

# Real-Time Depth-Based Hand Tracking for American Sign Language Recognition

Brandon Thomas Taylor

CMU-HCII-18-103

August 2018

Human-Computer Interaction Institute  
School of Computer Science  
Carnegie Mellon University  
Pittsburgh, PA 15213

**Thesis Committee:**

Daniel Siewiorek, Chair  
Anind K. Dey, Co-Chair  
Roberta Klatzky  
Carolyn Rosé  
Asim Smailagic

*Submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy.*

Copyright © 2018 Brandon Thomas Taylor

This work was supported by the National Science Foundation under grants IIS-1065336 and CNS-1518865 and by the NSF Quality of Life Technology Engineering Research Center under grant EEC-0540865. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and should not be attributed to Carnegie Mellon University or the funding sources.

**Keywords:** Sign Recognition, Hand Tracking

# Abstract

There are estimated to be more than a million Deaf and severely hard of hearing individuals living in the United States. For many of these individuals, American Sign Language (ASL) is their primary means of communication. However, for most day-to-day interactions, native-ASL users must either get by with a mixture of gestures and written communication in a non-native language or seek the assistance of an interpreter. Whereas advances towards automated translation between many other languages have benefited greatly from decades of research into speech recognition and Statistical Machine Translation, ASLs lack of aural and written components have limited exploration into automated translation of ASL.

In this thesis, I focus on work towards recognizing components of American Sign Language in real-time. I first evaluate the suitability of a real-time depth-based generative hand tracking model for estimating ASL handshapes. I then present a study of ASL fingerspelling recognition, in which real-time tracking and classification methods are applied to continuous sign sequences. I will then discuss the future steps needed to expand a real-time fingerspelling recognition to the problem of general ASL recognition.



# Acknowledgements

Thanks to my parents, Tom and Caroline, for always supporting my interests and letting me have the room to make my own decisions. Thanks to my sister, Brooke, for setting high bars for me to surpass.

Thanks to the many educators who have offered encouragement, opportunities, and support throughout my life. Chris Copeland pushed from a young age and showed me there's always more to learn. Lex Blue did math. The staff of Missouri Academy created a wonderful place that was exactly what I needed at the time. No thanks to the recent stewards of the state of Missouri who ensured that future students will not have the opportunities I did.

Thanks to the ionosphere research community. Trevor Garner gave me my first real research opportunity and defended me against the first grumpy old academic who challenged that research. Anthea Coster and the Haystack Observatory opened my eyes to how many research opportunities exist in the world. Hien Vo and the Arecibo Observatory let me take advantage of such opportunities in Puerto Rico.

Thanks to the Media Lab. Mike Bove gave me the chance to spend a very influential two years in the lab. Quinn Smithwick and Jeevan Kalanithi showed me the ropes and helped me start creating things for 'research'. Tak, Kyle, the Manas, Luis, Tai Sakuma, Jim Trankelson, and the Thirsty Ear crew inspired and conspired. Zoz and all the employees of Cannytrophic Design got me thinking about next steps.

Thanks to Masahiko Inami, Takeo Igarashi, and everyone at the Igarashi Design Interface Project. They gave me a chance to live abroad and tolerated my uninspiring contributions while in Japan. Thanks to Woohyoung Lee and everyone at Samsung Electronics for making me feel at home in Korea.

Thanks to everyone at CMU. Dan Siewiorek offered me the chance to come back to school and stuck with me as I took my sweet time figuring out what to focus on. I especially appreciate being given the chance to find my own way. Anind Dey welcomed me into the Ubicomp lab. I don't recall ever leaving a meeting with Anind without feeling better about my research than when I went in. Both Dan and Anind made it clear that they cared more about their students than their students' work.

Thanks to Trohoc. Dan, Tati, David, Anthony, Nikola, Jenny, and Chris were the best group I could hope to suffer with. Thanks to extended trohoc: Vi, Tycho, Annie, Kabir, Ryan, Caitlin, and Shuang for bearing with us. Thanks to many other HCII students who also offered support and advice.

Thanks to JoAnna for inspiring the work I did and tolerating my doing it. Lastly, thanks to Curie and her yet-unnamed sibling for making me realize there is a time to wrap things up.



# Contents

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Why ASL Recognition? . . . . .	1
1.2 Why Now? . . . . .	2
1.3 How to Move Forward? . . . . .	2
1.4 Document Overview . . . . .	2
<b>2 Defining the problem: What is ASL?</b>	<b>5</b>
2.1 Misconceptions . . . . .	5
2.1.1 ASL is Not Pantomime . . . . .	5
2.1.2 ASL is Not English . . . . .	5
2.1.3 ASL is Not Universal . . . . .	6
2.2 Linguistics of ASL . . . . .	6
2.2.1 Parts of Signs . . . . .	6
2.2.2 Constraints in Formalized ASL . . . . .	10
2.2.3 Grammatical Features . . . . .	14
2.3 Representations of ASL . . . . .	16
2.3.1 Glossing . . . . .	16
2.3.2 Transcription Systems . . . . .	16
2.3.3 The Movement-Hold Model . . . . .	19
2.3.4 Translation . . . . .	21
2.4 Variations in ASL . . . . .	21
2.5 ASL Recognition Requirements . . . . .	22
2.5.1 Sensing Requirements . . . . .	22
2.5.2 Sub-lexical Parameter Primes . . . . .	23
<b>3 Approaches to Sign Recognition</b>	<b>27</b>
3.1 Sensor Methods . . . . .	27
3.1.1 Wearable Sensors . . . . .	27
3.1.2 Vision-Based Approaches . . . . .	28
3.1.3 Depth Based Vision Approaches . . . . .	31

3.1.4	Adopted Approach . . . . .	32
3.2	Language Recognition Methods . . . . .	32
3.2.1	Handshape Detection . . . . .	32
3.2.2	Isolated Sign Detection . . . . .	33
3.2.3	Continuous Sign Recognition . . . . .	33
3.2.4	Non-Manual Features . . . . .	34
<b>4</b>	<b>Depth-Based Hand Tracking for Manual Parameters</b>	<b>37</b>
4.1	Real-Time Model Based Hand Tracking . . . . .	37
4.2	Handshape Classification . . . . .	39
4.2.1	Static Handshape Detection . . . . .	39
4.2.2	Depth Based Handshape Classifiers . . . . .	41
4.3	Handshape Study . . . . .	42
4.3.1	Study Methods . . . . .	42
4.3.2	Participant Validation . . . . .	45
4.3.3	Evaluation Criterion . . . . .	48
4.3.4	Tracking Issues . . . . .	52
<b>5</b>	<b>Continuous Sign Recognition</b>	<b>55</b>
5.1	Prior Work . . . . .	56
5.2	Study Methods . . . . .	56
5.2.1	Data Annotation . . . . .	57
5.2.2	Qualitative Evaluation . . . . .	57
5.3	Segmentation . . . . .	57
5.3.1	Movement-Hold Model for ABC signs . . . . .	60
5.3.2	Timing Parameters . . . . .	62
5.3.3	Hold Model Training . . . . .	63
5.3.4	Hold Model Evaluation . . . . .	65
5.4	Letter Classification . . . . .	66
5.4.1	Empirical Model . . . . .	67
5.4.2	Individual Idiosyncrasies . . . . .	67
5.5	Sequence Analysis . . . . .	73
5.5.1	String Analysis . . . . .	73
5.5.2	Optimal Segmentation . . . . .	74
5.5.3	Spell Checking . . . . .	74
5.6	Error Sources . . . . .	75
5.6.1	Addressing Entry Errors . . . . .	76
5.6.2	Addressing Classifier Errors . . . . .	78
<b>6</b>	<b>A Complete System</b>	<b>83</b>
6.1	Two Hand Tracking . . . . .	83
6.1.1	Requirements . . . . .	84
6.1.2	Implementation . . . . .	84
6.1.3	Challenges . . . . .	84



6.2	Face Tracking . . . . .	84
6.2.1	Requirements . . . . .	85
6.2.2	Implementation . . . . .	85
6.2.3	Challenges . . . . .	85
6.3	Body Tracking . . . . .	85
6.3.1	Requirements . . . . .	85
6.3.2	Implementation . . . . .	86
6.3.3	Challenges . . . . .	86
<b>7</b>	<b>Conclusions</b>	<b>87</b>
7.1	Depth-based Tracking offers Advantages . . . . .	87
7.1.1	Real-Time Performance . . . . .	87
7.1.2	Signer Independence . . . . .	88
7.1.3	Improvements . . . . .	88
7.2	Meaningful Measures of Performance . . . . .	88
7.3	Integration and Interfaces Are Under-Explored . . . . .	89
	<b>Appendices</b>	<b>91</b>
<b>A</b>	<b>ASL Handshape Primes</b>	<b>93</b>
<b>B</b>	<b>Fingerspelling Word Lists</b>	<b>107</b>
B.1	Proper Nouns . . . . .	107
B.2	Nouns . . . . .	108
B.3	Non-English Words . . . . .	109
<b>C</b>	<b>Data Collection Guide</b>	<b>111</b>
C.1	Linguistic Goals . . . . .	111
C.1.1	Static vs. Continuous Data . . . . .	111
C.1.2	Variations . . . . .	111
C.2	Datasets and Collection . . . . .	112
C.2.1	Datasets . . . . .	112
C.2.2	Collecting Data . . . . .	112



# List of Figures

2.1	Minimal pair signs distinguished only by handshape . . . . .	7
2.2	Minimal pair signs distinguished only by palm orientation . . . . .	8
2.3	Minimal pair signs distinguished only by location . . . . .	9
2.4	Minimal pair signs distinguished only by movement . . . . .	9
2.5	Minimal pair signs distinguished only by mouth shape . . . . .	10
2.6	Minimal pair signs distinguished only by facial expressions . . . . .	11
2.7	The set of seven unmarked handshapes . . . . .	12
2.8	Diagram of the Symmetry Condition . . . . .	13
2.9	Diagram of the Dominance Condition . . . . .	13
2.10	Examples of adverbs expressed via non-manual features . . . . .	15
2.11	A transcription using Stokoe notation . . . . .	17
2.12	A transcription using SignWriting notation . . . . .	17
2.13	A transcription using HamNoSys notation . . . . .	18
2.14	The sign <b>WEEK</b> . . . . .	19
2.15	Movement epenthesis and hold deletion . . . . .	20
2.16	Movement-Hold Models for the expression ‘Good Idea’ . . . . .	20
3.1	Example hand images from Pugeault dataset . . . . .	31
4.1	A kinematic hand model . . . . .	38
4.2	Geometric hand models from various hand tracking algorithms . . . . .	39
4.3	The ASL alphabet . . . . .	40
4.4	Handshape primes from The ASL Handshape Dictionary . . . . .	43
4.5	Effects of hand model calibration . . . . .	44
4.6	Static ASL alphabet sign classification . . . . .	46
4.7	Renderings of the average handshape poses across participants . . . . .	47
4.8	ASL alphabet naive Bayesian classification results . . . . .	49
4.9	ASL handshape naive Bayesian classification results . . . . .	50
4.10	ASL handshape SVM classification results . . . . .	51
4.11	Distribution of tracking errors across all handshape estimates . . . . .	52
4.12	Examples of hand pose estimates with different tracking errors . . . . .	53
4.13	Classifier performance across tracking error . . . . .	53
5.1	A browser-based tool for letter annotation . . . . .	58
5.2	A browser-based tool for pose evaluation . . . . .	59

5.3	Timing between letter signs . . . . .	62
5.4	Time between letters across trials . . . . .	63
5.5	Movement features for the word ‘Inglewood’. . . . .	64
5.6	Predicted and labeled hold frames for the word ‘Life’ . . . . .	65
5.7	Examples of the variations in handshapes for the sign <b>N</b> . . . . .	66
5.8	Coarticulation effect in the sequence <b>I-L-Y</b> . . . . .	66
5.9	Confusion matrix of signer independent naive Bayesian classifiers from continuous fingerspelling data . . . . .	68
5.10	Observed handshape variations . . . . .	69
5.11	Relationship between qualitative error frequency and hand distance from camera . . . . .	70
5.12	Comparison of double letter sliding and tapping. . . . .	71
5.13	Double <b>Z</b> performed by repetition of <b>Z</b> . . . . .	72
5.14	Double <b>Z</b> performed with a <i>Bent V</i> handshape . . . . .	72
5.15	An example of an orientation error . . . . .	77
5.16	Cumulative tracking error distribution . . . . .	79
5.17	The qualitative error rates per participant per letter . . . . .	80
5.18	A common finger extension tracking error . . . . .	81
5.19	The average letter classification rate across participants for qualitatively correct pose estimates . . . . .	82

# List of Tables

2.1	Non-manual grammatical markers . . . . .	14
2.2	Movement-Hold Model description of <b>WEEK</b> . . . . .	19
2.3	Comparison of location primes . . . . .	24
2.4	Stokoe movement primes . . . . .	25
4.1	The palm orientations of the ASL alphabet letters . . . . .	41
4.2	An overview of recent studies using depth cameras to classify ASL handshapes .	41
5.1	Fingerspelling study participants. . . . .	56
5.2	Movement-Hold Model for the letter <b>C</b> . . . . .	60
5.3	Movement-Hold Model for the sequence <b>C-A-T</b> . . . . .	61
5.4	Movement-Hold Model for the sequence <b>J-I-M</b> . . . . .	61
5.5	Hold predictions. . . . .	65
5.6	String level analysis of fingerspelling letter classifier performance . . . . .	74
5.7	Word prompts in the dictionary . . . . .	75
5.8	String analysis after spellchecking. . . . .	76
5.9	Recording error counts . . . . .	76
5.10	Rates of orientation errors compared to specific letter occurrence rates . . . . .	78
5.11	Comparison of character error rates and tracking error metrics . . . . .	79
5.12	Effect of ‘hidden’ classifiers of classifier accuracy . . . . .	81
6.1	ASL parameter recognition requirements and implementation status . . . . .	83
A.1	A listing of handshape primes recognized by different linguists. . . . .	105



# Glossary

**Chereme** Stokoe's term for an individual unit of a parameter. We will use the term prime.

**Designator** (dez) Stokoe's term for handshape parameters.

**Dominant/Non-dominant** Indication of the preferred (non-preferred) hand for motor tasks. Handedness does not impact meaning in ASL. Which hand is used in the performance of singlehanded signs can be left to the signer. For non-symmetric two-handed signs, dominant/non-dominant is used to distinguish the hands.

**Hand Configuration** A particular positioning of the fingers relative to the hand, commonly used in sign language linguistics literature.

**Hand Pose** A particular positioning of the fingers relative to the hand, commonly used in hand tracking literature.

**Manual/Non-Manual Features** Manual features are aspects of a sign language expressed by articulations of the hand. Non-manual features refer to relevant articulations expressed in any other manner (e.g., facial expressions or postures).

**Marked Handshapes** A set of more complex handshapes that are used in disproportionately few signs.

**Minimal Pair** A pair of signs that differ in only one aspect of their production.

**Parameter** The sublexical components of a sign. The five parameters in ASL are handshape, palm orientation, movement, relative position, and non-manual features.

**Posture** The absolute orientation of the hand. Posture is a superset of all possible orientations, of which there are a subset of postures recognized as meaningful Palm Orientation primes.

**Prime** A discrete member of a set which form the meaningful parameters of ASL. Referred to as a chereme in Stokoe's work.

**Signation** (sig) Stokoe's term for movement parameters.

**Tabula** (tab) Stokoe's term for location parameters.

**Unmarked Handshapes** A set of 7 handshapes, (/5, /B, /I or /G, /A, /S, /O, /C) recognized as the most natural or basic poses. The unmarked handshapes are used with a disproportionate frequency in forming signs.





# Chapter 1

## Introduction

### 1.1 Why ASL Recognition?

American Sign Language (ASL) is the primary language of an estimated 500,000 Deaf and hard of hearing individuals throughout North America [9]. For these individuals, most communication with the wider world necessarily takes place either through a third party interpreter or some form of written English. Neither solution is ideal. On the one hand, interpreters are costly, often scarce, and usually require advanced scheduling. Managing tasks with written English, on the other hand, can be tedious and error prone, as Deaf individuals in the United States typically achieve only a third or fourth grade English reading level by the age of 18 [100]. There simply is no good solution to help Deaf individuals to engage in non-essential or spontaneous communication with the wider non-signing population.

With the growing popularity of natural speech recognition interfaces and automatic machine translation services, it is a natural question to wonder why no technology exists to facilitate ASL to English translation. One major roadblock to automatically translating sign languages is that they do not have natural written forms. With written languages, one can compare literally millions of documents that have been written in one language and already translated into another. Computers can be trained to learn patterns between synonymous texts and infer translations. With ASL, the only texts that exist are transcriptions created by linguists to study the language. Before machine translation can begin in earnest, there needs to be a systematic way to generate representations of ASL. Automatic Sign Recognition (ASR) seeks to do just that.

Research into ASR dates back to at least the mid-nineties [89], but to date, no real-time recognition system that works beyond a very constrained vocabulary has been demonstrated. Glove-based approaches have been met skeptically by the deaf community, while accurate, real-time hand tracking has proven to be a very difficult problem for computer vision [110]. While there has been some examination of non-manual features of ASL, without the crucial real-time hand pose estimates, little progress has been possible.

## 1.2 Why Now?

The past few years have seen rapid developments in depth camera technologies. Since the widespread commercial availability of the Kinect in 2010, numerous depth cameras with a variety of spatial and temporal resolutions have been released. Not entirely coincidentally, there has been an emergence of commercially available virtual reality (VR) and augmented reality (AR) devices. As companies have made investments in AR and VR, there has been ongoing research into real-time hand tracking as a natural interface for these platforms.

While ASL (particularly the alphabet) often gets co-opted as a test case for research into hand pose classification, depth cameras have not been widely adopted into ASR research. Most existing sign corpuses have been recorded with standard RGB-video [61] or leveraged sign video recording from television broadcasts [27]. Analyzing and annotating sign data is a time consuming task which requires language-specific expertise. There is an existing inertia and substantial value in maintaining consistency in the corpuses that do exist, so the adoption of new technologies is not without cost. That said, should a new technology offer significant enough improvements in recognition, it would certainly be advantageous to the ASR community to consider widespread adoption in future data collection.

## 1.3 How to Move Forward?

This thesis is an exploration of state of the art in depth camera-based hand tracking applied to the task of real-time ASL recognition. Through this work we will demonstrate the following contributions to the field of ASR:

- A demonstration of the effectiveness of generative-model hand tracking approaches for the real-time recognition of ASL.
- An evaluation of current state-of-the-art depth-based generative hand tracking model fidelity in the task of ASL handshape detection
- The implementation and evaluation of a real-time, signer independent ASL fingerspelling recognition.
- The evaluation of specific system parameters against ASL-task specific outputs to provide guidelines for improved recognition rates for ASL tasks.
- An exploration of edge-cases and limitations of the current technologies with recommendations for future research.

## 1.4 Document Overview

This document is designed to provide sufficient context for the contributions above. Chapter 2 provides an overview of American Sign Language from a linguistic perspective in order to establish the technical requirements for sign language recognition. Chapter 3 reviews previous work in the field of sign recognition and establishes limitations and opportunities for exploration. Novel research contributions begin in Chapter 4 with a study conducted to evaluate the

effectiveness of a depth-based generative-model hand tracking algorithm at distinguishing ASL handshapes. Chapter 5 builds upon that work and applies the hand tracking algorithm in a real-time continuous fingerspelling recognition system. Chapter 6 then discusses modifications to the algorithm necessary to achieve the full set of recognition goals established in Chapter 2. Finally, the document is concluded in Chapter 7.



# Chapter 2

## Defining the problem: What is ASL?

This chapter focuses on explaining the structure of American Sign Language. Any effort to recognize the language ought to begin with an understanding of what the language is. Here we intend to establish the basic requirements that a complete ASL recognition system will necessarily include.

### 2.1 Misconceptions

In order to understand exactly what the challenges are in automatically recognizing ASL, it can be helpful to first dismiss some common misconceptions about ASL.

#### 2.1.1 ASL is Not Pantomime

To a non-signer it may seem as if ASL is essentially indistinguishable from a game of charades. Some signs are iconic and fluent signers very well may switch into using non-ASL gestures in order to be understood by non-signers. However, the language is not mere gesturing. There is consistency at the sign level and it has distinct grammatical structures. Many of these features will be discussed later.

#### 2.1.2 ASL is Not English

ASL is not a gestural coding of the English language. There is not a one-to-one mapping between signs and words. As such, the automatic translation from ASL to English is not just a matter of being able to recognize the signs. Even if one can flawlessly represent a signed sequence, there is an entirely separate step necessary to translate that sequence into grammatically correct English.

To make this matter more confusing, there is a gestural system for coding English known as Signed Exact English. Signed Exact English was created to directly map signs to each word of English, obviating the need for translation. The pedagogical intent and impact of Signed Exact English is beyond the scope of this work, suffice to say it is not the same as ASL. ASL does have borrowed words and is influenced by English, but is a distinct language.

### **2.1.3 ASL is Not Universal**

Across the globe there are 119 recognized sign languages, each with distinct lexicons and grammatical structures [31]. The languages are distinct and incompatible. For example, even British Sign Language and American Sign Language use completely different signs to represent the English alphabet.

The formation of a school for the deaf in Nicaragua in 1977 provided linguists with an opportunity to observe the conditions under which a sign language can arise [82]. Without a critical mass of Deaf individuals to interact with, most individuals will only create a limited set of gestures. However, with exposure at developmentally critical ages, grammatical structures will arise.

Unlike written languages, prior to the advent of film and television broadcasts, signed communication was restricted to face-to-face encounters. It was only with the relatively recent development of video telephony that remote, two-way signed communication was possible. As a result, sign languages tend to be more regionalized than most spoken languages.

ASL itself follows in the tradition of French sign language. A French monk, the Abbé Charles-Michel de l'Épée, made drawings of signs he observed in 18th-century Paris. From this, the first schools for the deaf were formed. The United States' first school for the Deaf was established in 1817 by Thomas Galludet, marking the formalization of American Sign Language.

Even so, ASL was not afforded much respect as a true language until recently. Tensions have existed over the degree to which Deaf education ought to focus on ASL versus English. Whole signing systems such as Signed Exact English were created as a pedagogical tool to provide a way to present English to Deaf individuals [32]. It was only recently that linguists began analyzing ASL and it was afforded wide recognition as a distinct language.

## **2.2 Linguistics of ASL**

Like any language, ASL is not a static set of prescribed rules. The language is in constant flux, evolving and adopting new signs and means of expression. Any language analysis will be necessarily incomplete and unable to account for all extent variations. Sign languages, without a written form and with high degrees of regional variation, are particularly tricky to analyze.

It is only within the past 60 years or so that ASL has truly been recognized as a distinct language. William Stokoe's seminal work examining the grammatical structure of ASL is often cited as the first rigorous effort to understand ASL in a linguistic fashion [91]. Since Stokoe's work, numerous linguists have seriously examined ASL, working to catalog and explain both the sublexical components and syntactical features. This section will focus on a subset of these linguistic features.

### **2.2.1 Parts of Signs**

One of the key aspects of understanding ASL as a language is to recognize the underlying parts of signs, known as parameters. In ASL, five parameters, handshapes, palm orientation, relative position, movement, and non-manual features, combine to uniquely define individual signs.

Much in the same way that words in spoken English can be broken down into a limited discrete set of phonemes, each of the ASL parameters contains a discrete, limited number of units, known as primes or cheremes.

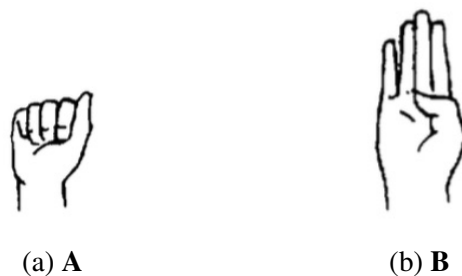
## Handshapes

As the name implies, handshapes refer to the particular configurations in which the fingers of a hand are positioned at a given instance of time. While the various joints of the hands and fingers can be flexed along a continuous range, the handshapes recognized as meaningful in ASL are a limited subset of the physically possible hand configurations. In the same way that different spoken languages can be formed by distinct sets of phonemes, the sets of handshapes used in different sign languages can vary.

In some linguistic work hand configurations are referred to as designators, whereas some fields related to hand tracking will refer to the same concept as the hand's pose. Here we will use *hand configuration* or *pose* to refer to any physically possible arrangement of fingers relative to the hand and *handshape* will specifically refer to a subset of hand configurations which are used in lexical units of signs.

Linguists differ in the count of unique handshape primes that comprise the set of meaningful handshapes in ASL. Stokoe's analysis only defined 19 distinct handshapes, whereas 40-50 handshape primes are recognized by most linguists. Regional variations, lack of formalization, and classification discrepancies across observers account for the range in handshape counts. A very descriptivist approach to handshape recognition recognizes upwards of 80 handshape primes [60].

Figure 2.1: Minimal pair signs distinguished only by handshape



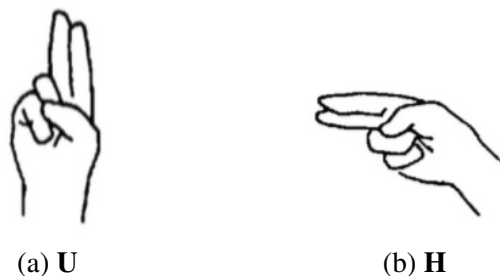
## Palm Orientation

Palm orientation was not indicated as a distinct parameter in Stokoe's notation, though he did recognize that hand posture played an important part in differentiating some signs. Subsequent linguists have sought to define a more robust set of parameters by highlighting signs that form minimal pairs [15, 44]. Minimal pairs are two signs that are only distinguished from one another by a single aspect of their formation. If a descriptive feature is the only distinction between the minimal pair, then that feature must represent a fundamental parameter of the language.

The signs for the letters **U** and **H** form a minimal pair that highlights the fundamental need to recognize the palm orientation. For both signs, the signer's dominant hand is held motionless

in the same position with no particular non-manual features. As can be seen in Figure 2.2, the handshapes are also identical, leaving only the different orientations of the hands to distinguish the signs. Without some method of recognizing the palm orientation, it would be impossible to distinguish **U** from **H**.

Figure 2.2: Minimal pair signs distinguished only by palm orientation



While absolute orientation of the palm varies continuously over three degrees of freedom, small rotational differences are not recognized as meaningful. As with the handshape primes, the exact number of palm orientation primes varies from linguist to linguist. However, accounting for the palm orientation of both dominant and non-dominant hands, there are on the order of 12-18 recognized palm orientation primes [11].

As with handshapes, in this text we will distinguish between references to the general unconstrained orientation of the hand, *hand posture*, and discrete set of postures which are recognized as lexical parameters, *palm orientation*.

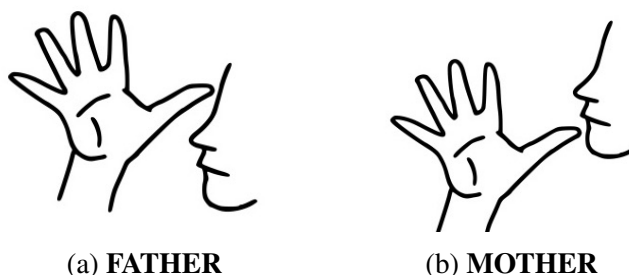
## Location

The location parameter describes where a hand is held in relation to the signer's body. Location can also account for contact points between a hand and another hand or another part of the body. As with palm orientation, the location is impacted less by precise distances and more by the body region the hand is occupying. Thus, changing the location by a few inches when the hand is held away from the body is unlikely to be meaningful, whereas changing contact points on one's face by a few inches can lead to completely different sign (see Figure 2.3).

As with other parameters, different linguists have created different sets of location primes. Stokoe's notation included 12 distinct location (referred to as *tabula*) symbols [91]. Liddell and Johnson took a different approach that broke down the definitions of locations into subgroups that formed what they referred to as *articulatory bundles* [53]. The articulatory bundle would define a particular location as a combination of parameters such as a primary body location and the hand's horizontal and vertical distance from that location. This approach introduces some redundancy to the description, but allows for the observance of variations in sign formation that may contain meaning.



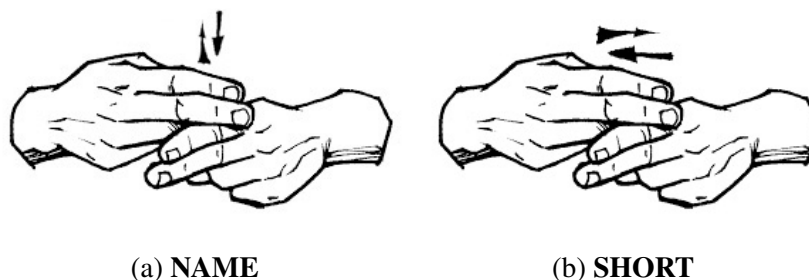
Figure 2.3: Minimal pair signs distinguished only by location



## Movement

Movements, as sub-lexical units (see Figure 2.4), can be split between total hand movements in which the position of the hand is varying and local movements such as rotations or waving fingers. Stokoe's notation provided some 26 annotations, referred to as signation, to describe all the different motions from directional movements to interactions between the hands [91]. In contrast, Liddell and Johnson treated the gross hand movements as a distinctly different aspect of the sign with only three movement types: straight, curved, or sharply angled [53]. In Liddell and Johnson's modeling approach, local finger movements are again included within the underlying articulatory bundles and additional movements can be added sequentially to describe more complex movements.

Figure 2.4: Minimal pair signs distinguished only by movement



However, movement in sign language is not only a phonetic component of the individual signs. How one moves a hand can also encode higher level semantic meaning. For example, to express that an action occurs quickly or slowly, the speed with which the relevant verb is signed is often critical. Thus, for complete ASL recognition, the movement primes alone, are not enough. Additional movement features need to be captured to represent adverbs. How to make a system capable of recognizing both the underlying movement primes necessary to distinguish individual signs (e.g., recognize the sign for 'run') and capable of recognizing specific performative movements that express higher level grammatic features (e.g., distinguish 'run quickly' from 'run slowly') is an open challenge.

## Non-Manual Features

There is some dispute amongst linguists about the exact degree to which non-manual features act as a lexical parameter. Stokoe’s original analysis of ASL did not recognize non-manual features (or orientation) as a parameter [91]. Even as the study of ASL linguistics became more commonplace, there was a resistance to the idea that mouth shapes were a part of the “real ASL”. Some argued that mouthing was just an artifact from previous efforts to insert English into ASL education [71] and others viewed it as merely an optional way of adding emphasis to manual signs [52].

However, non-manual features have generally become an accepted, if often unused, lexical component of the language and are taught as such [88]. For the sake of sign recognition, we can again defer to the existence of minimal pairs of signs distinguished only by mouth shapes and expressions to conclude that non-manual features are a necessary component of a complete ASL recognition system. In Figure 2.5, the signs **LATE** and **NOT-YET** are shown. The only difference between the two signs is the shape of the mouth during the signing. In Figure 2.6, the signs **SHOULD** and **HAVE-TO** are shown. The only difference between the two signs is the facial expression. While non-manual features have importance for other aspects of ASL as well (see Section 2.2.3, these examples demonstrate the necessity of including non-manual feature recognition in any complete ASL sign recognition system.

Figure 2.5: Minimal pair signs distinguished only by mouth shape. Images from LifePrint [103].



(a) **LATE**



(b) **NOT-YET**

### 2.2.2 Constraints in Formalized ASL

While sign languages exhibit a natural evolution, many signs are deliberately introduced. As new signs are introduced, a formalizing process has been observed which constrains the signs.

Figure 2.6: Minimal pair signs distinguished only by facial expressions. Images from LifePrint [103].



(a) **SHOULD**



(b) **HAVE-TO**

Many ASL signs have roots in mimetic gestures. Others have formed around fingerspelling English words. “When signs have changed, they have changed in ways that have made them more conventional in form and thus more arbitrary in meaning” [44].

This formalization process is beneficial for sign recognition as it reduces the feature space in which actual signs reside. For example, using the parameter counts provided by Robin Battison, there are about 45 different handshapes, 25 distinct locations, 12 types of movement and 12 orientations observed in ASL [11]. That is 162,000 ( $45 \cdot 25 \cdot 12 \cdot 12$ ) potential feature combinations. Add a second, independent hand and the possible sign features space exceeds 26 billion ( $162,000 \cdot 162,000$ ) discrete possibilities. And that’s not even accounting for double handshape signs or non-manual features!

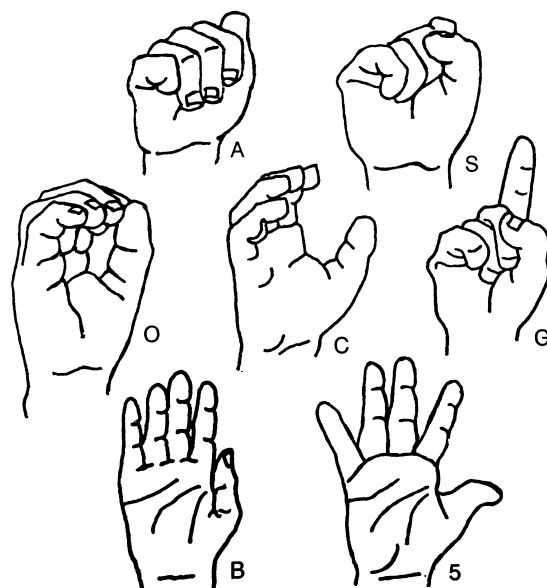
With a cataloged lexicon of approximately 6000 signs [104], such a vast feature space relative to the number of signs presents a number of challenges. Fortunately, signs are not uniformly distributed amongst the parameters and many interactions between the different parameters pare down the feature space of realized signs considerably. This section will discuss a number of observed constraints on ASL signs that make the recognition problem more tractable.

### **Marked and Unmarked Handshapes**

The unmarked handshapes are a set of seven handshapes that are considered to be the most basic handshapes (see Figure 2.7). These particular handshapes have been observed in all known sign languages and are typically among the first handshapes mastered by Deaf children [11]. They also represent a maximally distinct set of geometric shapes (excepting /A and /S, which are similar) and have been observed to exhibit a wider variety in how they contact the body or other

hand [11]. While some linguists have argued that the unmarked set of handshapes should be even more exclusive, there is consistent agreement that a small set of handshapes have out-sized importance in signed languages [34]. Marked handshapes include any handshape that is not one of the seven unmarked handshapes.

Figure 2.7: The set of seven unmarked handshapes.



### Double Handshape Signs

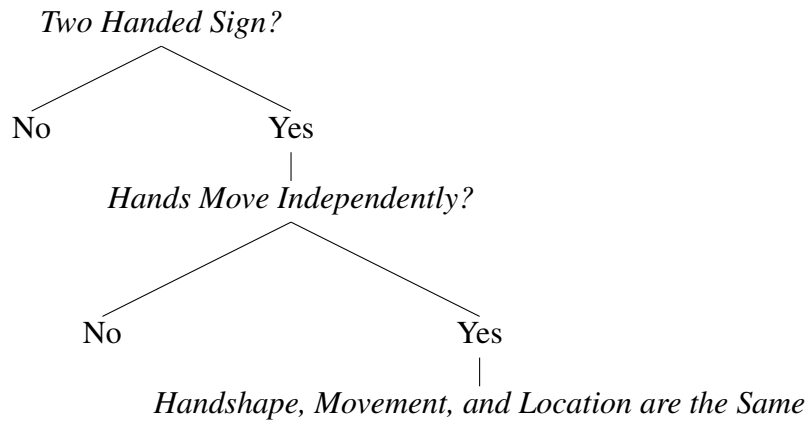
There are a subset of signs, known as double handshape signs, during which a single hand transitions from one handshape to another. It's been observed that in 87.7% of these signs at least one of the handshapes is an unmarked handshape. For a full 63.2% of these signs, both handshapes come from the unmarked set [11].

### Symmetry Condition

The symmetry condition (see Figure 2.8) describes a set of observed constraints that applies to two-handed signs. It was first described by Battison as follows, "If both hands of a sign move independently during its articulation, then both hands must be specified for the same location, the same handshape, the same movement (whether performed simultaneously or in alternation), and the specific orientation must either be symmetrical or identical" [11, p. 22].

This condition greatly reduces the independence of the two hands in ASL and the secondary hand does not result in an exponential increase in the space of 'valid' signs. For the purpose of recognition, two-handed signs with hand movements create a clear correlation between the two handshapes that can be expressed.

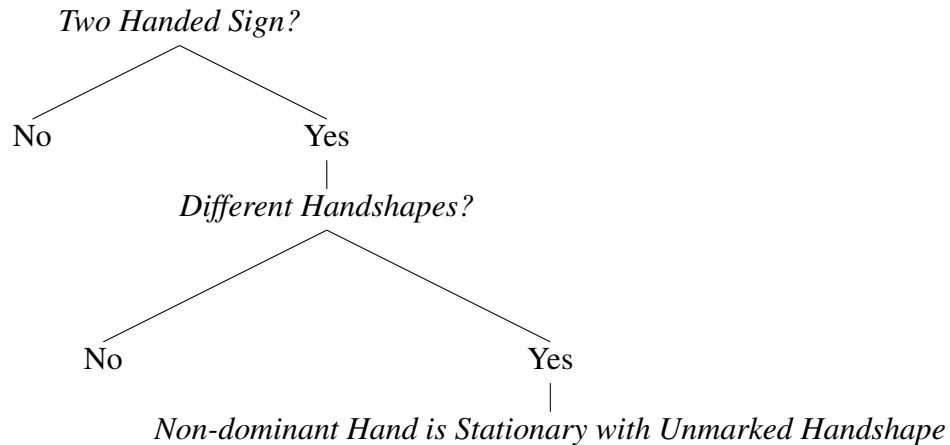
Figure 2.8: Diagram of the Symmetry Condition



### Dominance Condition

The dominance condition (see Figure 2.9) is complimentary to the symmetry condition. Per Battison’s description, “If the hands of a two-handed sign do not share the same specification for handshape (i.e., they are different), then one hand must be passive while the active hand articulates movement AND the specification of the passive hand is restricted to one of [the unmarked handshapes]” [11, p. 23]. That set of signs is the unmarked set described above. Here again, we see a cross-parameter dependence that significantly limits the space of observed ASL signs. In this case, for recognition, a clear indication of distinct handshapes dictates that the non-dominant hand remain stationary.

Figure 2.9: Diagram of the Dominance Condition



### Location and Handshape Interactions

Early observations of ASL signs indicated that marked signs were used with higher frequency in the Head and Neck (33.1% marked handshapes) region than other locations (24.1% marked

handshapes) [11]. It was proposed by Siple that the drop off in visual acuity and the fact that conversing signers fix their gaze on each other’s faces might explain the reduced variety of handshapes away from the head region [87]. Regardless of the reason, the observed interaction between location and handshape could be leveraged by a recognition system to adjust handshape probabilities according to location.

### 2.2.3 Grammatical Features

Individual signs represent the lexical units of ASL. In the same way that words alone do not comprise the entirety of the English language (!?), signs alone do not convey the entirety of meaning in ASL. This section will focus on a few grammatical features and highlight how they differ from the sign components discussed in Section 2.2.1.

#### Facial Signals

Facial signals are used to indicate a variety of grammatical markers. Scott Liddell cataloged a range of non-manual features, many of which had been previously observed [52]. Below, in Table 2.1, is a subset of grammatical topics with brief descriptions taken from his work.

Grammatical Function	Gloss	Expression Properties
Yes-No Questions	q	Brows raised, head forward, body forward
Negation	n	Side-to-side headshake, non-neutral expression
Topic Marker	t	Brow raised, head tilted slightly back
Assertion	hn	Slow head nod
Relative-Clause Marker	r	Brow raised, head tilted backward, cheek and upper lip raised
Adverb (Regularly)	mm	Lips together and push out without puckering, slight head tilt
Adverb (Carelessly)	th	Lips apart and pushed out, tongue protruding

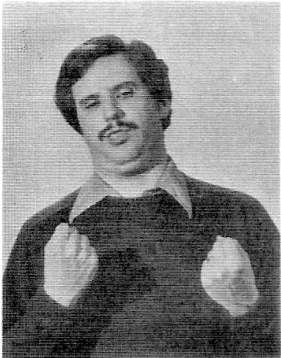
Table 2.1: Non-manual grammatical markers. The gloss symbol represents a convention for indicating grammatical functions when annotating signed sequences.

While some of the grammatical features (e.g., head shaking as negation) are straightforward in their meaning, others need context to understand. Figure 2.10 shows the expressions being described by the adverbs listed in Table 2.1. The important thing to note for the context of ASL recognition is the necessity of capturing such features in order to properly convey the syntax of ASL.

#### Body and Gaze Shifting

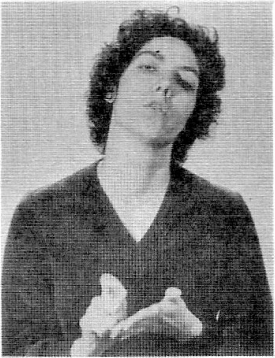
While body postures do play an important role in ASL, they represent grammatical, not lexical, features. That is to sign, body postures do not form a part of the individual signs, but do contribute to the meaning of a signed sequence. For example, to express a conversation in ASL, a signer can use a process called Direct Address [9]. In Direct Address, subjects can be introduced and

Figure 2.10: Examples of adverbs expressed via non-manual features. Images from Baker-Shenk [9].



th  
**DRIVE**

'drive without attending to what's going on'



th  
**WRITE**

'write carelessly'

(a) Carelessly



mm  
**DRIVE**

'drive along regularly'



mm  
**WRITE**

'write at a regular pace'

(b) Regularly

associated with a gaze direction and body position. By shifting between positions, the signer is in effect quoting the subject associated with a particular gaze direction.

## 2.3 Representations of ASL

While most spoken languages have a corresponding written form, there is no natural written form of ASL. While pictorial representations are frequently used, it is no simple matter to illustrate dynamic signs in a clear and concise manner. The need for reproducible representations of signs that do not rely on translated meanings has led to a number of approaches to representing signs.

### 2.3.1 Glossing

Glossing is a method of textually representing the meaning of a sign or sign sequence. It requires prior knowledge of the semantic meaning of signs in order to record them and does not capture the entirety of a signed expression. While this approach can work for record keeping or teaching, it does not provide a way to describe variations in how signs were performed (i.e., two different signs with the same English equivalent would be glossed the same way).

There are a number of conventions used in glossing (see [9] for more details). In this work, as is common, glosses will be presented as bold, capitalized words (see Figure 2.5a). If more than one English word is required to explain a sign, the words are hyphenated (see Figure 2.5b). If a word is fingerspelled, the gloss will be hyphenated between each letter. While not a standard part of glossing, we will represent handshapes using italics and a ‘/’ prefix. Thus, **A** would represent the sign for the English letter ‘A’, whereas */A* represents a specific hand configuration and no more.

### 2.3.2 Transcription Systems

As linguists began to seriously examine sign languages, it became necessary to descriptively annotate signs in a way that could allow for post-hoc analysis of the language. Stokoe notation, developed by William Stokoe, was the first scripting system for textually recording ASL [91]. Stokoe notation uses sets of ordered characters to represent handshapes, hand locations, and movements with subscripts to denote hand orientations. By linking the various sign parameters to characters, Stokoe demonstrated the phonetic structure of ASL and created a method for annotating signs without reliance on semantic meaning.

Since Stokoe’s groundbreaking work, a number of other notations systems have been developed to address various deficiencies in the Stokoe system. Signwriting, for example, was the first annotation system designed to represent the non-manual parameters that were overlooked by Stokoe notation [93]. However, while its spatial layout and iconography can be more visually intuitive to read than more linear representations, the sheer range of symbols renders it far more difficult to write. Designed to work across different Signed Languages, the most recent Unicode standard of Sutton SignWriting contains some 672 unique symbols [85].

The Hamburg Notation System (HamNoSys) was developed as an offshoot of the Stokoe notation with an aim of representing sign languages generally, rather than just ASL [33]. Since



Figure 2.11: An example of Stokoe notation transcribing a telling the story of Goldilocks in ASL. Each grouping of symbols represents an individual sign.

$B_a B_a^{z\sim} \quad \ddot{N} \ddot{N} \dot{a} \cdot \quad 3^\perp \quad [] \quad /C^\dagger /C_X^\vee \cdot \quad \} Y^\ominus \quad /G_\lambda <^{\vee} <$   
 $\bar{B}_a \quad \sqrt{B_\lambda} \quad \psi \quad G^\perp \quad B_\lambda^! B_\lambda^\ddagger \quad D \dot{A}^{\otimes x} \quad \underline{B}_D \quad B_D^\perp$   
 $G^\triangleright \quad \wedge \dot{5}^x \quad [] \quad /C^\dagger /C_X^\vee \cdot \quad X_1 X_1 \dot{a} \quad B_T \quad V_D^\vee \cdot$   
 $\bar{B}_a \quad L \# \cdot \quad X_1 X_1 \dot{a}$

being developed in the mid-80s, HamNoSys, has undergone multiple versions and been the de-facto sign annotation system for a number of substantial investigations of signed languages in Europe [1, 4, 5]. More recently, HamNoSys has been adapted into an XML based format known as Signing Gesture Markup Language [23, 24].

To understand how these transcription systems differ from one another, it can be helpful to compare transcriptions of the same source material. Figures 2.11, 2.12, and 2.13 each show a transcription of a signer beginning to tell the story of Goldilocks and the 3 Bears in Stokoe notation, SignWriting and HamNoSys, respectively. The first line in Figure 2.11 is a description of the signer providing the name of the story. Each grouping of symbols represents the phonetic structure of a different sign (the third symbol is the number 3 in the title). The second line sets the scene of a house somewhere in the woods. In Figure 2.12, SignWriting presents the same breakdown in two columns. In HamNoSys, shown in Figure 2.13, each sign is described on its own line, with facial expressions marked on the right-hand side.

Figure 2.12: The beginning of Goldilocks in SignWriting notation. The first column describes the name of the story and subsequent columns begin the tale. second column setting the opening scene.



Figure 2.13: The same passage of Goldilocks from Figures 2.11 and 2.12 transcribed in HamNoSys. The left column describes one sign per line while the right column indicates non-manual gestures.

Goldilocks & The Three Bears in HamNoSys		Susanne Bentele/10/10/1999
(written for a right handed signer)		
[I had a few difficulties not knowing the ASL citation forms; I might have transcribed unimportant features (movements, locations, etc.). I put facial expressions in a separate column. As of yet there is no standardized way of notating facial expressions; usually the movement of eyebrows or head is included in the movement section with the hands.]		
.. 𐀓𐀔𐀕𐀖𐀗𐀘	what	[ 𐀙 𐀚 ]
.. 𐀓𐀔𐀕 [ 𐀖𐀗𐀘 ]	quote	[ 𐀙 𐀚 ]
𐀓𐀔𐀕	three	[ [ 𐀖𐀗 ] [ 𐀙 𐀚 ] ]
.. 𐀓𐀔𐀕 X 𐀖𐀗𐀘 [ 𐀙 𐀚 ] +	bears	
𐀓𐀔𐀕 25 𐀖𐀗𐀘 [ 𐀙 𐀚 ] [ 𐀛 𐀜 ] [ 𐀝 𐀞 ] [ 𐀟 𐀠 ] [ 𐀡 𐀢 ] [ 𐀣 𐀤 ] [ 𐀥 𐀦 ]	Goldilocks	
𐀓𐀔𐀕 [ 𐀖𐀗𐀘 ]	somewhere wandering	[ 𐀙 𐀚 ]
: 𐀓 [ 𐀔𐀕𐀖 ] [ 𐀗𐀘𐀙 ] [ 𐀚𐀛 ] [ 𐀜𐀝 ] [ 𐀞𐀟 ] [ 𐀠𐀡 ] [ 𐀢𐀣 ] [ 𐀤𐀥 ]	deep forest	[ 𐀙 𐀚 ]
𐀓𐀔𐀕 [ 𐀖𐀗𐀘 ] [ 𐀙 𐀚 ]	somewhere wandering	
𐀓𐀔𐀕 [ 𐀖𐀗 ]	oh! look! there!	[ 𐀙 𐀚 ]
.. 𐀓𐀔𐀕 X 𐀖𐀗𐀘	house	
[ 𐀓𐀔𐀕𐀖𐀗 ] 𐀘 [ 𐀙 𐀚 ]	sitting on a hill	[ 𐀙 𐀚 ]
𐀓 [ 𐀔𐀕𐀖 ] 𐀗 [ 𐀘𐀙 ]	enter	[ 𐀖𐀗 ]
𐀓𐀔𐀕 𐀖	there (index)	[ 𐀙 𐀚 ]
𐀓𐀔𐀕 [ 𐀖𐀗 ] X +	papa	
.. 𐀓𐀔𐀕 X 𐀖𐀗𐀘 [ 𐀙 𐀚 ] +	bear	
.. 𐀓𐀔𐀕 X 𐀖 [ 𐀗𐀘𐀙 ]	open newspaper	[ 𐀙 𐀚 \ 𐀛 ]
[ 𐀓𐀔𐀕𐀖𐀗𐀘 ] 𐀙 [ 𐀚𐀛 ] +	read	[ 𐀙 𐀚 \ 𐀛 ]
[ 𐀓𐀔𐀕𐀖𐀗 ] 𐀘 [ [ 𐀙 X 𐀚 ] 𐀛 ] +	newspaper	
.. 𐀓𐀔𐀕 X 𐀖 [ 𐀗𐀘𐀙 ]	open newspaper	[ 𐀙 𐀚 \ 𐀛 ]

### 2.3.3 The Movement-Hold Model

The Movement-Hold Model was introduced by Liddell and Johnson in an effort to overcome some of the limitations of Stokoe’s parameter-based approach to labeling signs [53]. In particular, Stokoe’s notational approach is limited in describing sequences and transitions between parameters that may occur in a single sign.

As the name implies, the Movement-Hold Model analyzes signs as sequences of hold segments and movement segments. Signs can vary in the number of hold and movement sequences that comprise them and each hold and movement sequence has a set of articulatory features that closely resemble the sign parameters described in Section 2.2.1. In addition to the parameters previously described, the Movement-Hold Model accounts for hold durations, points of contact, local movements and descriptions of both hands.

For an example, consider the sign **WEEK** shown in Figure 2.14. The strong hand (typically the signer’s dominant hand) is held with the index finger extended and slid along the palm of the weak hand. Within the Movement-Hold Model, this sign has three units: An initial hold position, a direct movement, and a final hold position (see Table 2.2). Aspects which are not observed within a particular sign are left blank. While the articulatory features are presented descriptively in Table 2.2, Liddell and Johnson define a very precise taxonomy that can be used articulate features.

Figure 2.14: An illustration of the sign **WEEK**

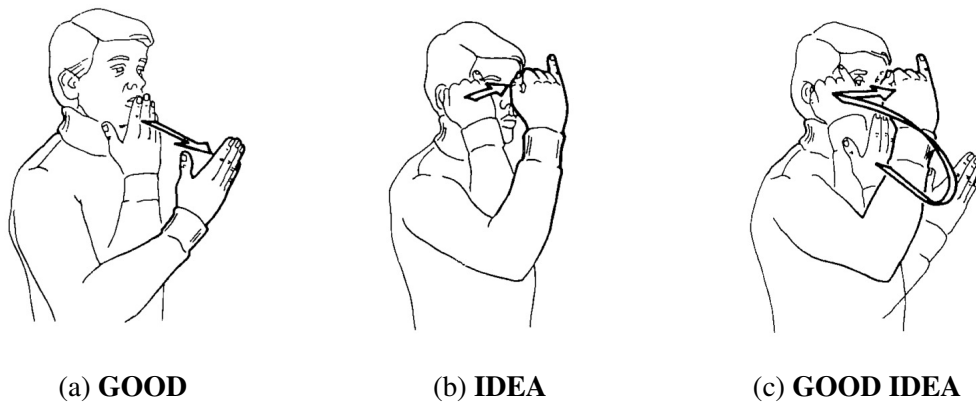


	Timing Unit	Short Hold	Movement	Long Hold
	Contour Contact Local Movement	+	+	+
Strong Hand	Handshape Location Orientation Non-manuals	1 Base of palm of weak hand Palm down		1 Fingertips of weak hand Palm down
Weak Hand	Handshape Location Orientation Non-manuals	B In front of torso Palm up		B In front of torso Palm up

Table 2.2: Movement-Hold Model description of **WEEK**

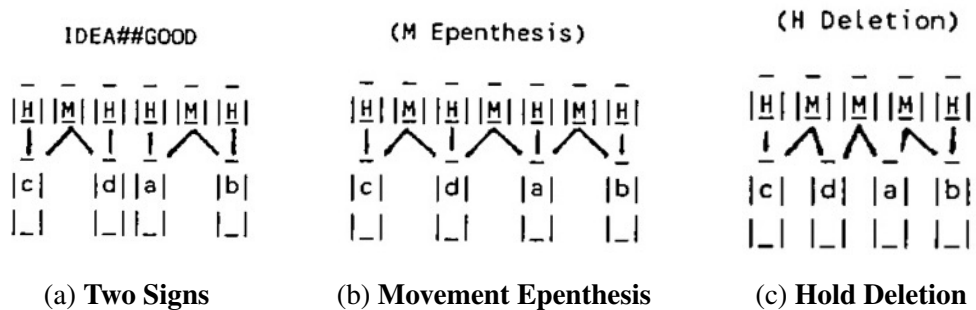
One of the benefits of the Movement-Hold Model is that it can account for continuous sequences of signs and coarticulation issues that arise during sign transitions. Take for example the short sequence **GOOD IDEA**. The individual sign **GOOD** is shown in Figure 2.15a and the individual sign **IDEA** is shown in Figure 2.15b. Both signs are formally performed by beginning in a specific hold, moving, then ending in a second hold (see Figure 2.16a for the models of these signs).

Figure 2.15: Movement epenthesis and hold deletion. The isolated signs **GOOD** and **BAD** are shown in Figures 2.15a and 2.15b. When performed sequentially, the hand must move from the end position of **GOOD** up to the forehead to begin **IDEA** as shown in Figure 2.15c. This additional, unavoidable movement is called a movement epenthesis.



In putting these two signs together to sign **IDEA GOOD**, additional movements are necessarily introduced to transition from the end of **IDEA** to the beginning of **GOOD**. This addition is known as movement epenthesis. Figure 2.15c shows the articulation of this sign sequence with Figure 2.16b highlighting how the model is modified to include this additional movement epenthesis.

Figure 2.16: Movement-Hold Models for the expression ‘Good Idea’



However, in practice, signers rarely articulate signs so completely. Often, rather than clearly articulating the final hold of **GOOD**, signers will immediately begin moving the hand towards the start of the following sign and transition to the second sign’s handshape in a single movement. This is referred to as Hold Deletion and can be modeled as seen in Figure 2.16c. All three

models shown in Figure 2.16 convey the same underlying idea (e.g., all three represent ways to sign **GOOD IDEA**). However, the models correspond to different articulations, with the version shown in Figure 2.16c being the more common articulation compared to the more formally ‘correct’ sequence of Figure 2.16a.

In terms of a sign recognition system, these model variations can be seen as a challenge left for sign translation. The goal of the sign recognition system would be to capture the sign as articulated, whether it be performed as Figure 2.16a or Figure 2.16c. Recognizing that these different articulations embody the same concept would be a challenge left for translation.

### 2.3.4 Translation

Automatic Machine Translation uses machine learning to model grammatical rules at the sentence level. For many written languages, there exist corpuses with millions of synonymous string pairs that can be used to train such models [46]. As a language without a natural written form, the availability of transcribed ASL data is very limited. Most ASL transcription corpuses that do exist focus on narrow contexts such as the weather [27]. The largest of these corpuses contains approximately 10,000 English-ASL sentence pairs [13]. This makes efforts toward automatic translation all the more difficult.

Creating a transcription corpus is a time consuming process. Someone trained in a particular transcription system must observe signed video and manually annotate numerous features. Automatic sign recognition could dramatically simplify the process of transcribing ASL data. If recognition is sufficiently accurate in classifying ASL parameters, it could produce the transcription and require only translated English sentences to provide training for automatic ASL-English translation.

## 2.4 Variations in ASL

It is important to remember that ASL was not created whole cloth with precise rules and regulations. Instead ASL evolved naturally, adopting signs from other languages and formalizing signs in ways that linguists have worked to catalog. However, the purpose of the language has always been to communicate.

Thus, intuitive human visual perception can inform assumptions within the language. People are not great judges of exact distances, so if two signs were identical except that one was held eight inches in front of the signer and the other was held six inches in front, they would often be confused. However, a sign that contacts the inner side of the eyebrow versus one that makes contact at the outer side of the eyebrow would not. Understanding how to convert between measurements provided by sensors into the categories that are meaningful to people is important.

Similarly, minor handshape variations are made less often as the hand is held away from the observer’s focus on the signer’s face (see Section 2.2.2). Instead, signs formed more distally from the signer more often rely on the more limited set of unmarked handshapes. Knowing how features interact within the context of the language is important for combining pieces of ASL recognition into something capable of working at the scale of the language.

Other issues to consider are the context and audience one is designing for. Consider some of the discrepancies reported by different linguists. Are there 40 handshape primes or 80? Well, if there are 80 that have been observed, but only 40 are used in 99% of communication, then for many purposes, 40 will do. Or perhaps, any one individual only uses 40, but individuals from different regions use different sets of 40 handshape primes. In that case, which sets are recognized will greatly impact the usability of the system.

ASL corpuses are, relative to other languages, small. Even the largest corpuses typically have fewer than a dozen signers [27, 61]. While reducing classification errors is important, it's critical to understand how limitations in data availability limit the generalization of classification techniques. This need for signer independent classification techniques has been noted as a research frontier for ASR [19].

Given the limitations of available data, it would be beneficial to not just try to minimize classification errors, but to focus on the types of errors being made. There are studies that can inform how people categorize continuous variations in hand poses into discrete handshape primes [8]. There are other studies of how observers misclassify handshapes in the presence of noise [44]. Designing systems that make errors in ways human observers make errors are more likely to benefit from the natural redundancies in the language.

## 2.5 ASL Recognition Requirements

Given the range of features that convey meaning in ASL, it is understandable that researchers limit the scope of their work to focus on particular aspects of the language. This section is designed to serve as a brief guide to the components of ASL so that researchers can understand the scope.

### 2.5.1 Sensing Requirements

In order to provide functional automatic translation of ASL, the following sensing requirements will be necessary. This listed is intended to be more of a guide to minimal necessary requirements rather than an exhaustive list of sufficient requirements.

- **Real-time Performance** - While offline sign recognition could provide sign transcriptions for automatic translation, any system designed for automatic interpreting would need response times comparable to a live interpreter.
- **Finger-level Resolution** - Hand poses must be recognized with sufficient details to distinguish minor differences in finger positions.
- **Two-hands** - The particular hand poses and movements of both hands, which can act independently and often strongly interact or occlude each other, must be recognized.
- **Full Body Field of View** - The area in which signs are performed, the sign space, typically ranges from the signer's waist and extends no farther than an arm's reach from their body.
- **Facial Features** - Facial features such as mouth shapes, eye gaze, and brow furrowing are necessary for complete lexical and grammatical expression in ASL.

## **2.5.2 Sub-lexical Parameter Primes**

In the 1970's Klima and Bellugi commented that "determining the precise number [of primes] depends on a more complete phonetic-level analysis than is now available and on resolving a number of descriptive problems" [44, p. 22]. While the phonetic-level analysis may be more complete today, disputes about parameter primes persists. Since no conclusive set of primes can be provided, we will provide necessary, if not exhaustive, lists of primes that will need to be included in any system hoping to recognize the entirety of the ASL language.

### **Handshapes Primes**

While Stokoe's original analysis only differentiated between 19 handshapes, most linguists consider 40-50 handshape primes necessary for fully expressing ASL. See Appendix A for a detailed listing of handshape primes provided by different linguists.

### **Palm Orientations**

While palm orientation is clearly an important lexical feature in ASL (see Section 2.2.1), it is not clear how distinctly different orientations must be articulated. While Battison remarks on 12-18 orientations (or combinations of two-handed orientations) begin observed in ASL, the distinct orientations are not enumerated [11, p. 15].

Fortunately, many hand tracking techniques provide an estimate of the hand's absolute orientation (or posture). Even without an articulated list of palm orientation primes, measures of absolute orientation can be used to train proper sign classification.

### **Locations**

Locations in ASL typically express where the hands are in relation to the body. In Table 2.3 we have listed the 12 location primes first designated by William Stokoe [90] and the 20 major body locations recognized by Liddell and Johnson. There is not a direct correspondence between the two sets, so we have grouped them roughly by body part. In addition to the regions shown, Liddell and Johnson further specify the location by indicating the side of the body and whether the contact is made at the top or bottom of the region [52].

### **Movements**

Hand movements have been treated quite differently by different linguists. For example, Liddell and Johnson, describe movements by their trajectory and treat them as a distinct component in their Movement-Hold Model [52]. In Table 2.4 we have presented the movement primes originally annotated by Stokoe.

### **Non-Manuals**

No comprehensive set of meaningful non-manual features has been well defined. At a minimum, the grammatical functions listed in Table 2.1 need to be incorporated into any complete recogni-

	Stokoe		Liddell & Johnson	
Body Part	Symbol	Description	Symbol	Description
<b>Head</b>				
	∅	face, or whole head	BH TH	back of head top of head
	∩	forehead, brow, or upper face	FH	forehead
	⊐	eyes, nose, or mid face	NS	nose
	U	lips, chin, or lower face	MO LP JW CN	mouth lip jaw chin
	3	cheek, temple, ear, or side face	SF CK ER	side of forehead cheek ear
<b>Body</b>				
	∏	neck	NK	neck
	⊐	torso, shoulders, chest, trunk	SH ST CH TR AB	shoulder sternum chest trunk abdomen
<b>Arms</b>				
	∂	non-dominant upper arm	UA	upper arm
	√	non-dominant elbow, forearm	FA	forearm
	α	inside of wrist		
	∂	back of wrist		
<b>Other</b>				
	∅	neutral location	LG	leg

Table 2.3: The location primes specified by Stokoe and Liddell & Johnson

tion system. Additionally, some measure of head orientation, gaze direction and body orientation are necessary.



Stokoe Symbol	Movement Description
D <sup>^</sup>	moving upward
D <sup>v</sup>	moving downward
D <sup>N</sup>	moving up and down
D <sup>&gt;</sup>	to the dominant side
D <sup>&lt;</sup>	to the non-dominant side
D <sup>≈</sup>	side to side
D <sup>T</sup>	toward the signer
D <sup>⊥</sup>	away from the signer
D <sup>↑</sup>	to and fro
D <sup>a</sup>	supinate (turn palm up)
D <sup>b</sup>	pronate (turn palm down)
D <sup>w</sup>	twist wrist back & forth
D <sup>η</sup>	nod hand, bend wrist
D <sup>l</sup>	open up
D <sup>#</sup>	close
D <sup>e</sup>	wriggle fingers
D <sup>@</sup>	circle
D <sup>)</sup>	approach, move together
D <sup>x</sup>	contact, touch
D <sup>‡</sup>	link, grasp
D <sup>+</sup>	cross
D <sup>o</sup>	enter
D <sup>:</sup>	separate
D <sup>'</sup>	exchange positions

Table 2.4: The signifiers, or movement primes, as annotated by Stokoe. In Stokoe's annotation system, the signifiers would be appended as a superscript to the symbol annotating the handshape (or designator) shown in this table as a 'D'.



# Chapter 3

## Approaches to Sign Recognition

Sign language recognition can be considered a complex form of gesture recognition. Sign languages are formalized with conventions that constrain a gesture set. Semantics of the language further constrain sequences of signs to meaningful expressions.

Standard definition video of a signer can be understood by other signers, providing proof that meaningful data about sign language can be adequately captured via video. However, just as speech recognition was not a trivial follow-up to having digital audio recordings, extracting the meaningful features of sign languages has proven far more complex than merely recording them.

In this chapter, we will review the relevant approaches taken for sign language recognition. We will begin with a focus on the different sensor modalities that have been explored and explain the motivation for the sensors used in our studies. We will then shift away from detecting *features* of signs and focus on techniques for *making sense* of signs in the context of the language.

### 3.1 Sensor Methods

There are two distinct sensing approaches which have both been explored in the context of sign recognition. The first, *wearable sensors*, uses some form of active sensing, such as a glove with sensors that detect joint angles, worn directly on the signer's body. The second approach, *vision-based*, relies instead on video feeds and computer vision techniques. More recently, as depth cameras have become more accurate and available, *depth-based vision* approaches have emerged as a subset. Obviously, these approaches are not exclusive and hybrid systems can be developed. In this section, we will highlight some of research that has been published using different sensing technologies.

#### 3.1.1 Wearable Sensors

With Wearable Sensor systems, users are required to wear specialized equipment that directly measures features such as hand locations and finger joint angles. For sign language recognition, active sensing systems typically include a glove device to capture hand articulation, and may or may not extend to full body tracking. The range of signal quality in active sensing systems can vary widely with approaches ranging from low-cost inertial measurement units (IMUs) embed-

ded in prototype gloves to film quality motion capture suits and glove systems that cost tens of thousands of dollars [20].

To date, glove-based sensing approaches, using Cybergloves to measure finger movements and Polhemus magnetic sensors for tracking relative hand locations have achieved the most impressive results in sign recognition. One study focused on isolated sign recognition (i.e., recognizing individually recorded signs isolated from surrounding context) of Chinese Sign Language, and achieved a recognition accuracy of 82.9% over a vocabulary of more than 5000 signs [29]. This work was later extended to apply to continuous sign data by clustering and modeling the transition between signs [26]. The model was trained on data from two participants recording 750 signed sequences containing a total of 4994 signs and then tested on a second set of the same 750 signed sequences. Signs were correctly classified 91.9% of the time in near real-time. It is unclear how well the clustering approach used in the work would apply to a signer-independent scenario or even one in which the precise test sentence were not in the training data. Nonetheless, the results presented on such a large vocabulary set are an encouraging indication that real-time classification approaches can be effective given accurate enough manual tracking.

There are a number of limitations to glove-based approaches. One obvious limitation to the approach described above is the high cost of the sensor systems used. While other researchers have explored similar approaches using lower cost sensors [51], the results were slightly reduced accuracy on a smaller vocabulary set. However, more limiting than the cost factor is the fact that glove-based systems alone offer are incapable of recognizing non-manual features. Without this capability, many aspects of ASL grammar and some sign parameters cannot be detected, meaning the approach cannot scale to recognize the entirety of ASL (see Section 2.2.1 for more details).

However, the most problematic issue of glove-based systems is whether the Deaf community would accept and use such an approach. In recent years, new glove-based systems have garnered significant press attention as ASL translation devices [6, 65]. These have met significant pushback from the Deaf community for overlooking necessary non-manual parameters and not accounting for the Deaf community’s preferences into account. In an open letter to the University of Washington’s Office of News and Information signed by 19 ASL instructors and linguists, the glove-based approach in the SignAloud project [6] was described as “a technological advance that places the burden upon the Deaf person”.

### **3.1.2 Vision-Based Approaches**

Some of the earliest explorations into sign language recognition (SLR) were conducted using computer vision techniques [89, 104]. However, even recent research has struggled to provide a robust, real-time solution that can adequately track handshapes against varying backgrounds and with occlusions [25]. While passive tags or colored gloves have been shown to improve tracking results [14, 106], the goal of tracking unmarked hands across standard video frames with sufficient accuracy for ASL recognition has yet to be achieved.

## Sign Video Corpuses

One reason that hand tracking from standard definition video streams is such an appealing idea is that such an approach could be applied to the largest existing corpuses of sign data. For example, Hamburg University has collected over 500 hours of video, spanning hundreds of individual users in their German Sign Language (Deutsche Gebrdensedsprache or DGS) corpus [1]. For American Sign Language, the largest known corpus, the ASL Lexicon Video Database (ASLLVD) [60, 62] contains video of some 3,300 distinct signs collected by half a dozen native signers. Developing recognition systems that work with existing data corpuses provides an opportunity for training and testing that would otherwise require massive data collection efforts to match.

Of course, collecting video sequences of signs is not the only step necessary for training and testing sign recognition algorithms. The sign sequences must also be labeled to enable machine learning approaches to learn aspects of the signs. While most sign corpuses provide annotation in one way or another, it often focuses on the sign-level meaning rather than underlying sign parameters. Thus, if a sign involves a transition from one handshape to another over the course of one second, the annotation is unlikely provide information about the handshape at the per frame level.

## Discriminative Approaches

Discriminative, or appearance-based classification techniques, work by training classifiers directly on images or regions of images without any intermediate hand modeling. Though most research has focused on handshapes and manual parameters, there are a few examples of discriminative approaches applied to facial expressions as well [58, 69, 72]

The biggest issue with discriminative approaches is the need for massive amounts of training data. Even cropped image regions contain a very large feature space, necessitating more training data. While sign corpuses exist, the amount of annotated data available is still relatively scarce for such training. A single image frame can contain millions of pixels, but an annotated dataset may only have on the order of a thousand labeled training examples [48].

The need for training data is also impacted by the number of output states desired. For example, different palm orientations can make a single handshape appear drastically different in an image. The only way to address is to provide sufficient training data for all combinations of palm orientation and handshape. Similar data requirements are necessary to overcome seemingly minor variations in lighting conditions, skin tones or background environments.

Lastly, training on a per image basis like this also eliminates the value provided by temporal continuity. For signs that involve hand rotation, a discriminative approach may lose sight of distinguishing features and misclassify the handshape depending on the orientation.

Given these factors, most discriminative approaches have been limited to fairly narrow results. In order for these approaches to be more effective, either much more data needs to be recorded and annotated, or the existing sign corpuses need a much finer level of annotation (i.e., labeling parameters at the per image frame level, rather than sign level labels).

The Deep Hand project focuses explicitly on the problem of extracting parameter data from sign level labeled videos [48]. In this work, a convolutional neural network is trained to classify hand configurations on a per frame basis in videos of sign sequences. The training data comes

from three separate corpuses of different sign languages and includes more than 1,000,000 total image frames [28, 50, 57]. However, the handshake labels are derived from sign level annotations. Thus, a video sequence labeled with a specific handshake will certainly contain the labeled handshake, but may also include transitions to other hand configurations depending on the sign performed. The resulting classifier is capable of running in real-time and was able to correctly label handshakes on 59.6% of the 3361 manually labeled handshakes images when the test data were independent of the training data.

While this approach represents an interesting way of making use of existing corpuses and demonstrates that underlying parameters can be recognized across sign languages, it has significant limitations. As mentioned above, the approach used in Deep Hand explicitly trains across variations in palm orientation, thus the results give no indication of the palm orientation. In order to be able to distinguish palm orientation as well, the amount of training data would need to exponentially increase along with the distinct combinations of handshakes and palm orientations used in the language. It is also unclear how well the approach would work on video sequences not recorded in a studio setting. Nevertheless, the work represents the most extensively trained discriminative classifier focusing solely on sign videos to date.

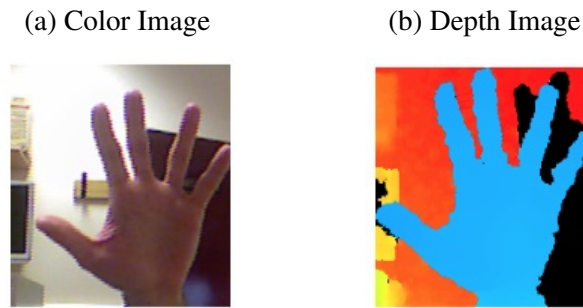
## **Generative Approaches**

In contrast to discriminative approaches, generative or model-based approaches rely on a constrained intermediate model which is fit to an image frame. Generative models are typically designed to run in real-time and make use of assumptions of temporal continuity which can increase reliability. Unlike discriminative approaches, models do not necessarily need extensive training, however, erroneous assumptions in the model (e.g., an incorrectly scaled model) can lead to poor fits. Beyond poor fitting, the typical limitation to generative models is the time it takes to evaluate the potential model space. Simple models can typically be evaluated quickly over a wider range of configurations, whereas complex models may fit data better but take longer to explore the space. Generative models require an initialization method and will often rely on some form of discriminative approach for this step.

Work by Dilsizian et al. explored a generative approach using synthetically rendered hand models to train a mapping between 2D images and 3D hand configurations across a variety of palm orientations and sign motions [21]. This mapping was then used to classify hand shapes across 100 signs taken from videos in the American Sign Language Lexicon Video Dataset (ASLLVD) [62]. The specific sign videos were chosen to include 77 different handshakes and about 40 of the videos involved a transition from one handshake to another. From this test set, the correct handshake was directly recognized 71.02% of the time. By applying linguistic knowledge about handshake transitions and probabilities, the results were boosted to 81.76%.

There is much to appreciate in this work. The classifier is trained on a handshake set large enough to cover the entire set of meaningful ASL handshakes and the testing is performed on samples taken from continuous sign sequences. However, the test set only included 100 signs. The set was designed to focus on spanning the set of handshakes, but it is unclear how the performance may be impacted by different palm orientations or movements in other signs. Additionally, the authors provide no information about processing times, leaving the feasibility of applying this approach in real-time unexamined.

Figure 3.1: Example hand images from the Pugeault dataset [74]



### 3.1.3 Depth Based Vision Approaches

One of the clearest advantages provided by depth cameras is to simplify the process of parsing subjects from the background. With the use case of ASL recognition, signers can be assumed to be fully visible in the frame without any objects between them and the camera. Assumptions about the closest region to the camera can greatly simplify and speed up hand tracking approaches. The availability of RGB feeds with most commercially available depth cameras also ensures that there is no loss in employing Depth cameras instead of standard video.

Over the past decade there has been significant exploration into using depth cameras to track hands [18]. Much of the work has focused on generalized hand tracking, with a priority on real-time processing and dealing with capture from arbitrary camera angles [83, 96].

The first generation Kinect was able to provide native skeletal tracking, but had a limited spatial resolution. This forced a choice between exploring hand or body tracking, as field of view and resolution make simultaneous finger and body tracking difficult. As a result, some researchers focused on signs with distinguishable movement parameters [14], whereas others ignored full body capture and focused solely on how depth data could improve hand tracking. Pugeault and Bowden were early adopters of the Kinect and provided a dataset of depth recordings that have been used to explore a number of classification techniques. Figure 3.1 shows an example color and depth image from Pugeault’s dataset.

Another depth camera, the Leap Motion Controller was introduced in 2012 with the specific goal of providing free-space hand tracking. While the device runs at a higher framerate than the Kinect (up to 200fps), it only has an effective sensing range of around two feet [2]. While there have been some explorations of the Leap’s capability for detecting sign languages [56], significant tracking errors were reported for particular hand rotations and finger poses [73]. Even if software refinements could improve the hand tracking reliability, the device’s limited sensing range severely limits its applicability for general sign recognition.

Since the original Kinect camera was released, many depth cameras have become commercially available offering incremental improvements in temporal and spatial resolutions. With the Kinect V2 or the Intel SR300, for example, researchers no longer need to choose between the field of view necessary for body tracking and the resolution necessary for hand tracking.

Depth cameras have quickly become the equipment of choice for general hand tracking applications [18]. With continued interest in natural hand interactions for augmented and virtual reality, it is reasonable to assume these techniques will continue to improve. However, there is

considerable cost to applying these techniques to ASL data. Depth-based approaches cannot be applied on the standard video corpuses, like ASLLVD. Thus, in order to leverage the improvements in real-time, signer independent hand tracking offered by depth cameras, researchers must conduct new user studies and build new corpuses of depth data.

### **3.1.4 Adopted Approach**

With the recent improvements in depth-camera technology and growing interest in direct hand tracking as an input modality for Augmented and Virtual Reality applications, it seems likely that depth-based hand tracking performance is likely to continue to improve. However, few of these more recently model-based approaches have been applied to the problem of ASL recognition. While such an approach requires the collection of wholly new datasets, the opportunity for real-time tracking is worth exploring. Given potential benefits of complementary IMU measurements from less obtrusive wearable devices such as smart watches, we have opted to explore additional wrist-worn devices as well.

## **3.2 Language Recognition Methods**

One of the challenges facing automatic ASL recognition is that there are many structural levels at which the language can be characterized (e.g., parameters, signs, sentences), but the interrelations between these levels is rarely explored. For example, parameter interactions observed at the sign level (see Section 2.2.2) may be useful in disambiguating handshapes, but require a system that recognizes multiple parameters and includes a model of sign structure. Similarly, sentence level structures such as the establishment of tense at the beginning of a sentence [9] would impact the likelihood of sign detection based on sequencing, but require both sign and sentence level modeling. These interactions between levels of information make it difficult to understand the impact that marginal improvements in one aspect of recognition (e.g., handshape recognition accuracy) have on the ultimate goal of ASL recognition.

To date, ASL recognition research has focused largely on solving pieces of the recognition problem, with little attention placed on putting the pieces together. This section will focus on ASL recognition research from the perspective of the ASL language components that have been explored in prior research. The review will highlight some of the research gaps that motivate some of our subsequent research.

### **3.2.1 Handshape Detection**

Handshape recognition is easily the most commonly explored aspect of ASL recognition research. Much of the research is not actually motivated by an effort to improve ASL recognition, though. Instead, ASL handshapes, particularly those of the 24 static ASL alphabet signs, are frequently used as an evaluation metric for validating a sensor or classification technique aimed at broader hand tracking or hand configuration recognition. Consequently, much of the research that purports to focus on ASL recognition explores only a subset of handshapes, conflates the



handshape and palm orientation parameters, and is rarely applied in the context of continuous signing by fluent signers.

There are a handful of exceptions to this that do seek to recognize handshapes in the context of language recognition. The aforementioned Deep Hand project [48] trained across the complete set of 60 handshapes used in Dutch Sign Language. An approach by Thangali et al., was trained on all 82 handshapes annotated in videos contained in the ASLLVD [61] achieving recognition accuracy of 32.1% [98]. Another approach by Dilsizian achieved 81.8% accuracy across 77 handshape classes using the data from the ASLLVD [21].

### 3.2.2 Isolated Sign Detection

Isolated sign recognition is focused on recognizing individual signs, forming a dictionary-style look up from a set of previously observed signs. While any approach to ASL recognition needs to be able to distinguish isolated signs, evaluations of isolated sign recognition performance do not necessarily imply that an approach is suitable to the broader goal of complete language recognition. Typically, research focusing on isolated sign recognition can only validate an approach for recognizing a specific class of signs over a limited vocabulary. For example, ASL alphabet classifiers are a form of isolated sign recognition that uses only handshape and palm orientation (and movement if **J** and **Z** are included). Just as the accuracy of an alphabet classifier does not indicate whether that approach can be applied generally, isolated sign recognition over a limited vocabulary does not necessarily indicate a scalable solution.

The largest issue with isolated sign recognition is that it does not directly apply to the real world scenario of recognizing natural signed communication. For one thing, sublexical features, such as directions indicated by finger pointing can have specific semantic meaning that is independent of the concurrent sign. Pronouns are often associated with a particular spatial location around the signer and referenced directionally. A system designed around isolated sign look up, may recognize a pronoun, but has no way to distinguish what it is referencing.

Another issue revolves around coarticulation. Just like the way abutting phonemes in speech can impact audible articulations, the precise articulation of a sign is impacted by the preceding and following signs. For a system trained on isolated sign performance, these variations in appearance can cause significant issues in recognition when performed continuously.

### 3.2.3 Continuous Sign Recognition

Recognizing signs performed in a continuous sign sequence is important for a number of reasons. As discussed in Section 2.3.3, coarticulation effects mean that the sequencing of signs can impact their individual articulation. Relying on methods that assume formalized performance of signs may not work in the context of actual sign performance.

#### Segmentation

There have been a few approaches to sign segmentation. Gao et al. explored an approach using dynamic time warping to train explicit recognition of sign transitions. They were able to identify approximately 90% of transitions performed by a single user using a glove-based approach [29].

Vogler and Metaxis were inspired by Liddell and Johnson's Movement-Hold Model to explore the application of hidden Markov models (HMMs) for segmenting continuous sign data [105]. Their work was restricted to a very limited dataset, but offered promising results.

## **Fingerspelling**

Fingerspelling, the act of sequentially signing English letters to spell words, represents a necessary component of American Sign Language (ASL). Fingerspelling offers a constrained set of signs, allowing many aspects of ASL to be disregarded while still providing a meaningful dataset. For example, in ASL all 26 alphabet signs are performed with a single hand and performed in the same location relative to the signers body. Only two alphabet signs (**J** and **Z**) involve movement, the other 24 letters are distinguished solely by varying handshapes (particular hand configurations) and palm orientations (global posture of the hand relative).

In addition to accurately recognizing the 26 alphabet signs, real-time fingerspelling recognition requires the detection of transitions between signs. In theory, any handshape recognition approach that can be calculated sufficiently quickly could be combined with a separate segmentation algorithm. For more detailed discussion of fingerspelling research, refer to Section 5.1.

### **3.2.4 Non-Manual Features**

It has long been noted that research into the automatic recognition of non-manual features has lagged behind work focused on manual recognition [19, 66]. However, as sign recognition research moves forward, non-manual features will need to be recognized. Opportunities to adopt approaches from relevant computer vision fields certainly exist.

## **Body Postures**

While body postures do play an important role in ASL, they represent a grammatical, not a lexical, feature (see Section 2.2.3). Thus to understand the meaning of body postures within the context of ASL, one needs to be able to recognize the concurrent signs. This need for a functioning recognition engine, upon which body posture recognition can be applied, is perhaps why little to no sign recognition research has incorporated body postures [19].

Even so, computer vision research has advanced to a point where existing approaches to skeletal tracking should be applicable within the context of sign recognition. When the Kinect SDK was released, it provided a depth-camera based skeletal tracking API and made depth cameras much more widely available. Since that time, research into depth-based approaches to human skeletal tracking has blossomed [110]. More recently, impressive real-time results have been achieved from standard RGB videos as well [16]. From a technical perspective, there is little reason these approaches could not be applied to sign recognition.

## **Facial Expressions and Mouth Shapes**

Facial expression recognition has received considerable research outside of the context of sign recognition. Broadly speaking, this research can be divided into two primary veins corresponding

to the level at which results are classified. The first approach seeks to recognize higher-level affective meanings of expressions, such as happiness or anger [45]. The other approach focuses on recognizing facial action units which are objectively defined by the underlying movements of the facial muscles [36, 81].

A survey presented by Antonakos shows that applications of facial expression recognition in the context of ASR have largely followed the same divisions, either treating the image region containing the face as a feature or seeking specific geometric measures of specific features such as mouth shape or eyebrow positions [7]. The most extensive studies have relied on the German Phoenix-Weather corpus [27] and focused specifically on detecting particular mouthshapes [47, 80].

There are a few recent examples applied to ASL. Within the context of continuous sign recognition, Prashar demonstrated that including features derived from the principal component analysis of the cropped facial image can improve recognition rates [69]. Nguyen et al., presented an approach to track geometric facial features to classify a set of grammatical six features [63]. Work by Neidle et al., demonstrated that that videos from the ASLLVD [61] could be used to achieve 95% accurate detection of negations and Wh-questions [60]. More recent work by Metaxes focuses largely on the problem of addressing occlusions which often occur in signing, but are often not handled well by facetracking approaches [58].



# Chapter 4

## Depth-Based Hand Tracking for Manual Parameters

Ensuring that each of the five ASL parameters can be detected is a necessary step toward building a system capable of recognizing ASL. Without this capability, no amount of degrees of accuracy or inference will be able to make up for the absence of critical information. From a sensing perspective, there is a significant amount of overlap between many of the parameters. In vision-based hand tracking, for example, palm orientation is a necessary prerequisite to hand pose estimation. Glove-based systems, on the other hand, can directly measure hand poses, but need an external reference to the body in order to measure palm orientation.

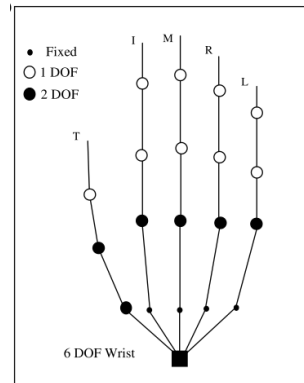
The system developed here, relies on a vision-based approach to hand tracking. This approach provides a camera-centric position and orientation measure of the hand along with an estimate of the hand's pose. As these features are calculated for every video frame, the palm orientation, handshape, and movement parameters are provided by a single algorithm. Separate face and body tracking algorithms, necessary for detecting non-manual parameters, can leverage the same video feed provided a proper camera field of view is available. The combination of hand, face and body positions can then be used to provide Relative Hand positions.

In this chapter, we will focus on the hand tracking algorithm used to capture manual parameters. This will begin with a general discussion of approaches to depth based hand tracking and the specific algorithm used in our system. We will then review previous approaches to handshape classification with a particular focus on prior work using depth data. The next section will describe the study used to validate this approach to manual parameter recognition. Following sections will present results and conclusions drawn from this study.

### 4.1 Real-Time Model Based Hand Tracking

There are many applications, particularly in virtual or augmented reality, where using the hand as a direct input modality is desirable. However, given the hand's high degree of articulation and frequent self-occlusions, accurate, real-time vision-based hand tracking has proven to be a difficult problem for the computer vision community to address [25]. In more recent years, the advent of commercially available depth sensors has fueled significant research into depth-

Figure 4.1: A kinematic or skeletal hand model taken from [25]



based hand gesture recognition [18, 110]. For this work, we will be focusing exclusively on full degree of freedom, model-based hand tracking from depth data. Recent publications in this vein have shown promising real-time tracking results with the potential to scale appropriately for sign recognition [84, 96, 99]. More extensive surveys on vision-based hand tracking [18, 25] and depth-based tracking [110] can be found elsewhere.

The basic approach to full degree of freedom, model-based hand tracking from depth data consists of finding the best match between a measured depth frame and rendered depth map created by a particular pose for a predefined articulated hand model. While specific approaches vary in how they render the hand model, explore the pose space, and evaluate the fit between prediction and measurement, the underlying approach remains similar.

To begin with, an underlying kinematic skeletal model is used to constrain the possible hand poses. Figure 4.1 shows an example of a 27-DOF model, though other variations such as constricting the thumb's MCP joint to a single degree of freedom or adding a wrist joint and forearm model have also been explored. In addition to constraining the kinematics of the hand, physical measurements can be used to place limits on the joint angles [75]. The kinematic model also contains information about the hand size by fixing the distances between joints.

On top of the kinematic skeletal model, a geometric model is attached to allow the rendering of a 3D hand corresponding to a particular pose. The complexity of the geometric model is essentially a trade-off between more accurately rendered hand models, which are more computationally complex and thus slower to render, and more quickly, but less realistically, rendered approximations. Figure 4.2 shows a set of different geometric models. The more realistic hand models, such as the mesh rendering in Figure 4.2b, are more likely to accurately match measured hand data, but quicker renderings, such as the sphere collection in Figure 4.2a allows for broader exploration of potential pose space.

Given a particular model, the basic process for hand tracking is similar across algorithms, though there are various ways to implement each step. The first step is to locate and segment the hand from the rest of the image. While this process can be more complex, constrained use cases where the hand is expected to be the closest object to the camera can allow for approaches as simple as setting a depth threshold which approximates the size of a hand [92]. An estimation of the hand's posture is then made based on prior estimations and/or a single frame pose estimation

Figure 4.2: A collection of geometric hand models from various hand tracking algorithms.



(a) Simple Sphere Model [75] (b) Complex Mesh Model [96] (c) Hybrid Sphere-Mesh Model [99]

used to initialize the algorithm [25]. The pose is then applied to the underlying kinematic model and an estimated depth map is rendered from the geometric model connected to the kinematic model. This rendered depth map is then compared to the measured depth stream and an iterative process updates the pose estimation to better fit the measured depth.

Throughout the process, error measures can represent the difference between the rendered hand model and the measured depth frame. The measures can give an indication of how well an estimated pose matches the underlying data. Minimizing this error is the goal of the iterative fittings steps. The number of iterations possible is typically constrained by the computational complexity of the geometric hand model and the rendering capabilities of the computer. Given the high degree of articulation of the hand, exhaustive searches of hand pose space are not possible in real-time, thus different algorithms use different approaches to explore the pose space.

Despite the interest in real-time hand tracking, most published work has not provided publicly available implementations. The work of Tkach et al. [99] is an exception which has allowed us to develop the system here for testing the applicability of real-time model-based hand tracking for recognizing ASL.

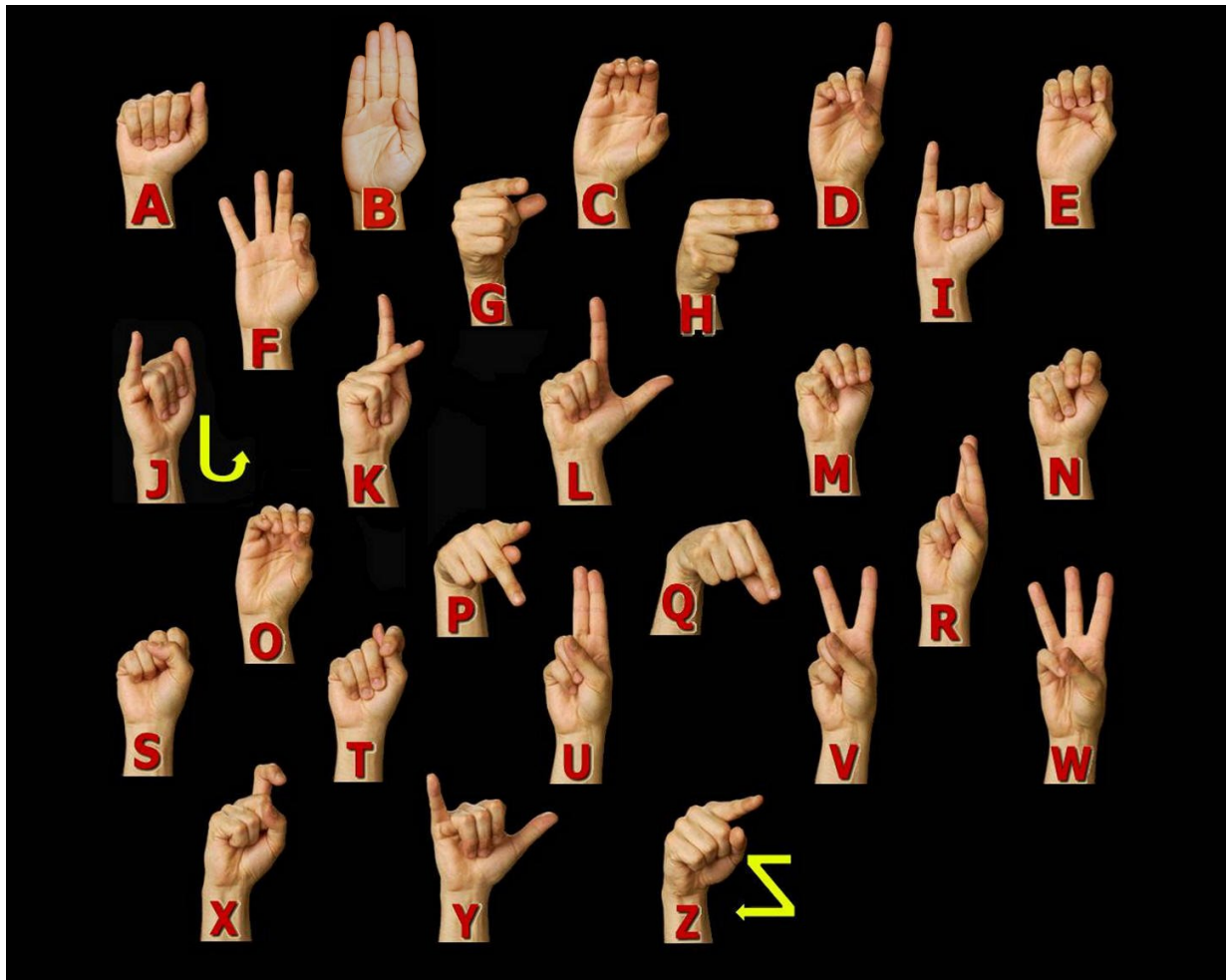
## 4.2 Handshape Classification

Whereas hand tracking seeks to model the hand configuration of the user through a continuous range of possibilities, handshape classification seeks to reduce the possible configurations to a predefined set of possibilities. This can be done using the hand configuration estimates of a hand tracking algorithm or directly from hand images.

### 4.2.1 Static Handshape Detection

While work on hand tracking and pose recognition generalizes beyond the set of handshapes used in ASL (or any sign language), the existing defined sets of poses used in sign languages make sign classification a popular performance metric. Rather than trying to create and explain a new set of hand poses, researchers can point to extant sign dictionaries to describe the target pose.

Figure 4.3: The ASL alphabet



The alphabet signs in ASL provide a particularly compelling test set due to their single-handed formation and the fact that 24 of the 26 letters are performed without any movement (**J** and **Z** being the exceptions). For these reasons, and the straightforward mapping of pose to letter, much of the work evaluating hand tracking and pose classifiers overlaps with a subset of basic ASL handshape recognition.

Unfortunately for the development of sign recognition, the 24 poses used in the static ASL alphabet are not the entire set of meaningful handshapes. In fact, as Table 4.1 shows, the alphabet set is not even a set of 24 unique handshapes. There is disagreement amongst linguists about the exact number of handshape primes used in ASL, with most counts coming in around 40 to 50 [11] and top end estimates ranging up to 80 [62]. Thus, to be able to scale to the entirety of the ASL language, a system needs to recognize at least 40 and potentially as many as 80 unique handshapes.

For researchers primarily interested in general hand poses, though, the ASL alphabet is much simpler to explain and label than the 40 plus handshapes that compose the language. As a re-



sult, many handshape classifiers are trained on 24 classes and make no distinctions about palm orientations.

Palm Orientation	Letter
Forward	A,B,D,E,F,I,K,L,M,N,R,S,T,U,W,X,Y
Right	C,O
In	G (/D), H (/U)
Down	P (/K), Q (/D)

Table 4.1: The palm orientations of the ASL alphabet letters. Most letters are formed with handshapes named for the letter (e.g., the sign for **A** is formed with the handshape /A) held with the palm facing outward. The letters **G**, **H**, **P**, and **Q** are the exceptions and have their corresponding handshapes in parenthesis. See Figure 4.3 for visualizations of the letters.

## 4.2.2 Depth Based Handshape Classifiers

Pugeault and Bowden were early adopters of the Microsoft Kinect for fingerspelling recognition [74]. The dataset used in their study was made available and has been used by a significant number of researchers to explore useful classification techniques [40, 49, 67, 70, 78, 86, 107]. A collection of these and other recent studies [22, 101] that use depth cameras are presented in Table 4.2 to give an indication of the current state of the art.

Of these studies, it's worth drawing attention to the work by Kang [37]. Kang's work collected a new dataset of depth images using the Creative Senz3D camera. The 31 classes for which data was collected include the 24 static alphabet signs (see Table 4.1) and seven numeral digit signs (the signs for **2**, **6**, and **0** were left out as they are redundant with the signs **V**, **W**, and **O**; respectively). To our knowledge, this work represents the most comprehensive depth-based classifier of ASL manual parameters to date. Looking beyond ASL classification, work by Takimoto et al., reported greater than 90% accuracy in classifying a set of 41 Japanese sign language handshapes [95].

First Author/Year	Accuracy	N	Classes
Pugeault 2011	47%	5	24
Pedersoli 2014	56%	5*	24
Kuznetsova 2013	57%	5* + 3	24
Dong 2015	70%	5	24
Uebersax 2011	76%	7	26
Rioux-Maldague 2014	77%	5*	24
Keskin 2012	84%	5*	24
Kang 2015	85%	5	31

Table 4.2: An overview of recent studies using depth cameras to classify ASL handshapes. N refers to the number of participants creating handshapes and \* indicates studies using the Pugeault and Bowden dataset.

There are a handful of studies based on standard video images (with no depth information) that are trained with more classes. A convolutional neural network was trained to detect different handshapes across millions of video frames from a number of different sign language corpuses, reporting up to 62.8% accuracy across 60 classes and multiple signers [48].

For ASL, the ASL Lexicon Video Dataset represents the largest publicly available corpus [61]. By using manually created annotations of hand position (thus eliminating the need for hand detection), one study was able to correctly classify 32.1% of 82 handshape classes from video frames [98]. Another study was able to distinguish 77 handshape classes with 71% accuracy and increase accuracy to 81% by using linguistic knowledge about handshape likelihoods. However, these approaches have not yet been shown to operate with sufficient speed to be employed in a real-time system.

## 4.3 Handshape Study

One advantage offered by generative model-based tracking over discriminative classification approaches (see Section 3.1.2) is that the model estimates provide salient features that can be used beyond handshape classification. For example, renderings of the hand model, which are used in the hand tracking algorithm can also provide an intuitive picture of how well the algorithm is performing. If the tracking is offline or not detecting the user’s hand, it is immediately apparent through the rendering.

The hand pose estimate can also be used for purposes beyond handshape recognition. In ASL, pronouns can be indicated by referencing particular locations in space. With a model of the hand configuration, directionality can easily be determined. Using a discriminative approach, however, no information about directionality is provided. To extract this additional information, an additional recognition method would need to be employed.

However, these advantages are not worth much if the model-based tracking cannot adequately distinguish the relevant handshapes. How well relevant ASL handshapes can be distinguished from the typical results of existing generative model-based hand tracking algorithms has yet to be established.

### 4.3.1 Study Methods

The primary goal of this first study was to measure the effectiveness of currently available, model-based tracking algorithms to capture the entire set of ASL handshape features. To answer this question, we felt it was important to focus not only on the ASL alphabet, but on the entire set of meaningful handshapes used in ASL. While the encompassing set of ASL handshapes is somewhat disputed amongst linguists [44, 102], we chose to include the 40 handshapes defined by the American Sign Language Handshape Dictionary [97] as shown in Figure 4.4. In addition to these 40 handshapes, we additionally recorded samples of the letters **P** and **Q** to allow more direct comparison to previous studies that focused on the ASL alphabet.

We recruited 12 participants for our study. Each participant was briefed on the study procedures and asked to wear a provided yellow wristband on their left wrist. Participants were seated at a desk in front of a computer monitor with an Intel SR300 depth camera placed on top of it.

Figure 4.4: Handshape primes from The ASL Handshape Dictionary [97]

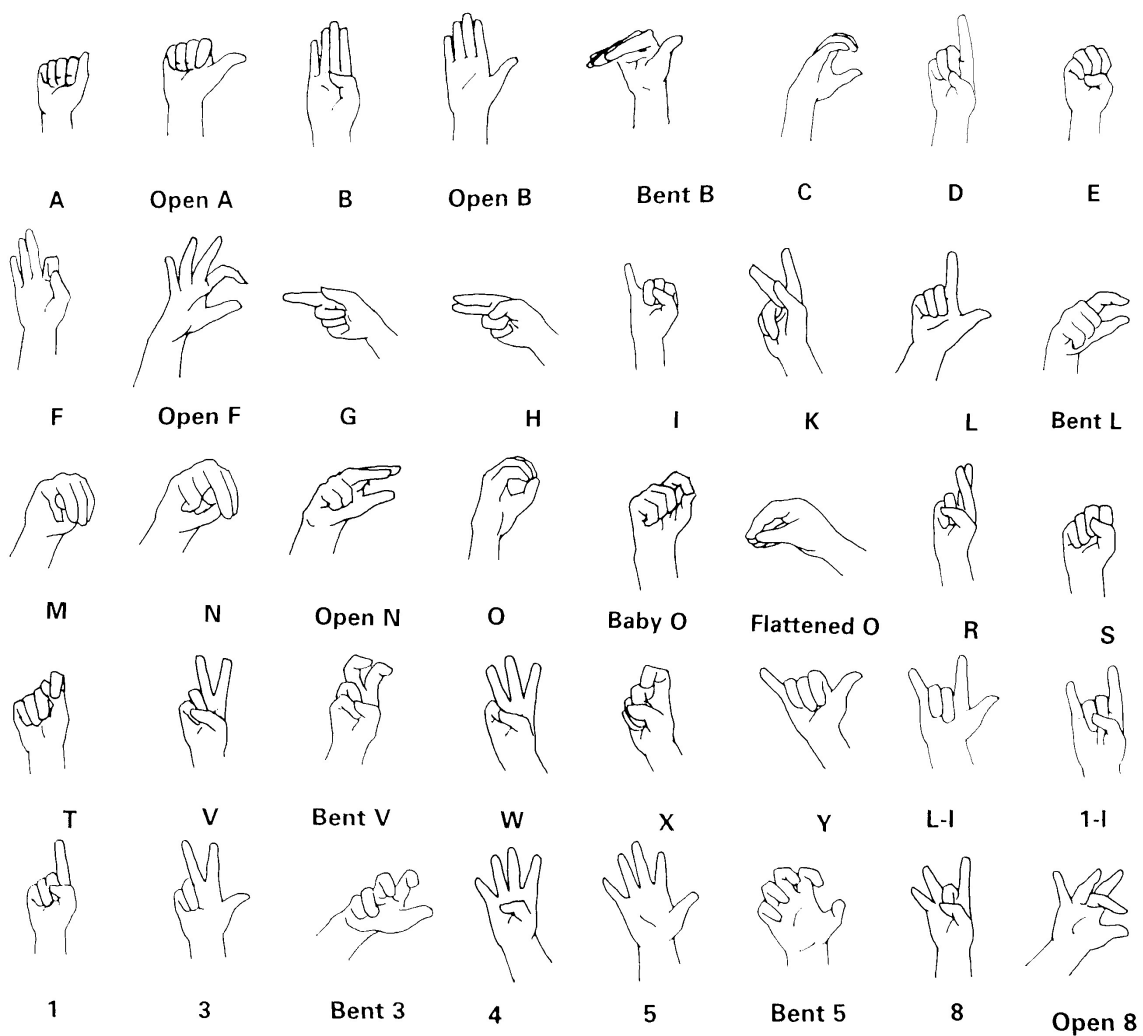
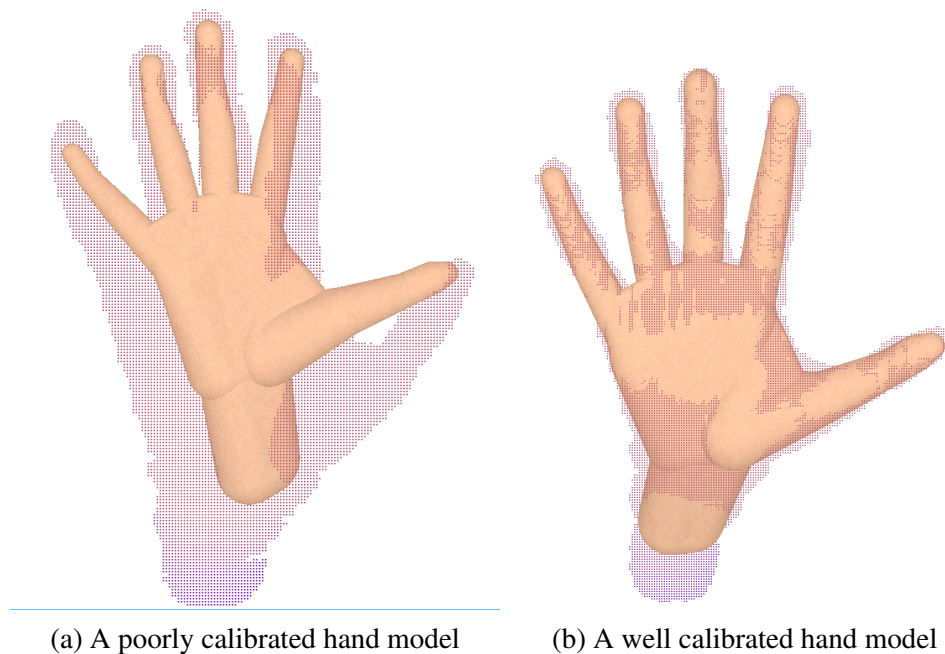


Figure 4.5: Proper hand model calibration greatly improves results. Figure 4.5a is the default, uncalibrated hand model overlaid on a silhouette of measured depth data. In Figure 4.5b the sizing parameters of the hand model have been adjusted to match the measured data for a given participant.



We ran Tkach et al.’s open-source Sphere Mesh [99] algorithm to collect the data. Each participant was given a yellow wristband to place on their left wrist as required by the algorithm and a manual calibration of the hand model’s size parameters was performed. This process consisted of adjusting three parameters (scale, width, thickness) up or down by 5% at a time until the rendered hand model aligned well with the hand silhouette measured by the depth camera (see Figure 4.5b for an uncalibrated and calibrated example). The calibration parameters were recorded for each participant (for further discussion of model calibration see Section 4.3.1). The model visualization was present on the screen for the first three sets of handshape prompts.

After calibrating the hand model, participants were presented with a series of 40 handshape prompts on the monitor. Participants were instructed to begin each collection sequence by spreading their fingers and directing their palm toward the camera (as shown in Figure 4.5b). When they were confident they understood the handshape presented in the prompt, participants initiated recording by pressing the space bar. They were then asked to move their hand from the initialization pose to the prompted handshape and designated orientation. Participants pressed the space bar again to end the recording sequence at which point the next prompt was presented automatically. After collecting a set of 40 handshapes in a given orientation, the process was repeated for another orientation. Data was collected for three palm orientations in the following sequence: PalmTowardCameraFingersUp, PalmRightwardFingersUp, PalmInwardFingersRightward (For examples, consult the following handshapes in Figure 4.4: /Open B, /C, and /G).

Following the third set of handshapes, three additional prompts were given: the ASL alphabet

signs for **P** and **Q**. These letters are designated by redundant handshapes (/K and /G respectively) at palm orientations that were not collected in our three sets. These signs were collected separately for the sake of comparisons to prior work.

Finally, one last set of handshape sequences was recorded in the PalmTowardCameraFingersUp orientation. This time, the hand tracking visualization was removed from the monitor. A monitor visible to the researcher allowed for verification that the participant remained within the cameras field of view. This set was collected to provide additional data and to examine whether visual feedback of the hand tracking algorithm impacted how participants moved from the initialization pose to the prompted handshape.

During data collection, estimates of the hand poses were generated and rendered in real-time. The underlying sensor streams (a 320x240 pixel 8-Bit RGB image and a 320x240 pixel 16-Bit depth image), from which the hand poses could be fully recreated, were recorded at 60fps. Preserving the sensor streams allowed us to reprocess sequences using different hand model parameters or initialization frames. None of the offline processing that we performed altered the algorithm in a way that would have reduced its performance below real-time response.

The pose estimates generated by Tkach et al.’s algorithm are defined by  $\Theta$ , a vector of 28 values representing the 28 degrees of freedom in the hand model first presented by Tagliasacchi [94]. The first three values represent the global location (x,y,z) of the hand in the camera space. The next three values represent the hand orientation. The rest of the values correspond to various individual finger joints. For each frame of depth data recorded,  $\Theta$  is fit to the data providing a measure of the hand pose at that frame. This measure, along with the hand model, can be used to render a visual representation of the user’s hand or to define the instantaneous pose. Every recorded sequence began with the participant’s hand in the same pose and ended in the prompted handshape and palm orientation. Since we were not interested in the transition for this analysis, we simply select the final hand pose,  $\Theta$ , as descriptive of the prompted pose.

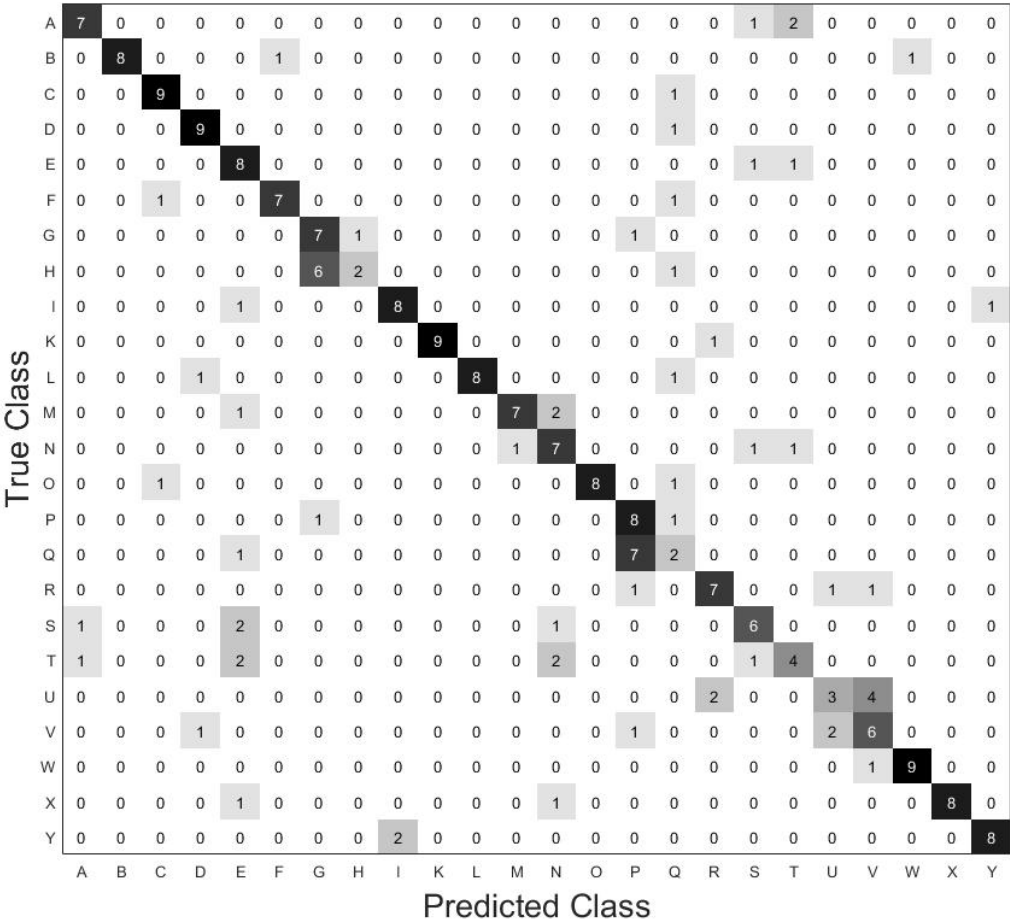
### 4.3.2 Participant Validation

As mentioned in Section 4.1, one advantage of using model-based hand tracking techniques is that the hand pose representation they provide is largely user agnostic. While signing styles may still result in individual variations, our classifier need not consider differences such as skin tone, hand size, and lighting conditions.

To demonstrate this, we collected a separate set of handshapes from two expert signers (one professional ASL interpreter, and one deaf ASL teacher). These experts followed a similar procedure to the standard participants except that rather than vary palm orientations, each collected three sets of 43 handshapes (the 40 distinct handshapes from [97] along with the letters **U**, **P**, and **Q**). The experts were at liberty to rotate their hands as needed to ensure the real-time model rendering accurately represented the prompted handshape. Due to a recording error, one data set was lost, leaving a total of five exemplar sequences for each handshape.

For a direct as possible comparison with previous work (see Table 4.2) we first trained a naive Bayesian classifier on the 24 static ASL alphabet poses presented by the two expert signers. We used only the finger joint angles from  $\Theta$  (ignoring global position and orientation) produced by the algorithm as the features to classify. The tracking was visualized and the pose parameters were recorded only when the rendered representation satisfied the expert signer. Five sets of

Figure 4.6: A confusion matrix showing the static ASL alphabet sign classification results. A naive Bayesian model was trained using data from the expert signers and tested on a single instance of each ASL alphabet sign from each participant. Data collection errors reduced the test set for the letters F, G, H, and U by one each. In total 165 of 236 (69.9%) sample poses were correctly classified.

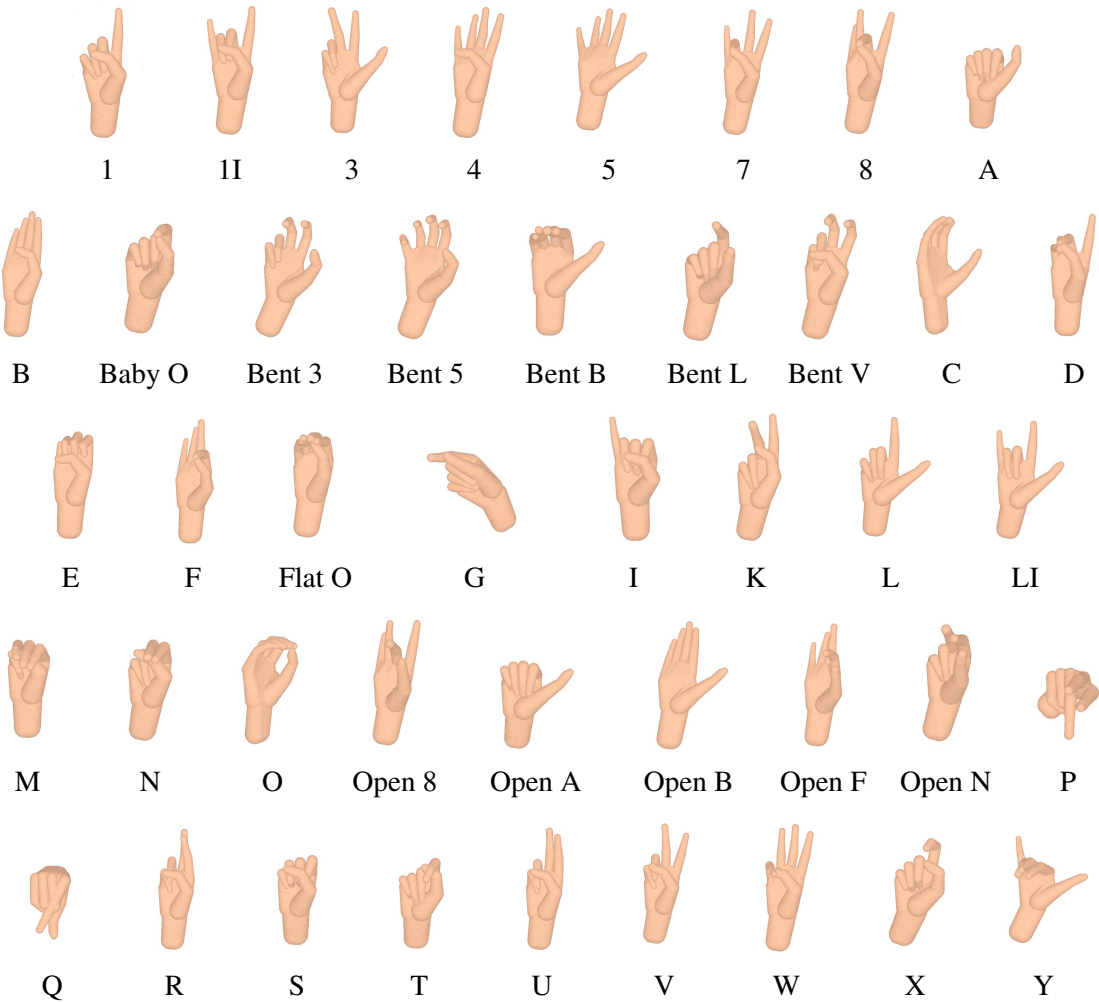


the 40 handshapes were recorded for both experts to provide a measure of variability and both experts' data was used to train the classifier.

The naive Bayesian classifier was then tested across the ten non-signer participants. The confusion matrix is shown in Figure 4.6, which shows the results across all participants. The average classification accuracy is 69.9%. This is an encouraging indication that the non-expert signers are accurately forming the handshapes and that the hand tracking algorithm is properly representing the handshapes (see Table 4.2 for a comparison with other research).

To give a clearer picture of how the hand tracking algorithm represents the poses of the handshapes, consider Figure 4.7 which shows a rendering of the average pose across the two experts for each handshape class that was recorded. The renderings are clearly recognizable and distinct as compared to handshape representations found in ASL instruction materials such as Figure 4.4.

Figure 4.7: Renderings of the average handshape poses across participants



### 4.3.3 Evaluation Criterion

One of the limitations of many of the static handshape studies that evaluate hand pose classifiers using ASL handshapes is the use of average classification accuracy (precision) as the default metric. While high precision is certainly useful, the frequency with which handshapes occur in signs is highly asymmetric. For example, the seven unmarked handshapes are reported to account for 70% of all ASL signs [44]. There are also correlations between the sign’s relative location and handshape frequencies, with more complex, marked signs occurring more often in positions around the signer’s face where the observer’s visual attention tends to focus [11].

Loeding and Sarkar’s review of ASLR approaches noted the reliance on recognition rates as a metric and advocated “a need to build consensus regarding meaningful measures of performance from a communication point of view” [55]. Recognizing that other factors such as hand location or the handshape of the second hand may greatly influence the prior probability of a specific handshape, we began to focus on the clustering of handshape classes from our classification techniques. Borrowing from Klima and Bellugi’s studies of sign confusion in the presence of video noise, we present measures of hierarchical clustering of the sign classes [44].

To analyze a larger dataset, data from both PalmTowardCameraFingersUp conditions (with and without the model visible, see Section 4.3.1 for further details) for non-signers and data from both experts were combined into one set. One model was trained on the entire set of data and used to generate the Hierarchical Clustering shown in Figure 4.8a. This clustering diagram presents the relative distances (a normalized euclidean distance of the finger joint angles) between the classes. The y-axis is a relative distance measure that indicates how close handshapes (or groups of handshapes) are to one another in this space. The ordering of the handshape classes has been aligned according to this hierarchical clustering.

A leave-one-participant-out training and testing approach was then used to evaluate the classification accuracy. Unlike Figure 4.6, the confusion matrix in Figure 4.8b presents the average class probability rather than the predicted class count. This view better represents how class confusion occurs in the model. The dashed line in Figure 4.8a is an arbitrary marker selected to highlight clusters withing Figure 4.8b. The two groups of handshapes underneath the dashed line, ‘*R, /U, /V*’ and ‘*/A, /T, /M, /N*’ are outlined in Figure 4.8b. Unsurprisingly, there is higher confusion amongst these ‘close’ classes.

For the complete handshape analysis, we again trained naive Bayesian classifiers using a leave-one-participant out approach using PalmForwardFingersUp conditions. Figure 4.9 shows the clustering and confusion matrix.

While naive Bayesian classifiers are simple, there are reasons to believe they may not be the best classifier for handshapes. For one thing, the naive Bayes classifier assumes the features are independent. However, this is not the case. For example, the flexion of the final finger joint (distal interphalangeal) is constrained by the flexion of the proximal interphalangeal joint and metacarpal joint flexion impacts metacarpal abduction [25].

To explore how other models may account for some of these relationships, we also trained a signer independent multi-class support vector machine classifier (Figure 4.10) which improved the average classification accuracy from 69.2% (for a signer independent naive Bayesian classifiers) to 76.3%.





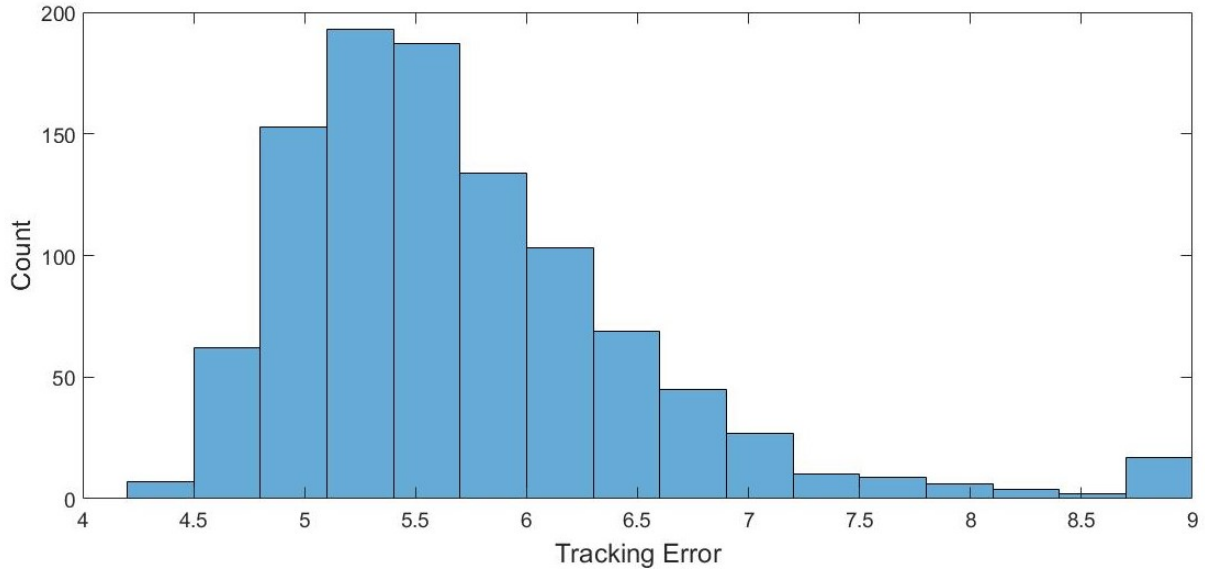




### 4.3.4 Tracking Issues

The hand tracking algorithm provides tracking error metrics that represent the amount of misalignment between the measured depth data and the position of the rendered hand model. While the absolute value of the metric is not particularly meaningful (see [99] for specific metric details), relative changes in the error metric for a calibrated hand model can be indicative of how well a particular pose estimation aligns with the underlying data.

Figure 4.11: Distribution of tracking errors across all handshape estimates. All tracking errors greater than 8.7 have been included in the final bin.

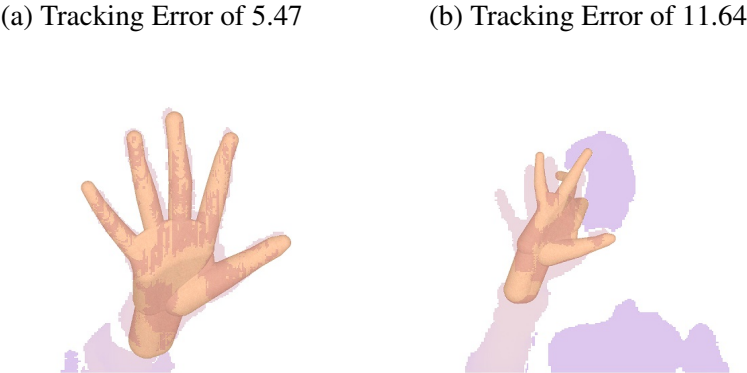


Tracking errors are reported with pose estimation. Figure 4.11 shows the distribution of tracking errors across all handshape poses for all participants. In our dataset the tracking errors ranged from a minimum of 4.27 to 14.45. To give a sense of the variation in accuracy, consider Figure 4.12. In Figure 4.12a, the hand model aligns with the underlying data and it clearly recognizable as an /5 pose. In Figure 4.12b, however, the hand is distorted into an unnatural position which clearly does not align with the /5 pose visible in the depth data.

If we aggregate classification performance by ranges of tracking error, as in Figure 4.13, we can see a general downward trend in classification accuracy as the reported tracking error increases. Both the class precision and the true class's predicted probability decline as tracking errors increase. In fact, for tracking errors over 8.4, no class is correctly classified. Of the 1028 sample pose estimates, 19 (1.85%) have an error greater than 8.4.

In addition to the occasionally extreme tracking error metrics, there are a number of artifacts of the tracking approach that can be observed in the pose estimates. One issue is that the joints of extended fingers tend to lock into the values imposed by the joint limits. While this is not necessarily a problem, it results in very limited variations for some joints of some classes. For classifiers such as naive Bayesian which rely on variance measures, this can skew the poses in which joints are largely extended toward classes that have partially flexed joints since they have

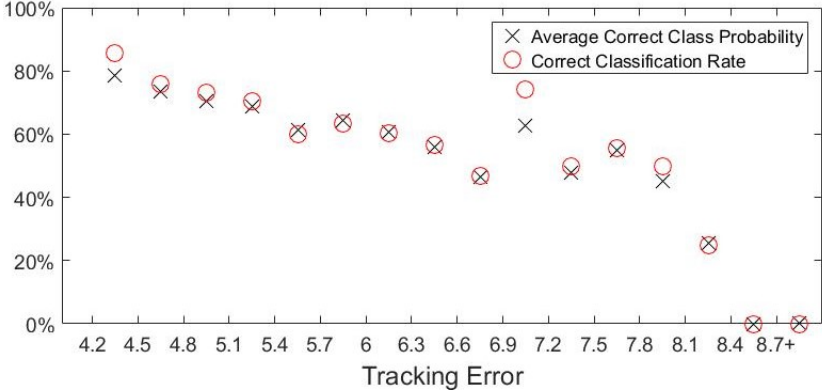
Figure 4.12: Examples of hand pose estimates with different tracking errors. The dotted images indicate the recorded depth map for the frame and the rendered hand indicates the pose estimated for that frame.



a wider joint angle distribution.

The uniform approach to hand fitting also raises some problems. Since the hand model fitting could only be adjusted with broad scaling features, individual hand variations could not be accounted for. One consistently observed result was a mismatch between the length of an individual finger for a particular participant and the model finger length. If the participant’s pinky was longer than the model’s, for example, the pose estimates would consistently bend the distal pinky joint. While this bend would better conform with the depth map, it often produced impossible hand configurations.

Figure 4.13: Classifier performance across tracking error. Bin widths are set to .3 with the lowest edge at 4.2





# Chapter 5

## Continuous Sign Recognition

Chapter 4 showed that real-time model-based hand tracking provides sufficient fidelity to adequately discriminate between the hand configurations that form ASL handshapes. However, by using participants with no signing experience and allowing them to form the hand poses at their own pace, the findings do not necessarily extrapolate to real-world signing scenarios. In this chapter, we will explore the performance of the model-based hand tracking when used by *fluent* signers in a *natural signing task*.

In order to make the problem more tractable, we will constrain the problem we focus on in this chapter to recognizing ASL fingerspelling. Fingerspelling is the act of sequentially signing English letters to spell words and represents a necessary component of American Sign Language (ASL). Primarily used to represent proper nouns and technical terms which lack formalized signs, linguists have estimated that as much as 35% of signed communication consists of fingerspelling [68].

In terms of recognition, fingerspelling offers a number of constraints which make it an appealing early test case for online ASL recognition. For example, in ASL all 26 alphabet signs are performed with a single hand and performed in the same location relative to the signer's body. Only two alphabet signs (**J** and **Z**) involve movement, the other 24 letters are distinguished solely by varying handshapes (particular alignment of finger position) and palm orientations (global alignment of the hand relative to the body). Additionally, since fingerspelling is by definition used to spell non-ASL words, linguistic parameters about letter sequences can be derived from available language corpuses.

This chapter will make a number of important contributions. First and foremost, we have collected the largest corpus of ASL fingerspelling data known. From this dataset we present an analysis of the impact of a number of system features (i.e., tracking model parameters, additional sensors) and provide guidance about subject and camera relationships for similar hand tracking implementations. We will also present a higher-level evaluation approach that focuses on the goal of ASL recognition (e.g., word level recognition) rather than focusing solely on the underlying feature recognition. Finally, we will present the real-time, signer independent ASL fingerspelling recognition system developed from this work. To our knowledge, this is the first such fully functioning ASL recognition system.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Hard of Hearing/Deaf		X	X	X		X				X
Right Handed	X	X	X	X	X	X	X		X	X

Table 5.1: Fingerspelling study participants.

## 5.1 Prior Work

In theory, any handshape recognition approach that can be calculated sufficiently quickly could be combined with a separate segmentation algorithm. However, in practice, few such systems have been explored. For example, while Pugeault’s work [74] was designed for real-time usage, the dataset used excluded the non-static **J** and **Z** signs and was collected one letter sign at a time, rather than in a natural fingerspelling sequence. Other work [30, 54, 77] has explored fingerspelling sequences, but has been constrained to limited vocabularies.

The most relevant work consists of a series of studies by Kim et al. [41, 42, 43]. These studies have relied on 60 fps videos of two to four native signers fingerspelling 300 words designed to explore coarticulation between alphabet signs [39]. The earlier studies examined different letter recognition approaches and achieved nearly 90% letter recognition rate, but relied on signer-dependent training [42]. A completely signer-independent approach using a deep neural network applied to the same dataset, however, achieved only roughly 40% letter accuracy. This number could be boosted to nearly 70% using weak word-level supervision, in which the word being signed was known though no specific frames were labeled, or to roughly 80% accuracy by adapting the model to the individual using manually annotated ground truth data.

However, the approaches described are applied to standard RGB video and provide no indication of the feasibility of applying these approaches in real-time. For the sake of comparison, we have adopted the word lists used in these studies; however, to make use of depth cameras, a wholly new dataset was collected.

## 5.2 Study Methods

At the heart of our data collection system is the the Sphere-Mesh hand tracking algorithm presented by Tkach et al. [99]. While we necessarily modified the code to incorporate the Myo Armband and customize the data collection process, we did not make any significant changes to the hand tracking algorithm presented. For more details of the hand tracking algorithm, see Section 4.1.

For this study, we recruited 10 experienced signers to collect samples of fingerspelling sequences. We specifically targeted the Deaf community in our recruitments but specified only fingerspelling proficiency in recruitment materials. Half of the participants self-identified as Deaf or Hard of Hearing. Two participants (P2 and P9) expressed reservations about their signing proficiency.

Each signer was equipped with a Myo armband and colored wristband on their signing arm. Participants were seated approximately 100 cm in front of a monitor within the view of an Intel SR300 depth camera. Images and depth maps were recorded at a 320x240 pixel resolution at 60



fps. Gyroscope, accelerometer and EMG data were also recorded from the Myo armband. Prior to any prompts, the Myo Armbands were calibrated and a coarse manual adjustment was made to the Tkach hand model to fit the participant’s hand [99].

The participants recorded three sets of 100 words each from the word lists used in [39]. The three word lists are comprised of proper nouns, nouns, and non-English words. The word prompts were displayed in a pre-defined randomized order (see Appendix B for details) with participants completing one entire list before moving on to the next in the following order: proper nouns, then nouns, then non-English terms. Participants started and stopped recording by pressing the space bar on a keyboard in front of them and were allowed to proceed at their own pace. Participants were allowed to press the back key to re-record any word. Doing so would not overwrite the original data collection.

Signers were not given any specific instructions about how to sign. If the participant asked, they were instructed to sign as they would to a novice signer. The choice of how to represent repeated letters in a word was left up to the individual.

### **5.2.1 Data Annotation**

Ground truth data about the fingerspelling sequences were created by viewing standard RGB video frame by frame and labeling the beginning and end of each letter. A simple, interactive browser-based viewer was designed to present the frame sequences cropped around the relevant hand. Since the word being spelled was contained in the file name, annotations could be performed by simply clicking the first and last frame corresponding to each letter. Figure 5.1 shows the browser-based viewer and the annotations for one participants spelling of the name ‘Joe’.

Often the precise beginning and ending of a letter is unclear. However, in the Movement-Hold Model, the precise moment of transition is not critical. As long as there is even a single frame that can be recognized as a clear hold from which the appropriate articulatory features can be determined, then the label is adequate. In practice, we found the 60 fps video rate to be adequate to capture distinct holds for most participants the vast majority of the time.

### **5.2.2 Qualitative Evaluation**

Renderings of each estimated hand pose were produced for every sign sequence. A simple browser-based viewer was developed to display and qualitatively evaluate estimated poses. See Figure 5.2 for an example of the tool.

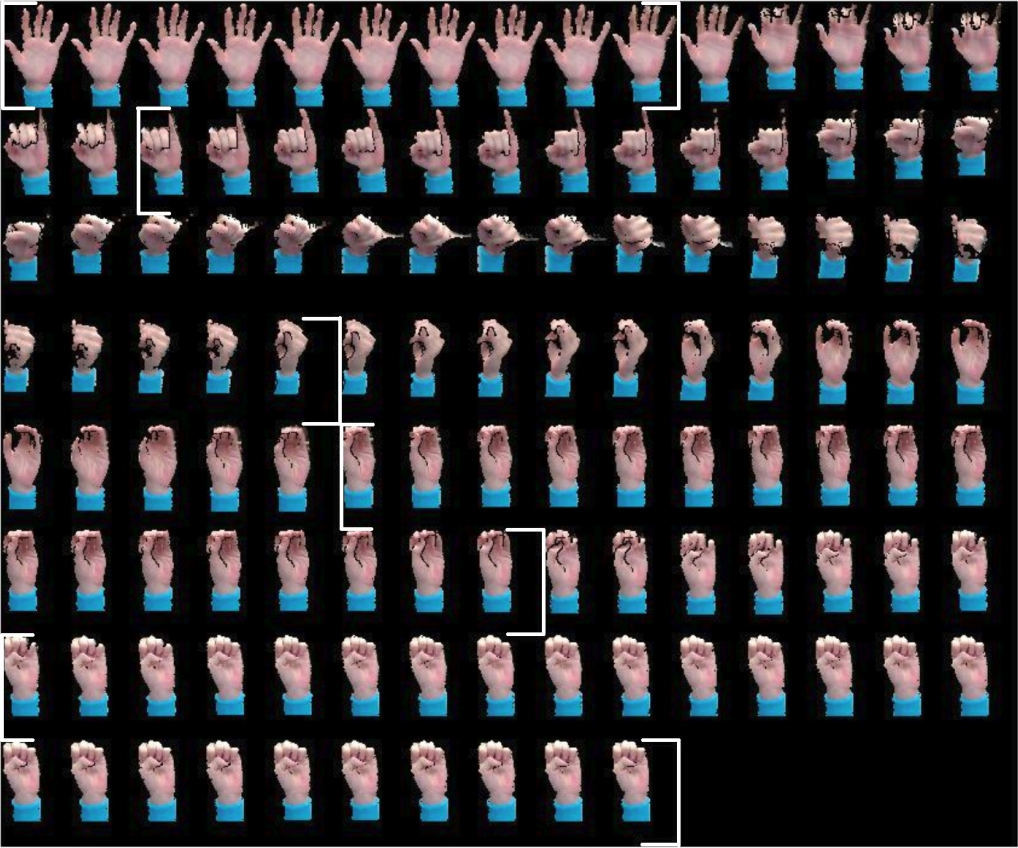
Using the tool, annotations about different poses could be made. Alternate handshapes could be indicated as a separate class. Gross tracking errors could be noted as well as more specific errors tied directly to a specific digit. These annotations allowed for much of the subsequent analysis evaluating specific issues related to the Hand Tracking algorithm.

## **5.3 Segmentation**

Unlike speech, which is composed of a series of articulations separated by periods of silence, signs have no clear ‘off’ condition. Hands remain visible and continue to occupy some con-

Figure 5.1: A snapshot of the browser-based frame viewing tool with annotations for the name 'Joe' marked. Participants began each sequence with an /5 handshape. The annotated frames are indicated by white brackets around the hand images and the corresponding letters and frame numbers are shown at the bottom of the viewer. Notice that the performance of the letter **J** follows the Movement-Hold Model with a stable // handshape at the beginning and several frames of a stable pose at the end. Unbracketed frames are considered samples of transitory movement epenthesis.

Start with 5



**5: 1-10, J: 18-50, O: 66-83, E: 91-115**

Figure 5.2: A browser-based viewer for evaluating pose estimates. The first example shows a pose estimated for an /Bent M handshape in which the ring finger is not correctly extended. The second example is a correctly estimated /Bent M example. The third example shows a gross orientation misalignment.

**Selected Letter:**

Letter:  Participant:  Set Size:  Set Number:

M/F15 amy 059			<input checked="" type="checkbox"/> Alternate HS <input type="checkbox"/> Orientation <input type="checkbox"/> Thumb <input type="checkbox"/> Index <input type="checkbox"/> Middle <input checked="" type="checkbox"/> Ring <input type="checkbox"/> Pinky <input type="checkbox"/> Other
M/F15 cameroon 073			<input checked="" type="checkbox"/> Alternate HS <input type="checkbox"/> Orientation <input type="checkbox"/> Thumb <input type="checkbox"/> Index <input type="checkbox"/> Middle <input type="checkbox"/> Ring <input type="checkbox"/> Pinky <input type="checkbox"/> Other
M/F15 powazaniem 134			<input type="checkbox"/> Alternate HS <input checked="" type="checkbox"/> Orientation <input type="checkbox"/> Thumb <input type="checkbox"/> Index <input type="checkbox"/> Middle <input type="checkbox"/> Ring <input type="checkbox"/> Pinky <input type="checkbox"/> Other



figuration regardless of whether or not a signer is actively signing. This presents an additional challenge to ASR: not only must the underlying parameters be accurately recognized, it must be determined whether or not the parameters are part of a meaningful expression at any given instant. Even with perfect detection of all five ASL parameters, extracting individual signs from a continuous stream of signing remains a significant challenge [19]. In this section, we will focus on an approach to segmenting sign sequences into individual signs from which they can be classified.

### 5.3.1 Movement-Hold Model for ABC signs

Our approach to modeling sign articulation was to follow the Movement-Hold Model proposed by Liddel and Johnson (See Section 2.3.3 for details). The Movement-Hold Model is a descriptivist approach that posits that all signs and sign sequences can be described as sequences of stationary poses separated by transitional moves. Individual signs can be as simple as a single Hold or formed by a more complex series of holds and moves. Transitions between one complete sign to the next, known as movement epenthesis, can also be modeled, though they do not form a part of any individual sign. The five ASL parameters, described in Section 2.2.1, are contained in the individual segmental and articulatory features, corresponding to movements and holds respectively.

For most alphabet signs, nothing more than a single hold with specific articulatory features describing the hand shape and palm orientation is necessary to define the sign. An example of the Hold-Model for the letter **C** is illustrated in Table 5.2.

<b>Hold</b>
Handshape
/C
Palm Orientation
Palm Out, Fingers Up
Location
Neutral
Non-Manual
N\A

Table 5.2: Movement-Hold Model for the letter **C**

Fingerspelling ‘cat’ would necessarily introduce transitions between the letters, thus the sequence **C-A-T** would modeled as shown in Table 5.3. The transitory movements indicated in the ‘Move’ columns of Table 5.3 are necessary to fully describe the sequence but do not provide lexical information to the individual signs. That is, the sign **A** is defined entirely by the stationary hand pose and the movements leading into and out of the sign **A** do not alter the meaning.

However, this does not mean that movements can be disregarded. For the signs **J** and **Z** (and most signs, generally), holds alone do not define the signs. Compare the models for the signs **I** and **J**. Both begin with the same hold configuration, however, **J** is distinguished by the rotational

<b>Hold</b>	<b>Move</b>	<b>Hold</b>	<b>Move</b>	<b>Hold</b>
Handshape	Movement	Handshape	Movement	Handshape
/C	HS Transition	/A	HS Transition	/T
Palm Orientation		Palm Orientation		Palm Orientation
Palm Out, Fingers Up		Palm Out, Fingers Up		Palm Out, Fingers Up
Location		Location		Location
Neutral		Neutral		Neutral
Non-Manual		Non-Manual		Non-Manual
N\A		N\A		N\A
<b>C</b>		<b>A</b>		<b>T</b>

Table 5.3: Movement-Hold Model for the sequence **C-A-T**. The columns represent Hold or a Movement segments which occur in the sequence presented left to right. Under each Hold segment are the articulatory features that define that particular Hold. Under each Movement segment are the segmental features that define that particular Movement. For more details about articulatory and segmental features see [53].

movement (wrist supination) which leads to another hold with the same handshape at a different palm orientation. Fingerspelling the name ‘Jim’ would thus be modeled as shown in Table 5.4.

As discussed in Section 2.3.3, the Movement-Hold models are intended to represent a signed sequence as it was performed. A sign, or sequence of signs, may be articulated in different ways,

<b>Hold</b>	<b>Move</b>	<b>Hold</b>	<b>Move</b>	<b>Hold</b>	<b>Move</b>	<b>Hold</b>
HS	Move	HS	Move	HS	Move	HS
/I	Wrist Supination	/I	Wrist Pronation	/I	HS Transition	/M
P.O.		P. O.		P. O.		P.O
Palm Out, Fingers Up		Palm In, Fingers Out		Palm Out, Fingers Up		Palm Out, Fingers Up
Location		Loc		Loc		Loc
Neutral		Neutral		Neutral		Neutral
N.M.		N.M.		N.M.		N.M.
N\A		N\A		N\A		N\A
<b>J</b>				<b>I</b>		<b>M</b>

Table 5.4: Movement-Hold Model for the sequence **J-I-M**. Unlike the other letter signs presented, **J** is composed of a Hold-Movement-Hold sequence. Notice that the first Hold segment of the sign **J** is identical to the Hold segment of the letter **I**.

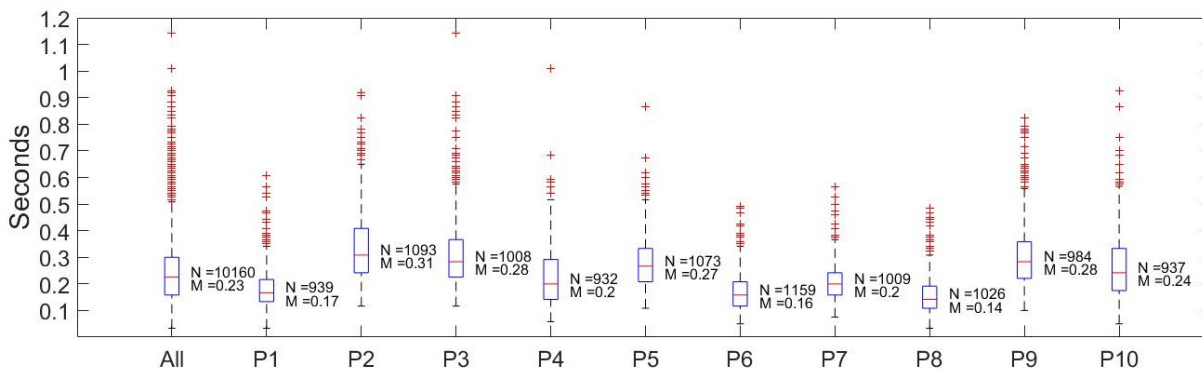
resulting in different Movement-Hold Model representations of the same sign. However, the underlying structure of the Movement-Hold Model remains the same. If the holds and movements can be accurately detected and the underlying articulatory and segmental features can be recorded, then the sign sequence is being effectively transcribed.

### 5.3.2 Timing Parameters

For each set of frames corresponding to a labeled letter, we marked the time of the middle frame. We then measured the time between subsequent letters in each sign sequence. The time from the beginning of the sequence to the formation of the first letter was disregarded to avoid variances caused by delays between the beginning of recording and starting the sign sequence. Similarly, the time between the penultimate and final letters in each sequence was discarded since the final letters were held for varying amounts of time until the recording was stopped.

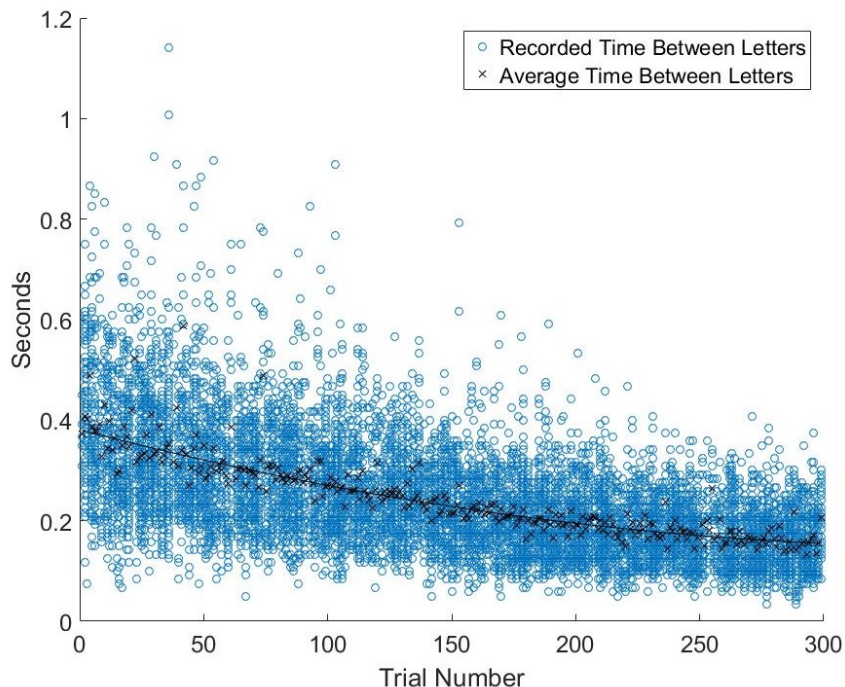
The box plots in Figure 5.3 show the range of times between letters for the entire data set as well as the individual participants. As can be seen, there is significant variation in the signing pace of different participants. For example, participant 8 transitions between letters at more than twice the pace of participant 2. There is no obvious relationship between signing time and participants self-identifying as hearing or Deaf (See Table 5.1). The two participants who expressed reservations about their signing abilities, P2 and P9, did tend to sign slower than most, but not dramatically so.

Figure 5.3: Box plots of the time between letters across the entire dataset and for each individual. The boxes represent the 25th to 75th percentile of times. The bar within the box indicates the median time between letters and the whiskers extend 1.5 times the range of the box beyond the quartiles. Individual measurements that fall beyond the whiskers are indicated by a ‘+’.



It is likely that rate of signing is as much a matter of choice than an indication of ability (at least for the participants in this study). However, there was noticeable change in signing rate as the study progressed. Figure 5.4 shows the measured times between letters for all participants across the 300 sequential trials. By fitting a curve to the data and modeling the timing as a function of the trial number, we see nearly a quarter second reduction in time between letters at the end of the study compared to the beginning.

Figure 5.4: The time between letters presented across the order of recorded sequences for all participants. Each participant recorded a total of 300 word sequences. The time between each letter for each trial is presented as a circle. The ‘x’s indicate the average time between letters for each trial. The line represents the second order polynomial fit of the time between letters across trials.



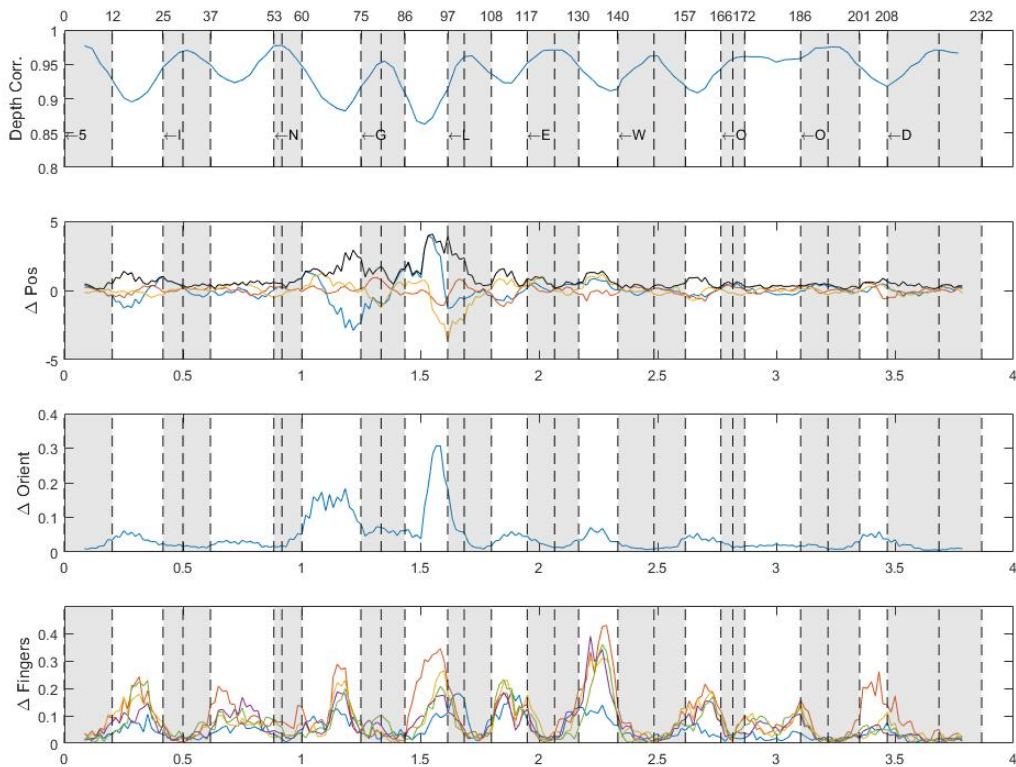
It’s not entirely surprising that participants sped up as the study progressed. Fingerspelling a sequence of 300 words is not a natural signing task and even though they were free to take breaks throughout the study, participants clearly did tire of the process as the study progressed.

### 5.3.3 Hold Model Training

While the Movement-Hold Model theorizes that signs are composed of distinct static poses with all movements occurring between holds, in practice, recognizing the beginning and end of a Hold is as much art as science. Liddell and Johnson explicitly acknowledge practical variations in signs, describing how hold deletions and sign modifications that are dependent upon sequencing can be described using their approach [53]. Ultimately, the Movement-Hold Model is designed to be descriptive of a sign’s performance, with the possibility of multiple distinct model descriptions applying to the same sign.

Fingerspelling, with its limited opportunities for global hand movements, provides a useful test case for exploring the Movement-Hold Model as a method for segmenting sign data. To do so, we calculated the following features related to manual movements. First, we derive the frame to frame correlation in the depth map segmented around the hand regions. This provides a snapshot of how much hand-localized movement is occurring between two frames. The top

Figure 5.5: Movement features for the word ‘Inglewood’. The first chart, **Depth Corr.** represents the frame to frame correlation of the depth data. The second chart,  $\Delta$  **Pos**, shows change in global position of the hand in x, y, z and absolute change. The third chart,  $\Delta$  **Orient**, shows change in absolute hand orientation. The final chart,  $\Delta$  **Fingers** shows the aggregate change in joint flexion across each finger.

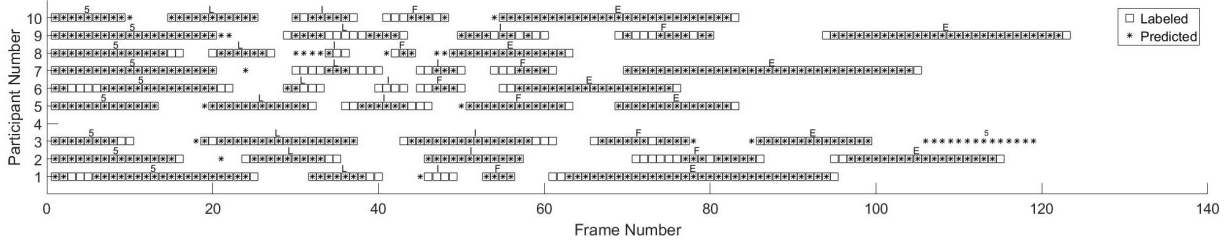


row of Figure 5.5 shows the frame-to-frame correlations smoothed across a five frame window. The gray regions indicate the annotated frames corresponding to the specific letter with an arrow pointing to the beginning frame. The vertical dashed lines are indicative of the apogee, or peak articulation, of each letter.

While the first row depends solely on the sensor data, the remaining features are derived from the hand tracking model and are thus vulnerable to any tracking errors that may arise. The second row charts the change in estimated global hand position. The change in (x,y,z) positions are presented along with an magnitude change in black. In the example in Figure 5.5, there is generally very little global movement of the hand with most corresponding with the rotation between **N** and **G** and back to **L**. The next row displays the change in global orientation of the hand. Similar to the position change, hand rotation is primarily limited to the transition to and from **G**, as would be expected. The final row displays changes in finger joint angle for each finger and an aggregated value.



Figure 5.6: Predicted and Labeled Hold frames for the word ‘Life’. Each participant has begins with their hand in the /5 handshake then moves through four signs to spell life. Though the timings are different, the labels show five distinct holds with gaps representing movement segments between them (participant 3 returns to the /5 pose at the end of the word creating a 6th hold). No data is present for participant 4 due to a recording error.



All	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
94.7%	81.7%	100%	99.4%	94.3	99.8%	90.6%	92.1%	94.2%	97.5%	97.2%

Table 5.5: Hold predictions.

In keeping with Liddell and Johnson’s Movement-Hold Model, there is a general pattern of movements stabilizing around the annotated frames. To see how well the annotations could be modeled, we trained a support vector machine with a linear kernel to distinguish hold from movement segments. In order to limit any errors that may arise from boundary decisions or poor tracking, we trained the model on the frame midway between each boundary removing any frames that had a 2D tracking error above 6 (see [99] for explanation of the error metric). Data from the letters **Z** and **J** were also removed as well as any instances of double letters as they often contain additional movements.

### 5.3.4 Hold Model Evaluation

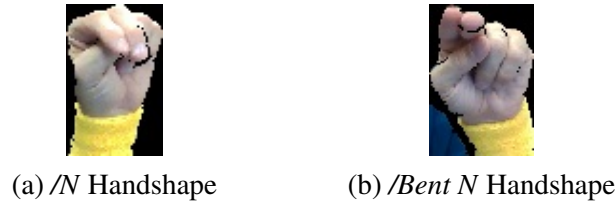
The model predicts the labeled frames with a total accuracy of 78.5%. However, correctly identifying the entire range of frames that express the hold segment of a letter is less important than being able to identify at least one frame within that range (see Figure 5.6 for an illustration). If we instead look at the rate at which there is overlap between predicted hold segments and labeled hold segments, which is necessary if we hope to classify a hold, the accuracy increases to 94.7%. Table 5.5 shows the rate of overlap between labeled and predicted holds across participants. As can be seen, there is significant variation in performance across participants. It’s important to keep in mind that if the model fails to overlap with a hold it is effectively missing a letter. Thus, Participant 1 is already facing a 19% letter error rate even if every identified letter could be perfectly classified. We will further explore variations across participants in Section 5.4.2.

## 5.4 Letter Classification

The study in Chapter 4 presented a very idealized approach to collecting handshape data. Participants were directly imitating a presented configuration. There was a knowledgeable observer present to correct any errors. There was no hurry or need to transition to any follow-up hand configuration. In practice, sign sequences are not so clean and clearly articulated.

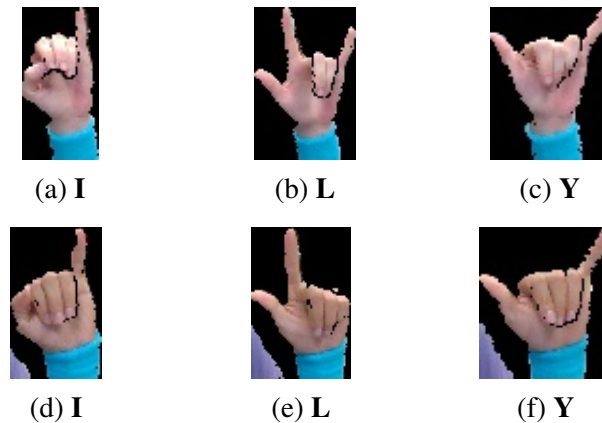
For one thing, signers do not strictly adhere to a single, formalized prescriptive approach to sign articulation. Regional variations like dialectics introduce variations, and individuals have unique idiosyncrasies that provide an additional challenge at the level of individual sign recognition. Even within the constrained context of fingerspelling, there are variations on English alphabet signs. For example, the sign **N** can be articulated using the standard */N* handshape or the */Bent N* handshape, as shown in Figure 5.7.

Figure 5.7: Examples of the variations in handshapes for the sign **N**.



Coarticulation, the situation in which speech sounds can be influenced by the preceding or following sounds, also has an analog in signing [39]. Within fingerspelling, particular orientations and even specific hand configurations can be influenced by surrounding letters. Figure 5.8 shows an example from the signing of **F-A-M-I-L-Y**. Both the letters **I** and **Y** are expressed with an extended pinky. In this example, the signer retains the extended pinky throughout the sign for the letter **L**.

Figure 5.8: Two examples of the sequence **I-L-Y**. The first example shows a coarticulation effect impacting the formation of the letter **L** between two letters formed with an extended pinky. The second example shows another participant without a coarticulation effect.



In this section, we will begin by exploring the limitations of a naive empirical approach that does not explicitly account for sign variations. We will then address a number of variations observed in our study data and evaluate approaches to modifying classification approaches to account for these situations.

### 5.4.1 Empirical Model

To begin, we simply adopted the approach to building a handshape classifier from Chapter 4. Each of the 26 letters and the /5 initialization pose were treated as a distinct class. The middle frame of each manually labeled letter sequence was selected as training data. A leave-one-participant-out approach was used, so each participant's data was evaluated using a unique naive Bayesian classifier trained on the other nine participants. The aggregate results are shown in Figure 5.9.

Across some 17,000 letter examples, this simple approach achieves 55.5% signer independent letter classification accuracy. Individual signer classification accuracies varied from 42 - 76%. However, this approach does not solve the problem of continuous sign recognition, since it required the manual selection of frames. Rather, it provides us a first look into how continuous sign performance impacts the accuracy of handshape classification. In Section 4.3.2, a similar classification approach achieved 68.8% accuracy under more controlled circumstances.

As can be seen, there is significant variation in individual letter classification accuracies and misclassification rates. The remainder of this section will explore various factors that may impact the classification rates.

### 5.4.2 Individual Idiosyncrasies

In this section we will examine variations between participants and discuss how such variations impact the relative performance of the system.

#### Handshape Variations

Handshape variations can be tied to particular individuals or to particular signs. Not all individuals have the same flexibility and range of motion and may articulate some signs slightly differently as a result. However, some signs have multiple accepted articulations and a given signer may alternate between sign articulations even within a single sequence.

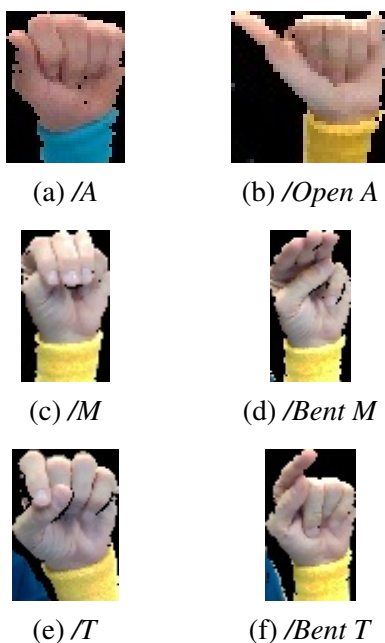
In addition to the variation in expressing **N** shown in Figure 5.7, we observed distinct articulations of **A**, **M**, and **T**. Each of the variant articulations employed the corresponding 'open' or 'bent' handshape (see Figure 5.10).

For the most part, these handshape variations were used interchangeably by multiple participants. The alternate **A** was rarely used by any participant other than participant 1, but was the exclusive articulation of **A** used by participant 1. This was not the only example of a participant-specific articulation. Participant 10, for example, had a tendency to express **Y** with the palm oriented downward and the index, middle, and ring fingers extended straight downward. Participant 3 tended to express **X** with far less index extension than other participants.

Figure 5.9: A confusion matrix presenting the aggregate class probabilities of 10 leave-one-participant-out, empirically trained naive Bayesian classifiers. Data samples were taken from the mid-frame of each manually annotated letter sequence.

True Class	#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	5	%	
A	1704	1084	5	34	1	32	1	20	27	3	18	0	9	2	16	14	2	1	2	277	119	2	0	5	18	2	0	10	64	
B	292	1	247	2	0	16	1	0	6	0	0	3	1	0	0	0	0	0	0	2	1	0	1	0	7	1	0	0	3	85
C	500	0	19	292	1	37	45	2	11	0	0	0	6	0	0	24	7	2	0	1	0	1	0	7	6	4	2	33	58	
D	389	1	4	4	213	4	0	1	13	0	0	5	0	0	0	3	8	16	23	3	4	18	7	11	29	1	21	0	55	
E	1624	4	200	45	3	856	3	12	26	53	10	1	5	6	6	18	2	2	29	177	0	11	2	10	117	7	5	14	53	
F	359	0	2	17	0	9	308	0	5	1	0	1	0	1	0	2	1	1	0	0	0	1	0	0	0	2	0	8	86	
G	402	22	0	2	2	7	0	183	70	0	29	1	0	0	4	1	25	15	1	9	10	0	0	0	8	0	9	4	46	
H	265	3	5	1	0	4	2	31	139	0	6	5	3	1	3	4	27	10	1	3	0	3	1	1	3	1	3	5	52	
I	1285	128	4	20	2	73	6	8	8	636	70	0	1	1	7	43	4	7	2	122	49	0	2	7	10	63	4	8	49	
J	161	13	0	1	0	2	0	7	1	14	70	0	0	3	1	5	2	7	1	17	9	0	0	1	3	0	3	1	43	
K	371	3	5	0	2	1	0	3	4	2	1	207	2	1	8	0	25	4	4	6	23	7	21	14	15	1	8	4	56	
L	838	80	14	18	6	35	3	18	21	0	4	19	390	0	0	0	11	42	11	16	8	17	4	2	75	3	11	30	47	
M	402	15	5	1	0	7	0	2	14	0	1	6	0	117	34	22	24	11	11	80	31	4	8	6	0	0	3	0	29	
N	1101	88	8	9	2	27	0	10	9	2	4	27	1	83	276	14	62	6	9	227	214	6	0	0	7	0	8	2	25	
O	1007	0	5	27	7	28	0	6	30	11	8	0	0	44	38	657	6	16	0	90	7	1	0	1	15	1	8	1	65	
P	282	8	1	2	1	0	0	14	26	0	9	8	0	0	3	0	144	21	4	1	1	19	5	2	2	5	4	2	51	
Q	145	0	0	0	1	0	0	11	3	0	2	0	1	0	1	2	28	69	8	2	4	0	1	0	1	4	6	1	48	
R	794	4	40	15	61	76	1	5	24	1	2	40	11	7	4	1	18	6	285	6	3	86	17	23	23	1	27	7	36	
S	733	18	1	4	4	30	0	1	2	4	5	0	0	6	31	17	1	1	2	560	41	0	0	1	2	1	1	0	76	
T	744	104	8	5	2	10	0	3	7	5	2	6	1	9	77	11	2	4	1	144	331	2	1	1	2	0	2	4	44	
U	566	6	19	2	41	10	0	9	25	0	5	37	8	5	0	0	42	7	71	10	2	101	90	29	18	2	26	1	18	
V	241	0	1	1	4	3	0	3	5	0	0	4	0	0	0	0	1	2	1	1	0	5	159	36	7	2	6	0	66	
W	194	0	6	1	2	8	0	1	3	0	0	0	0	0	0	1	0	0	1	0	0	1	8	162	0	0	0	0	84	
X	241	5	1	2	22	32	0	6	4	2	7	0	3	1	2	3	4	2	3	37	15	2	0	1	79	1	5	2	33	
Y	371	67	1	7	3	1	0	4	5	42	11	0	2	0	0	1	16	35	0	5	2	0	6	0	0	160	0	3	43	
Z	241	4	2	1	13	13	0	2	4	2	10	3	2	5	4	7	24	12	9	25	24	6	1	2	13	1	50	2	21	
5	2525	6	31	50	12	88	30	1	12	2	1	3	98	0	1	3	4	2	1	4	1	1	4	13	26	25	9	2097	83	

Figure 5.10: Other examples of handshape variations collected during fingerspelling.



How well a classification approach works generally (across a population) and how much individualized training needs to be performed remains an open question. Most studies, including this one are limited to relatively few participants, so it's difficult to generalize about the rate of alternate articulation in the wider signing population. However, it's no coincidence that the letters for which we observed alternate articulations were correctly classified at lower than typical rates (See Figure 5.9).

### Hand Size vs. Distance

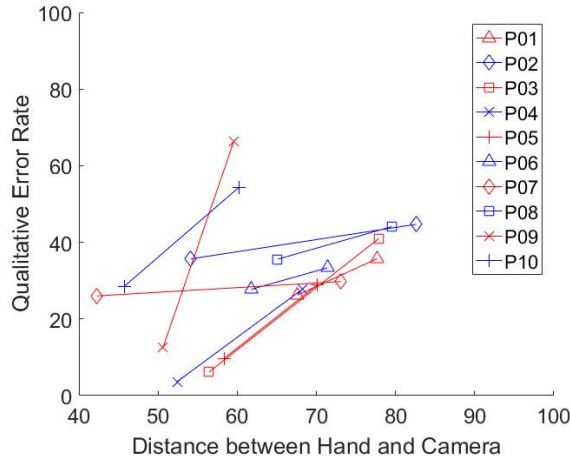
To date, most vision-based ASR studies rely on data collected in fairly standardized studio environments [61]. Signers are typically wearing solid print clothing that contrasts strongly with their skin tone and placed in front of a solid backdrop at a fixed distance. Our approach, with its use of depth cameras is less susceptible to interference from background imagery. However, the depth sensor's spatial resolution requires the participant to be seated closer to the camera than most standard RGB video approaches.

The Intel SR300 depth camera that we used to record data has a recommended operating range of 20-150cm. At 45cm distance, the camera has a point density of 1 point per  $\text{mm}^2$ ; this drops off to about 1 point per  $5 \text{ mm}^2$  at a distance of 1 meter [17]. The camera has a field of view width of about  $72^\circ$ , so a horizontal movement of 10cm at a depth of 45cm has an effective angular change of about  $12.5^\circ$ .

As a result, a participant's precise location relative to the camera can have a significant impact on the quality of data recorded. Every labeled frame for each participant was evaluated for notable tracking errors or digit misalignments. We then ran a regression of the qualitative tracking errors against the measured distance from the camera for each participant. The linear fit of

these regressions is presented in Figure 5.11.

Figure 5.11: Simple linear regressions between the rate of qualitative errors observed in the pose estimates and the distance of the hand position from the depth camera. For each participant there is a positive correlation between the rates of errors and the hand’s distance from the camera.



The precise fit of the regressions are probably not particularly meaningful given the likely non-linear relationship between the diminishing density associated with distance and the tracking algorithm’s accuracy. However, there is a clear increase in within-subject, qualitative errors as the participants’ tracked hand recedes from the camera.

If we look across subjects, we can see how hand size also plays a role. Participant 1, who had frequent qualitative finger tracking issues also had a relatively small hand (finger width of about 12.5mm) and sat slight farther back than most participants (average distance between hand and camera of 71.6cm). Given the characteristics of the SR300 we can calculate the average point density per finger width using Equation 5.1. This gives an average of 7.7 points per finger width for participant 6. Compared to participant 8 (16mm finger width, 65cm average distance, 10.9 points per finger width) and that is only 70% of the point density per finger.

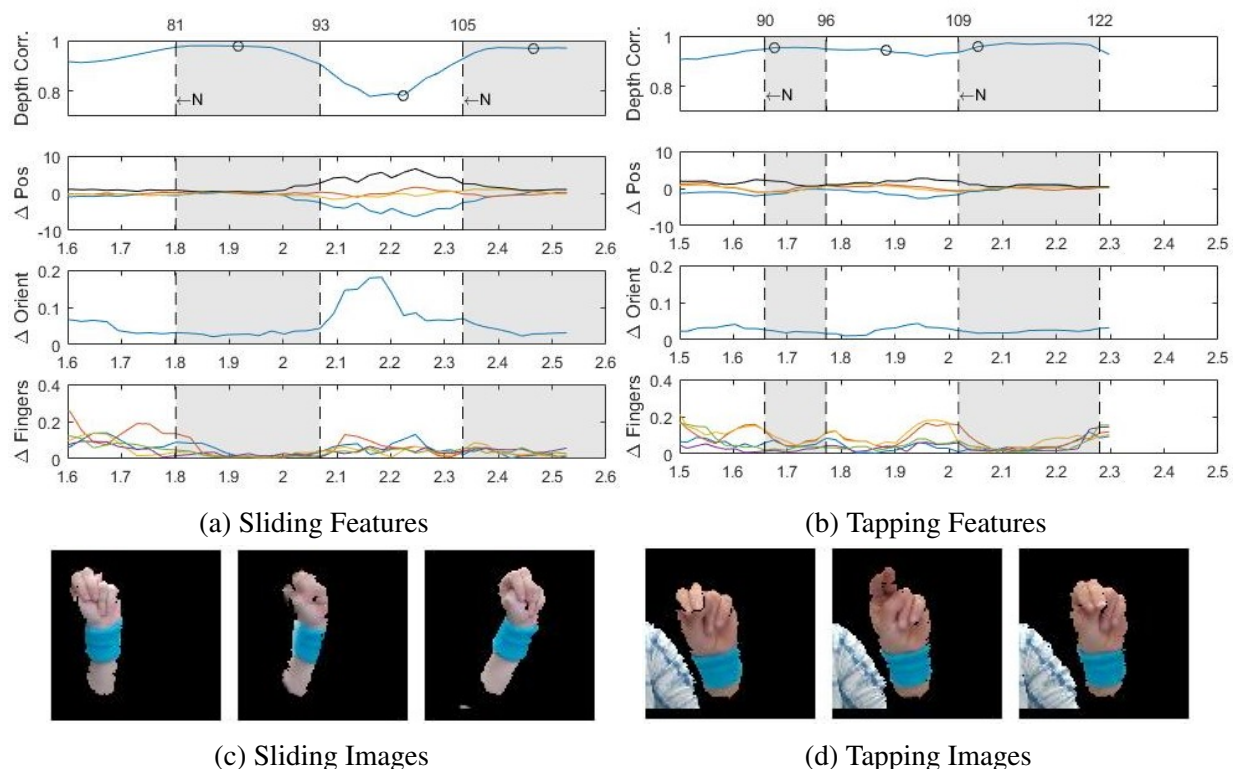
$$Pixels\ per\ Finger = \frac{Finger\ Width * 360px}{Z * \cos(36^\circ)} \quad (5.1)$$

While it is perhaps unsurprising that moving away from the camera results in more tracking errors, the fact that such effects occur so consistently even across relatively short distances is important to note.

## Double Letters

When fingerspelling a double-letter word (i.e., one that has the same letter twice in a row), there are two primary ways to indicate the letter repetition. Sliding is the more formal approach and is performed by holding the handshape of the letter steady and sliding the hand slightly to the side. Tapping is performed by slightly relaxing the hand from the letter’s handshape and then

Figure 5.12: Comparison of double letter sliding and tapping.



reforming the sign. Figure 5.12 shows an example of sliding and tapping performed by different participants.

On the left, participant 7, performs double-*n* by sliding. The hand's position shifts significantly as can be seen in the image sequences. The movement data clearly shows a shift in both global position and orientation of the hand and a significant dip in the depth data correlations between the two letters.

On the right, participant 8, performs a double-letter tap which consists of briefly opening from a *N* handshape, then closing again. The orientation, position and depth data correlations barely shift during the process, though the tracking does register changes in the finger joints between the two letters.

## Double **Z**

Unlike other double letter situations in which a slide or tap is used to provide separation between two instances of the same sign, double **Z** can be presented using a distinct handshape from the typical **Z** sign. Instead of the */D* handshape, the */Bent V* handshape is used with the same movement as the **Z** sign. Figure 5.13 shows a participant repeating two **Z** signs in a row. Figure 5.14, in contrast, shows the alternate */Bent V* handshape used to represent double **Z**.

Figure 5.13: Double Z performed by repetition of Z. The frame number for each image is marked in the top right corner.

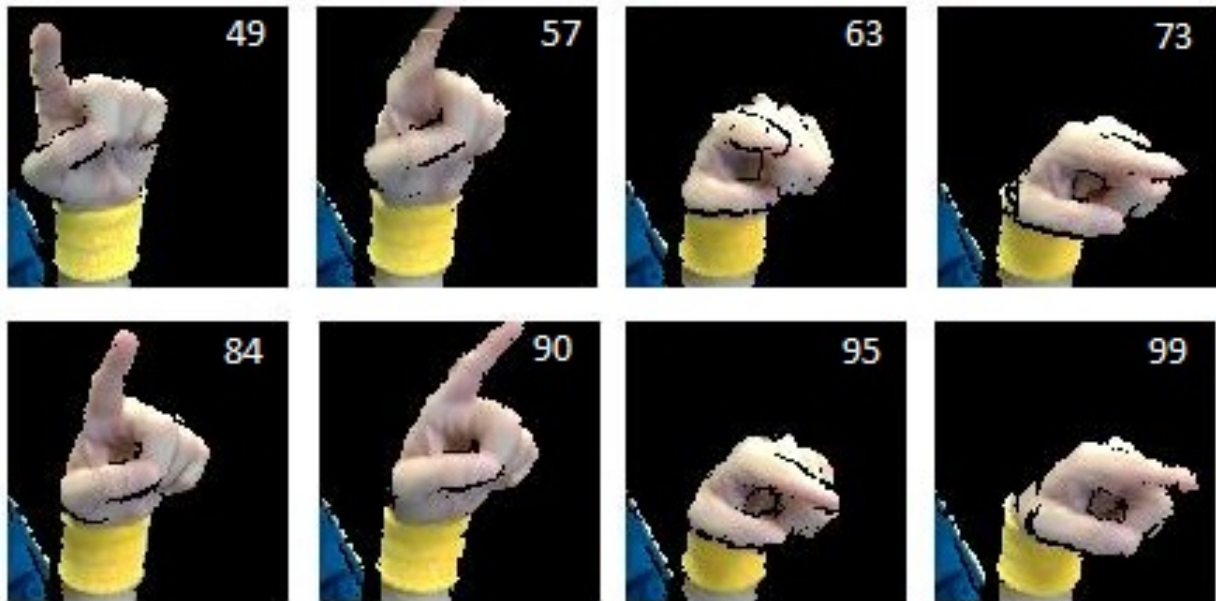


Figure 5.14: Double Z performed with a Bent V handshape. The frame number for each image is marked in the top right corner.





## 5.5 Sequence Analysis

While accurate detection of underlying ASL parameters is a necessary precondition of ASL recognition, the motivation for such systems (e.g., translating sign sequences to English) usually lies at a semantic, rather than phonetic level. Only focusing on the phonetic parameters overlooks how the structure of the language and interdependencies within the parameters can influence understanding.

There are limited examples in which the structure of ASL has been leveraged to improve classification. The relationship between handshapes at the beginning and end of a single sign have been leveraged to improve handshape classifications [98]. Accounting for simultaneous non-manual features have also been shown to improve the handshape classification rates [69]. As recognition tasks move from underlying parameters to higher levels of meaning, language structures will need to be accounted for whether explicitly or implicitly through representative training data.

Unfortunately, language modeling for ASL is no simple task. Without a standard written form, there is no easy way to mine massive datasets of exemplar speech in ASL. While efforts to create such datasets have been initiated [61], there is simply nothing on the scale of Google Books N-gram viewer [59].

Conveniently for the case of fingerspelling, the sign sequences being conveyed are not ASL.

### 5.5.1 String Analysis

For a fingerspelling system, the ultimate goal is to support accurate detection at the word level. How that goal is impacted by underlying letter recognition is dependent on both the classifier and structure of the language. For example, all else being equal, a classifier that frequently misclassifies the letter **Z** would likely be more useful than one that frequently misclassifies **E** based solely on letter frequencies. To take it a step further, a classifier that misclassifies **Q** but accurately recognizes **U** may be able to compensate for many errors based on letter sequences.

In this section, we will consider fingerspelling recognition as a form of text entry. Here we will explore how techniques developed for optical character recognition and soft-keyboard text entry can be adapted to improve the word-level recognition of our fingerspelling system.

In the field of optical character recognition, the character error rate is a common approach for evaluating results. There are generally three types of entry errors that can occur: insertions, deletions, and substitutions. The error types are straightforward with insertion errors indicating an additional letter added to the correct sequence, deletion errors indicating a correct letter omitted from the correct sequence, and substitutions indicating the transposition of an incorrect letter in place of a correct letter.

The character error rate for a word is then calculated as the sum of the three error types divided by the number of letters in the correct sequence (see Equation 5.2).

$$CER = (i + s + d)/n \quad (5.2)$$

There is no limit to the number of insertion errors that could be added to a string, so the character error rate can exceed 100%. There is, however, a known algorithm for calculated

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Signer Independent										
CER	59%	63%	46%	27%	44%	46%	37%	53%	54%	55%
WER	98%	99%	95%	79%	93%	93%	86%	96%	96%	96%
Signer Dependent										
CER	41%	32%	23%	15%	21%	31%	24%	42%	27%	29%
WER	90%	81%	70%	49%	67%	80%	69%	91%	67%	78%

Table 5.6: String level analysis of fingerspelling letter classifier performance. CER indicates the average number of character error types across the total number of characters as indicated in Equation 5.1. WER indicates the rate at which individual words contained at least one error. Signer Independent models were trained and tested using a leave-participant-out approach across the entire dataset. Signer depended models were trained and tested using a leave-one-sequence out approach across each individual participant.

the minimal edit distance between two strings [109]. By calculating the minimal edit distance between the detected string and ground truth, we can evaluate the results at the word level.

### 5.5.2 Optimal Segmentation

Selecting frames for letter classification from labeled letter sequences, provides the best case scenario for segmentation. Effectively, this removes any deletion or insertion errors, forcing classification at manually annotated points in the sign sequence. For this analysis we produced two sets of letter classifiers. The first is a set of wholly *signer-independent* naive Bayesian classifiers trained on the other nine participants for each participant. The second set is a *signer-dependent* set of naive Bayesian classifiers, trained using a leave-one-out approach at the word level. Thus each word for each participant has a unique classifier trained on the other word samples for that participant.

Table 5.6 shows the character error rate and rate of words for which each letter was correctly classified. There is a wide variation of performance across participants. Participant 4 is noteworthy with perfect fingerspelling recognition over 20% of the time using a signer-independent classifier. As would be expected, classifiers customized to the individual participants perform significantly better. Participant 4 again leads the way with over 50% of words being recognized correctly.

### 5.5.3 Spell Checking

The approach taken to analyze fingerspelling sequences thus far does not leverage any details about English. However, the advantage of access to large databases of English is that dictionaries of known words can be used. An open source spellchecking algorithm [10] based on an open source Word Frequency List [35] was applied to the predicted letter sequences. This spellchecking algorithm considers the three error types described above as well as transpose errors, in which subsequent letters are swapped. As implemented, the algorithm compares entered letter

Proper nouns absent from dictionary
skokie, rangerover, alcapulco, flossmor
Proper nouns omitted due to space character
el salvador, oak park, san francisco
Nouns absent from dictionary
appetizers, fanbelt, firewire, jawbreaker, softserve, twizzlers, xmen
Non-English Words in dictionary
ole, cie, rado, mina, hol, missa, axon, usta, itt

Table 5.7: A list of the seven proper and seven nouns not found in the dictionary. All other words on the ‘Proper Nouns’ and ‘Nouns’ lists from Sections B.1 and B.2 were present. Only nine words from the ‘Non-English Words’ list in Section B.3 were present in the dictionary. They are also listed.

sequences to all known words with an edit distance of no more than two, returning the sequence with the lowest edit distance or highest word frequency if multiple options have the same edit distance.

Of the 300 words used in the study, 195 were present in the open source dictionary (see Appendix B for the complete word lists and Table 5.7 for details on the words included in the dictionary). Table 5.8 shows the character error rate and word error rate for each participant for the set of words contained within the dictionary. As compared to Table 5.6, there is a much larger improvement in Word Error Rates than Character Error Rates. The decrease in WER clearly shows that the algorithm does eliminate errors. However, at times the algorithm will occasionally introduce more character level errors by correcting to an incorrect, but more frequently used word. For example, if the sequence **T-E-A** were recognized as **T-E-H**, the spellchecking algorithm as implemented would return the sequence **T-H-E** since the word ‘the’ is a single transpose error away and is more frequently used than the word ‘tea’, with its single substitution error.

There is a clear opportunity to improve spellchecking for fingerspelling systems by better modeling fingerspelling errors. One issue is that transposition errors, which occur quite frequently on two-handed keyboards, are far less likely given the sequential entry method of fingerspelling. Removing transposition errors from consideration would likely improve the spellchecking performance. Beyond that, one could adapt approaches used on small touchscreen keypads by considering likely entry errors. For small touchscreens, keys located near each other on the QWERTY layout are far more likely to be substituted for each other by mistake. In our fingerspelling system, similar handshapes are far more likely to be substituted for each other (see Figure 4.9a). By accounting for error likelihoods rather than just making selections based on word frequency, further gains are probable.

## 5.6 Error Sources

In order to improve the effectiveness of our approach, it is important to understand the sources of errors. In this section we will focus on where our approach fails to accurately recognize letters and what the options are for improving our approach.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Signer Dependent										
CER	37%	27%	21%	10%	16%	26%	22%	37%	23%	22%
WER	67%	51%	40%	19%	35%	51%	44%	63%	37%	45%

Table 5.8: String analysis after spellchecking.

### 5.6.1 Addressing Entry Errors

During data collection, participants did not always fingerspell sequences flawlessly. Our collection approach allowed participants to re-record sequences if they felt they made an error. Table 5.9 lists the number of times that participants felt they made an error and chose to re-record a sequence.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Second Recording	19	16	12	7	4	4	4	31	22	20

Table 5.9: Instances of repeated recordings by participant. Each participant recorded a total of 300 samples.

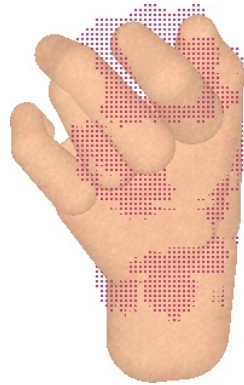
While it is doubtful that the rate of errors in our test conditions are indicative of fingerspelling error rates in typical conditions, they do highlight the fact that people make mistakes. In addition to the instances of re-recording, there were also occasional instances where the participant either started a sequence again without re-recording or omitted or misformed letters while recording. The participants either did not notice these errors or simply opted not to repeat the task.

For the purposes of our study, sequences with entry errors were simply ignored. However, for an actual sign recognition system, error handling would be a necessary component. To date, very little consideration has been given to the design of sign recognition interfaces. Pugeault’s work is the only example, that we are aware of, of an online system that explored how the participant might interact with a fingerspelling recognition system [74].

One of the advantages of the generative hand model approach that we use is that it provides real-time renderings of the hand tracking estimates. This gives participants an indication of how the system is performing. Though we disabled the rendering display during data collection, it provides an intuitive indication of when the pose estimates are failing.

Extending this interface to provide information about letter recognition and offer methods for interacting with results is an interesting avenue for future work. At a minimum, the system should display alternate dictionary options and provide a method for making selections. We are in the process of implementing our system on a large area touchscreen. We believe that the mid-air gestural interfaces (as proposed by Pugeault [74]) would lead to mode-switching confusion given the gestural nature of ASL. The touchscreen provides a clear distinction between system interactions and sign performance without requiring the use of a peripheral (e.g., mouse or keyboard).

Figure 5.15: An example of an orientation error. The participant was signing the letter **M** with palm facing forward, but earlier in the sign sequence the pose estimate rotated to face the other direction.



### Addressing Tracking Errors

There are numerous factors that impact tracking performance. In this section, we will discuss factors that contribute to tracking errors and approaches that might mitigate them.

### Participant Positioning

As noted in Section 5.4.2, there is a correlation between the distance from the camera and the qualitative error rate in the hand tracking estimates. A more rigorous study examining tracking performance against hand sizes and the depth camera's point density at various ranges would be helpful for establishing more exact guidelines.

With such guidelines, it would be simple to instruct participants to position themselves within an appropriate range. It would likely be worth exploring the value of providing some sort of visual indicator to the participant to let them know if their hand has exceeded the ideal depth range. Modifying the background color of the hand tracking rendering display would be a simple way to provide participants with a sense of how their positioning may be affecting tracking accuracy.

### IMU data

One issue with the hand tracking approach that we use is that pose estimates can get broadly misaligned with the participant's correct orientation. If the tracking algorithm estimates that the hand is oriented opposite of its true orientation, it often maintains that error for a significant time period. See Figure 5.15 for an example of misaligned orientation.

When the hand estimate is rendered, these errors are immediately obvious and participants can realign the hand tracking algorithm by holding a simple, stable pose (e.g., /5) for a short period of time. However, it would obviously be preferable to avoid such errors in the first place.

In looking at the instances of orientation errors noted in our examination of qualitative errors (see Section 5.2.2), it was quickly apparent that orientation errors were associated with wrist

movements and alternate palm orientations. Table 5.10 shows the rate at which specific letters were present in words during which orientation errors were observed.

	G	H	P	Q	J	Z	None
Orientation Error Rate	37.3%	40.5%	23.7%	13.8%	9.7%	9.4%	3.5%
Word List Rate	13.3%	10.3%	13%	7%	8.7%	9%	50.7%

Table 5.10: The rate at which specific letters were present in a word in which an orientation error was observed. For comparison, the rate at which those letters were present in words in the data set are also presented. The final column indicates the rate of errors and occurrence of word that do not contain any of the preceding letters. There is a clear over-representation of orientation errors in letters articulated with movement and alternate palm orientations.

A simple method for addressing this would be to incorporate wrist-worn accelerometers and gyroscopes which are available in many smartwatch devices. We collected data from forearm-worn accelerometers and gyroscopes in the Myo armband during this study. Unfortunately, the independent rotation of the wrist meant that the motions that most frequently resulted in the orientation tracking errors were not captured in the data.

## Hardware Improvements

Since collecting data for this study, a new generation of Intel depth cameras have been released. With a higher spatial resolution and slightly wider field of view, the D435 offers the same point density at 69 cm depth that the SR300 offers at 45 cm. The wider field of view offers a horizontal range of 1.28 m at that density, doubling the SR300’s 64 cm span. Additionally, the Intel D435 depth camera can operate at 90 fps compared to the SR300’s 60 fps.

Additional sensor resolution is not a total panacea for hand tracking. The generative modeling approach we are using works by exploring permutations of potential handposes and comparing them to the measured data. Increasing the point density increases the computational costs of comparing the potential pose space. Increasing the sensor’s temporal resolution reduces the amount of time available for iterating across fitting functions.

Of course, increases in computation power available on CPUs and GPUs help to offset these additional computational requirements and sensor data can always be downsampled. The PC used to collect and process the data for this study had a single NVidia GTX 1080 GPU. Since collecting that data, we have acquired another PC with dual NVidia GTX 1080 GPUs.

How these hardware changes impact tracking performance needs to be explored. Needless to say, the current pace of hardware improvements is outpacing the rate at which data can be collected and analyzed, and is a promising development for the potential of this approach.

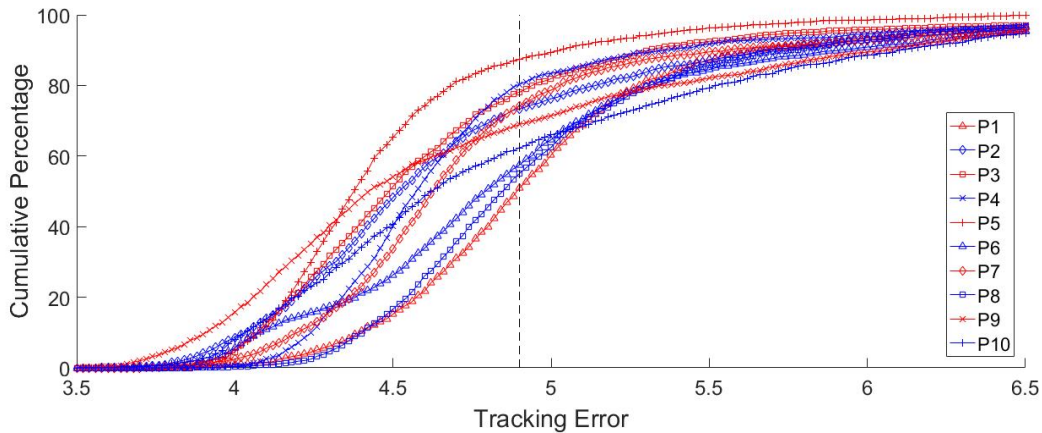
### 5.6.2 Addressing Classifier Errors

Classifier errors do not necessarily directly correlate with lower level tracking errors. Classifiers can adapt to consistent tracking errors in ways that may or may not be ultimately beneficial. Here will discuss approaches to addressing errors observed at the classifier level.

## Tracking Error Effects

To examine how tracking errors impact higher level classification results, we first examined the distribution of tracking across individuals. Figure 5.16 shows the cumulative distribution of tracking errors across all frames for each individual. While the range of errors is not huge in absolute terms, there are notable differences amongst participants.

Figure 5.16: Cumulative tracking error distribution per participant. The individual curves indicate the percentage of frames recorded by an individual participant with a tracking error under a given value. The line at 4.9 is drawn to highlight the difference between participants relative error as shown in Table 5.11



The cumulative rate of errors diverges most across participants most significantly between about 4 and 5.5. In Table 5.11 we compare the fingerspelling character error rates presented first in Table 5.6 with the percentage of frames with a tracking error under 4.9 for each participant. The tracking error rate is clearly not the only factor involved in accurate character recognition; P5 clearly has the best tracking error distribution in Figure 5.16, yet P4 has the lowest character error rate. However, there is a clear inverse correlation between the frequency of tracking errors under 4.9 and classified character error rate. Compare the error rates of the two best performing participants, P4 and P5, with those of the worst two, P1 and P8. It stands to reason that efforts to improve tracking accuracy discussed in Section 5.6.1 will likely have beneficial effects on classification as well.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Signer Dependent										
CER	41%	32%	23%	15%	21%	31%	24%	42%	27%	29%
Tracking Error < 4.9	51%	73%	78%	80%	87%	57%	74%	55%	69%	62%

Table 5.11: A comparison of character error rates from fingerspelling recognition and lower level tracking error metrics. There is a strong inverse correlation between the rate with which a participant records a tracking error under 4.9 and their classified character error rate.

## Filtering Suboptimal Data

One of the issues in having multiple layers of feature detection is that interactions between the layers can behave in unexpected ways that mask underlying issues. For example, if our hand tracking approach produces errors that are consistent for a particular class of handshapes, a letter classifier may be able to consistently predict the class based on the erroneous pose estimate. Whether or not this is a problem, depends on the task at hand. For fingerspelling recognition, as long as the letter is correctly recognized, it does not matter if the underlying hand pose is correctly estimated or if the hand pose has a consistent error.

However, classifiers that are built compensating for errors run the risk of failing to perform correctly in the future if underlying estimates are improved. Similarly, if tracking errors are not consistent across user populations, then a classifier will not generally perform well. While additional user data would certainly mitigate such issues, we can analyze the qualitative errors observed across our ten participants to see if errors are consistent.

Figure 5.17: The qualitative error rates per participant per static letter sign.

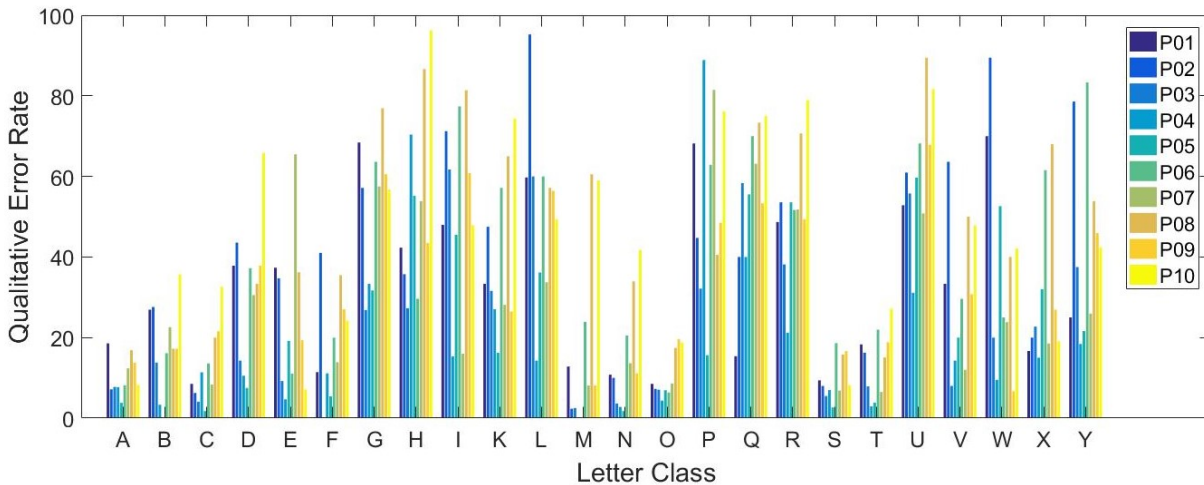


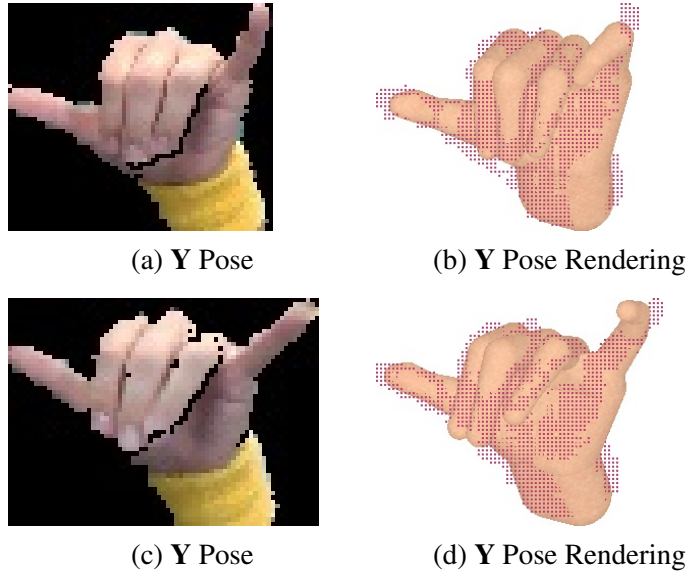
Figure 5.17 shows the qualitative error rate across classes and participants. Generally, open handshapes have higher observed error rates given the opportunity for correctly extended fingers to be transposed by adjacent fingers. Closed handshapes are simply more constrained by the hand kinematics. Looking closely, one will notice that error rates for particular classes vary significantly from participant to participant. In practice this means that common tracking errors are not consistent across participants. This contributes to a poorer performance in signer-independent classification rates since there are participant-specific errors that crop up at different rates for different classes.

Consider participant 6's poor qualitative tracking performance for the letter **Y**. Figure 5.18b shows an example of participant 6 forming the letter **Y** and the rendering of the corresponding pose estimate. In the rendering, the pinky finger is not extended fully. This is a common qualitative tracking error, but one that is unusually common for participant 6, occurring over 80% of the time. If a simple Bayesian classifier is trained on participant 6, these erroneous pose estimate



will be classified correctly, whereas the more accurately estimated pose shown in Figure 5.18d will be misclassified as the letter **I**.

Figure 5.18: An example of participant 6 performing the sign for **Y** and the corresponding pose estimate rendering. Notice that the pinky is not fully extended in the rendering.



Disregarding poses with observed qualitative errors improves the signer-independent classification accuracies. Figure 5.19 compares the aggregate classification accuracies of signer-independent classifiers for each participant trained on the entire data set and the filtered data without observed qualitative errors.

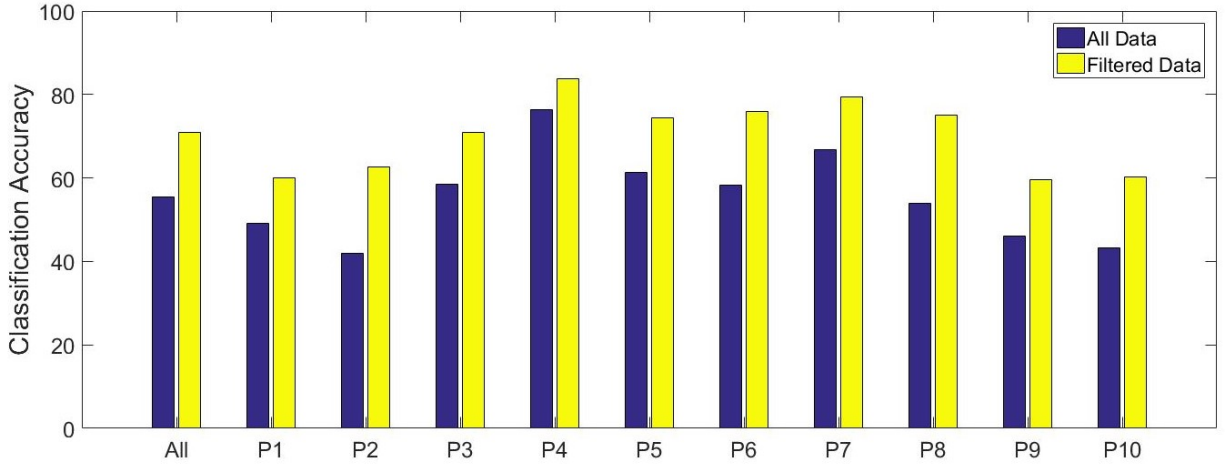
### Recognizing Additional Hidden Classes

As noted in Section 5.4, there are handshape variations used to express a number of letter signs. By creating additional classes for these alternate handshapes, we can distinguish distinct handshape variations. Whereas previously, we had empirically trained a single class for each letter, we now add additional ‘hidden’ classes for /*Open A*, /*Bent M*, /*Bent N*, and /*Bent T*. These classes are trained on the data marked as alternate handshapes as noted in Section 5.2.2.

	All	A	M	N	T
No Hidden Classes	76.6%	80%	41%	37%	52%
Hidden Classes	77.3%	84%	57%	36%	52%

Table 5.12: Classification accuracies comparing a naive Bayesian classifier trained with additional ‘hidden’ classifier to one without. The data was restricted to poses without observed qualitative errors. Accuracies presented here are resubstitution rates for training data across all participants.

Figure 5.19: The average letter classification rate across participants for the entire dataset and for data without an observed qualitative error. Unsurprisingly, filtering out examples with clear tracking issues improves classifier performance. However, the gains are mediated somewhat by the fact that errors are not entirely independent from the letter class. For many letters with extended fingers, finger transposition was a common error (e.g., a ring finger extended in place of the pinky). These consistent errors could at times be modeled by the classifier resulting in better performance than one might expect given the rate of observed qualitative error.



The classes are ‘hidden’ because when evaluating the classifier performance, predictions for these classes get added back to the baseline letters. Thus, poses that are classified as *Bent M* ultimately get recognized as **M**. As can be seen in Table 5.12 the addition of hidden classes has a slight positive impact on the aggregate classification rate. Improvements are not uniform across all classes with alternate handshapes, however.

# Chapter 6

## A Complete System

While the fingerspelling system described in Chapter 5 has been designed to operate in real-time, it does not satisfy all the requirements of a complete ASL recognition system as described in Chapter 2. Table 6.1 shows the ASL parameters that a complete system needs to capture and the parameters recognized by the system described in Chapter 5.

Parameter	Implementation	Details
Handshapes	X	Single Hand
Palm Orientation	X	Single Hand
Location	O	Absolute, not relative
Movement	O	Absolute, not relative
Non-Manual		Not implemented

Table 6.1: Complete ASL parameter recognition requirements and implementation status. An ‘X’ indicates the system can currently recognize the particular parameter. An ‘O’ indicates the system partially recognizes the parameter. Currently, handshapes and palm orientations are recognized in real-time for only a single hand. The Location and Movement of the hand is recognized in absolute space, but not currently relative to the body as the parameters are typically described by linguists.

In this chapter, we will discuss the limitations of the fingerspelling system as implemented and explore modifications necessary for the system to be fully capable of capturing all aspects of ASL.

### 6.1 Two Hand Tracking

Accurate, real-time hand tracking is a difficult and computationally intense problem for a single hand. For wholly independent gestures, tracking a second hand is theoretically straightforward and simply requires a second model fitting pipeline. As the hands begin interacting and occluding each other, however, the complexity of the modeling problem increases dramatically.

### 6.1.1 Requirements

Essentially, everything that was implemented for a single hand needs to be repeated for a second hand. While there are certain language constraints that reduce the independence of the two hands (see Section 2.2.2), it is not entirely clear how those constraints should be applied in practice.

### 6.1.2 Implementation

Modifying the original sphere-mesh hand tracking algorithm [99] to track both hands was fairly straightforward. Wristbands of distinct color were used so that color filtering could be used to identify both wrists. The depth images were then segmented into two continuous regions connected to each wristband. From there, the original modeling fitting algorithm was simply run separately on a Left and Right hand model.

The additional hand that must be tracked necessarily reduces the number of iterations that can be run for each hand's pose estimate, in turn reducing the tracking accuracy. However, hardware improvements are continuously mitigating the impact of the additional computational needs. Since the time of our original handshape classification study described in Chapter 4, we have acquired a computer with more than twice the GPU cores available to it. While the algorithm's performance may not scale with GPU cores in a precisely linear fashion, it does observably improve performance. A more precise benchmarking, as demonstrated in the original Sphere Mesh paper would provide a clearer measure of the impact of tracking a second hand.

### 6.1.3 Challenges

The biggest limitation of the current algorithm's approach to two-handed tracking is that it assumes the hands are free in space. As the hands come in contact with each other (or other body parts) the approach to depth data segmentation breaks down and the tracking accuracy falls off precipitously.

The current approach simply segments the depth data into contiguous regions from a seed point provided by color filtering. A more sophisticated approach to segmenting the depth data into distinct hand regions will be necessary for representing many of the common interactions that occur in ASL. Incorporating simple body and head tracking may be sufficient for distinguishing the hands from the body for simple, single point contacts (for example the signs in Figure 2.3). However, for signs in which the hands interact and occlude each other more strongly, an alternate approach to fitting both hands simultaneously may be necessary.

## 6.2 Face Tracking

While face tracking has received significant research attention generally, applications specifically adapted to sign recognition have been limited [7]. We have implemented a face tracking approach that can run alongside the fingerspelling system describe in Chapter 5, however, we have not yet evaluated its performance in real-time signing tasks.

### **6.2.1 Requirements**

Non-manual features, such as facial expressions and mouth shapes, are a necessary lexical component for some ASL signs (see Section 2.2.1). These features also play an important role in many aspects of ASL grammar. To date, studies have demonstrated that non-manual features can improve the classification rates of other parameters [69]. They have also been used to recognize a number of grammatical features including Yes/No questions, hypothetical conditionals, Wh-questions, assertions and negations [12, 64].

Despite the clear need non-manual parameters, the underlying features needed to recognize grammatical features are not clearly defined.

### **6.2.2 Implementation**

A open-source, deformable face tracking approach created by Saragih [79] was integrated along with the Sphere-Mesh hand tracking algorithm. The face tracking algorithm fits a model of 68 landmark facial features to the standard RGB video stream supplied by the Intel SR300 depth camera. Validation of this face tracking approach by implementing previously reported non-manual ASL feature detection algorithms is planned. Additionally, benchmarking studies to determine the impact of the additional computational requirements of the face tracking algorithm on the existing hand tracking algorithm are planned.

### **6.2.3 Challenges**

Many face tracking algorithms operate on the assumption that the face being tracked is clearly visible within the scene. While this is a valid assumption for many situations in which a person is being videotaped, it is not always the case in signing performance. As noted by Metaxas et al., partial facial occlusions occur with regularity in typical signing scenarios and typical face tracking approaches are not prepared to handle this situations [58].

Depth data offers a clear opportunity for detecting partial facial occlusions. Whereas with standard RGB video, the limited contrast between hands and faces make occlusion detection difficult, depth data provides a more straightforward method for determining when standard face tracking algorithms are likely to fail. An exploration of how depth data can be leveraged to improve detection of non-manual ASL features is planned for future work.

## **6.3 Body Tracking**

Body tracking in the context of sign recognition has received little attention to date. Open source skeletal tracking algorithms do exist, but have not yet been incorporated into our system.

### **6.3.1 Requirements**

Body tracking is a necessary component of determining location parameters. While hand tracking algorithms provide a global word position (assuming a fixed camera), the salient location

feature of many signs is actually the position of the hands relative to other body parts. For this, a skeletal model of the body would be useful. There are also ASL signing situations in which the signer's body orientation conveys meaningful information.

### **6.3.2 Implementation**

The Kinect camera was released with a native skeleton tracking algorithm. Various open source implementations have since followed suit, perhaps most notably the Open NI Tracker [3].

No body tracking algorithm has been implemented alongside the fingerspelling system described in Chapter 5.

### **6.3.3 Challenges**

Little work has been done to evaluate body tracking in the context of sign recognition. Implementing such a system will offer opportunities to explore the value of body tracking in the context of sign recognition.

# Chapter 7

## Conclusions

This document has presented an overview of the challenges of developing an automatic ASL recognition system. It has also detailed prior work in this field in order to frame the studies conducted described in Chapters 4 and 5. In this chapter, we will highlight broad findings from those studies and the development of an ASL recognition system building on their findings.

### 7.1 Depth-based Tracking offers Advantages

As demonstrated in this thesis, using depth cameras and generative-model-based hand tracking algorithms is a promising approach to ASL recognition. The results presented in Chapter 4 demonstrate that existing open-source algorithms [99] are currently capable of providing pose estimates with sufficient fidelity to achieve comparable results with other state of the art approaches to Handshape recognition [21, 48].

The combination of depth cameras and generative model-based hand tracking also offer the following advantages:

- Real-Time Performance
- Signer Independent Classification
- Pathway for Improved Performance

#### 7.1.1 Real-Time Performance

The work presented in this thesis was conducted with an explicit aim toward developing a sign recognition engine capable of running in real-time. Parameters in the hand tracking algorithm were set such that the results could be produced during live operation. The classification techniques explored do not require significant computational overhead.

While reducing the constraint that all pose estimation and classification be performed at the rate at which sensor data is acquired could improve results, we felt it was important to operate under conditions in which such a system would be utilized.

## 7.1.2 Signer Independence

One advantage of depth-camera approaches is that they are not susceptible to lighting conditions and skin tones in the same way that standard RGB video is. This fact alone eliminates one of the challenges to producing signer independent vision-based approaches.

By using a hand tracking approach, our system performs classification on higher level representations of the hand rather than underlying sensor data. Individual variations, such as hand size are then handled separately from the sign classification. While this does not entirely eliminate signer-dependent aspects (see Section 5.4.2), it does allow independent development of parameter classification and hand recognition.

## 7.1.3 Improvements

Depth camera-based hand tracking solutions have received a lot of research attention in recent years [76, 83, 96]. With significant commercial interest in AR and VR platforms, improvements in hand tracking are likely to continue.

By focusing on general models of hand configuration (see Section 4.1) as the underlying feature of our classifications, the approach we have adopted will be able to apply future hand tracking improvements with minimal modification. In terms of vocabulary size, glove-based approaches from more than a decade ago [29] still outperform modern vision-based approaches. However, glove-based systems have not seen wide-spread commercial adoption and have generally been dismissed by the Deaf community.

We feel that broader trends, including improvements in depth-camera hardware and the interest in natural hand gesture interfaces for VR and AR, position depth-camera based approaches to improve in the near future. While approaches reliant on depth information cannot make use of existing databases of ASL sign videos [61], we feel that the benefits of current real-time performance and the likely improvements in accuracy make the collection of new datasets a worthwhile pursuit.

## 7.2 Meaningful Measures of Performance

Research that focuses only on the accuracy of detecting and classifying underlying sign features does not necessarily indicate the utility of that approach to the ultimate goal of sign recognition. ASL evolved naturally within the constraints of human perception and studies have demonstrated that higher-level language models can improve lower-level feature recognition [69, 98]. To understand how well a technique works for sign recognition, evaluation approaches need to better capture the ultimate goal of the system.

Building on the real-time focus of Chapter 4, we explicitly applied our recognition system to the task of ASL fingerspelling in Chapter 5. Fingerspelling is a constrained, but still necessary, aspect of ASL. By focusing our second study on fingerspelling, we could analyze how systemic errors in our hand tracking impacted classification recognition rates at the word level (see Section 5.6.2). Utilizing English dictionary information, we demonstrated that higher-level error



correction approaches can overcome some limitations in the detection of underlying features (see Section 5.5.3).

Given that the largest available databases of signer data are still limited to no more than a dozen signers [27, 61], fixating on recognition rates seems premature. This is not to say that recognition rates are unimportant, only that the results may not generalize. Understanding how errors occur, both in human perception of signing [44] and in an approach to sign recognition, is important to understand how well an approach might apply generally.

## 7.3 Integration and Interfaces Are Under-Explored

There are still significant limitations to the capabilities of real-time hand tracking approaches. Occlusions and interactions between hands are not well-handled.

Little development has been put into interfaces for ASR. How are results presented to users? Generative models have the advantage of providing representations of detected hand configurations to the user. This can provide immediate indication when models have converged to an incorrect pose and allow the user to intervene (e.g., by returning to easily recognizable hand configuration). How this feedback can inform the signer and how they respond to it is not clear. We found anecdotal evidence that participants do respond to tracking errors (see Section 4.3.2), but how to effectively leverage these responses has not been studied.

How to edit and correct recognition errors at the semantic level is also unclear. Mobile phones have led to the integration of auto-correct into keyboards, but still provide users with the ability to override entries manually. How do signs get represented as they are detected and how does a user adjust or edit the input?

These are the types of questions that have not been explored because the underlying systems necessary to make them possible did not exist. With the development of the real-time sign recognition system we have built, we will be able to explore a rich area important for the progress of sign recognition.

The complete system proposed in Chapter 6 will greatly expand how different aspects of ASL can be recognized and how users can engage with a recognition system. We have begun to incorporate large screen touch interface to with our sign recognition system to allow for direct user input. Touch input seems to offer a promising approach as it does not require an additional peripheral device but retains a clear separate interface signing. With the system established through our work presented in Chapters 4 and 5, we have the pieces necessary to explore and create novel interfaces for sign recognition.



# Appendices

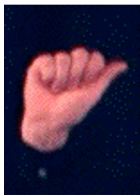













# Appendix A





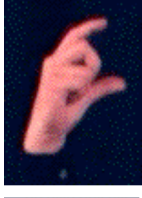
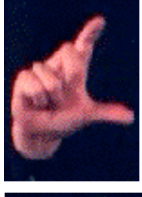


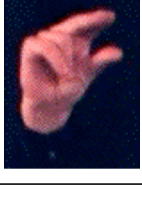

## ASL Handshape Primes


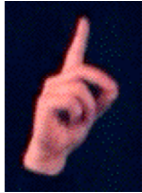









Different linguists have arrived at different counts of meaningfully distinct sets of handshape primes. In this appendix we will present a comprehensive overview of handshapes as recognized by different sources. Since Boston University’s American Sign Language Lexicon Video Dataset recognizes the most extensive set of distinct ASL handshapes available, we will present examples and labels provided by their documentation [61] for each handshape prime. Next we match the 40 handshapes from the American Sign Language Handshape Dictionary [97] to the Boston University (BU) primes and present model renderings from of the average pose collected for each handshape in our studies. Also included are the designator symbols and primes presented in Stokoe’s Dictionary of American Sign Language [90] and the collection of handshape primes presented by Klima and Bellugi [44].

In aligning sets of handshape primes from different sources there are necessarily some discrepancies. The alignments presented here are merely judgments. For more details on how specific handshape primes are defined by various linguists, we refer you to the cited sources.


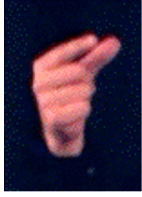


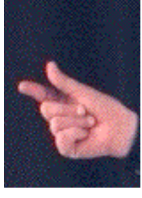
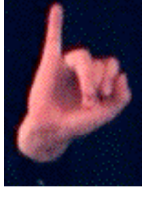

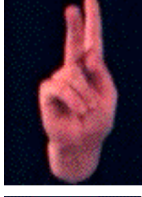

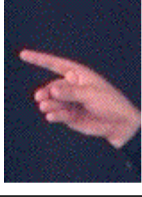
BU Gloss	Image	ASL Dictionary	Rendering	Stokoe	Klima
/A		/A		A	A
/B		/B		B	B



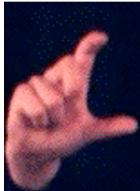

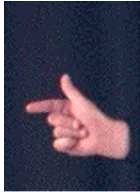
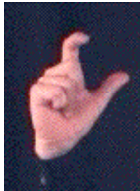


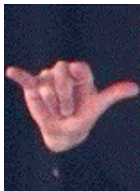
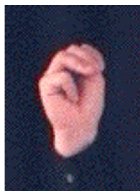

BU Gloss	Image	ASL Dictionary	Rendering	Stokoe	Klima
<i>/crvd-B</i>					
<i>/crvd-sprd-B</i>					
<i>/B-xd</i>					B <sub>b</sub>
<i>/flat-B</i>					
<i>/crvd-flat-B</i>					
<i>/B-L</i>		<i>/Open B</i>			$\dot{B}$
<i>/bent-B</i>					






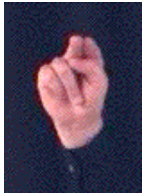

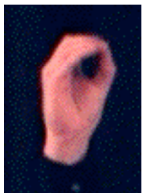

BU Gloss	Image	ASL Dictionary	Rendering	Stokoe	Klima
<i>/bent-B-L</i>		<i>/Bent B</i>			Ê
<i>/C</i>		<i>/C</i>		C	C
<i>/sml-C/3</i>					
<i>/lrg-C/3</i>					
<i>/flat-C</i>					
<i>/tight-C</i>					
<i>/tight-C/2</i>		<i>/Bent N</i>			

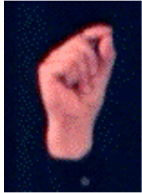






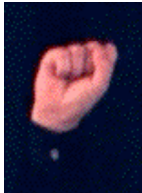

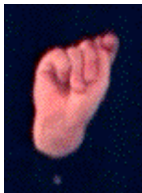

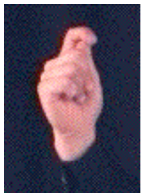
BU Gloss	Image	ASL Dictionary	Rendering	Stokoe	Klima
<i>/C-L</i>					
<i>/D</i>		<i>/D</i>			<b>G<sub>d</sub></b>
<i>/E</i>		<i>/E</i>		<b>E</b>	<b>E</b>
<i>/loose-E</i>					
<i>/F</i>		<i>/F</i>		<b>F</b>	<b>F</b>
<i>/cocked-F</i>					
<i>/open-F</i>		<i>/Open F</i>			



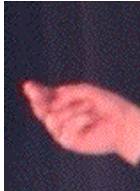
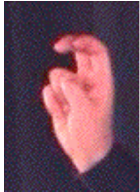
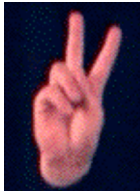





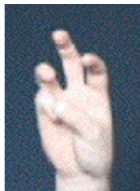


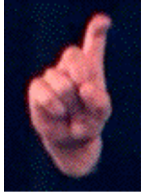





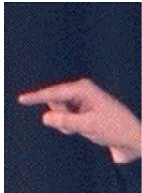




BU Gloss	Image	ASL Dictionary	Rendering	Stokoe	Klima
<i>/flat-F</i>					
<i>/G</i>		<i>/G</i>		G	G
<i>/flat-G</i>					
<i>/alt-G</i>					G <sub>g</sub>
<i>/I</i>		<i>/I</i>		I	I
<i>/K</i>		<i>/K</i>		K	K
<i>/alt-P</i>					













BU Gloss	Image	ASL Dictionary	Rendering	Stokoe	Klima
/L		/L		L	L
/crvd-L		/Bent L			Ĭ
/bent-L					
/L-X					
/I-L-Y		/L-I			Ȳ
/bent-I-L-Y					
/M		/M		M	M










BU Gloss	Image	ASL Dictionary	Rendering	Stokoe	Klima
<i>/bent-M</i>					
<i>/full-M</i>					
<i>/alt-M</i>					
<i>/N</i>		<i>/N</i>			
<i>/bent-N</i>					
<i>/alt-N</i>					
<i>/O</i>		<i>/O</i>		<i>O</i>	<i>O</i>

BU Gloss	Image	ASL Dictionary	Rendering	Stokoe	Klima
<i>/baby-O</i>		<i>/Baby O</i>			bO
<i>/flat-O</i>		<i>/Flat O</i>			ô
<i>/flat-O/2</i>					
<i>/R</i>		<i>/R</i>		R	R
<i>/S</i>		<i>/S</i>			A <sub>s</sub>
<i>/T</i>		<i>/T</i>			A <sub>t</sub>
<i>/x-over-thumb</i>					

BU Gloss	Image	ASL Dictionary	Rendering	Stokoe	Klima
/U		/H		H	H
/bent-U					
/crvd-U					Ü
/V		/V		V	V
/crvd-V		/Bent V			ÿ
/W		/W		W	W
/crvd-W					

BU Gloss	Image	ASL Dictionary	Rendering	Stokoe	Klima
/X		/X		X	X
/Y		/Y		Y	Y
/I		/I			G <sub>1</sub>
/bent-I					
/Horns		/I-I			⌣
/O/2-Horns					
/bent-Horns					

BU Gloss	Image	ASL Dictionary	Rendering	Stokoe	Klima
/3		/3		3	3
/U-L					
/crvd-3		/Bent 3			3̂
/4		/4			5 <sub>4</sub>
/5		/5		5	5
/crvd-5					5̂
/5-C		/Bent 5			5̂

BU Gloss	Image	ASL Dictionary	Rendering	Stokoe	Klima
<i>/5-C-tt</i>					
<i>/5-C-L</i>					
<i>/6</i>					
<i>/7</i>					
<i>/8</i>		<i>/8</i>			8
<i>/cocked-8</i>					
<i>/open-8</i>		<i>/Open 8</i>		⌘	⌘



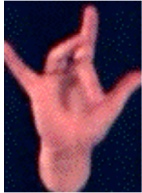





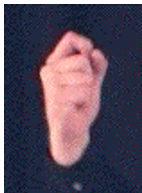

BU Gloss	Image	ASL Dictionary	Rendering	Stokoe	Klima
<i>/25</i>					
<i>/9</i>					
<i>/open-9</i>					
<i>/10</i>		<i>/Open A</i>			A*
<i>/fanned-flat-O</i>					
<i>/cocked-S</i>					
<i>/cocked-U</i>					

Table A.1: A listing of handshape primes recognized by different linguists.



# Appendix B

## Fingerspelling Word Lists

The word list presented below were taken directly from [38]. The same sets of words were used in fingerspelling studies by Kim et al. [41, 42, 43]. The words are listed here by column in the order they were presented to participants.

### B.1 Proper Nouns

- beijing
- afghanistan
- matt
- josh
- aberdeen
- everglades
- gary
- toby
- exxon
- jason
- naperville
- jimmy
- flossmoor
- owen
- tallahassee
- kelly
- yellowstone
- libya
- venice
- caribbean
- ann
- joe
- danny
- franklin
- yosemite
- angelica
- amy
- xavier
- venezuela
- alcapulco
- africa
- leo
- inglewood
- fred
- xerox
- mediterranean
- el salvador
- alan
- pam
- camilla
- paraguay
- zoe
- botswana
- william
- don
- quentin
- tanzania
- tom
- cleveland
- quincy
- moscow
- francesca
- columbus
- excel
- himalaya
- viv
- tiffany
- sam
- tobias
- mississippi
- zack
- san francisco
- nic
- atlantic
- alexander
- carl
- debbie
- kate
- russ
- quotation
- rangerover
- cameroon

- tokyo
- izzy
- will
- george
- greg
- skokie
- mary
- naomi
- bea
- giordano
- mongolia
- lexus
- apraxia
- mexico
- finn
- jacqueline
- chris
- sara
- john
- mia
- rita
- bill
- scotland
- oak park
- enrique
- felix
- mauritania
- gayle

## B.2 Nouns

- axis
- juice
- xylophone
- strawberry
- basil
- neighborhood
- earthquake
- lamb
- pony
- firewire
- bass
- turquoise
- waffle
- zebra
- claw
- cliff
- tulip
- executive
- beef
- equal
- wing
- life
- staff
- oxen
- liquid
- dogfight
- luggage
- dinosaur
- fir
- expectation
- seed
- appetizers
- rest
- jade
- flour
- fern
- twizzlers
- van
- cabin
- quiz
- quarter
- notebook
- oxygen
- axel
- taxi
- fanbelt
- furniture
- xenophobia
- weed
- headlight
- vacuum
- quicksand
- square
- sun
- flea
- quilt
- jawbreaker
- cliffhanger
- glue
- quarry
- expo
- mustang
- queen
- xenon
- box
- ink
- jewelry
- sauce
- squirrel
- question
- silk
- grape
- oval
- quantity
- spice
- material
- carp
- gravity
- asphyxiation
- riddle
- findings
- stool
- boo
- yard
- instrument
- aquarium
- fanny
- softserve
- sequel
- family
- deck
- spruce
- cadillac
- xmen
- expert
- herb
- mitten
- campfire
- ataxia
- windshield

## B.3 Non-English Words

- kerul
- feleseg
- informacja
- huomenta
- egyenesen
- hei
- kto
- piec
- belyeg
- elnezeit
- procvicovat
- zopakovat
- felkelni
- hyvaa
- juna
- mina
- tancolni
- nogi
- viisi
- neni
- jsou
- hogyan
- zгода
- maanantai
- surgos
- yksi
- sina
- jegy
- kuusi
- dlaczego
- zdrowie
- prekladatel
- huone
- miluji
- ferfi
- zizen
- ole
- penzvaltas
- powazaniem
- potrebuji
- hlad
- powaznie
- nigdy
- igek
- dnia
- zgubilam
- pocalujmy
- missa
- opravdu
- igen
- dekuji
- moc
- czesc
- puhu
- zyc
- vitej
- tuhat
- kde
- anteeksi
- onko
- rado
- nerozumim
- zobaczenia
- kahdeksan
- cie
- blahopreji
- axon
- voitte
- hol
- rano
- rendorseg
- usta
- toistekan
- wlosy
- navstivil
- nelja
- mennyibe
- palyadvar
- itt
- pospeste
- kaksi
- ahoj
- daj
- siusiu
- korhaz
- chvilke
- spotykac
- szia
- lentokentta
- vcera
- rakastan
- paljonko
- kolik
- utca
- koszi
- csokifagyit
- fiu
- nowych
- przepraszam
- pojd



# Appendix C

## Data Collection Guide

The completion of this work led to many practical lessons for working towards ASL recognition and translation. In this section, we will highlight useful information and procedures useful for initiating investigations into sign recognition.

### C.1 Linguistic Goals

A common issue in research related to ASL recognition is that language recognition is actually an incidental goal. Researchers are interested in hand tracking or a particular classification technique and ASL merely provides a defined sets of poses with a straightforward justification. This is why the 24 static ASL alphabet signs get tested so often despite being an insufficient set for any ASL parameter. In Section 2.5 and Appendix A we have articulated feature requirements for ASL sign recognition and compiled lists of the sublexical parameters recognized by different linguists. While not all the parameters are comprehensively defined, these sections provide guidance about the scope at which research aimed at pushing forward sign recognition should be conducted.

#### C.1.1 Static vs. Continuous Data

There is a large and important distinctions between the data collected in the handshape study in Chapter 4 and the fingerspelling study in Chapter 5. While both used a hand tracking approach designed to work in real-time, the nature of the handshape study did not really provide information about the utility of the approach for real-time sign recognition. The only way to do so, is to record natural signing sequences as was done in the fingerspelling study. This way, the hand tracking algorithm can be evaluated in its ability to respond sufficiently to naturally transitions.

#### C.1.2 Variations

Another advantage to recording data sequences (as opposed to isolated signs) is that real-world coarticulation effects can be recorded. Even amongst a set of signs as constrained as ASL fingerspelling, we observed variations in sign formations. Section 5.4 discusses handshape variations

and coarticulation effects that were observed in our study. ASL has many regional sign variations and most research has focused on relatively limited datasets. Appropriately interpreting individual and regional signing variations will be a necessary step for recognition systems.

## **C.2 Datasets and Collection**

This work focused on the use of depth data for sign recognition which necessitated the collection of new data. However, approaches designed to work with standard definition video can benefit from a number of databases of signing data.

### **C.2.1 Datasets**

For ASL, Boston University’s American Sign Language Lexicon Video Dataset [62] is the largest dataset available. Purdue University also has a collection of sign videos available in their RSL-SLLL American Sign Language Database [108]. Amongst other languages, Hamburg University likely has collected the most extensive database with some 500 hours of German Sign Language (Deutsche Gebrdensprache or DGS) videos [1].

### **C.2.2 Collecting Data**

Collecting sign data can be challenging. The open source hand tracking approach developed by Tkach et al. was invaluable to our work. A follow up algorithm for automatically fitting hand models (see Section 4.3.4) was recently released and looks to improve results significantly [76]. A discussion of the technical setup used in our data recordings is provided in Section 5.2. Issues relating to specific participants is discussed in Section 5.4.2 and general hand tracking error sources are discussed Section 5.6.1.



# Bibliography

- [1] DGS-Corpus. URL <https://www.sign-lang.uni-hamburg.de/dgs-korpus/>. 2.3.2, 3.1.2, C.2.1
- [2] Leap Motion. 3.1.3
- [3] OpenNI Programmer's Guide. URL <https://structure.io/openni>. 6.3.2
- [4] VISICAST Home Page. URL <http://www.visicast.co.uk/>. 2.3.2
- [5] eSIGN. URL <https://www.sign-lang.uni-hamburg.de/esign/>. 2.3.2
- [6] SignAloud, 2016. URL <http://www.washington.edu/news/2016/04/12/uw-undergraduate-team-wins-10000-lemelson-mit-student-prize-for-gloves>. 3.1.1
- [7] Epameinondas Antonakos, Anastasios Roussos, and Stefanos Zafeiriou. A survey on mouth modeling and analysis for Sign Language recognition. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015*, 2015. doi: 10.1109/FG.2015.7163162. 3.2.4, 6.2
- [8] Stephanie a Baker, William J Idsardi, Roberta Michnick Golinkoff, and Laura-Ann Petitto. The perception of handshapes in American sign language. *Memory & cognition*, 33(5):887–904, 2005. ISSN 0090-502X. doi: 10.3758/BF03193083. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2730958&tool=pmcentrez&rendertype=abstract>. 2.4
- [9] Charlotte Baker-Shenk and Dennis Cokely. *American Sign Language*. Gallaudet University Press, Washington, DC, 1980. 1.1, 2.2.3, 2.10, 2.3.1, 3.2
- [10] Barrust. PySpellChecker. 5.5.3
- [11] Robbin Battison. Analyzing Signs. In *Lexical Borrowing in American Sign Language*, pages 19–58. Linstok Press, Silver Spring, MD, 1978. 2.2.1, 2.2.2, 2.2.2, 2.2.2, 2.2.2, 2.2.2, 2.5.2, 4.2.1, 4.3.3
- [12] C. Fabian Benitez-Quiroz, Kadir Gökgöz, Ronnie B. Wilbur, and Aleix M. Martinez. Discriminant features and temporal structure of nonmanuals in American sign language. *PLoS ONE*, 9(2):25–27, 2014. ISSN 19326203. doi: 10.1371/journal.pone.0086268. 6.2.1
- [13] Mary Elizabeth Bonham. *English to ASL Gloss Machine Translation*. M.a., Brigham Young University, 2015. 2.3.4
- [14] Helene Brashear, Valerie Henderson, Kwang-Hyun Park, Harley Hamilton, Seungyon Lee, and Thad Starner. American sign language recognition in game development

- for deaf children. *Proceedings of the 8th international ACM SIGACCESS conference on computers and accessibility*, pages 79–86, 2006. doi: 10.1145/1168987.1169002. URL <http://ncsu.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwVZ0xDgIxDAQjehokqPlApMRxEqdGnHjAfcBOnPIq{ }q{ }zIQp4gaudXVle2b.3.1.2, 3.1.3>
- [15] Diane Brentari. Modality differences in sign language phonology and morphophonemics. In *Modality and Structure in Signed and Spoken Languages*, pages 35–64. 2002. ISBN <http://dx.doi.org/10.1017/CBO9780511486777.003>. 2.2.1
- [16] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. 2016. ISSN 10636919. doi: 10.1109/CVPR.2017.143. URL <http://arxiv.org/abs/1611.08050>. 3.2.4
- [17] Monica Carfagni, Rocco Furferi, Lapo Governi, Michaela Servi, Francesca Ucheddu, and Yary Volpe. On the Performance of the Intel SR300 Depth Camera: Metrological and Critical Characterization. *IEEE Sensors Journal*, 17(14):4508–4519, 2017. ISSN 1530437X. doi: 10.1109/JSEN.2017.2703829. 5.4.2
- [18] H Cheng, L Yang, and Z Liu. A Survey on 3D Hand Gesture Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1, 2015. ISSN 1051-8215. doi: 10.1109/TCSVT.2015.2469551. 3.1.3, 3.1.3, 4.1
- [19] Helen Cooper, Brian Holt, and Richard Bowden. Sign language recognition. *Visual Analysis of Humans*, (231135):539–562, 2011. doi: 10.1007/978-0-85729-997-0. URL <http://dx.doi.org/10.1007/978-0-85729-997-0{ }27>. 2.4, 3.2.4, 3.2.4, 5.3
- [20] CyberGloveSystems. CyberGlove Systems, 2015. URL <http://www.cyberglovesystems.com/sites/default/files/CyberGloveIII{ }MoCap{ }Glove{ }System{ }Brochure.pdf>. 3.1.1
- [21] Mark Dilsizian, Polina Yanovich, Shu Wang, Carol Neidle, and Dimitris Metaxas. A New Framework for Sign Language Recognition based on 3D Handshape Identification and Linguistic Modeling. pages 1924–1929, 2013. 3.1.2, 3.2.1, 7.1
- [22] Cao Dong, Ming C Leu, and Zhaozheng Yin. American Sign Language Alphabet Recognition Using Microsoft Kinect. pages 44–52, 2015. 4.2.2
- [23] R. Elliott, J. R. W. Glauert, J. R. Kennaway, and I. Marshall. The development of language processing support for the ViSiCAST project. *Proceedings of the fourth international ACM conference on Assistive technologies - Assets '00*, pages 101–108, 2000. doi: 10.1145/354324.354349. 2.3.2
- [24] Ralph Elliott, Javier Bueno, Richard Kennaway, and John Glauert. Towards the integration of synthetic SL animation with avatars into corpus annotation tools. *4th Workshop on the ...*, 2010. URL <http://www.researchgate.net/publication/228963928{ }Towards{ }the{ }Integration{ }of{ }Synthetic{ }SL{ }Animation/file/79e4150bb4df676713.pdf>. 2.3.2
- [25] Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, and Xander Twombly.

- Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 108(1-2):52–73, 2007. ISSN 10773142. doi: 10.1016/j.cviu.2006.10.012. 3.1.2, 4.1, 4.1, 4.1, 4.3.3
- [26] Gaolin Fang, Wen Gao, and Debin Zhao. Large-vocabulary continuous sign language recognition based on transition-movement models. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 37(1):1–9, 2007. ISSN 10834427. doi: 10.1109/TSMCA.2006.886347. 3.1.1
- [27] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater, and Hermann Ney. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. *International Conference on Language Resources and Evaluation*, pages 3785–3789, 2012. URL <http://www.lrec-conf.org/lrec2012>. 1.2, 2.3.4, 2.4, 3.2.4, 7.2
- [28] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the Sign Language Recognition and Translation Corpus. *Lrec 2014*, (May):1911–1916, 2014. URL <http://www.lrec-conf.org/proceedings/lrec2014/summaries/585.html>. 3.1.2
- [29] Wen Gao, Gaolin Fang, Debin Zhao, and Yiqiang Chen. Transition Movement Models for Large Vocabulary Continuous Sign Language Recognition. 2004. 3.1.1, 3.2.3, 7.1.3
- [30] Paul Goh and Eun-J Holden. Dynamic Fingerspelling Recognition Using Geometric and Motion Features. In *IEEE Conference on Image Processing*, pages 2741–2744, 2006. ISBN 1424404819. 5.1
- [31] Raymond G. Gordon, editor. *Ethnologue: Languages of the World*. SIL International, 15th edition, 2005. 2.1.3
- [32] Gerilee Gustason. Signing Exact English. In *Manual Communication: Implications for Education*, pages 108–128. 1990. 2.1.3
- [33] Thomas Hanke. HamNoSys - Representing sign language data in language resources and language processing contexts. *LREC 2004, Workshop proceedings: Representation and processing of sign languages.*, pages 1–6, 2004. URL <http://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt{ }pdf/HankeLRECSLP2004{ }05.pdf>. 2.3.2
- [34] Jonathan Henner, Leah C Geer, and Diane Lillo-martin. Calculating Frequency of Occurrence of ASL handshapes. pages 1–4, 2013. 2.2.2
- [35] Hermitdave. FrequencyWords. 5.5.3
- [36] Takeo Kanade and Jeffrey F Cohn. Comprehensive database for facial expression analysis. *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53, 2000. ISSN 0-7695-0580-5. doi: dx.doi.org/10.1109/AFGR.2000.840611. URL <http://www.pitt.edu/{~}jeffcohn/biblio/Cohn-Kanade{ }Database.pdf>. 3.2.4
- [37] Byeongkeun Kang. Real-time Sign Language Fingerspelling Recognition using Convolutional Neural Networks from Depth map. pages 136–140, 2015. 4.2.2

- [38] Jonathan Keane. *Towards An Articulatory Model of Handshape: What Fingerspelling Tells Us about the Phonetics and Phonology of Handshape in American Sign Language*. PhD thesis, The University of Chicago, 2014. B
- [39] Jonathan Keane, Diane Brentari, and Jason Riggle. Coarticulation in ASL fingerspelling. *Proceedings of the North East Linguistic Society, No 42.*, 2013. URL <http://pubs.jonkeane.com/pdfs/Keane2012aa.pdf>. 5.1, 5.2, 5.4
- [40] Cem Keskin, Furkan Kiraç, Yunus Emre Kara, and Lale Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7577 LNCS(PART 6):852–863, 2012. ISSN 03029743. doi: 10.1007/978-3-642-33783-3\_61. 4.2.2
- [41] Taehwan Kim, Karen Livescu, and Gregory Shakhnarovich. American sign language fingerspelling recognition with phonological feature-based tandem models. *2012 IEEE Workshop on Spoken Language Technology, SLT 2012 - Proceedings*, pages 119–124, 2012. doi: 10.1109/SLT.2012.6424208. 5.1, B
- [42] Taehwan Kim, Greg Shakhnarovich, and Karen Livescu. Fingerspelling recognition with semi-markov conditional random fields. *Proceedings of the IEEE International Conference on Computer Vision*, pages 1521–1528, 2013. ISSN 1550-5499. doi: 10.1109/ICCV.2013.192. 5.1, B
- [43] Taehwan Kim, Weiran Wang, Hao Tang, and Karen Livescu. Signer-Independent Fingerspelling Recognition with Deep Neural Network Adaptation. pages 6160–6164, 2016. 5.1, B
- [44] Edward Klima and Ursula Bellugi. *The Signs of Language*. Harvard University Press, 1979. 2.2.1, 2.2.2, 2.4, 2.5.2, 4.3.1, 4.3.3, 7.2, A
- [45] Byoung Ko. A Brief Review of Facial Emotion Recognition Based on Visual Information. *Sensors*, 18(2):401, 2018. ISSN 1424-8220. doi: 10.3390/s18020401. URL <http://www.mdpi.com/1424-8220/18/2/401>. 3.2.4
- [46] Philipp Koehn. Europarl : A Parallel Corpus for Statistical Machine Translation. *MT Summit*, 11:79—86, 2005. ISSN 9747431262. doi: 10.3115/1626355.1626380. URL <http://mt-archive.info/MTS-2005-Koehn.pdf>. 2.3.4
- [47] Oscar Koller, Hermann Ney, and Richard Bowden. Read my lips: Continuous signer independent weakly supervised viseme recognition. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8689 LNCS(PART 1):281–296, 2014. ISSN 16113349. doi: 10.1007/978-3-319-10590-1\_19. 3.2.4
- [48] Oscar Koller, Hermann Ney, and Richard Bowden. Deep Hand : How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3793–3802, 2016. 3.1.2, 3.2.1, 4.2.2, 7.1
- [49] Alina Kuznetsova, Laura Leal-Taixe, and Bodo Rosenhahn. Real-Time Sign Lan-

- guage Recognition Using a Consumer Depth Camera. *2013 IEEE International Conference on Computer Vision Workshops*, pages 83–90, 2013. doi: 10.1109/ICCVW.2013.18. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6755883>. 4.2.2
- [50] Gabriele Langer, Reiner Konrad, Thomas Hanke, Gabriele Langer, Thomas Troelsgård, and Jette Kristoffersen. Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing. *LREC - 7 th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, (may):143–152, 2016. 3.1.2
- [51] Kehuang Li, Zhengyu Zhou, and Chin-hui Lee. Sign Transition Modeling and a Scalable Solution to Continuous Sign Language Recognition for Real-World Applications. *ACM Transactions on Accessible Computing*, 8(2):1–23, 2016. ISSN 19367228. doi: 10.1145/2850421. URL <http://dl.acm.org/citation.cfm?doid=2878628.2850421>. 3.1.1
- [52] Scott K. Liddell. *American Sign Language Syntax*. Mouton, 1980. ISBN 9027934371, 9789027934376. 2.2.1, 2.2.3, 2.5.2, 2.5.2
- [53] Scott K. Liddell and Robert E. Johnson. American Sign Language: The Phonological Base. *Sign Language Studies*, 1064(1):195–277, 1989. ISSN 1533-6263. doi: 10.1353/sls.1989.0027. URL <http://muse.jhu.edu/content/crossref/journals/sign{ }language{ }studies/v1064/64.liddell.html>. 2.2.1, 2.2.1, 2.3.3, 5.3, 5.3.3
- [54] Stephan Liwicki and Mark Everingham. Automatic recognition of fingerspelled words in british sign language. *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, (iv):50–57, 2009. ISSN 2160-7508. doi: 10.1109/CVPR.2009.5204291. 5.1
- [55] Barbara Loeding and Sudeep Sarkar. Progress in Automated Computer Recognition of Sign Language. *Lecture Notes in Computer Science*, (July):1079–1087, 2004. ISSN 0302-9743. doi: 10.1007/978-3-540-27817-7. 4.3.3
- [56] Rajesh B Mapari. Real Time Human Pose Recognition Using Leap Motion Sensor. *IEEE Conference on Research in Computer Intelligence and Communication Networks*, pages 323–328, 2015. doi: 10.1109/ICRCICN.2015.7434258. 3.1.3
- [57] David McKee, Rachel McKee, Sara Pivac Alexander, Lynette Pivac, and Mireille Vale. Online Dictionary of New Zealand Sign Language. 23(July 2012):500–531, 2012. URL <http://nzsl.vuw.ac.nz>. 3.1.2
- [58] Dimitris Metaxas, Bo Liu, Fei Yang, Peng Yang, Nicholas Michael, and Carol Neidle. Recognition of Nonmanual Markers in American Sign Language ( ASL ) Using Non - Parametric Adaptive 2D - 3D Face Tracking. *Language Resources and Evaluation (LREC 12)*, pages 2414–2420, 2012. 3.1.2, 3.2.4, 6.2.3
- [59] Jean-Baptiste Michel. Quantitative Analysis of Culture Using Millions of Digitized Books. 5.5
- [60] C Neidle, N Michael, and J Nash. A method for recognition of grammatically significant

- head movements and facial expressions, developed through use of a linguistically annotated video corpus. *Proc. of 21st ESSLLI ...*, 2009. URL [http://128.197.26.3/asllrp/papers/Neidle\\_{\\_}ESSLLI\\_{\\_}2009.pdf](http://128.197.26.3/asllrp/papers/Neidle_{_}ESSLLI_{_}2009.pdf). 2.2.1, 3.1.2, 3.2.4
- [61] Carol Neidle. SignStream Annotation : Conventions used for the. *Gesture*, (11), 2002. 1.2, 2.4, 3.2.1, 3.2.4, 4.2.2, 5.4.2, 5.5, 7.1.3, 7.2, A
- [62] Carol Neidle, Ashwin Thangali, and Stan Sclaroff. Challenges in Development of the American Sign Language Lexicon Video Dataset ( ASLLVD ) Corpus. 2012. 3.1.2, 3.1.2, 4.2.1, C.2.1
- [63] Tan Dat Nguyen and Surendra Ranganath. Recognizing continuous grammatical marker facial gestures in sign language video. *Lecture Notes in Computer Science (including sub-series Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6495 LNCS(PART 4):665–676, 2011. ISSN 03029743. doi: 10.1007/978-3-642-19282-1\_53. 3.2.4
- [64] Tan Dat Nguyen and Surendra Ranganath. Facial expressions in American sign language: Tracking and recognition. *Pattern Recognition*, 45(5):1877–1891, 2012. ISSN 00313203. doi: 10.1016/j.patcog.2011.10.026. URL <http://dx.doi.org/10.1016/j.patcog.2011.10.026>. 6.2.1
- [65] Timothy F. O’Connor, Matthew E. Fach, Rachel Miller, Samuel E. Root, Patrick P. Mercier, and Darren J. Lipomi. The Language of Glove: Wireless gesture decoder with low-power and stretchable hybrid electronics. *PLoS ONE*, 12(7):1–12, 2017. ISSN 19326203. doi: 10.1371/journal.pone.0179766. 3.1.1
- [66] S.C.W. Ong and Surendra Ranganath. Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):873–891, 2005. ISSN 0162-8828. doi: 10.1109/TPAMI.2005.112. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1432718>. 3.2.4
- [67] K. Otiniano-Rodriguez and G. Camara-Chavez. Finger Spelling Recognition from RGB-D Information using Kernel Descriptor. In *Graphics, Patterns and Images*. IEEE, 2013. doi: 10.1109/SIBGRAPI.2013.10. 4.2.2
- [68] Carol Padden and Darline Clark Gunsauls. How the Alphabet Came to Be Used in a Sign Language. *Sign Language Studies*, 4(1):10–33, 2003. ISSN 1533-6263. doi: 10.1353/sls.2003.0026. 5
- [69] Ayush S Parashar. Representation and interpretation of manual and non-manual information for automated American Sign Language recognition. pages 1–70, 2003. doi: OCLC52832761. 3.1.2, 3.2.4, 5.5, 6.2.1, 7.2
- [70] Fabrizio Pedersoli, Sergio Benini, Nicola Adami, and Riccardo Leonardi. XKin: An open source framework for hand pose and gesture recognition using kinect. *Visual Computer*, 30(10):1107–1122, 2014. ISSN 01782789. doi: 10.1007/s00371-014-0921-x. 4.2.2
- [71] Roland Pfau and Josep Quer. Nonmanuals: Their grammatical and prosodic roles. *Sign Languages*, pages 381–402, 2010. doi: 10.1017/CBO9780511712203.018. 2.2.1

- [72] Tomas Pfister, James Charles, and Andrew Zisserman. Large-scale Learning of Sign Language by Watching TV. *Proceedings of the British Machine Vision Conference 2013*, pages 20.1–20.11, 2013. doi: 10.5244/C.27.20. URL <http://www.scopus.com/inward/record.url?eid=2-s2.0-84898465878&partnerID=tZOtx3y1>. 3.1.2
- [73] Leigh Ellen Potter, Jake Araullo, and Lewis Carter. The Leap Motion controller. *Proceedings of the 25th Australian Computer-Human Interaction Conference on Augmentation, Application, Innovation, Collaboration - OzCHI '13*, (February 2016):175–178, 2013. doi: 10.1145/2541016.2541072. URL <http://dl.acm.org/citation.cfm?id=2541016.2541072>. 3.1.3
- [74] Nicolas Pugeault and Richard Bowden. Spelling it out: Real-time ASL fingerspelling recognition. *Proceedings of the IEEE International Conference on Computer Vision*, pages 1114–1119, 2011. ISSN 1939-3539. doi: 10.1109/ICCVW.2011.6130290. 3.1, 4.2.2, 5.1, 5.6.1
- [75] Chen Qian. Realtime and Robust Hand Tracking from Depth. *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014. 4.1, 4.2a
- [76] Edoardo Remelli, Anastasia Tkach, Andrea Tagliasacchi, and Mark Pauly. Low-Dimensionality Calibration through Local Anisotropic Scaling for Robust Hand Model Personalization. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob(2):2554–2562, 2017. ISSN 15505499. doi: 10.1109/ICCV.2017.277. 7.1.3, C.2.2
- [77] Susanna Ricco and Carlo Tomasi. Fingerspelling recognition through classification of letter-to-letter transitions. In *Asian Conference on Computer Vision*, volume 5996 LNCS, pages 214–225, 2009. ISBN 3642122965. doi: 10.1007/978-3-642-12297-2\_21. 5.1
- [78] L Rioux-Maldague and P Giguere. Sign Language Fingerspelling Classification from Depth and Color Images Using a Deep Belief Network. *Computer and Robot Vision (CRV), 2014 Canadian Conference on*, pages 92–97, 2014. doi: 10.1109/CRV.2014.20. 4.2.2
- [79] Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. Subspace Constrained Mean-Shift. *Proceedings of the IEEE International Conference on Computer Vision*, (CIm): 1–19, 2009. ISSN 1550-5499. doi: 10.1109/ICCV.2009.5459377. URL [http://ieeexplore.ieee.org/xpls/abs/\\_all.jsp?arnumber=5459377](http://ieeexplore.ieee.org/xpls/abs/_all.jsp?arnumber=5459377). 6.2.2
- [80] Christoph Schmidt, Oscar Koller, Hermann Ney, Thomas Hoyoux, and Justus Piater. Using Viseme Recognition to Improve a Sign Language Translation System, 2013. 3.2.4
- [81] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unied Embedding for Face Recognition and Clustering. In *Computer Vision and Pattern Recognition*, pages 815–823, 2015. ISBN 9781467369640. doi: 10.1109/CVPR.2015.7298682. 3.2.4
- [82] Ann Senghas and Marie Coppola. Children Creating Language: How Nicaraguan Sign Language Acquired a Spatial Grammar. *Psychological science*, 12:323–328, 2001. 2.1.3
- [83] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David

- Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, Daniel Freedman, Pushmeet Kohli, Eyal Krupka, Andrew Fitzgibbon, and Shahram Izadi. Accurate, Robust, and Flexible Real-time Hand Tracking - Supplemental Material. *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 3633—3642, 2015. doi: 10.1145/2702123.2702179. 3.1.3, 7.1.3
- [84] Toby Sharp, Duncan Robertson, Jonathan Taylor, and Jamie Shotton. Accurate , Robust , and Flexible Real-time Hand Tracking : Supplementary Material. *CHI '15*, 2015. 4.1
- [85] Sutton Signwriting and Sutton Signwriting. Sutton SignWriting. 2016. 2.3.2
- [86] Samira Silva, William Robson Schwartz, and C Guillermo. Spatial Pyramid Matching for Finger Spelling Recognition in Intensity Images. pages 629–636, 2014. 4.2.2
- [87] Patricia Siple. Visual Constraints for Sign Language Communication. *Sign Language Studies*, 1019, 1978. 2.2.2
- [88] Cheri Smith, Ella Mae Lentz, and Ken Mikos. *Signing Naturally*. DawnSignPress, 2008. 2.2.1
- [89] Thad Eugene Starner and Alex Pentland. Visual Recognition of American Sign Language Using Hidden Markov Models. *Media*, pages 189–194, 1995. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.51.6538&rep=rep1&type=pdf>. 1.1, 3.1.2
- [90] William Stokoe, Dorothy Casterline, and Carl Croneberg. *No Title*. Gallaudet College Press, Silver Spring, MD, rev. editi edition, 1965. 2.5.2, A
- [91] William C. Stokoe and Marc Marschark. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of Deaf Studies and Deaf Education*, 10(1):3–37, 2005. ISSN 10814159. doi: 10.1093/deafed/eni001. 2.2, 2.2.1, 2.2.1, 2.2.1, 2.3.2
- [92] Jesus Suarez and Robin R Murphy. Hand Gesture Recognition with Depth Images: A Review. *IEEE RO-MAN*, 2012. ISSN 1944-9445. doi: 10.1109/ROMAN.2012.6343787. 4.1
- [93] Valerie Sutton. SignWriting For Sign Languages. URL <http://signwriting.org/>. 2.3.2
- [94] Andrea Tagliasacchi, Matthias Schroder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. Robust Articulated-ICP for Real-Time Hand Tracking. *Computer Graphics Forum*, 34(5):101–114, 2015. ISSN 14678659. doi: 10.1111/cgf.12700. 4.3.1
- [95] Hironori Takimoto, Seiki Yoshimori, Yasue Mitsukura, and Minoru Fukumi. Classification of hand postures based on 3D vision model for human-robot interaction. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pages 292–297, 2010. ISSN 1944-9445. doi: 10.1109/ROMAN.2010.5598646. 4.2.2
- [96] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, Arran Topalian, Erroll Wood, Sameh Khamis, Pushmeet Kohli, Shahram Izadi, Richard Banks, Andrew Fitzgibbon, and Jamie Shotton. Efficient and Precise Interactive Hand Tracking Through



- Joint, Continuous Optimization of Pose and Correspondences. *ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques*, 2016. ISSN 15577368. doi: 10.1145/2897824.2925965. URL <http://dx.doi.org/10.1145/2897824.2925965>. 3.1.3, 4.1, 4.2b, 7.1.3
- [97] Richard A. Tennant and Marianne G. Brown. *The American Sign Language Handshape Dictionary*. Gallaudet University Press, 2 edition, 2010. ISBN 978-1-56368-444-9. 4.3.1, 4.4, 4.3.2, A
- [98] Ashwin Thangali, Joan P Nash, Stan Sclaroff, and Carol Neidle. Exploiting phonological constraints for handshape inference in ASL video. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 521–528, 2011. ISBN 9781457703942. doi: 10.1109/CVPR.2011.5995718. 3.2.1, 4.2.2, 5.5, 7.2
- [99] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. Sphere-Meshes for Real-Time Hand Modeling and Tracking. *Proceedings of SIGGRAPH Asia, ACM Transactions on Graphics*, 35(6), 2016. 4.1, 4.2c, 4.1, 4.3.1, 4.3.4, 5.2, 5.2, 5.3.3, 6.1.2, 7.1
- [100] Carol Bloomquist Traxler. The Stanford Achievement Test , 9th Edition : National Norming and Performance Standards for Deaf and Hard-of-Hearing Students. *Journal of Deaf Studies and Deaf Education*, 5(4):337–348, 2000. 1.1
- [101] Dominique Uebersax, Juergen Gall, Michael Van Den Bergh, and Luc Van Gool. Real-time sign language letter and word recognition from depth data. *Proceedings of the IEEE International Conference on Computer Vision*, pages 383–390, 2011. doi: 10.1109/ICCVW.2011.6130267. 4.2.2
- [102] Clayton Valli, Ceil Lucas, and Kristin J. Mulrooney. *Linguistics of American Sign Language*. Press, Gallaudet University, 4th edition, 2005. 4.3.1
- [103] William Vicars. ASL University. 2.5, 2.6
- [104] C. Vogler and D. Metaxas. Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods. *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, 1:156–161, 1997. ISSN 1062-922X. doi: 10.1109/ICSMC.1997.625741. 2.2.2, 3.1.2
- [105] Christian Vogler and Dimitris Metaxas. Toward scalability in ASL recognition: Breaking down signs into phonemes. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1739:211–224, 1999. ISSN 16113349. doi: 10.1007/3-540-46616-9\_19. 3.2.3
- [106] Robert Y. Wang and Jovan Popović. Real-time hand-tracking with a color glove. *ACM Transactions on Graphics*, 28(3):1, 2009. ISSN 07300301. doi: 10.1145/1531326.1531369. 3.1.2
- [107] C. S. Weerasekera, M. H. Jaward, and N. Kamrani. Robust ASL Fingerspelling Recognition Using Local Binary Patterns and Geometric Features. *2013 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, 2013. doi: 10.1109/DICTA.2013.6691521. URL <http://www.scopus.com/inward/record.url?eid=2-s2.0-84893266000&partnerID=tZ0tx3y1>. 4.2.2

- [108] Ronnie Wilbur and Avinash C Kak. Purdue RVL-SLLL American Sign Language Database. Technical Report September, Purdue University, W. Lafayette, IN, 2006. C.2.1
- [109] Jacob O Wobbrock and Brad A Myers. Analyzing the Input Stream for Character- Level Errors in Unconstrained Text Entry Evaluations. 13(4):458–489, 2006. 5.5.1
- [110] Mao Ye, Qing Zhang, Liang Wang, and Jiejie Zhu. A Survey on Human Motion Analysis. pages 149–187, 2013. 1.1, 3.2.4, 4.1