.2015.7131735`.

133. **Poier, G.**, **M. Opitz**, **D. Schinagl**, and **H. Bischof** (2019). Murauer: Mapping Unlabeled Real Data for Label Austerity. *In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE.

134. **Poier, G.**, **K. Roditakis**, **S. Schulter**, **D. Michel**, **H. Bischof**, and **A. A. Argyros** (2015). Hybrid One-Shot 3D Hand Pose Estimation by Exploiting Uncertainties. *In* **X. Xie**, **M. W. Jones**, and **G. K. L. Tam** (eds.), *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*. BMVA Press, `doi:10.5244/C.29.182`.

135. **Pouget, A.**, **J. C. Ducom**, **J. Torri**, and **D. Bavelier** (2002). Multisensory Spatial Representations in Eye-Centered Coordinates for Reaching. *Cognition*. ISSN 00100277, `doi:10.1016/S0010-0277(01)00163-9`.

136. **Powers, D. M.** (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*.

137. **Prasad, M. R.**, **M. Manivannan**, **G. Manoharan**, and **S. Chandramohan** (2016). Objective Assessment of Laparoscopic Force and Psychomotor Skills in a Novel Virtual Reality-Based Haptic Simulator. *Journal of surgical education*, **73**(5), 858–869.

138. **Qian, Y. Y.** and **R. J. Teather** (2017). The Eyes Don'T Have It: An Empirical Comparison of Head-based and Eye-based Selection in Virtual Reality. *In Proceedings of the 5th Symposium on Spatial User Interaction*, SUI '17. ACM, New York, NY, USA. ISBN 978-1-4503-5486-8, `doi:10.1145/3131277.3132182`.

139. **Rad, M.**, **M. Oberweger**, and **V. Lepetit** (2018). Feature Mapping for Learning Fast and Accurate 3D Pose Inference From Synthetic Images. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

140. **Raghu Prasad, M. S.**, **S. Purswani**, and **M. Manivannan** (2013). *Modeling of Human Hand Force Based Tasks Using Fitts's Law*, 377–386. Springer India, India. ISBN 978-81-322-1050-4. `doi:10.1007/978-81-322-1050-4_30`.

141. **Rizun, P. R.**, **P. B. McBeth**, **D. F. Louw**, and **G. R. Sutherland** (2004). Robot-Assisted Neurosurgery. *In Seminars in laparoscopic surgery*, volume 11. Sage Publications Sage CA: Thousand Oaks, CA.

142. **Romero, J.**, **D. Tzionas**, and **M. J. Black** (2017). Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, **36**(6).

143. **Ross, L. M.** and **E. D. Lamperti** (2006). *Thieme Atlas of Anatomy: General Anatomy and Musculoskeletal System*. Thieme.

144. **Roy, K.**, **A. Mohanty**, and **R. R. Sahay** (2017). Deep Learning Based Hand Detection in Cluttered Environment Using Skin Segmentation. *In IEEE International Conference on Computer Vision Workshops*. ISBN 978-1-5386-1034-3, `doi:10.1109/ICCVW.2017.81`.

145. **Ryf, C.** and **A. Weymann** (1995). The neutral zero method—a principle of measuring joint function. *Injury*, **26**, 1–11.

146. **Sagayam, K. M.** and **D. J. Hemanth** (2017). Hand Posture and Gesture Recognition Techniques for Virtual Reality Applications: A Survey. *Virtual Reality*, **21**(2), 91–107.

147. **Sarlegna, F. R.** and **R. L. Sainburg** (2007). The Effect of Target Modality on Visual and Proprioceptive Contributions to the Control of Movement Distance. *Experimental Brain Research*. ISSN 00144819, `doi:10.1007/s00221-006-0613-5`.

148. **Satava, R. M.**, **D. Stefanidis**, **J. S. Levy**, **R. Smith**, **J. R. Martin**, **S. Monfared**, **L. R. Timsina**, **A. W. Darzi**, **A. Moglia**, **T. C. Brand**, *et al.* (2019). Proving the Effectiveness of the Fundamentals of Robotic Surgery (FRS) Skills Curriculum: A Single-Blinded, Multispecialty, Multi-Institutional Randomized Control Trial. *Annals of surgery*.

149. **Schechter, M.** (1998). Sensitivity, Specificity, and Predictive Value. *In Surgical Research*, 257–269. Springer.

150. **Sears, A.** and **J. A. Jacko** (2007). *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications (Human Factors and Ergonomics Series)*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA. ISBN 0805858709.

151. **Secretary, I. C.** (2012). *9241–411 Ergonomics of human-system interaction–Part 411: Evaluation methods for the design of physical input devices*. Standard, International Organization for Standardization, Geneva, CH.

152. **Shannon, C. E.** (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, **27**(3), 379–423.

153. **Sharp, T.**, **C. Keskin**, **D. Robertson**, **J. Taylor**, **J. Shotton**, **D. Kim**, **C. Rhemann**, **I. Leichter**, **A. Vinnikov**, **Y. Wei**, *et al.* (2015). Accurate, Robust, and Flexible Real-Time Hand Tracking. *In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.

154. **Shen, D.** and **Y. Zhou** (2017). Effect of VR interactive design system on visual feedback and operational experience. *Proceedings - 2017 10th International Symposium on Computational Intelligence and Design, ISCID 2017*, **2**, 357–360, `doi:10.1109/ISCID.2017.190`.

155. **Simon, D.** and **J. R. Boring III** (1990). Sensitivity, Specificity, and Predictive Value. *In Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition*. Butterworths.

156. **Simon, M.**, **K. Amende**, **A. Kraus**, **J. Honer**, **T. Samann**, **H. Kaulbersch**, **S. Milz**, and **H. Michael Gross** (2019). Complexer-Yolo: Real-Time 3D Object Detection and

Tracking on Semantic Point Clouds. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

157. **Simon, T.**, **H. Joo**, **I. Matthews**, and **Y. Sheikh** (2017). Hand Keypoint Detection in Single Images Using Multiview Bootstrapping. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

158. **Slater, M.** and **S. Wilbur** (1997). A Framework for Immersive Virtual Environments (Five): Speculations on the Role of Presence in Virtual Environments. *Presence: Teleoperators & Virtual Environments*, **6**(6), 603–616.

159. **Smith, M. W.**, **J. Sharit**, and **S. J. Czaja** (1999). Aging, Motor Control, and the Performance of Computer Mouse Tasks. *Human Factors*, **41**(3), 389–396, `doi:10.1518/001872099779611102`. PMID: 10665207.

160. **Sober, S. J.** and **P. N. Sabes** (2005). Flexible Strategies for Sensory Integration During Motor Planning. *Nature Neuroscience*. ISSN 10976256, `doi:10.1038/nn1427`.

161. **Soukoreff, R. W.** and **I. S. MacKenzie** (2004). Towards a Standard for Pointing Device Evaluation, Perspectives on 27 Years of Fitts' Law Research in HCI. *International journal of human-computer studies*, **61**(6), 751–789.

162. **Spurr, A.**, **U. Iqbal**, **P. Molchanov**, **O. Hilliges**, and **J. Kautz** (2020). Weakly Supervised 3D Hand Pose Estimation via Biomechanical Constraints. *In* **A. Vedaldi**, **H. Bischof**, **T. Brox**, and **J.-M. Frahm** (eds.), *Computer Vision – ECCV 2020*. Springer International Publishing, Cham. ISBN 978-3-030-58520-4.

163. **Sridhar, S.**, **F. Mueller**, **A. Oulasvirta**, and **C. Theobalt** (2015). Fast and Robust Hand Tracking Using Detection-Guided Optimization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **07-12-June-2015**, 3213–3221. ISSN 10636919, `doi:10.1109/CVPR.2015.7298941`.

164. **Sridhar, S.**, **A. Oulasvirta**, and **C. Theobalt** (2013). Interactive Markerless Articulated Hand Motion Tracking Using RGB and Depth Data. *In Proceedings of the IEEE international conference on computer vision*.

165. **Srinivasan, M. A.** (1995). What Is Haptics? *Laboratory for Human and Machine Haptics: The Touch Lab, Massachusetts Institute of Technology*, 1–11.

166. **Steinicke, F.**, **G. Bruder**, **J. Jerald**, **H. Frenz**, and **M. Lappe** (2008*a*). Analyses of Human Sensitivity to Redirected Walking. *In Proceedings of the 2008 ACM Symposium on Virtual Reality Software and Technology*, VRST '08. ACM, New York, NY, USA. ISBN 978-1-59593-951-7, `doi:10.1145/1450579.1450611`.

167. **Steinicke, F.**, **G. Bruder**, **T. Ropinski**, and **K. Hinrichs** (2008*b*). Moving Towards Generally Applicable Redirected Walking. *In Proceedings of the Virtual Reality International Conference (VRIC)*. IEEE Press.

168. **Stenger, B.**, **P. R. S. Mendonca**, and **R. Cipolla** (2001). Model-Based 3D Tracking of an Articulated Hand. *In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2.

169. **Strauss, G.**, **M. Hofer**, **W. Korb**, **C. Trantakis**, **D. Winkler**, **O. Burgert**, **T. Schulz**, **A. Dietz**, **J. Meixensberger**, and **K. Koulechov** (2006). [accuracy and precision in the evaluation of computer assisted surgical systems. a definition]. *HNO*, **54**, 78–84.

170. **Sun, X.**, **J. Shang**, **S. Liang**, and **Y. Wei** (2017). Compositional Human Pose Regression. *In Proceedings of the IEEE International Conference on Computer Vision*.

171. **Sun, X.**, **Y. Wei**, **S. Liang**, **X. Tang**, and **J. Sun** (2015). Cascaded Hand Pose Regression. *In Proceedings of the IEEE conference on computer vision and pattern recognition*.

172. **Sunil, T.** (2004). Clinical Indicators of Normal Thumb Length in Adults *The Journal of Hand Surgery*, **29**(3), 489–493. ISSN 0363-5023, doi:https://doi.org/10.1016/j.jhsa.2003.12.016.

173. **Tagliasacchi, A.**, **M. Schröder**, **A. Tkach**, **S. Bouaziz**, **M. Botsch**, and **M. Pauly** (2015). Robust Articulated-ICP for Real-Time Hand Tracking. *In Computer Graphics Forum*, volume 34. Wiley Online Library.

174. **Tang, D.**, **H. J. Chang**, **A. Tejani**, and **T.-K. Kim** (2016). Latent Regression Forest: Structured Estimation of 3D Hand Poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**(7), 1374–1387.

175. **Taylor, J.**, **L. Bordeaux**, **T. Cashman**, **B. Corish**, **C. Keskin**, **T. Sharp**, **E. Soto**, **D. Sweeney**, **J. Valentin**, **B. Luff**, *et al.* (2016). Efficient and Precise Interactive Hand Tracking Through Joint, Continuous Optimization of Pose and Correspondences. *ACM Transactions on Graphics (TOG)*, **35**(4), 1–12.

176. **Teather, R. J.** and **W. Stuerzlinger** (2011). Pointing at 3D Targets in a Stereo Head-Tracked Virtual Environment. *In 2011 IEEE Symposium on 3D User Interfaces (3DUI)*. doi:10.1109/3DUI.2011.5759222.

177. **Teather, R. J.**, **W. Stuerzlinger**, and **A. Pavlovych** (2014). Fishtank Fitts. *Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems - CHI EA '14*, 519–522. ISSN 10960384, doi:10.1016/0003-9861(76)90239-3.

178. **Thayananthan, A.**, **B. Stenger**, **P. H. Torr**, and **R. Cipolla** (2003). Learning a Kinematic Prior for Tree-Based Filtering. *In BMVC*, volume 2. Citeseer.

179. **Tompson, J.**, **M. Stein**, **Y. Lecun**, and **K. Perlin** (2014). Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks. *ACM Transactions on Graphics*, **33**.

180. **Torrey, L.** and **J. Shavlik** (2010). Transfer Learning. *In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, 242–264. IGI global.

181. **Valve** (2003). SteamVR. https://www.steamvr.com/.

182. **Van der Meijden, O. A.** and **M. P. Schijven** (2009). The Value of Haptic Feedback in Conventional and Robot-Assisted Minimal Invasive Surgery and Virtual Reality Training: A Current Review. *Surgical endoscopy*, **23**(6), 1180–1190.

183. **Vollmer, J.**, **R. Mencl**, and **H. Mueller** (1999). *Improved laplacian smoothing of noisy surface meshes*, volume 18-3. Wiley Online Library.

184. **Volpe, A.**, **K. Ahmed**, **P. Dasgupta**, **V. Ficarra**, **G. Novara**, **H. van der Poel**, and **A. Mottrie** (2015). Pilot Validation Study of the European Association of Urology Robotic Training Curriculum. *European urology*, **68**(2), 292–299.

185. **Wan, C.**, **T. Probst**, **L. V. Gool**, and **A. Yao** (2018). Dense 3D Regression for Hand Pose Estimation. ISBN 9781538664209. ISSN 10636919, doi:10.1109/CVPR.2018.00540.

186. **Wan, C.**, **T. Probst**, **L. V. Gool**, and **A. Yao** (2019). Self-Supervised 3D Hand Pose Estimation Through Training by Fitting. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

187. **Wang, G.**, **X. Chen**, **H. Guo**, and **C. Zhang** (2018). Region Ensemble Network: Towards Good Practices for Deep 3D Hand Pose Estimation. *Journal of Visual Communication and Image Representation*, **55**, 404–414.

188. **Wang, R. Y.** and **J. Popović** (2009). Real-Time Hand-Tracking With a Color Glove. *ACM transactions on graphics (TOG)*, **28**(3), 1–8.

189. **Wania, C. E.**, **M. E. Atwood**, and **K. W. McCain** (2006). How Do Design and Evaluation Interrelate in HCI Research? *In Proceedings of the 6th conference on Designing Interactive systems*. ACM.

190. **Wedmid, A.**, **E. Llukani**, and **D. I. Lee** (2011). Future Perspectives in Robotic Surgery. *BJU international*, **108**(6b), 1028–1036.

191. **Westebring-van der Putten, E. P.**, **R. H. Goossens**, **J. J. Jakimowicz**, and **J. Dankelman** (2008). Haptics in Minimally Invasive Surgery–a Review. *Minimally Invasive Therapy & Allied Technologies*, **17**(1), 3–16.

192. **Wilson, G.**, **M. McGill**, **M. Jamieson**, **J. R. Williamson**, and **S. A. Brewster** (2018). Object Manipulation in Virtual Reality Under Increasing Levels of Translational Gain. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–13, doi:10.1145/3173574.3173673.

193. **Xie, X.**, **Q. Lin**, **H. Wu**, **G. Narasimham**, **T. P. McNamara**, **J. Rieser**, and **B. Bodenheimer** (2010). A System for Exploring Large Virtual Environments That Combines Scaled Translational Gain and Interventions. *In Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, APGV '10. ACM, New York, NY, USA. ISBN 978-1-4503-0248-7, doi:10.1145/1836248.1836260.

194. **Xiong, F.**, **B. Zhang**, **Y. Xiao**, **Z. Cao**, **T. Yu**, **J. T. Zhou**, and **J. Yuan** (2019). A2J: Anchor-to-joint regression network for 3D articulated pose estimation from a single

depth image. *Proceedings of the IEEE International Conference on Computer Vision*, **2019-October**, 793–802. ISSN 15505499, `doi:10.1109/ICCV.2019.00088`.

195. **Xu, C.** and **L. Cheng** (2013). Efficient Hand Pose Estimation From a Single Depth Image. *In Proceedings of the IEEE international conference on computer vision*.

196. **Yeo, H.-S.**, **B.-G. Lee**, and **H. Lim** (2015). Hand Tracking and Gesture Recognition System for Human-Computer Interaction Using Low-Cost Hardware. *Multimedia Tools and Applications*, **74**(8), 2687–2715.

197. **Yoo, C.-H.**, **S. Ji**, **Y.-G. Shin**, **S.-W. Kim**, and **S.-J. Ko** (2020). Fast and Accurate 3D Hand Pose Estimation via Recurrent Neural Network for Capturing Hand Articulations. *IEEE Access*, **8**, 114010–114019.

198. **Yuan, S.**, **Q. Ye**, **B. Stenger**, **S. Jain**, and **T.-K. Kim** (2017). Bighand2. 2m Benchmark: Hand Pose Dataset and State of the Art Analysis. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

199. **Zeng, X.**, **A. Hedge**, and **F. Guimbretiere** (2012). Fitts' Law in 3D Space with Coordinated Hand Movements. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, **56**(1), 990–994, `doi:10.1177/1071181312561207`.

200. **Zhai, S.** (2004). Characterizing computer input with fitts' law parameters - The information and non-information aspects of pointing. *International Journal of Human Computer Studies*, **61**(6), 791–809. ISSN 10715819, `doi:10.1016/j.ijhcs.2004.09.006`.

201. **Zhai, S.**, **P. Milgram**, and **W. Buxton** (1996). the Influence of Muscle Groups on Performance of Multiple Degree-of-Freedom Input. *In CHI*, volume 96. Citeseer. ISBN 0897917774, `doi:10.1145/238386.238534`.

202. **Zhou, X.**, **Q. Huang**, **X. Sun**, **X. Xue**, and **Y. Wei** (2017). Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach. *In Proceedings of the IEEE International Conference on Computer Vision*.