Towards the Automatic Translation of American Sign Language

Brandon Taylor
Ph.D. Thesis Proposal
November 29th, 2016
Human-Computer Interaction Institute, Carnegie Mellon University

Committee:
Dan Siewiorek (Chair)
Anind Dey
Carolyn Rose
Roberta Klatzky
Asim Smailagic
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 benefitted greatly from decades of research into speech recognition and Statistical Machine Translation, ASL’s lack of aural and written components have limited exploration into automated translation of ASL.

Previous research efforts into sign language detection have met with limited success primarily due to inaccurately tracking handshapes. Without this vital component, research into ASL detection has been limited to focusing on isolated components of ASL or restricted vocabulary sets that reduce the need for accurate handtracking. However, improvements in 3D cameras and advances in handtracking techniques provide reasons to believe some of the technical sensing limitations may no longer exist. By combining state of the art handtracking techniques with ASL language modeling, there is an unexplored opportunity to develop a system capable of fully capturing ASL.

In this work, I propose to develop the first ASL translation system capable of detecting all five necessary parameters of ASL (Handshape, Hand Location, Palm Orientation, Movement, and Non-Manual Features). This work will build on existing handtracking techniques and explore the features that are best capable of discriminating the 40 distinct handshapes used in ASL. An ASL language model will be incorporated into the detection algorithm to improve sign detection. Finally, the system will output a form of transcribed ASL that will allow for the separation of sign detection and ASL-to-English language translation.
Table of Contents

1. Introduction................................................................................................................. 4

2. Background .................................................................................................................. 6
   2.1 Approaches to Sign Language Recognition ......................................................... 6
   2.2 Fingerspelling and Handshapes ........................................................................... 7
   2.3 ASL Linguistics ..................................................................................................... 8
   2.4 ASL Transcription ............................................................................................... 11
   2.5 ASL Transcription Corpuses .............................................................................. 13
   2.6 Sign Language Machine Translation .................................................................. 13

3. Proposed Work .......................................................................................................... 14
   3.1 Sensor Exploration for Hand Tracking ................................................................. 15
   3.2 Complete Handshape Classifier ......................................................................... 15
   3.3 Complete Five Parameter Model ....................................................................... 17
   3.4 Language Modeling ........................................................................................... 18
   3.5 Transcribed ASL Output .................................................................................... 19

4. Schedule ................................................................................................................... 20

5. Expected Contributions ............................................................................................ 21

6. References ................................................................................................................. 22
1. Introduction

Sign languages are a form of communication in which visual and gestural modalities are used to convey symbolic meanings. Despite the existence of 121 cataloged sign languages throughout the world [1], it was not until the 1960s that American Sign Language was widely recognized to be a linguistically distinct language [2]. Since that time, surveys have indicated that upwards of 500,000 individuals primarily communicate using ASL [3]. For many of these individuals, day-to-day interactions require a mixture of gestures and written communication in their non-native language.

For situations in which precise communication is required, the assistance of licensed interpreter is often required. As video capabilities on mobile devices have increased, the availability of Video Relay Services, mandated by the FCC, has provided far greater flexibility in accessing interpreter services [4]. However, at more than $5 per minute, these services do require substantial public funding [4]. Interpreters and relay services also introduce potential privacy and availability issues that could be mitigated by an Automated Sign Language Translation (ASLT) system.

There are two distinct challenges that would need to be solved for functional ASLT system. The first is a matter of accurately detecting signs. At a minimum this would require a system capable of detecting all five of parameters that form and distinguish signs in ASL: handshape, movement, location, orientation and non-manual signals. The system would need to operate continuously in near-real time and, ideally, it would generalize across a population of users. The second challenge would be a language translation problem to properly generate English sentences from signed input. Unfortunately, unlike most spoken languages, ASL does not have a natural written component, which has made data-intensive statistical machine translation approaches difficult. However, linguists have developed a variety of notation systems for ASL [5] and work on machine translation system has been explored using limited corpuses.

In this work, I propose to focus primarily on the problem of ASL sign detection. Technological advancements such as higher resolution 3D cameras and wearable EMG sensors have brought about improvements in real-time handtracking performance [6]–[8] that have yet to be applied to sign language translation efforts. By adopting these methods and focusing specifically on maximizing the discriminability of the finite set of handshapes used in ASL, it is likely that improved sign classification can be achieved.

In addition to technological sensor improvements, there are new ways to leverage linguistic information about ASL to improve modeling. It has been noted in multiple surveys of SLR studies that non-manual parameters have been under-examined [9], [10]. Work focusing on German Sign Language recognition has found that incorporating facial expression modeling improved sign detection accuracies [11].
Applying a similar approach to ASL detection may well offer improved accuracies as well as a necessary step towards completeness. Similarly, applying information about ASL’s structure offers potential avenues for improving performance. A simple example would be to incorporate an English language corpus to adjust sign probabilities when detecting sequences of fingerspelling.

Lastly, by generating an output not in English, but as an existing ASL notation form, such as ASL gloss, the system could leverage independently developments in ASL-English translation models. At present, ASL-English statistical machine translation models have been limited by the lack of robust corpuses of ASL annotations. However, focusing on an annotated ASL output could simplify the process of generating corpuses that would improve such models.
2. Background
As automatic speech recognition (ASR) systems have become increasingly robust and available, more efforts have been made to apply the successful techniques developed for ASR to the problem of automatically translating signed languages. However, a number of challenges, both technical and cultural, have prevented Sign Language Recognition (SLR) from progressing to a mature solution.

2.1. Approaches to Sign Language Recognition

Glove-based Systems
Glove-based systems have achieved the most impressive SLR results in terms of vocabulary size, with over 90% accuracy being obtained in continuous sign detection across more than 5000 Chinese signs [12]. However, such systems are both expensive and require the user to wear unnatural devices. Additionally, glove based systems are unable to capture non-manual parameters, thus preventing such a solution from ever scaling to entirely represent ASL.

Computer Vision
Some of the earliest explorations into SLR were conducted using computer vision techniques [13], [14]. However, even recent research has struggled to provide a robust, real-time solution that can adequately track handshapes against varying backgrounds and occlusions [15]. By using colored gloves or tags on subjects’ hands, researchers have been able to improve vision-based hand tracking accuracy [16], [17].

Depth Cameras
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 arbitrary camera angles [6], [7].

Sign Modeling
Given the success that Hidden Markov Models (HMM) have had in the field of Automatic Speech Recognition, it’s understandable that the approach would be adopted for SLR. Indeed, some of the most promising early work achieved results at the sentence level [13]. However, whereas speech consists of a single linear signal, sign languages include multiple signals (two hands, facial expressions, postures) operating simultaneously. This poses a problem. HMMs cannot simply be scaled to accommodate the phoneme structure of sign languages [10].

A number of approaches have been explored to get around this problem, but no clear consensus as to which holds the most promise has formed. Some of the approaches, such as instance based learning using k-Nearest Neighbors were shown to work in a particular case [19], but are unlikely to generalize well. Vogler and
Metaxas presented an approach using Parallel Hidden Markov Models that could be independently calculated and combined into a final output probability later [20]. Other variants on HMMs have been shown to work using glove-based systems across large vocabularies of thousands of signs [12], [21].

Evaluation Metrics
To date, however, most efforts towards SLR have focused on isolated signs or very restricted sets of sentences [10]. Metrics have typically focused on Word Error Rates or classification accuracy, though both have been criticized as not necessarily being a good metric of communication performance [22]. Often this is a direct result of the limitations of the sensing technology [23]-[25] preventing scalable solutions from being a possibility. In other cases, it’s a result of focusing on a subset of the problem (fingerspelling) that disregards the need for language understanding.

2.2. Fingerspelling and Handshapes
Fingerspelling is the process by which the alphabet of a spoken language is represented via signs. Primarily used for proper nouns and technical terms, fingerspelling is a commonly used aspect of ASL [26]. However, the broader utility of fingerspelling recognition alone is restricted due to the limitations of English fluency throughout much of the deaf community [27].

In ASL, 24 letters of the English alphabet are represented by static handshapes while two letters (‘z’ and ‘j’) involve dynamic gestures. Given the straightforward concept of representing English letters and the range of handshapes used in doing so, fingerspelling is an inviting test case for sign language recognition systems. Across sensing modalities (e.g. gloves, computer vision) fingerspelling detection is the most commonly explored aspect of sign language. While some research is explicitly focused on the problem of sign language recognition, often, fingerspelling alphabets are used as test cases for the accuracy of general hand tracking algorithms [28]. As a result, compared to other aspects of sign language recognition, there is a relative wealth of previous work with which to compare the effectiveness of a new approach [29].

Fingerspelling also provides researchers with an avenue to explore sign recognition without the difficulty of modeling, or even necessarily understanding, the grammar of ASL. By constraining sign sequences to fingerspelling of a known vocabulary, researchers have been able to improve the accuracy of individual alphabet signs [30], [31]. However, prior work has also shown accuracy rates that inversely correlate with vocabulary size. A system that could accurately detect a sequence of static, single-handed fingerspelling across an extensive English language corpus would be a valuable contribution.

While the English fingerspelling alphabet is a necessary subset of ASL signs, classifying the handshapes used to represent alphabet alone are not sufficient to recognize the entirety of ASL. For example, a classifier capable of detecting all the
handshapes of the English alphabet would be able to detect the number ‘9’ which happens to use the same handshape as the letter ‘f’, but many other digits would require additional classifiers. The exact number of handshapes necessary to represent all ASL signs is somewhat contested with researchers interested in variations in ASL and other sign languages sometimes marking up to 80 different handshapes [32]. Most estimates of handshapes typically used in ASL come in between 40 and 50 handshapes [33], with the Gallaudet University Press’s The American Sign Language Handshape Dictionary recognizing 40 basic handshapes [34].

2.3. ASL Linguistics

Non-manual Parameters
While the majority of work on SLR systems has focused on manual parameters, facial expressions and body postures are a necessary element of understanding ASL. Surveys of SLR research have noted that non-manual parameters have often been overlooked by researchers [9], [10]. However, available computer vision approaches to facial tracking are capable of recognizing the relevant features. How to make sense of the temporal nature of the features that may not align with manual features is being explored [35].

Mouth Morphemes
Mouth shapes are an important non-manual feature. The shape of the mouth can be the only distinguishing feature between signs, as with the signs for ‘need’ and ‘should’. As with many aspects of ASL linguistics, the precise categories of mouth morphemes and their exact meanings are somewhat disputed. However, there is no disputing that information conveyed by mouth shapes and movements are important for understanding both individual signs and higher level grammatical concepts [36].

Sentence Types
Facial expressions and head movements often distinguish different types of sentences. These meaningful expressions can also be transcribed in various ways, typically for studying or teaching ASL. The following meanings and transcriptions are adopted from Baker and Cokely’s text on ASL grammar [3].

Questions can be indicated by raising or furrowing the brow and marked by a ‘q’. Negation is typically indicated by shaking of the head and can be marked as ‘neg’. Direct eye contact with a possible frown or forward head movement can indicate a command, indicated by a ‘*’. Assertions can be made by nodding and tightening the lips, transcribed as ‘nod’. These expressions can be also be used in combinations, such as a negative question.

Table 1 demonstrates how different meanings can be ascribed to the same set of signs depending entirely upon the corresponding non-manual actions. The second column describes the signs using an Baker and Cokely’s ASL gloss notion. In each
case, the sign for ‘home’ (transcribed as 'HOME') is followed by the sign for ‘you’ (transcribed as ‘PRO.2’). The bracket indicates the timing of the non-manual action and the symbol following the bracket indicates a particular class of non-manual gesture. As can be seen, furrowing the brow or shaking one’s head can significantly alter the meaning of a sign.

<table>
<thead>
<tr>
<th>Type</th>
<th>Gloss form</th>
<th>English meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Declarative</td>
<td>{HOME PRO.2}nd</td>
<td>You are home.</td>
</tr>
<tr>
<td>Yes-No Question</td>
<td>{HOME PRO.2}q</td>
<td>Are you going home?</td>
</tr>
<tr>
<td>Negation</td>
<td>{HOME PRO.2}neg</td>
<td>You weren’t home.</td>
</tr>
<tr>
<td>Command</td>
<td>{HOME PRO.2}*</td>
<td>Go home.</td>
</tr>
</tbody>
</table>

Table 1. Non-manual effects on sentence types. The symbols describing the non-manual gesture follow the brackets, which indicate their duration. The ‘q’ symbol indicates a ‘yes-no question’, ‘nd’ indicates an ‘assertion’, ‘neg’ indicates a ‘negation’, and ‘*’ indicates ‘emphasis’.

**Inflecting**

Often the intensity of an adjective is expressed through exaggerated facial expressions or movements. For example, the difference between ‘tall’ and ‘very tall’ does not involve an additional sign for ‘very’. Instead, ‘very’ is indicated by a more dynamic gesture and intense facial expression.

![Image of modifying adjectives](image)

**Adverbs**

While some adverbs, such as indications of an action’s speed, are expressed through corresponding modification of sign movements, facial expressions are another key way to convey how actions take place. Figures 2 and 3 each show two people signing ‘Drive’. However, in Figure 2, the individuals are expressing the “mm” mouth shape, which denotes a ‘regular’ course of action. In Figure 3, the “th” mouth shape indicates something done carelessly.
Handshape Restrictions
Given that sign languages naturally developed, it is not surprising to observe that signs tend to accommodate the dynamics of the human body and visual system [3]. Rather than consisting of arbitrary movements, linguists have observed certain general rules that govern the formation of signs.
One example is the symmetry condition which posits that two handed signs, where the hands move independently, will usually have the same handshape, location and type of movement [33]. Another is the dominance condition which observes that if two hands have different handshapes, the non-dominant hand will usually be stationary and its handshape will typically be one of limited set of seven 'unmarked handshapes' [3].

2.4 ASL Transcription
While glossing (see Table 1) provides a way to represent sign language textually, it requires a prior knowledge of the semantic meaning of signs in order to record them. 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).

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 [5]. 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 [37]. However, while it's 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 [38].

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 [39]. Since being developed in the mid-80's, HamNoSys, has undergone multiple versions and been the defacto sign annotation system for a number of substantial investigations of signed languages in Europe [40]–[42]. More recently, HamNoSys has been adapted into an XML based format known as Signing Gesture Markup Language [43], [44].

To understand how these transcription systems differ from one another, it can be helpful to compare transcriptions of the same source material. Figures 4-6 each show a transcription of a signer beginning to tell the story of Goldilocks and the 3 Bears in Stokoe notation, SingWriting and HamNoSys, respectively [37]. The first
line first line in Figure 4 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 5, SignWriting presents the same breakdown in two columns. In HamNoSys, shown in Figure 6, each sign is described on its own line, with facial expressions marked on the right-hand side.

![Figure 4](image1.png)

Figure 4. An example of Stokoe notation transcribing a telling the story of Goldilocks in ASL. Each grouping of symbols represents an individual sign.

![Figure 5](image2.png)

Figure 5. The beginning of Goldilocks in SignWriting notation. The first column here describes the name of the story, with the second column setting the opening scene.

![Figure 6](image3.png)

Figure 6. The same passage of Goldilocks from Figures 4 and 5 transcribed in HamNoSys. The left column describes one sign per line while the right column indicates non-manual gestures.
2.5 ASL Transcription Corpuses
Though there are more than a hundred sign languages that have been recognized across the globe, political and social recognition of sign languages as true languages is a relatively recent phenomenon [1]. Linguists only began studying ASL in earnest in the 1960s [45]. Unlike written languages, which have been mass distributed for centuries, it was not until the 20th century that film video made indirect signed communication possible. Even then, it wasn’t until more recent advances in video compression and internet technology that made individual video communication widely available [46].

The ephemeral nature of signed communication prior to video recording technology has left the study of signed languages to lag behind other languages. Another direct result of the lack of distributed communication is that ASL naturally tends to have more regional variations. While a number of transcription systems have been developed to record sign languages for studying, they have typically found little use outside of academic circles. As a result, the available corpuses of transcribed sign data are severely restricted in comparison to most written languages.

Recently, though, various academic groups have proposed and begun collecting video corpuses to aid the study of signed languages [40], [32], [47], [48]. Often these corpuses include gloss annotations or translations of the data. While these corpuses do not offer direct benefits to Sign Language Recognition efforts, they are an important step towards building a better understanding of sign language linguistics. They also provide an avenue for independently developed sign language translation.

2.6 Sign Language Machine Translation
Even with the academic efforts aimed at collecting sign language corpuses, sources of sign language transcriptions are woefully lacking. This poses a challenge for researchers hoping to apply machine translation techniques to the problem of sign language translation. Open source machine translation systems (e.g. Moses [49]) have been used to model ASL to English translation [50]. However, unlike most written languages, which can provide millions of sentences for model training, the largest single corpus of transcribed sign language sentences consists of some 14,000 pairs [51].

There have been calls for standardization and centralizing repositories to address the limited amount of data available for Sign Language Machine Translation (SLMT) [52]. However, there are numerous challenges. To begin with, there are only a handful of research groups that have explored the SLMT [50]. Many of these groups are not working in the same languages. While techniques may be adapted from one sign language to another, data corpuses cannot be. Even when groups are working with the same language sets though, transcription systems are often vary from group to group, making cooperative efforts all the more difficult [53].
3. Proposed Work

The ultimate goal of this work is a complete ASL recognition system. The system diagram seen in Figure 7 shows the basic flow of the proposed system. Within the proposed system there are a number of subsystems that will be explained in more detail in the following sections.

![Figure 7. Proposed system diagram.](image)
3.1 Sensor Exploration for Hand Tracking
Choosing an appropriate set of sensors is an important aspect of any sign language recognition system. While glove-based systems have been shown to be capable of handling very large vocabulary [12], [21], sensor-equipped gloves are costly, constrict signer mobility, and provide no way capturing important non-manual gestures [10]. To avoid these problems, I intend to implement a handtracking algorithm using a combination of passive, color-coded gloves, the Kinect V2 depth camera and a Myo armband.

The Kinect V2 provides an a wide enough field of view to fully capture relevant gestures and postures, while still providing enough depth resolution to distinguish individual fingers [6], [7]. The Myo armband provides acceleration and gyroscopic data along with surface Electromyography (sEMG) measures from forearm muscles. These signals are complimentary to the Kinect data and have been shown to be useful in the classification of ASL [23], [54]. The use of a specially colored glove can help disambiguate fingers and provide more robust hand tracking [17].

While a purely vision-based solution would be preferable to requiring the user to wear either an armbands or even a passive glove, the purpose of using multiple sensing modalities in the early stages of this work is to analyze the tradeoffs between sensor modes. It's possible that specific ASL parameters can be more reliably detected using one sensing approach or another. Understanding the limitations of different sensors will be useful in system design decisions.

![Image of hand signs 'H' and 'U']

**Figure 8.** The signs for the letters 'h' and 'u' are distinguishable only by palm orientation

3.2 Complete Handshape Classifier
The next study I propose is to measure how handshape classification accuracies vary across the entire set of ASL handshapes. While many previous studies have focused on the subset of handshapes that compose the English alphabet, detection of the alphabet alone is not sufficient for capturing the entirety of ASL. In fact, the alphabet signs contain a number of redundant handshape configurations (e.g. ‘u’ and ‘h’) distinguished instead by palm orientation (See Figure 8).
Thus, instead of focusing on the 20 unique handshapes represented in the fingerspelling alphabet, I will design a classifier that can distinguish the 40 distinct handshapes in ASL (See Figure 9) [34]. To date, no handshape classifier has incorporated the entire set of ASL handshapes, thus no extant classifier is capable of recognizing the entirety of the ASL lexicon. By collecting data on all handshapes used in ASL, this work would provide an approach to handshape classification that can scale to the entirety of the language and better expose handshapes that have low levels of discriminability.

Figure 9. The entire set of distinct handshapes from which ASL is composed.

As with the previous section, the data collected during this study will include depth images, RGB images with colored gloves, sEMG signals and accelerometer and gyroscopic data from an armband. Such a data set will allow for cross-comparison of features to better understand which sensors provide information about different handshapes. For example, the handshapes for the English letters ‘m’ and ‘n’ are similar with a slight variation in thumb location (see Figure 10). It’s possible that colored finger tags would clearly distinguish the signs by the ordering of the color values. However, depth images alone may not be able to reliably discriminate between these similarly shaped signs.

Figure 10. Signs for the letters ‘M’ and ‘N’.

Measuring the relative discriminability of the entire set of ASL handshapes across different sensing modalities would provide valuable information about the reliability of the sensing modes. It would also provide insights into which technological improvements might offer greater or lesser gains.
3.3 Complete Five Parameter Model

To date, one of the biggest challenges yet to be addressed by SLR research is the synthesis of multi-modal information [9], [10]. Unlike speech recognition, which has a single channel of data, SLR requires the integration of at least 3 independent streams of data (dominant hand, non-dominant hand, non-manual gestures). Adding to the difficulty, these streams of data need not be synchronized [22] and can affect different aspects of communication (e.g. syntax or grammar) [2].

For example, concepts like ‘should’ and ‘must’ can depend on the co-articulation of manual gestures and facial expressions. Both signs are expressed by the same manual gesture (see Figure 11 & 12). While the concepts of ‘should’ and ‘must’ are similar, they are clearly distinct. However, an SLR system that does not take facial expressions into account will be entirely incapable of representing these as distinct concepts.

![Figure 11. The sign for ‘should’](image1)

![Figure 12. The sign for ‘must’](image2)
Despite the fundamental importance of non-manual gestures, no SLR system developed to date incorporates all five of the sign language parameters necessary for representing ASL. Some, such as glove-based systems, are fundamentally restricted by their sensor technologies [12], [55]. However, even amongst approaches that in theory could detect all parameters, no system that we are aware of has incorporated all five parameters.

In this work, incorporating all sign languages parameters will be a key contribution. In addition to making it possible to distinguish signs such as ‘should’ and ‘must’, the incorporation of five sign language parameters will make sure that my SLR approach is scalable to the entirety of ASL.

3.4 Language Modeling
Unlike spoken language, which is inherently a continuous temporal variation, not all signs are dynamic. For a subset of signs, a snapshot of an instant in time can fully capture the sign. Even signs for which hand movements are an integral part, the basic structure of the sign can typically be described by a single handshape coupled with a general motion (see Figure 13)

![Figure 13. On the left is the sign for 'B', a static pose. On the right is the sign for 'Blue', a dynamic gesture with the same hand shape as 'B'.](image)

However, signs that have changes in handshapes, double-handshape signs, are relatively rare and never have more than two different handshapes [33]. Thus, within a single sign, handshapes can be treated as isolated, static instances with transition periods during which the exact hand pose does not contain meaningful information.

In a way, this is similar to the co-articulation problem in speech recognition where particular speech sounds are changed by the transition to the next sound. However, unlike co-articulated speech phonemes, the degrees of freedom involved in transitioning from one handshape to another make the modeling of transitions untenable [55]. Instead, a common approach is to follow the studies of Liddell and Johnson [56] and split the signs into pauses and motions.
Additional linguistic modeling can provide prior information about the likelihood of various signs given a particular scenario. For an example, consider the double-handshape signs. Double-handshape signs are seemingly complex set of individual signs that include a transition between two handshapes. According to Battison’s research, there are 155 such signs in ASL [33]. Of these double-handshape signs, though, are majority are composed of transitions between a set of only seven handshapes [2]. Thus, when one of those seven handshapes is detected, its transition to another handshape is more likely to be included within double-handshape sign, whereas many other handshape transitions can be safely assumed to indicate the transition from one sign to another.

3.5 Transcribed ASL Output

In this work, I propose to develop a SLR system that separates the sign recognition problem from the language translation problem. In the simplest cases, such as isolated word detection or fingerspelling sequences, ASL has a near one-to-one correspondence with English. However, as sign sequences become more complex, the grammatical differences between ASL and English make direct translation more difficult. Thus, rather than trying to predict the equivalent English meaning of a sign sequence, I will adapt approaches from sign language transcription systems to record the meaningful structure of the signs. The translation between the transcribed signs and English can then be left as a secondary problem.

While statistical approaches to sign language translation have only been explored by a handful of researchers [50], it is reasonable to assume ASL to English translation will improve as the field matures. By far the biggest limitation to date in ASL to English translation is a lack of transcribed ASL corpuses. Given that there is no universal standard for sign language transcription, it’s not surprising that there are limited sets of annotated signs for use in statistical machine translation (SLMT) systems. Whereas most written languages can provide millions of exemplar sentences, the largest single corpus of annotated sign language to date consists of some 14,000 sentences of religious text translations produced by the LDS church [51]. Other extant corpuses focus on narrow domains such as the weather [53] or airport terminology [50].

By separating the sign language recognition problem from the sign language understanding problem, though, their individual developments could be mutually beneficial. Should recognition accuracy be advanced, it could greatly assist in the collection of annotated sign language for corpus development. At the same time, should further research into SLMT progress, it will be possible to incorporate a translation model onto the SLR system to provide direct ASL to English translation.
4. Schedule

December 2016: Finish handtracking and data collection algorithms.

Jan – Feb 2017: Run study on entire ASL handshape set. Determine accuracy and discriminability across sensor modalities (depth, color, sEMG). Run study on English word model effects on fingerspelling sequences.

April 2017: Submit Assets paper(s).

May 2017: Explore MOSES statistical machine translation. Collect ASL gloss corpus from other studies. Define ASL gloss output format to be used.

June-Aug 2017: Build detection model that incorporates all 5 ASL parameters and outputs a transcribed form of ASL that can be used with an ASL-English language translation model. Test effects of ASL language model on detection accuracy. Test population vs. trained user.

September 2017: Submit CHI paper(s).


Dec 2017: Defend.
5. Expected Contributions

This work will provide the first hand shape classifier trained to distinguish all 40 handshapes used in ASL. This is a necessary step to build a scalable system capable of representing the complete ASL lexicon.

This work will provide the first ASL recognition system that incorporates all five parameters of ASL. This will allow the system to offer complete phonetic representation of ASL signs.

This work will include an ASL language model capable of adjusting prior probabilities of both signs and sign components (e.g. specific handshapes) based on context.

This work will use an ASL transcription system output to separate the Sign Language Recognition problem from ASL to English translation problem. Such an approach will allow for the independent development of language translation models.
6. References


[34] R. A. Tennant and M. G. Brown, The American Sign Language Handshape


2011.
