Technical Report
1233

American Sign Language Recognition and Translation Feasibility Study

K. Brady
M.S. Brandstein
J.T. Melot
Y.L. Gwon
J. Williams
E. Salesky
M.T. Chan
P.R. Khorrami
N. Malyska

20 August 2018

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

This material is based upon work supported by the Department of the Air Force under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001.

This report is the result of studies performed at Lincoln Laboratory, a federally funded research and development center operated by Massachusetts Institute of Technology. This material is based upon work supported under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force.

Massachusetts Institute of Technology
Lincoln Laboratory

American Sign Language Recognition and Translation Feasibility Study

K. Brady
M.S. Brandstein
J.T. Melot
Y.L. Gwon
P.R. Khorrami
N. Malyska
Group 52

M.T. Chan
Group 45

J. Williams
University of Edinburgh

E. Salesky
Carnegie Mellon University

Technical Report 1233

20 August 2018

Lexington Massachusetts

This page intentionally left blank.

ABSTRACT

The development of a system for automatically and robustly translating between American Sign Language (ASL) and spoken English in real time on mobile devices holds the promise of enabling natural and spontaneous communication between Deaf ASL signers and English speakers anywhere and anytime. One key component of such a system is the automatic recognition of ASL signs, which is an active area of research in the academic community. A number of other system challenges remain in order to support deploying this technology on mobile devices that include addressing compute limitations and recognition robustness for acquired signals that are highly variable (e.g., sign variation, apparent pose angle) and poorly matched to existing training corpora. While several commercial companies have pursued the development of mobile translation systems, none of them has successfully commercialized such a system to date. This report investigates the technical feasibility of performing real-time translation between ASL and spoken English on mobile devices. An important aspect of this investigation is the identification of the key technical challenges in developing such a system and the development of a roadmap for addressing these challenges. In order to support the feasibility study detailed in this report, an extensive literature search has been completed, a number of rigorous experiments have been performed to characterize state of the art performance, and a prototype system has been developed.

This page intentionally left blank.

EXECUTIVE SUMMARY

This report summarizes a 12-month effort to evaluate the near-term feasibility of a portable device that will enable spontaneous, fluid communication between a Deaf American Sign Language (ASL) signer and an English speaker. Such a system would include components that are well understood (e.g., speech recognition and speech synthesis) and other components that are still largely in the active research domain. The focus of this report will be the unique ASL-related challenges fundamental to this system that are currently areas of active research, namely signing recognition, sign-sequence to/from English text mapping (i.e., ASL-English translation), and the problem of an ASL signing avatar.

The desired system would ultimately address the following requirements:

- Operate in a user-independent fashion across a broad range of environments
- Accommodate natural continuous signing and user/dialect variability
- Be applicable to a general range of conversational content as well as specific scenarios
- Provide easy-to-use interactivity in real time via a portable, unobtrusive device platform

Of course, the state of knowledge and technology will constrain these performance ideals in practice. However, with these overarching goals in mind, this report analyzes the current state of linguistic understanding of the ASL recognition and translation problems, the availability of appropriate data and software resources, the performance limits of the required machine learning algorithms, and the computational capacity and sensor capability of today's hardware technology. The functional aspects of the problem that are feasible are then outlined as well as a roadmap for the eventual deployment of an operational ASL-English communication system consistent with this overall vision.

In parallel with this overall feasibility analysis, the study included an additional effort to utilize currently available ASL resources to rapidly develop proof-of-concept systems with basic capabilities. Designed to illustrate the scope of present technology, three prototypes were produced and evaluated, each within a 3-month time frame that demonstrate fundamental elements of the overall concept. Specifically, the sub-problems addressed were:

- **Small-Vocabulary, Isolated-Signing:** Recognizes a set of 50 ASL signs produced by the user one at a time.
- **Continuous Fingerspelling:** Recognizes words spelled out by the user via a sequence of ASL alphabetic signs. The dictionary utilized incorporated over 10,000 entries.

- **Scenario-Specific, Continuous-Signing:** Recognizes sequences of ASL signs in the context of strong vocabulary and syntax constraints.

These prototypes employed commercial hardware and software supplemented by MIT Lincoln Laboratory-produced algorithms to achieve reasonable performance in near real-time conditions. While limited in functionality, they successfully applied published methods acquired from the scientific literature and leveraged existing technologies developed in other contexts. For example, these prototypes:

- Adapted deep neural network (DNN) image models trained for gesture recognition to effectively discriminate ASL handshapes
- Exploited dynamic recognition and language modeling methods from the acoustic speech recognition field to identify ASL sign sequences
- Utilized a commercially available image-depth camera to acquire video signals and track user body parts over time

The prototypes demonstrate the current (and limited) state of the field. In order to achieve the more ambitious requirements of the envisioned device, there are a number of significant theoretical and technical challenges that must first be addressed. These challenges are summarized as follows:

ASL Sign Recognition: There is a direct correspondence between the problems of sign recognition and speech recognition. While the signal modalities are clearly different (visual versus acoustic), once an appropriate feature parameterization is produced, sign recognition may be approached in a manner similar to speech recognition. As their input, speech recognizers employ a time-varying stream of feature vectors representative of the acoustic signal's spectral content. With sign recognition, the feature data is descriptive of the hands' shape, location, and motion and is supplemented by information representative of facial cues and other body gestures. Once a representative feature set is determined, the subsequent recognition methodology is agnostic to the underlying signal format. The three prototypes successfully leveraged this paradigm. In each case, the user's hands and face were tracked in the video stream and the hands converted to a feature vector using a DNN pre-trained for gestures. The face was not used for inferring the ASL sign or grammar, though this signal will be important for future systems. The pattern recognition component was coopted directly from speech recognizer technology.

Speech recognition has been studied for 50+ years (Juang et al 2005) and has only recently developed to the point where commercial systems are capable of accurately discerning casual, fluid speech in a user-independent setting. While improvements continue to be made, issues such as dialect variation, poor signal quality, and unexpected vocabulary still constitute active research areas. The dramatic advances seen today are the direct result of two factors: increased computational resources and a dramatic growth in the availability of appropriate training and evaluation data. The DNNs that are being employed to great effect rely on dedicated machinery and hundreds of hours of transcribed speech captured in environments

representative of their eventual use cases and operating conditions. The research community has addressed this data sufficiency challenge through concerted data collection efforts over several decades.

The path to utilizing speech recognition's algorithmic methods for the sign recognition problem is technically straightforward, but is severely hampered by a lack of data for training and evaluation. There is a relative paucity of available ASL datasets, and what does exist is poorly matched to the requirements of the desired communications system. These datasets are typically limited in their vocabulary range and subject sizes, and consist of isolated signs acquired under controlled (studio) conditions. They have proven useful for the purposes of developing these prototypes, but are not suitable for applications involving continuous signing, generalized content, user variability, or ambient environments.

One technical option to address recognition performance with limited available data is to leverage synthetic ASL datasets for augmenting the availability of training material. If successful, this would greatly reduce data collection requirements to that required for evaluation and seeding the synthesis process. Additionally, this low resource approach will have significant impact across a broad range of technical challenges faced by the USG that require the rapid response to potential threats where well-annotated datasets of sufficient size to support traditional approaches simply do not exist.

ASL-English Translation: The problem of converting a sequence of recognized ASL signs to an appropriate English sentence (and its inverse), would appear to be a relatively straightforward application of machine translation (MT) technology. Modern MT systems typically rely on either statistical or neural net-based methods trained on a corpus consisting of parallel phrases (usually in the tens to hundreds of thousands) in each language. In this case, the availability of usable data is more encouraging, as a number of video broadcasts include ASL interpretation along with the closed-captioning text. While these streams are not as closely coupled as typically found in the bilingual text case, it does offer a promising approach to developing an effective ASL-English translation in the near to mid-term time frame.

ASL Signing Avatar: Several incipient ASL signing avatars are available commercially and via academic institutions. While these vary in quality and scope, the underlying animation technology does exist and could be matured and adopted into the desired system (in some form) in the near-term.

Developing the envisioned ASL-English communication system would have been impractical even a few years ago. Recent breakthroughs in the speech processing and computer vision domains, combined with newly available mobile devices that increasingly include advanced imaging sensors and high-performance onboard computational ability, offer significant promise in addressing several important technical challenges. However, many technical and data-related issues remain and will need to be addressed in order to achieve the desired system functionality.

This page intentionally left blank.

ACKNOWLEDGEMENTS

The authors would like to thank our sponsor Dr. Patricia O'Neill-Brown and Group Leader Dr. Joseph Campbell for their support and extensive feedback.

This page intentionally left blank.

TABLE OF CONTENTS

This page intentionally left blank.

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS
(Continued)

LIST OF TABLES

1. OVERVIEW

This report details a 12-month study for exploring the feasibility of deploying a mobile real-time smartphone-based system to support a fluid conversation between a Deaf American Sign Language (ASL) signer and an English-speaking individual. The effort has the following components:

- A pre-development literature search to leverage best practices and relevant corpora in the design of a prototype system.
- Development of proof-of-concept prototype capabilities.

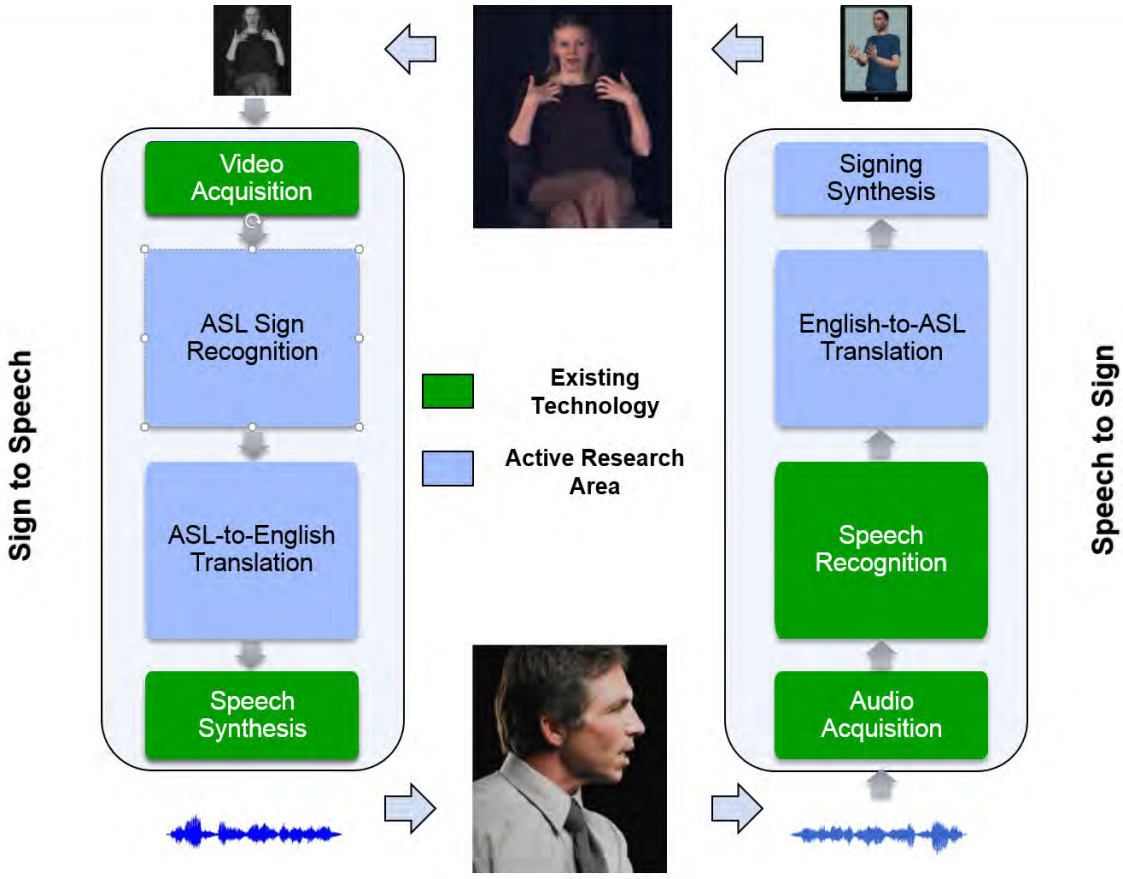

Figure 1. Nominal ASL-English translation system architecture. Photos courtesy of BU ASLLVD dataset.

Figure 1 highlights the various components of the envisioned fully-interactive system. The Sign-to-Speech track consists of modules for video signal acquisition, sign recognition, translation from sign sequences to English text, and finally, speech synthesis. The parallel Speech-to-Sign track reverses this process with audio capture and speech recognition, followed by English to sign translation and synthetic signing via an avatar. In practice, the system could function with a subset of the modules outlined below, using text-only output for instance. The green modules in the outline represent mature technologies (e.g., speech recognition, speech synthesis) which would be incorporated into the communication system, but will be treated as commodities for the purposes of this report.

Section 2 provides a brief background of the ASL challenge, including highlights from the literature search component and a discussion of ASL sign composition and variation. Section 3 provides an overview of the lessons-learned from the development of the proof-of-concept prototype capabilities as they pertain to the feasibility of deploying a real system. This will include supporting experimental results and a discussion of potential mitigation approaches. Section 4 describes a systematic data-driven approach for supporting the mitigation approaches outlined in Section 4. Section 5 provides some concluding remarks and a roadmap for future work, and Section 6 provides an extensive bibliography aggregated during the course of this work.

2. BACKGROUND

A broad range of estimates exist for the number of ASL signers in the United States that range from 250K to 500K (Mitchell et al 2005) to 500K to 2 million (Lane et al 1996). In spite of the sheer number of ASL signers, it is not clear that a commercial system for translating between ASL and spoken English in real time on mobile devices has been successfully brought to market. It is not clear whether this is due to the technical challenges of the problem or the inability to find a potentially profitable market. For an excellent review of the state of the art (circa 2014) of automatic sign language technology, there is a comprehensive review article (Sahoo et al 2014).

A number of companies have pursued the development, and in some cases the marketing, of ASL recognition capabilities.

- MotionSavvy (http://www.motionsavvy.com/) is ostensibly moving toward a commercial release of its tablet-based system that leverages a Leap Motion detector (two visible and three infrared sensors). The Leap Motion-based system does support a level of mobility that is not available with the other discussed systems. It is not clear, though, whether the company is still actively working towards bringing their product to market.

- SignAll (http://www.signall.us/) is developing a prototype for the commercial market, though it is not yet available for purchase, and few technical details are known beyond the fact that it uses three stationary standard video cameras, a depth camera, and a personal computer (SignAll 2018).

- KinTrans (http://www.kintrans.com/) is a Kinect-based system being developed for two-way communication between English speaking and ASL signing individuals.

- Microsoft Research (https://www.microsoft.com/en-us/research/blog/kinect-sign-language-translator-part-1/) was actively involved with developing a Kinect-based ASL translator (Chai et al 2013, Chen et al 2013) with its Chinese academic partners in 2013. There is no evidence that Microsoft has actively pursued this area since then, though their Chinese academic partners from this earlier work have continued to publish in this technical area (Chai et al 2016, Lin et al 2014, Wang et al 2015, Wang et al 2016, Yin et al 2015, Zhou et al 2016).

While some of these products have been marketed, it is not clear that there is an actual product for sale at the time this report was written for any of these systems.

ASL is an independent language that uses spatial handshapes, hand positions and trajectories, palm orientation, and side information in the face and head (Benitez-Quiroz et al, Nguyen et al 2012) to communicate. To be clear, it is not a gesture-based version of spoken English, and therefore the technical challenges are analogous to translating across distinct languages where one of them happens to be gesture-based. Structurally, it has greater similarity to Topic-Description languages such as Tagalog (Tatman 2017)

than to English's Subject-Verb-Object structure. It also has its own unique structures such as mouth morphemes (Bickford et al 2006) and spatial pronouns (Huenerfauth et al 2010).

Several video-based ASL 2D video corpora have been collected. The American Sign Language Lexicon Video Dataset (ASLLVD) (Athitsos et al 2008, Neidle et al 2008) has collected and annotated (e.g., start and stop times and handshapes) for an extensive set of about 3000 ASL signs from the Gallaudet Dictionary (Valli et al 2006). Four cameras (two frontal, one side view, and a frontal view of the face) were used for the data collections of one to six signers (single-session) per sign.

Another video corpora was collected at Purdue's Robotic Vision Laboratory (RVL) (Martinez et al 2002). This Purdue RVL-SLLL contains fourteen signers of data collected under two lighting conditions. The content includes fingerspellings of English letters and numbers (1-20), as well as a number of handshapes, signs in isolation, and several paragraphs.

An ASL motion capture dataset (Huenerfauth et al 2010, Lu et al 2012) has been collected, annotated, and analyzed to support technology improvements for generating ASL animations. The dataset supports the analysis of pronominal spatial reference points and inflected verbs. It includes extensive unscripted single-signer passages from nine native signers. An extensive set of motion capture sensors are used for the data collect, including CyberGloves, a head-mounted eye tracker, an inertial/acoustic tracker, and a bodysuit. Another sensor being explored for translating sign language is an armband (Metz 2016) that can "read" nerve impulses sent to the hand which can be used for inferring finger configuration.

There are not many canonical datasets available that are standards for evaluating systems on a defined protocol. The only one that has been found to date is the ASL Finger Spelling Dataset (Pugeault et al 2011) collected in the UK. It consists of 2D and depth images of static fingerspellings (no "j" or "z") collected by a Kinect sensor where the hands have been detected and extracted. Five users were collected for the primary dataset, and a secondary dataset of nine users was also collected for the depth sensor. Unfortunately, the hand images are of relatively low resolution. A number of other investigators have also baselined their performance against this dataset (Dong et al 2015, Garcia et al 2016, Keskin et al 2011, Kuznetsova et al 2013). A similar video image dataset of fingerspellings has been collected at Massey University (Barczak et al 2011) for five subjects under varying illumination, with more subjects potentially being released in the future. It is not as widely used as the UK dataset.

Another noteworthy dataset is the RWTH PHOENIX Weather (https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX/) dataset (Koller et al 2015) that contains continuous German sign language content with annotations collected on German public television over a three-year period (2009-2011) pertaining to weather.

The computer vision domain has increasingly been dominated by performance breakthroughs leveraging deeply learned models (Krizhevsky et al 2012) since 2012. Combined with the availability of large annotated datasets such as ImageNet (Deng et al 2009) and large compute capabilities, these deep neural network (DNN) systems have become the standard tool for feature extraction in image and video.

Many of these DNN models are available publicly. For example, Deep Hand (Koller et al 2016a), a DNN trained on a million hand images for recognizing handshapes for Danish Sign Language, is publicly available.

There are a number of other technical challenges relating to the automatic recognition of sign language. This includes the coarticulation problem (Channer 2012, Keane et al 2012) that occurs while transitioning from one sign to the next, and is analogous to the similar challenge in speech processing. Other specific problems include the handling of non-manuals (e.g., eyebrows, mouth morphemes, head tilt/squint), directional verbs, and positional signs (Cooper et al 2011, Tatman 2017). A number of excellent overviews of ASL linguistics are available (Stokoe 1960, Neidle et al 2000, Valli et al 2011).

In many ways the automatic sign language recognition (ASLR) domain challenges has many similarities to challenges encountered during earlier periods of the automatic speech recognition (ASR) domain. ASR technology has actually been around for decades, since Bell Labs researchers built a speaker-dependent digit recognizer in the 1950s (Juang et al 2005). Significant technical progress did not occur until the 1970s and 1980s, and the first successful commercial systems such as Dragon Dictate weren't released until the 1990s. During the 1970s and 1980s, investigators improved the core technology to address issues such as larger vocabulary sizes, speaker independence, conversational-style speech, and noisy acoustic environments—which are analogs to many of the core ASLR challenges. An important driver of performance was the development of benchmark evaluations by NIST in partnership with DARPA (Pallet et al 2003).

2.1 ASL SIGN VARIATION

Information is conveyed in ASL through a number of channels:

- Hand shape
- Hand orientation
- Hand location
- Hand movement
- Mouth shape and facial expression

The first four channels are generally considered phonemic in that changing them changes the word/concept being signed. For example, the signs “good” and “bad” differ only in the hand orientation used during part of the sign:

Figure 2. ASL signs for “good” and “bad” differ only in hand orientation. Photos courtesy of SigningSavvy.

The signs “paper” and “cheese” have the same handshape and orientation, but a different hand movement:

Figure 3. ASL signs for “paper” and “cheese” differ only in hand movement. Photos courtesy of SigningSavvy.

The signs “mother” and “father” have the same handshape, orientation, and movement, but a different location relative to the body:

Figure 4. ASL signs for “mother” and “father” differ only in sign location relative to the body. Photos courtesy of SigningSavvy.

Finally, the signs “white” and “like” differ only in their handshape:

Figure 5. ASL signs for “white” and “like” differ only in handshape. Photos courtesy of SigningSavvy.

To reliably distinguish signs, the ASL recognition system must be sensitive to differences in each of these phonemic features, yet be robust to signer and session variability. The system must also be robust to the narrative context of a sign, as a sign may be produced differently (with more dramatic movements, a different amount of repetitions, etc.) depending on exactly what is being said. Furthermore, there is a significant amount of dialectical variation in ASL, with some vocabulary items being signed completely differently in different cities. To be able to accurately perform recognition on ASL signers from different backgrounds, the ASL recognition system will eventually have to recognize multiple ways of signing the same word.

Mouthshape and facial expression also convey a great deal of information. In some cases, a certain mouthshape is an intrinsic part of the sign, but does not necessarily convey any special meaning. In other cases, mouthshape is used in conjunction with a sign to convey other information, such as adverbial information (indicating that something is happening to a great extent, that it is being done sloppily, etc.), or information such as size and distance. A few examples follow:

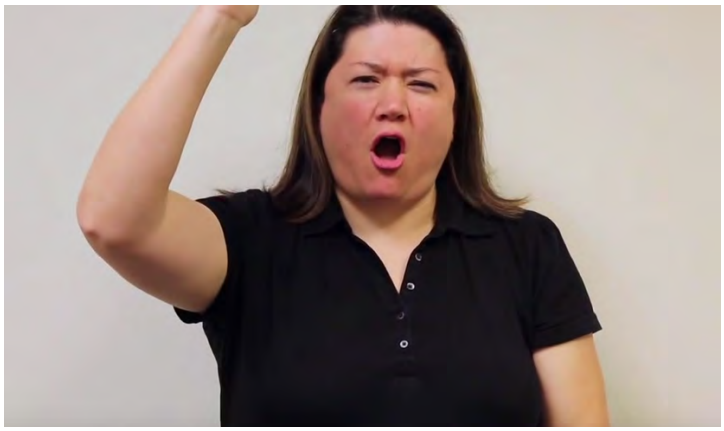

Figure 6. ASL "cha" mouthshape sometimes indicates that something is very big or tall. Photo courtesy of Ecampus ASL.

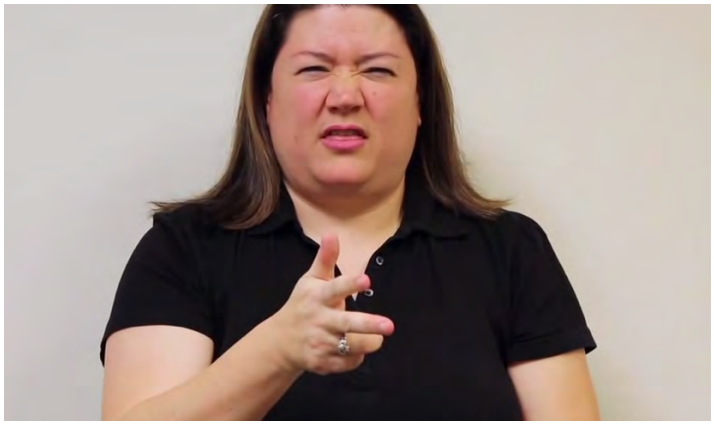

Figure 7. ASL "th" mouthshape can indicate that something is done clumsily or carelessly. Photo courtesy of Ecampus ASL.

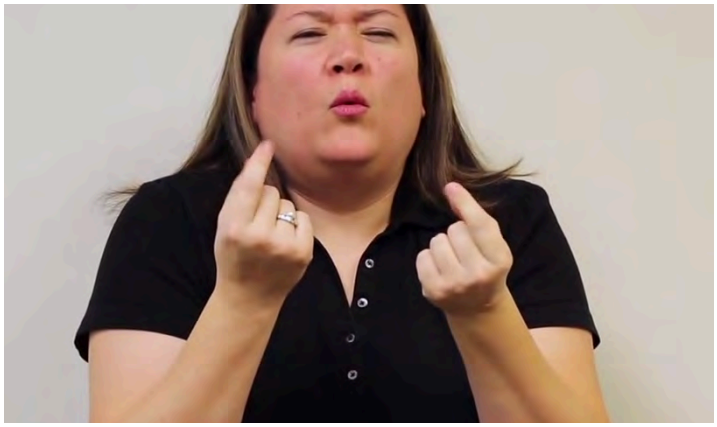

Figure 8. ASL "oo" mouthshape can be used to convey that something is very small. Photo courtesy of Ecampus ASL.

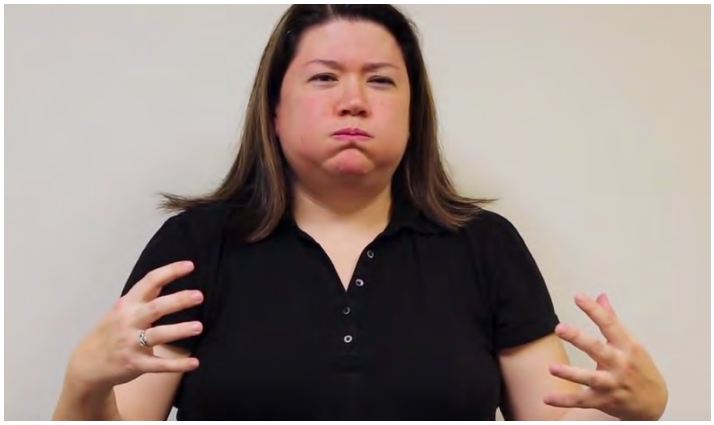

Figure 9. ASL "puffed" mouthshape sometimes means very fat, many, or long ago. Photo courtesy of Ecampus ASL.

Mouthshapes did appear in the vocabulary list and training data. After further progress with continuous sign recognition, integrating mouthshape recognition could eventually help us provide improved translations, although distinguishing meaningless from meaningful mouthshapes could be challenging.

3. PROTOTYPE SYSTEMS DEVELOPMENT AND LESSONS LEARNED

Over the past year, several ASL to English recognition prototypes have been developed, resulting in the proof of concept of a number of enabling technologies. This section provides an overview of these prototypes, reviews relevant performance results, and highlights promising paths forward for addressing technical gaps.

3.1 SIGN TO ENGLISH

In this subsection, an overview of the prototypes developed under this effort is provided, followed by a discussion of additional ASL experimental results, and finally several alternative technical approaches are highlighted, including the exploitation of synthetic ASL data and a DNN-based approach to the translation problem.

3.1.1 Prototypes Overview

The prototype hardware consists of a PC laptop and a Kinect 2.0 sensor. The Kinect includes both a high-quality imaging camera and a depth sensor. In addition to providing a 30 Hz video stream, the device outputs tracking data (specifically, 28 skeletal nodes) for up to six individuals. The laptop is integrated with the Kinect for real-time acquisition of the video-tracking data and performs all subsequent signal processing and user interface functions. For reasons of rapid development and testing, the prototype software is based on Matlab.

Isolated ASL Sign and Fingerspelling Prototype Overview

The user interface offers two distinct modes: discrete-word and fingerspelling sign recognition. The "Words" mode is illustrated in Figure 10. The left-hand panel displays buttons for the 50 signs currently in the prototype's vocabulary. By clicking one of these buttons, a video sample of the corresponding sign is displayed in the lower-mid "Demo" window. The intention of this functionality is twofold. For non-signers, it introduces users to the mechanics of the sign. With experienced users, it is designed to enforce a modicum of uniformity as to how the specific sign is performed when a number of options/variants are available. The upper-mid window displays the live video along with status labels. "Tracking" indicates that the Kinect has recognized the user and is monitoring skeletal data. The user initiates recording by raising his hands to the shoulders and then lowering them to the waist. The sign is then performed and the recording stops when the hands are again lowered. The "Control" panel will then offer several options: "Preview" the recording, "Save" the recording, or "Recognize" the intended sign. The "Status" window shows the results of the recognition procedure with ranking scores delineated by signing dynamics and hand-shape estimation criteria. The best-guess word has its corresponding button highlighted.

Figure 10. Prototype "Words" mode. Bottom picture courtesy of BU ASLLVD dataset.

With "Spell" mode (illustrated in Figure 11), the functionality is similar. While the left-hand panel has been replaced by letter illustrations, the procedure to initiate and stop recording is the same, as well as the resulting post-recording control functionality. Given the sequence of estimated letters, the recognizer now ranks the most likely words included in its 10,000+ entry lexicon. The results are displayed in the "Status" panel.

We now summarize the two recognition methods utilized in the Prototype and present some analysis of their effectiveness.

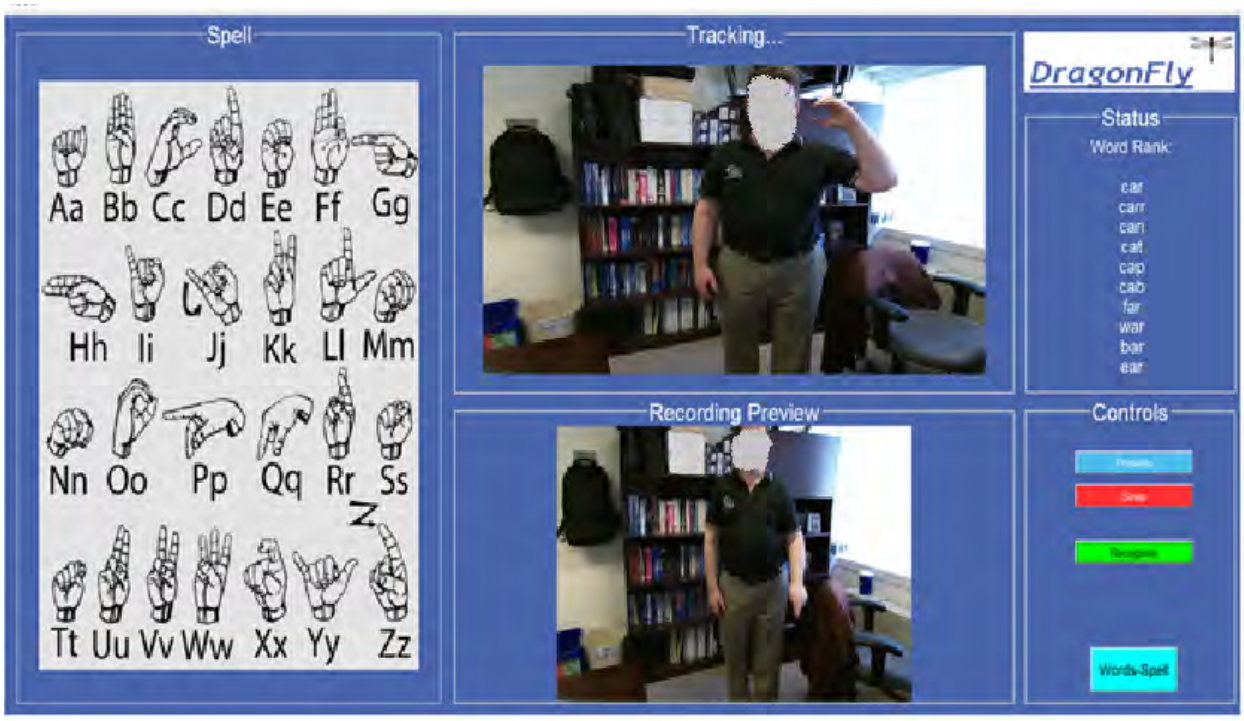

Figure 11. Prototype "Spell" mode.

Prototype's recording capability. Specifically, five users performed each of the 50 signs two times, for a total of 100 recordings per signer and 500 recordings in total. For each recording, the Kinect skeletal detail associated with the individual image frames was preserved. Word estimation was performed via a template matching procedure, i.e., an unknown test recording was compared to a set of pre-recorded examples of the 50 words to be discriminated. The test-template scoring is based upon the fusion of two distinct criteria: The time-varying similarities of the user's body configuration and the apparent hand gestures. The former utilizes the Kinect skeletal data and computes a similarity score by comparing the test and template elbow and hand positions over the course of the recording intervals. The final score is derived from a dynamic programming (DP) procedure. The hand gesture component is evaluated via a similar DP procedure, but with the inter-frame cost derived from the outputs of the Deep Hand DNN (Koller 2016a). Deep Hand is a publicly available neural network that has been trained (via ~1 million images) to recognize a set of 60 distinct Danish sign-language gestures. It was adapted for these purposes without separate training by extracting intermediate data, rather than the final classifications, from the network. The test-template comparison is illustrated in Figures 12 and 13. In each of these, the top plots show the respective signers and their skeletal tracks. The bottom-left plot shows the optimal path through the Skeleton similarity matrix (i.e., Signer 1, frame i vs. Signer 2, frame j), the bottom-middle plot illustrates the path through Deep Hand

space, and finally the bottom-right plot is a fusion of the two similarity measures. In the first figure, the signers are producing the same word: “Beautiful.” This results in high similarity scores for both the skeletal and Deep Hand measures, as well as their fusion. In the second set of plots, the signs being generated are different: “Cheap” and “Cannot.” Correspondingly, the skeletal and Deep Hand similarities are significantly lower, indicating a poor match.

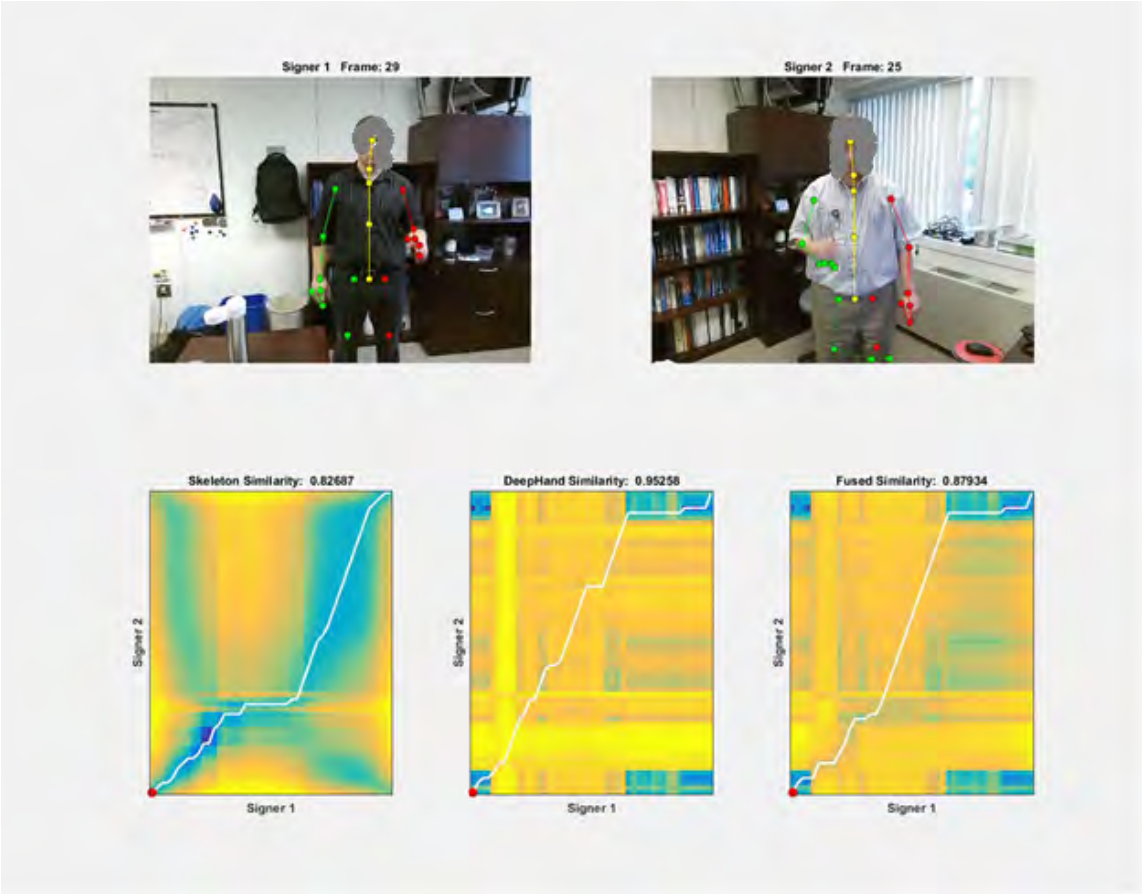

Figure 12. Word Scoring: “Beautiful” vs. “Beautiful.”

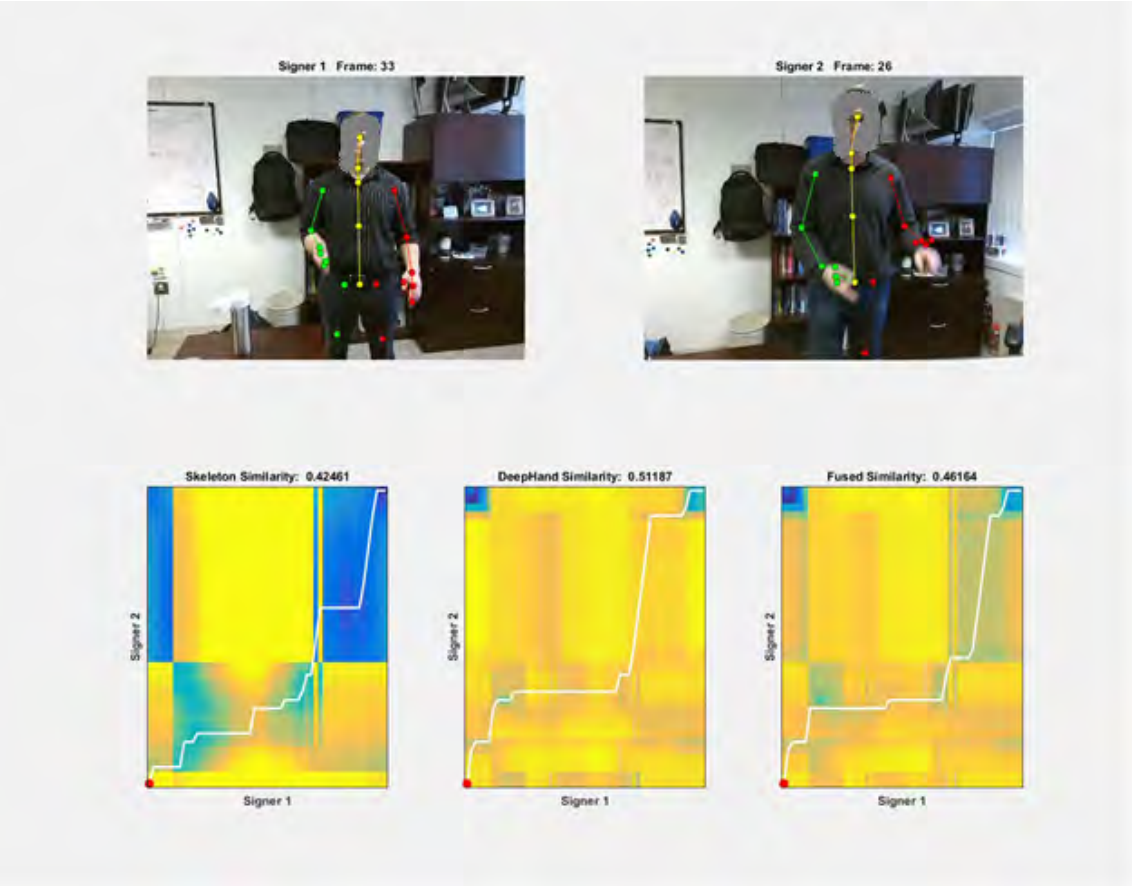

Figure 13. Word Scoring: "Cheap" vs. "Cannot."

Employing the 50 words recorded from each of the five users, a series of experiments were performed to evaluate the effectiveness of the prototype test-template scoring procedure detailed above. A sub-sample of the results are provided in Table 1. In each case, a test signer's videos were compared to various combinations of template videos. The "Top Score" column is the percentage of times the test signer's true word was identical to the word associated with the best scoring template. "Top 5 Scores" indicates that the correct word was included in the top 5 ranked template scores. The results indicate that a distinct signer dependency is present. The prototype performs better with some signers compared to others, and when an individual's first set of recordings are compared to his second, there is near perfect accuracy. In general, the scores improved when the template set was expanded to include all the available (non-tester) templates. Under these conditions roughly 60% of the test videos were identified correctly (top 1) while roughly 85% of the time a candidate video's sign was in the top 5 scores.

Table 1

Word Test-Template Scoring Analysis

Note that random performance for the Top Scores is 2% and for the Top 5 Scores is 10%.

Test Signs	Template	Top Score	Top 5 Scores
User 1	User 1 (Cuts 2)	98%	100%
	User 2	52%	76%
	User 3	50%	80%
	All Others	64%	92%
User 2	User 1	54%	80%
	User 2 (Cuts 2)	92%	100%
	User 3	42%	74%
	All Others	56%	82%
User 3	User 1	46%	82%
	User 2	44%	80%
	User 3 (Cuts 2)	96%	98%
	All Others	60%	88%

These results are very encouraging, especially since strictly isolated word recognition is rarely performed in practice. Usually, the word scoring procedure is incorporated into a language modelling scheme whereby estimation accuracy is greatly aided by a priori statistics derived from context and grammar. In light of these considerations, the "Top 5" scoring results are particularly relevant. Note that the random performance for the "Top Score" is 2% and for the "Top 5 Scores" is 10%. This means that the performance for the prototype system is significantly above random performance, and demonstrates the proof of concept for using the Deep Hand handshape recognition approach and the hand [trajectory] gesture as input features.

The Prototype currently employs the scoring procedure outlined above with a 500-recording template set.

Continuous Fingerspelling Prototype Overview

Spelling Recognition: Once again, a small set of each user's signs was collected using the prototype. In this case, four signers provided input to the prototype while sequencing through the fingerspelled ABC's. Each was input three times from slightly different orientations relative to the Kinect image camera. The videos were hand labelled and the Deep Hand features were computed for the individual frames. Two sets of input per signer were assigned to the Training set, while the third was designated to the Test set. These partitions resulted in roughly 8,000 training samples (225-390 samples/letter) and 3,000 testing samples (70-160/letter). The recognizer feature vector was derived from the three internal, fully-connected layers of the Deep Hand network along with hand position and motion data. After principle component analysis (PCA), the final feature vectors was roughly 300 dimensions. A number of classification techniques were explored, but in the end, a simple linear discriminator was found to be most effective with this data set.

Figure 14 offers the Confusion Matrix associated with the Fingerspelling recognizer performance for each user. In this matrix, the columns represent the true letter (i.e., Target Class) while the rows correspond the recognizer's estimate (i.e., Output Class). An ideal recognizer would therefore have a diagonal Confusion Matrix. The overall test accuracy is 75.7%, with a sizeable variation among letters. For instance, "B", "V", and "W" display 100% recognition accuracy while "P" is at only 34%. Where there are confusions, they tend to be intuitively reasonable (e.g., "M" is frequently confused with "S" and "N" which are very close in structure).

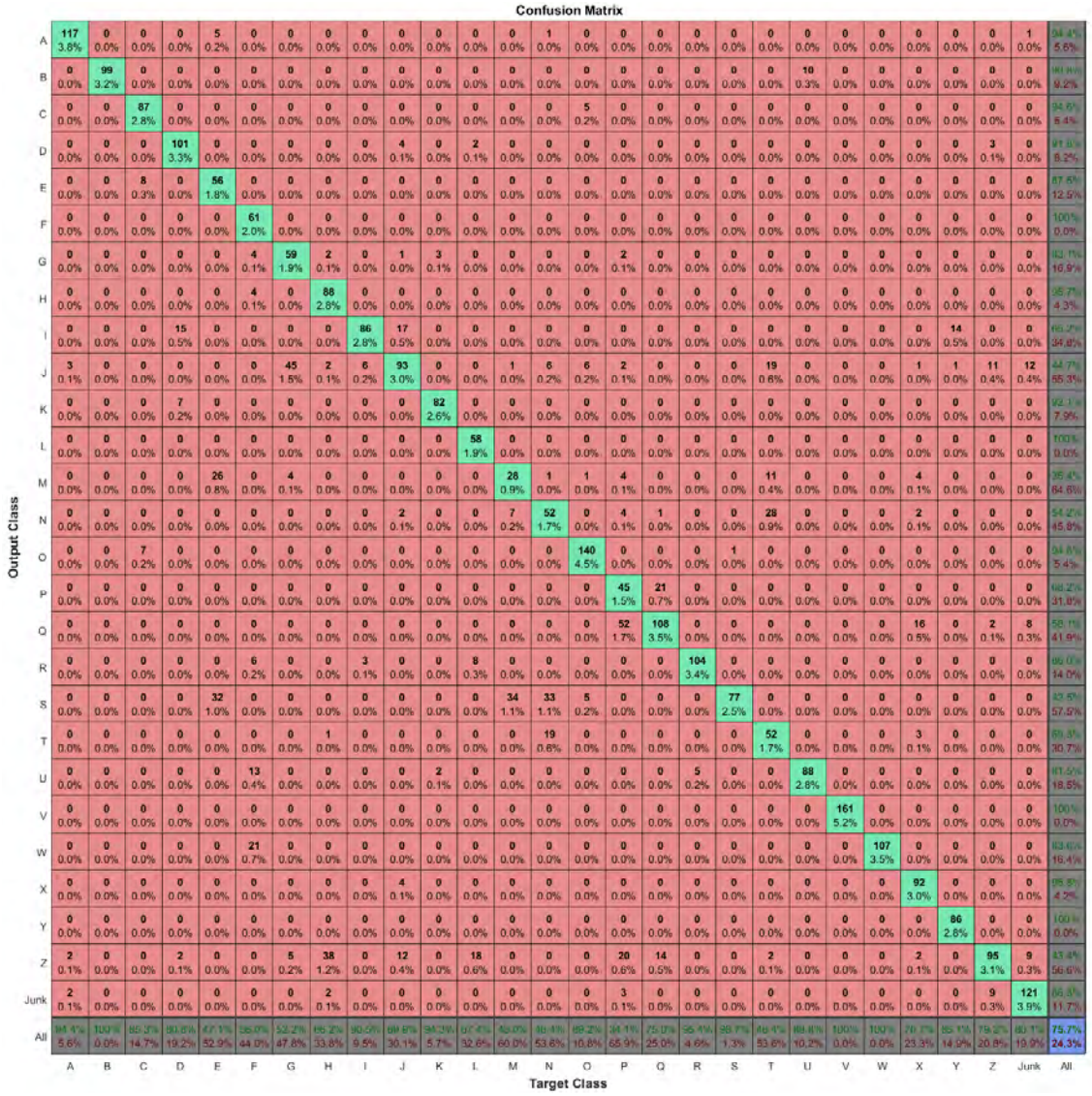

Figure 14. Finger Spelling Recognizer Confusion Matrix.

Fortunately, when combined with a word lexicon, the 75% accuracy of the prototype fingerspelling recognizer for individual letters is not a hindrance to identifying specific words. To quantify this, a series of Monte Carlo experiments was conducted using a dictionary of the top-5000 most frequent words in English. The words were perturbed via the recognizer's confusion matrix to model practical observations and then likelihood-ranked against the words in the lexicon. The top table in Figure 15 illustrates some examples of this process. Each row shows the original word, its perturbed observation, the most-likely

lexical entry associated with that observation, and finally, the rank of the true word. The lower table shows the overall results of this simulation. For three-letter words, the observed letter string is mapped to the true word 92.8% of the time and is in the top 5 words 99.6%. These values rise as the word length increases. With six-letter words, the word recognition rate is greater than 98%. Clearly, the illustrative examples in the top table (each with word rank > 1) are anomalous.

True Word	Observed	Most-Likely	Word Rank
But	BBJ	Bug	2
Him	HIN	Hit	3
Might	NIGZJ	Night	4
Treat	JRMAT	Great	2
Winner	WITTER	Winter	2

Word Length	Total Words	Top 1 Accuracy	Top 2 Accuracy	Top 5 Accuracy
3	276	92.8% (0.4%)	97.1% (0.7%)	99.6% (1.8%)
4	823	88.5% (0.1%)	93.0% (0.2%)	99.6% (0.6%)
5	844	96.6% (0.1%)	98.7% (0.2%)	99.9% (0.6%)
6	792	98.9% (0.1%)	99.8% (0.2%)	100% (0.6%)
7	704	99.4% (0.1%)	99.8% (0.3%)	100% (0.7%)
8	547	99.8% (0.2%)	99.9% (0.4%)	100% (0.9%)

Figure 15. Letter Recognizer vs. Lexicon Monte Carlo Simulation Results. The values in parenthesis correspond to random performance.

However, the situation is made somewhat more complicated by the fact that, in addition to incorrect letter observations, it is undetermined how many letters are being signed. In practice, both the letter quantity and content must be estimated simultaneously. To achieve this, the fingerspelling recognizer embedded in the prototype system applies a decoding approach that first smooths the frame-based observations by emphasizing motion-stable regions and then partitions and scores the estimates assuming a given word length. This procedure is made more robust by limiting the word-length possibilities to those consistent with the rank edit (Levenshtein) distance of the observed letters relative to the lexicon entries.

The results in this section demonstrates the potential impact of a language [spelling] model has in addressing system performance for ASL recognition by exploiting the correlations across temporally adjacent content. In the above example, the fingerspelling had modest performance of about 75% accuracy. Leveraging a spelling model (as seen in the Figure 15), though, results in word accuracy rates significantly above 90%.

Restaurant Prototype Overview

A task-specific prototype was created for demonstrating automatic ASL recognition capabilities at the PacRim Conference (Leang et al 2017). The nominal setting for this prototype is a restaurant where two signing individuals—a waiter and a diner—engage in a dialogue with a constrained set of material. The dialogue options for the diner and waiter are as follows:

Diner Dialog Options (FS corresponds to fingerspelling)

I WANT WATER OJ(FS) MILK SODA BEER TEA COFFEE WINE
READY
MORE TIME
I WANT SALAD FRENCH FRIES CHEESE BURGER STEAK FISH CHICKEN
WITHOUT ONION TOMATO LETTUCE SALT PEPPER
MORE KETCHUP MUSTARD (FS) MAYO(FS) DRESSING PICKLES
NO MORE
YES ALLERGIC PEANUTS DAIRY(FS) EGGS GLUTEN(FS)
AWFUL BAD SO-SO GOOD VERY_GOOD NOT_GOOD TASTY EXCELLENT

CHECK PLEASE CREDIT_CARD CASH
CALL-WAITER-OVER
THANK-YOU
FINE NO-PROBLEM

Waiter Dialog Options (FS corresponds to fingerspelling)

I WANT WATER
I WANT OJ(FS)
I WANT MILK
I WANT SODA
I WANT BEER
I WANT CHEESEBURGER FRENCH FRIES
I WANT CHEESEBURGER WITHOUT ONION FRENCH_FRIES
ALLERGIC PEANUTS DAIRY
ALLERGIC EGGS ALLERGIC
I WANT SALAD WITHOUT DRESSING MORE MAYO

This prototype leveraged context and a language model to improve system performance. Performance was significantly better on enrolled individuals than it was on unenrolled individuals.

3.1.2 Additional ASL Classification Experiments

During the period of performance, two ASL classification sub-problems were investigated using the ASLLVD dataset (Athitsos et al 2008, Neidle et al 2008):

1. 26-class fingerspelling recognition
2. 86-class word recognition

In machine learning, the performance of a classification algorithm depends on the quality of features. Classifiers trained directly on the raw data tend to perform poorly. Thus, high-performance classification pipelines usually include a separate stage dedicated to learn and extract features from the raw data. Feature engineering refers to the process of learning high-dimensional, characteristic representations for a discriminative task such as classification by running (unlabeled and unbiased) data examples.

The goals of the ASL feature engineering include:

1. Detailed micro hand gestures from frames of images;

2. Geometric alignment related to location of hands with respect to signer's anchor body part (e.g., face); and

3. Dynamic changes over frames.

Overall, the lack of relevant, labeled image frames makes the ASL classification a hard problem. As a remedy, a semi-supervised approach was used. In this semi-supervised approach, a large number of unlabeled images was used to learn the feature representation before training the ASL recognizers with labels. This approach relieves the small labeled dataset problem to a certain degree. However, in order to learn the ASL hand features from unlabeled examples, it require a good hand detector/tracker.

We process ASL image frames using deep learning and body geometry measurements. First, a general-purpose deeply learned visual features is extracted using a pre-trained convolutional neural network (CNN). In particular, the 19-layer CNN trained by Oxford's Visual Geometry Group (VGG-19) for ImageNet challenge (Simonyan et al 2014) is used. Each image containing an ASL signer is resized for the VGG-19 input specification and runs the feedforward path. The 1024-dimensional penultimate layer activation values serve as the feature for an image frame. These activations are grouped over multiple frames in time representing a sign (e.g., letter, word) by a technique called pooling before its application to classifiers. This general-purpose deep learning approach for ASL feature processing is illustrated in Figure 16.

Figure 16. General-purpose deep learning unit to extract visual features. Photos are courtesy of BU ASLLVD dataset.

Figure 17 illustrates an alternative deep learning scheme for ASL feature processing. A deep CNN specialized for hand gesture recognition is used, namely Deep Hand (Koller et al 2016a), developed by Human Language Technology & Pattern Recognition Group at Aachen University, Germany. Feature processing with Deep Hand is similar to that with VGG-19. The only difference is the leveraging of an automatic hand tracker to crop the hand region of a whole image frame. The cropped signing hand is then applied to the Deep Hand feedforward path. A 4096-dimensional penultimate layer activation of the Deep Hand CNN is used as the feature before pooling and classifiers.

Note that the Deep Hand CNN (not to be confused with DeepHand, which is a pose estimation approach from Purdue University), was derived from a CNN trained (Szegedy et al 2014) on the ImageNet corpus. Deep Hand was retrained on a set of over 1 million hands corresponding to Danish Sign Language and this data has recently become available. There are a number of opportunities to improve on the Deep Hand model. This includes retraining a system more closely matched to ASL and smartphone-style use cases. This includes the inclusion of handshapes better matched to ASL as well as retraining the CNN to optimize computational performance. In addition, the 3D component of the signal could be included in the Deep Hand model to further improve performance. Finally, the nominal hand resolution into Deep Hand (227x227) is of much higher resolution than the hands in the available ASLLVD and RVL-SLLL datasets, where hands are seldom more than about 40 pixels across.

Figure 17. Special-purpose deep learning unit to extract visual features focused on hands. Photos are courtesy of BU ASLLVD dataset.

In Figure 18, the ASL geometric feature extraction is depicted. The baseline geometric features are extracted from the face-hands triangular measurements. The angles are computed by the law of sines from the centroids of the bounding boxes that face and hand trackers locate in an image frame. Alternatively, a Microsoft Kinect is used, which is a motion sensor specialized for measuring human gestures and depth information.

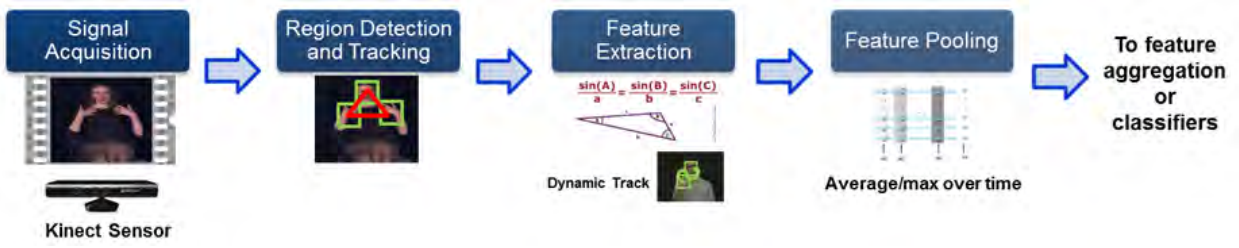

Figure 18. Face-hands triangular and other fine-grain body skeletal geometric measurements. Photos are courtesy of BU ASLLVD dataset.

Key principle for ASL classification is to use lightweight classifiers. Most of computational complexity is pushed to feature processing based on deep learning. Another principle is to support multiple classifier algorithms. The following algorithms are considered:

- Softmax regression
- Support vector machine (SVM)
- Dynamic time warping (DTW)

Softmax regression is a generic multi-class classification algorithm popularly implemented at the output layer of neural nets. Softmax computes 1-vs.-all class likelihood ratios:

$$P(y = j|\boldsymbol{x}) = \frac{e^{\boldsymbol{w}_j^{\mathrm{T}}\,\boldsymbol{x}}}{\sum_{k=1}^{K} \boldsymbol{w}_k^{\mathrm{T}}\,\boldsymbol{x}}$$

where there are K total classes, w_j is the trained softmax weight for class j. Therefore, the position with highest P( ) value gives the predicted class of a given input.

Support vector machines (SVM) seek a hyperplane that maximally separates data points of different classes. Fundamentally, SVM is a binary classification framework. Similar to softmax, a 1-vs.-all binary SVMs for multi-class support is trained.

Dynamic time warping (DTW), by itself, is not a classification algorithm. It gives a measure of similarity between two temporal sequences that vary in speed. DTW is popular in temporal pattern matching such as comparing video frames, audio clips, and other time-series data. For ASL classification, a DTW is used to compare templates that comprise sequence of features for a given letter or word. When an unknown input is acquired, the DTW cost is computed against all templates, and the predicted letter or word will be the one that matches a particular template with the smallest cost.

Additionally, class discriminating subspace approaches are used. In particular, principal component analysis (PCA) and Fisher's linear discriminant analysis (LDA) are considered. Strictly speaking, PCA and LDA are not classification algorithms, but can be made helpful for discrimination. Feature vectors after the PCA and LDA subspace transforms often train a classification algorithm such as SVM more effectively than non-subspace transformed feature vectors.

Our feasibility study includes two evaluation scenarios. In Scenario I, the classification performance of 26-way letter recognition is evaluated. In Scenario II, the classification performance of 86-way ASLLVD demo word recognition is evaluated.

For the task of fingerspelling recognition, the Purdue University RVL-SLL dataset (Martinez et al 2002) is used. The dataset contains video clips of fingerspelling done by 14 different signers for each English letter. Every signer has at least two sessions of complete 26-letter recordings. Due to the lack of frame-level annotations, the dataset was manually labeled frame-by-frame. Despite being laborious, this process is an important step in annotating the data, including identifying null states during the transitions between fingerspelling signs.

To extract per-frame visual features, a pretrained deep convolutional neural networks (CNNs) was used. After experimenting with VGG and other popular CNN models, Deep Hand was chosen as the ideal option. The hand tracker is used to identify a cropped signing-hand subimage that is resized to 227x227 according to the Deep Hand spec and compute its feedforward path to extract either 61x3 final softmax or 4096-dimensional fully-connected layer output. After average pooling of the feature vectors, letter recognizers in 26-way softmax regression or 1-vs.-all linear SVMs are trained.

We have cut the dataset into five folds. For softmax, rank-1 best-fold accuracy of 41% (random performance would be ~1%) was achieved. Linear SVMs have achieved 63%. With LDA over the CNN feature vectors the classification accuracy could be further improved by 4–7%. The use of Deep Hand has improved the previous accuracy result of 20% with VGG-19. Note that the use of a language (e.g., word spelling model) would further improve performance.

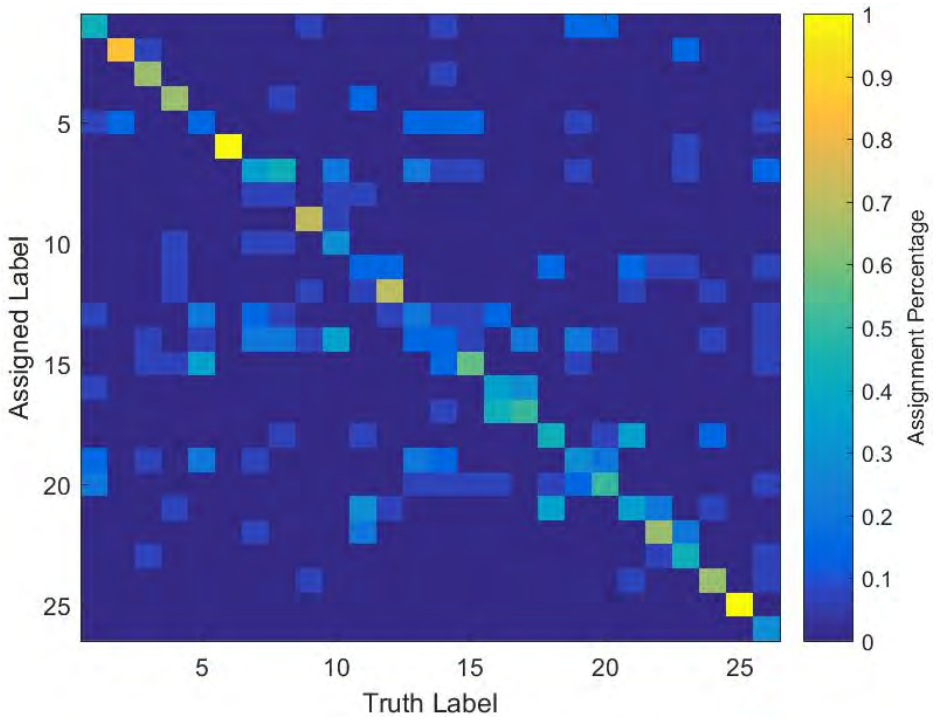

Figure 19. Confusion matrix for letter recognition task.

Table 2 lists vocabulary for the word recognition task. This is an excerpt from the Boston University ASLLRP dataset (Athitsos et al 2008, Neidle et al 2008). 50 words were originally selected by the sponsor. Due to inconsistent data availability, the following expanded set of 86 words listed in the table was used:

Table 2

ASLLVD Words

afraid	again	airplane+	all	answer	any	appointment
beautiful	blue	bore	boss	brown	buy	can
cannot	center	chat++	cheap	chemistry+	coat	conflict-intersection
court	cruel	dark	deaf	deposit	depress	disappoint
divorce	down	dress-clothes	drink	drunk	dry	earth

east	eat+noon	embarrass+	enough+	everyday	excited+	family
fly-by-plane	football+	forget	four	free	freeze	Friday+
friend	girl+friend	give-up	graduate	grandfather	grandmother	green
grow	happen	hard-of-hearing	have	head	heavy	hello
high	home	include-involve	keep	know+-1h-neg	learn	leave-there
left	letter-mail	like+-1h-neg	look	lousy	marry	medicine
miss-assume	price	stand-up	take-up	ns-Boston	ns-Chicago	ns-Detroit
ns-England-English						

Our experimental methodology is similar to Scenario I. Once again, Deep Hand was used to extract visual features from hand-tracked frames. Average-pooled feature vectors are applied to train 1-vs.-all linear SVMs. The dataset was partitioned into five cross validation folds. For the best case, an accuracy of higher than 80% was achieved for all positive and negative examples on average. Note that a portion of the dataset for each isolated word is heavily unbalanced because it is one word against 85 others. This makes positive detection rate the most important metric. For positive examples, a rank-1 average accuracy of 37% is achieved. As previously discussed, a language model leveraging the correlations with neighboring signs would further improve this performance.

This subsection has demonstrated the significant performance impact both by using a specialized DNN handshape classifier (Deep Hand), as opposed to a more general image classifier (VGG). It has also demonstrated the impact that various machine learning approaches have in further improvements to the baseline Deep Hand performance. As discussed, the Deep Hand approach, while making significant contributions to the effort, has the potential of further improvements.

3.1.3 Translation of ASL to English

General Discussion

While the recognition component of the Sign to English processing chain will provide useful output to an English speaker, ASL and English are distinct languages, so having a MT capability for translating this content will provide improved communication. The technical details of a Direct DNN approach are discussed below. There are two available dataset approaches for supporting the translation problem. One option is to leverage the inclusion of closed-caption and ASL content in multimedia content such as in movies and on the Sign Language Channel (DPAN.TV: https://dpan.tv/). Another option is to leverage

parallel corpora that annotate both the English words and ASL glosses at the phrase level (Othman et al 2012). These datasets, as well as the technical discussion below, support translation between ASL and English in both directions.

Direct DNN Translation of ASL to English

Sequence-to-sequence models were introduced relatively recently (2014) as a powerful new tool for end-to-end machine translation, quickly becoming state-of-the-art on many tasks (Cho et al 2014, Luong et al 2015). Building off of this success, the encoder-decoder model has been adapted and applied to a number of other tasks, including speech recognition (Chan et al 2015, Zhang et al 2016) image to markup language "translation" (Kanervisto et al 2016), and most recently speech-to-text translation (Kanervisto et al 2016). These papers have each contributed a number of new developments (convolutional LSTMs, multi-task learning, and more), requiring additional research to compare architectural choices on other domains and datasets. This approach would not necessarily require an intermediate recognition component to support translation capabilities.

In neural machine translation, a bidirectional recurrent neural network (RNN), known as an encoder, is used by the neural network to encode a source sentence for a second RNN, known as a decoder, which is used to predict words in the target language. Attention allows the model to learn which source word(s) to translate at any step in the process, creating soft alignments between source and target that can be learned and modified as the model is trained.

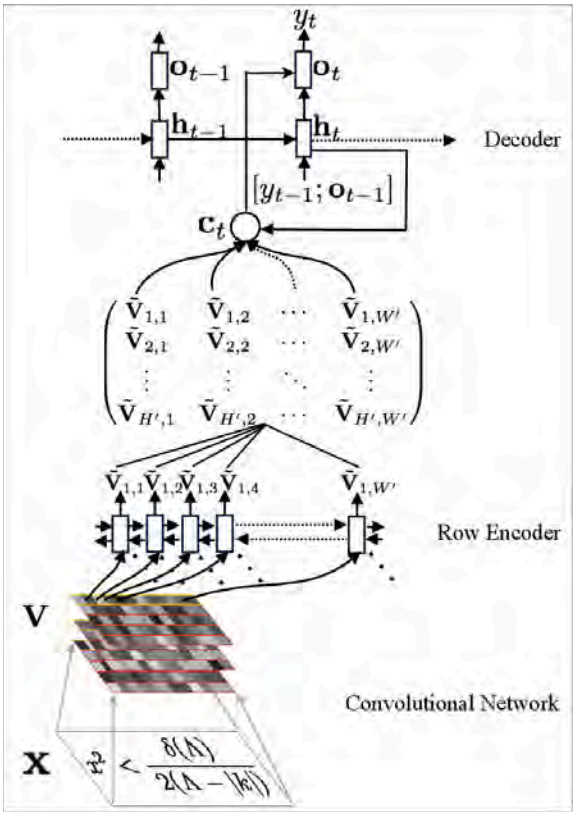

Figure 20. Notional RNN architecture for direct translation of continuous ASL signing. Photo courtesy of Harvard natural language processing (NLP) github website.

To translate American Sign Language (ASL) using neural machine translation architecture combines elements from speech and image-to-markup translation. Image-to-markup is a translation task using images as input, but without an element of temporal change, similar to text-to-text translation. Speech translation and recognition capture this element. This changes the nature and expected size of input sequences. Where in neural machine translation, input sequences are measured at the word-level and are typically up to 50 words, speech and sign language require windowed input sequences at the frame-level. A sentence sequence may be an order of magnitude longer, which may require more data to model. To reduce temporal resolution while maintaining generalizability, (Weiss et al 2017) experimented with initial convolutional layers and striding on the speech translation task. These are additionally used by (Deng et al 2016) for image-to-markup translation.

Different encoder-decoder architectures have been tested for the tasks of speech recognition and translation. In particular, (Chan et al 2015, Weiss et al 2017, and Zhang et al 2016) have each contributed

different layer architectures designed to capture structural information about the input signal, while also reducing the amount of time and data needed to train systems. These include pyramidal encoders, convolutional and convolutional LSTM layers, residual connections within networks, and more. An initial exploration of these architectures has begun using encoder-decoder models with attention on a constrained, clean speech dataset (the Wall Street Journal corpus) where expected performance is well known, making it clear how well the models are doing. The next step is to apply them to Sign Language datasets where the input is sequences of image frames rather than speech frames.

3.2 ENGLISH TO ASL

An ideal option for the English to ASL communication channel is to include an ASL-signing avatar in addition to, or instead of, an English text translation of the spoken English content. To date, an OpenSource avatar option has not been identified. There are commercial (ProDeaf, http://prodeaf.net/en-us) and academic (DePaul University, http://asl.cs.depaul.edu/) avatars, though no supported APIs (Application Programming Interface) are available for either system. The ProDeaf system is an active commercial product that was originally developed for Brazilian Sign Language, though ASL capabilities have since been added. It is unclear whether the DePaul system is anything more than a canned demonstration capability.

3.3 MOBILITY-RELATED ISSUES

There are several engineering and technical challenges to moving to a mobile ASL solution deployed on a smartphone. These include:

- Signal acquisition
- Computational
- Broader variation in acquisition environment

3.3.1 Smartphone-Based Signal Acquisition

A few recently released smartphones have sensor configurations similar to the Microsoft Kinect sensor used for this effort. These sensor packages include a lower resolution 3D camera and a higher resolution 2D camera. The 3D signal is useful for the high confidence detection of the body configuration, particularly the position and orientation of the signer's hands. The higher resolution 2D signal is important for supporting the DNN-based handshape recognition process. While the 3D signal was only used for body/hand detection, and not for sign recognition, for the MIT LL effort, it does have the potential for improving recognition performance, particularly for the mobile applications being considered here where increased viewing angle variation would be expected.

The iPhone X, released in November 2017, includes a TrueDepth Camera and Sensor System that leverages infrared (IR) structured illumination to attain 3D images that they use for their internal face

recognition-based authentication process. Samsung Galaxy phones also have 3D image capability that leverages Google's ARCore (replaces their discontinued Tango platform) that are primarily targeting the augmented reality domain. Intel also developed a smartphone with similar capabilities, but has discontinued those efforts in 2017. These smartphone-based capabilities can be repurposed to leverage these 3D capabilities, as well as the onboard high-resolution 2D cameras, to support the ASL signal acquisition challenge.

3.3.2 Smartphone-Based Compute

The ASL to English recognition process is computationally expensive, particularly for the DNN-based handshape recognition process. There are several options for addressing these compute issues:

- Retrain the DNN handshape recognizer to optimize effective computation.
- Leverage AI chipsets that are increasingly available on smartphones for doing onboard computation
- Leverage available communication channels to offload the extracted signal (or extracted features) to perform expensive computation operations on an external platform.

The datasets that support the training of the Deep Hand DNN are now publicly available, which enables the retraining of the DNN. This retraining could be extended to include the 3D signal component with a smartphone-based data collect to further improve handshape recognition performance.

We have investigated the GPU speedup option using a networked GPU. The first goal, with respect to processing time, was that each image takes less than 20 ms (0.020 seconds) to classify via deep neural network architecture. This is a first-step toward an ASL system that is operational in real time. This processing goal was achieved using a GPU-based remote server. The result is that a GPU remote server speeds processing time significantly compared to using CPUs local to the demo device (i.e., personal laptop) and this time can be further improved by distributing workload among multiple GPUs on a single server.

One option for operating ASL technology on a small electronic device, such as a smartphone, tablet, or smartwatch, is that the computational processing is offloaded from the user device onto a remote server. This is because small electronic devices are not equipped with enough memory or processing speed to handle image classification in real time at scale. The best solution, and one that is often referred to in industry practice (such as with Google Translate or Siri), is to use an HTTP service. A CPU-based and GPU-based HTTP web service was constructed. With this webservice, each image is transmitted to a remote server that is hosting the DNN classifier. The image is processed and a result is returned to the client. The client could be a miniature electronic device, for example.

Table 3

Characterization of CPU and GPU Performance

	Technique	CPU Per Image	CPU Full Set	GPU Per Image	GPU Full Set
PyCaffe (Python)	Non-parallel List comprehension	0.119	26.67	**0.025**	**5.76**
	Non-parallel Multithreading	0.114	25.56	0.840	189.96
	Parallel	0.117	26.36	0.595	133.38
	HTTP Server/Client	0.119	26.70	**0.025**	7.0
MatCaffe (MATLAB)	Non-parallel loop	0.116	26.13	--	--
	HTTP Server/Client	0.125	28.05	0.030	6.9

Table 3 indicates various parallelization methods that were experimented with for the Python and MATLAB versions of Deep Hand image classification. The table shows time profiles for CPU and GPU on a per-image basis as well as a full-set of 400 images. It was found this task is not well suited to the parallelization libraries of Python and MATLAB because built-in parallelization in both languages adds significant overhead. While it may seem counterintuitive that parallelization libraries cause processing time to increase, it is a commonly observed phenomenon in computer science and occurs when the individual task itself is very fast and not memory intensive, but highly repetitive. The task of image classification is not memory intensive and is highly repetitive. Therefore, faster processing time are observed when each image is processed in a serial manner. Likewise, it is possible to induce parallelization at a systems level by setting up multiple GPUs on a server with a front-end load balancer to distribute images between multiple GPUs. For example, by sending 50% of the images to GPU #1 and simultaneously sending 50% of the images to GPU #2, the overall time for processing a set of images will be halved.

This section has demonstrated how GPUs can be leveraged to address the computational challenges to the ASL problem. This challenge has been particularly acute for the Deep Hand feature extraction process. Going forward, technology options include retraining the Deep Hand model to reduce its computational demands, leveraging the GPU compute capabilities resident on the smartphone, or leveraging a GPU that is networked to the smartphone.

3.3.3 Smartphone-Based Signal Variation

A handheld smartphone-based signal acquisition environment will result in a broader variation of sign variation than the controlled laboratory-style signal acquisition environments used by the ASL corpus collections done to date, including those done by MIT LL in the course of this work. This variation includes illumination environment, image backgrounds, potential occluders, and viewing angle. It is expected that the DNN-based handshape recognition approach should be robust to the illumination environment and image background.

Occluders are always an issue for vision-based systems. Approaches to addressing these issues include developing approaches to recognize ASL content with partial information. A more promising approach, particularly for the short-term, is the use of a quality metric for ensuring the collected signal does not include occluded content.

Viewing angle issues can be mitigated through the exploitation of the 3D signal and broader datasets matched to these conditions that may include newly collected data and synthetic data. Since there is minimal existing off-angle ASL content in existing corpora, use case-relevant off-angle smartphone content needs to be collected to seed a synthesis effort. This issue will be discussed in greater detail in the next section.

This page intentionally left blank.

4. DATA-DRIVEN SYSTEM DEVELOPMENT

The mobile smartphone-based ASL challenge is inherently a data-driven problem necessitating data collection matching the intended use cases for supporting performance evaluation. Additionally, other data sources such as other ASL content that is not well matched to the intended use cases can be used to support the development process, namely in increasing the effective amount of available training material. In cases where additional natural data cannot easily be collected, synthetically-augmented datasets are another promising approach, where millions of image variants can be created from seed images and parametric descriptions.

4.1 SYSTEM CONCEPT

A primary challenge to developing a mobile smartphone-based communication system between ASL and English is that there is no known dataset that matches these use cases. This is important since it is impractical to measure the performance of a prototype system on data not representative of its intended use to determine whether it is ready for technical transition to a deployed system. It also means that the numerous system design decisions made during the technology development process to address performance shortfalls are not practical without such a representative dataset. It is, for example, impractical to enlist users to evaluate systems through each step of the development process without having them on constant call. This approach also has the drawback of a system "tuned" to whatever user was used during the development and evaluation process. This was, for example, the performance challenge of the prototype system development process discussed in the previous section that had strong performance results for the system developer though significantly weaker performance for novel users.

For the reasons stated above, many system development efforts leverage a development process that resembles Figure 21. During the technology development process (top box on right) development data is leveraged to internally train and evaluate both system components and the overall system. The development data that is used for evaluation should be matched to the intended use cases of the effort—unconstrained continuous ASL collected from a smartphone. All of the development data that is used for training does not necessarily need to be matched to the intended use cases. In the ASL case, for example, this train material could include synthetic data or ASL content that is leveraged from related corpora. Performance improvements are achieved through an iterative process of system design choices, retraining, and evaluation. Well-defined evaluation criteria can also support an informal apples-to-apples comparison of technology performance in published work. One risk of this approach is that over time, system performance can become overly tuned to the development dataset—particularly the development partition used for internal evaluation. Unfortunately, this performance frequently does not necessarily transfer to the operational domain. Ways to mitigate this issue include frequent refreshes of the development dataset with the infusion of new data, the intelligent partitioning of the development set leveraging techniques such as cross-validation, and the use of formal evaluation as discussed in the next paragraph.

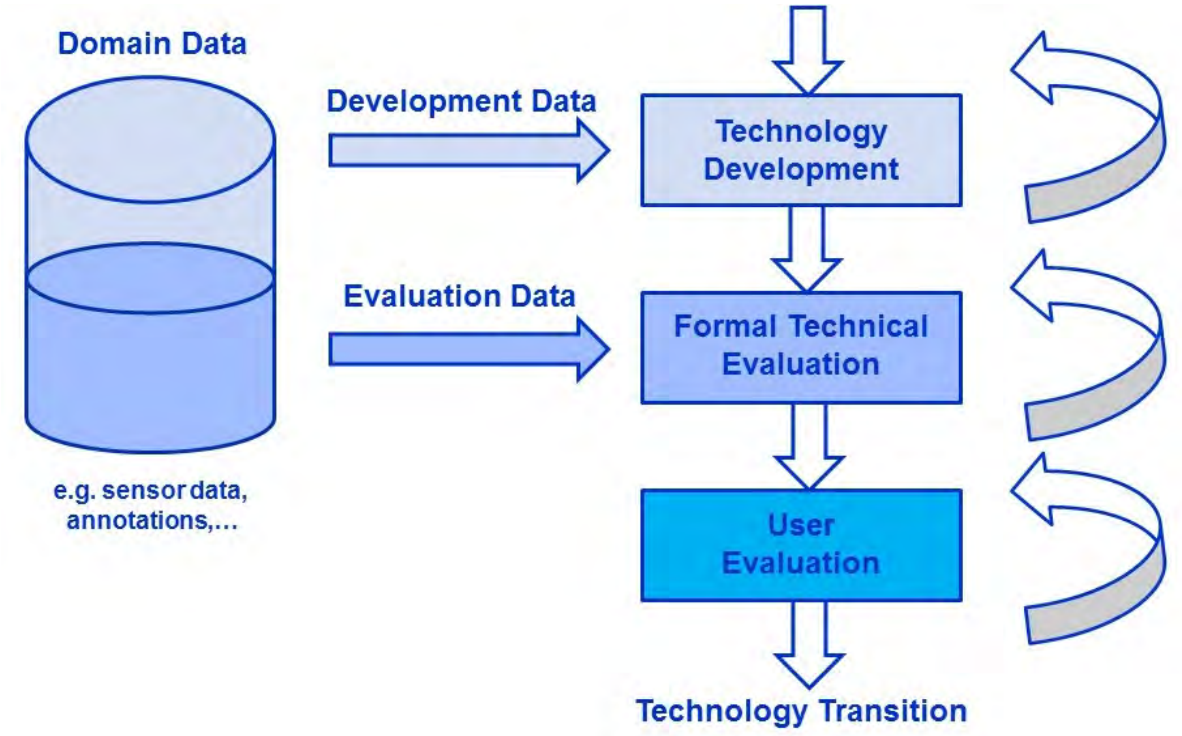

Figure 21. Train-test partitioning of dataset.

Formal technical evaluation can take many forms, but frequently involves the evaluation of systems on "freshly" revealed data. It supports an apples-to-apples performance comparison across many different systems being evaluated, as well as against the performance of previously evaluated systems. Formal evaluation also supports a decision process to determine whether the technology performance is sufficient to consider technology transition and user evaluation. Sometimes the evaluation is self-evaluated where the answer key annotations are shared with evaluation participants. More frequently, the results are evaluated by an evaluation coordinator where either results are submitted and evaluated or software is submitted, run, and the results evaluated. The results submission is popular due to its balance between maintaining the integrity of the evaluation and minimizing the labor involved with the evaluation.

4.2 DATA ACQUISITION AND ANNOTATION

Ultimately, the most valuable set of data for a data-driven approach is the data required for evaluating system performance. This data is indispensable both for supporting the regular "informal" internal evaluations of various system design decisions, as well as "formal" evaluations to determine whether the

technology performance has matured sufficiently to support technology transition and user evaluation. As previously discussed, no ASL datasets matching the mobile smartphone-based sensor collection model has been identified to date. Leveraging ASL datasets that do not match these use cases may positively impact the performance for a mobile ASL recognition or translation system in some cases. Best practice is to collect a small set of matched data which allows testing of how well machine learning approaches translate to real-world conditions.

The data acquisition and annotation process includes several tasks, including designing the actual data collect, building data collect tools, collecting data, annotating data, cleaning up and storing data. These processes are labor intensive in nature, with some modest costs also required for the necessary sensors (smartphones) and storage. Much of the cost (labor) is driven by the data collection and annotation tasks.

A data collection methodology is suggested to mitigate the effective cost the data collect by mitigating the labor required for data collection and annotation. A smartphone-based app can be developed that will prompt the subject to sign a specified ASL phrase. The signer will then sign the prompted phrase that the smartphone will collect and automatically send to a central data repository with the associated annotation relating to the signer and phrase. Such a capability minimizes the primary costs to data collect—collection and annotation.

A partnership with a University campus that has a strong Deaf community such as Gallaudet or the Rochester Institute of Technology (RIT) would provide an ideal venue for rapidly collecting content across a significant number of ASL signers. Leveraging these campuses may also significantly reduce the expenses for compensating subjects. These smartphones can be distributed to on-campus signers without the need for the data collection team to monitor the data collection. The actual data collection should occur across a broad range of indoor and outdoor collection conditions such as hallways, dorm rooms, and classrooms.

Suggested phrases for the data collection are subject-verb-object sentences. These sentences will need to be developed as part of the data collection design to cover the ASL vocabulary targeted for the data collection.

4.3 DATA PARTITIONING AND SUPPORTING EXPERIMENTS

The purpose of the data collection is to support a development process that includes an evaluation component to ensure the effectiveness of the completed system. An annotated data collect matching the use cases is necessary to support both the technology development and formal technical evaluation components. A single data collect would support the development of both of these datasets. Prior to this data collect, it is suggested that a smaller subset of the data collect is first done to work out the data collection kinks and to determine requirements for the data collect. This pilot corpus could, for example, be limited to a subset of the ASL vocabulary. A discussion of several experiments to support this specific investigation is discussed below.

Evaluation against the formal technical evaluation content should be done sparingly to reduce the risk of over-tuning the system to this dataset. Alternatively, this material can be further partitioned with portions of the data revealed to the system developers over time. Older data can be pulled into the technology development dataset. Another alternative is to do further data collect to refresh the formal technical evaluation dataset to refresh the dataset and perhaps adapt to an evolving set of use cases. The training partition of the technology development dataset can be potentially augmented with synthetic data and/or other non-matched sign language datasets.

Several experiments done on a smaller pilot data collect can refine the data requirement needs of the effort and identify promising paths for system development. Below is a list of experiments to consider. The first three, and particularly the first one, are important for determining the data requirements to developing an operational system:

- **Core data requirements experiment**: A core question for system development is how much data is required to effectively train and evaluate an ASL system. Specifically, how many exemplars are required per sign. A performance tradeoff can be done to determine how many training examples per sign are necessary to achieve the required system performance.

- **Enrollment experiment:** The data requirements can be potentially reduced through leveraging signer enrollment, though this process does place an added burden to the initial use of a fielded system. To support this experiment multiple sessions of ASL content will need to be collected for a meaningful subset of subjects. The sessions for a given signer should be done in different physical environments and on different days. Comparative performance results between using and not using one of these subjects' sessions can be performed to determine the impact of subject enrollment.

- **Synthetic experiment:** Leveraging synthetic ASL content has the potential of mitigating the data requirements for training, though not evaluating, an ASL system. An experiment is suggested that augments the training material with synthetic content to determine whether performance is meaningfully improved on the withheld evaluation content. Methods such as GAN and VAE, as discussed in the previous section, are well matched to generating synthetic material to improve recognition performance. It may be helpful to collect ASL content simultaneously from multiple smartphones from different viewing angles to support this experiment.

- **Neural translation experiment:** Leveraging DNN approaches to generate translations from ASL to English using low-resource data (e.g., closed captioning) has the potential of mitigating data requirements, as discussed in the previous section. One challenge is that this content, largely captioned TV shows and movies (https://dpan.tv/) is not well matched to the mobile ASL concept. An experiment is suggested that leverages such content to determine the potential impact of this approach to moving directly to a translation system.

- **ASL translation experiment:** Since ASL is a distinct language, there is a translation problem in addition to the recognition problem. An experiment is suggested that characterizes how well ASL translation can be performed given the accurate (though perhaps noisy) recognition of ASL glosses.

4.4 LEVERAGING SYNTHETIC DATA

One of the most pernicious challenges to the ASL challenge is acquiring sufficient data matched to the expected use cases. One potential approach to addressing this challenge is by leveraging synthetic data. Given the availability of state-of-the-art algorithms, a data-driven approach is likely more promising than a physics-engine approach. Two promising data-driven DNN approaches are Generative Adversarial Networks (GAN) (Antoniou et al 2017) and Variational Autoencoders (VAE) (Kingma et al 2014, Rezende et al 2014).

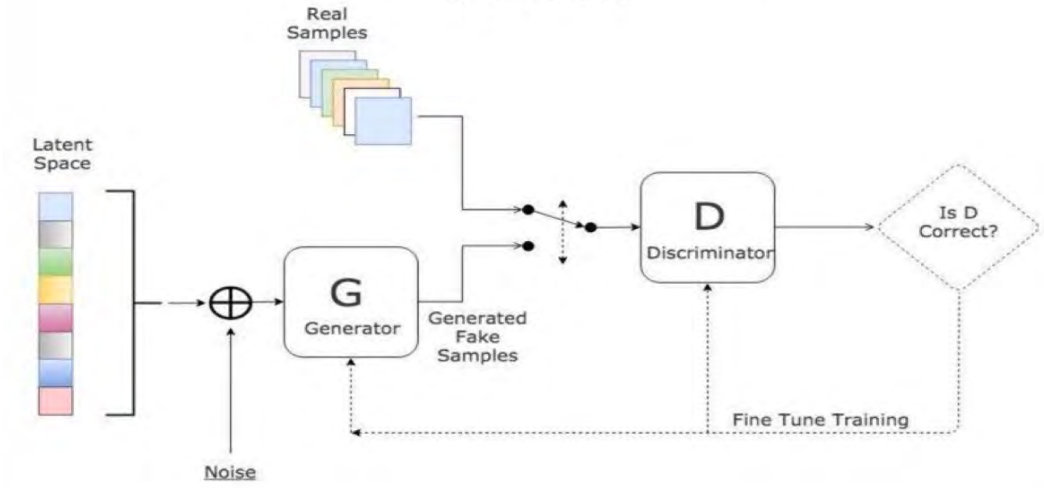

Figure 22. Generative Adversarial Network (GAN) architecture. Diagram courtesy of KDnuggets.

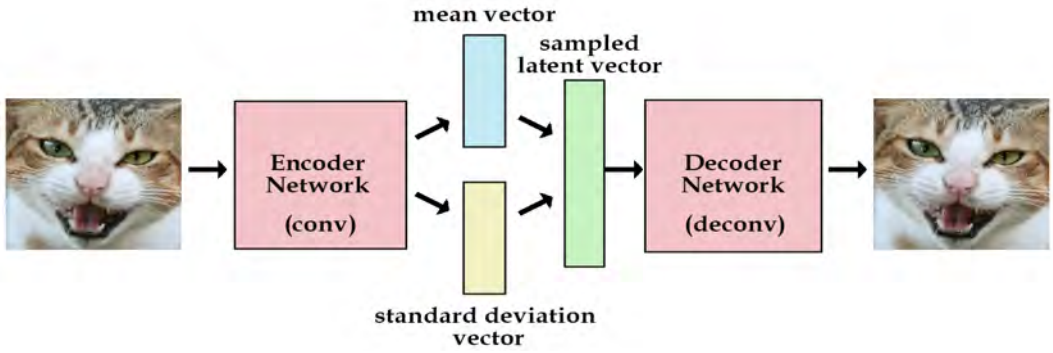

Figure 23. Variational Autoencoders (VAE) architecture. Diagram courtesy of Kevin Frans (kvfrans.com).

In both of these approaches, both the synthesis of novel content and the training of the classifier model are generated concurrently. The objective is to generate models for "unseen" examples, though these examples need to be consistent with the available training material. For example, if a face model is generated from content of various faces at various pose angles, the system would begin learning how to repose a given face to poses corresponding to other faces. If, however, only male faces are included in the training set, this approach will have problems accurately modeling female faces. In the ASL case, this means that this approach would need to be seeded with sufficient training material that is consistent with the type and quality of data expected in the operational system. Examples include the various observation angles, the sensor signal type (2D/3D?), and the expected handshape sequences and trajectories expected for the target vocabulary.

These synthesis approaches are consistent with similar approaches where collected data signals are perturbed or noise added to signals to extend the availability of training material (Lawson et al 2009, Harper 2015). They provide a potential opportunity to increase the dynamic range of available data for a given ASL sign or broaden the acquisition environment conditions.

4.5 CHALLENGE PROBLEMS

The recognition and translation problems between ASL and English are challenging and multi-faceted, requiring expertise in sensors, signal processing, machine learning, linguistics, and software engineering. For these reasons, it does lend itself to a public challenge problem, which is a good means of leveraging the efforts and expertise of a broad community of researchers for the modest cost of organizing and releasing ASL data content. The synthesis problem discussed above could be an interesting challenge problem in itself.

One promising path for a challenge problem is to have an associated workshop at a prestigious conference in a related field. Several conferences to consider are:

- IEEE Computer Vision and Pattern Recognition (CVPR)
- ACM Multimedia (http://www.acmmm.org)
- Interspeech
- International Conference on Language Resources and Evaluation (LREC)
- Speech and Language Processing for Assistive Technologies (SPLAT) (http://www.slpat.org/events.php)
    - Often hosted at Interspeech or LREC

Additionally, the SCALE (Summer Camp for Applied Language Exploration) program (https://hltcoe.jhu.edu/research/scale/), annually hosted by the Johns Hopkins University (JHU) Center of Excellence (COE), may be a good forum to consider for hosting a challenge where investigators from across the United States Government (USG), the National Laboratories, and academia can deeply engage with the problem domain.

This page intentionally left blank.

5. CONCLUDING REMARKS AND FUTURE ROADMAP

This report summarizes a 12-month effort to evaluate the near-term feasibility of a portable device to enable spontaneous, fluid communication between a Deaf ASL signer and an English speaker. While such a capability would have been unthinkable only a few years ago, recent technical breakthroughs such as DNN-based object recognition and smartphones equipped with 2D/3D sensors and AI chipsets, provide the promise of successfully developing and deploying such a capability in the near future. Based on an extensive literature review, and experiments and prototypes developed during the course of this effort, a number of key findings regarding the feasibility of such technology were found as follows:

- Recent DNN breakthroughs can effectively model handshapes that, coupled with hand trajectories, achieve encouraging isolated ASL sign and fingerspelling performance. The Deep Hand DNN model can be further improved to better match it with the considered use cases.

- Language-model exploitation can significantly improve system performance over isolated sign recognition. Similar to the Automatic Speech Recognition (ASR) problem, the correlations between a signal and the temporally adjacent signals can be exploited to improve recognition performance.

- A 2D/3D sensing capability adds significant performance improvements over a 2D sensing capability. The 3D component of the signal enables the robust detection and tracking of the hand and other body parts far superior to the 2D signal. This is a crucially important step for successfully extracting the hand subimage from the collected image prior to the subimage's submission to the DNN handshape recognizer. While the 3D signal was not exploited for handshape recognition since a legacy DNN (DeepHand) system was exploited that was based on 2D signal exploitation, it is expected that a retrained system additionally leveraging the 3D signal would have significant performance improvements. While a Microsoft Kinect sensor was used for the 2D/3D sensing capabilities, a number of smartphones such as the iPhone X have cameras with similar sensing capabilities. Additionally, the iPhone X has an AI chip that may support the requisite computational loading for achieving real-time or near real-time translation capabilities. Significant engineering is likely required to integrate the smartphone sensors and compute infrastructure.

- While the English to ASL channel output could be potentially supported by a text output, an ASL signing avatar may be an attractive alternative for this communication channel. Several existing capabilities in academia and industry were discussed, though their current availability and performance in practice is uncertain.

Going forward, the primary challenge is to attain suitable ASL recognition performance corresponding to the relevant use cases. While there are a number of technology gaps to address, the most

immediate challenge is the lack of annotated data matched to these use cases. The traditional approach to solving this challenge, as historically done with ASR, is the formal collection and annotation of datasets. One technical approach discussed was leveraging low-resource (e.g., closed-captioned) ASL content for bootstrapping a translation system. An alternative approach under consideration for continued investigation, as discussed in the Introduction and Section 5.4, is to leverage synthesized data to augment the availability of ASL data for the training process. If successful, this would greatly reduce the amount of data needed to be collected to only that data required for evaluation and for seeding the synthesis process. More importantly, the success of this approach would have great impact across a broad range of USG low-resource applications with limited data availability.

5.1 TECHNOLOGY CAPABILITIES AND DEPENDENCIES

The following tables provide a high-level summary of technical capabilities and their interdependencies required for implementing a mobile ASL recognition and translation capability. This includes considered technical approaches and potential dependencies and requirements, as well as references to more detailed discussions later in the document. There are separate tables for the Sign to Speech, Speech to Sign, and Mobility-related hardware challenges.

Table 4

Sign to Speech

Capability	Technical Approach	Dependencies/Requirements
Face/Head Extraction	Onboard skeletonization	• Leverage onboard capabilities if available. • Face detection capabilities are broadly available.
	Face detector	
Hand Extraction	Onboard skeletonization	
	Hand detector	• Evaluate and train hand detector (see performance-related capability issues below). • 3D component of signal is important for effective performance.
Feature Extraction	DNN-based model (See Section 3.1)	Compute (see mobility-related issues)
Recognition (See Section 3.1)	Sign model	Performance degradation
	Language model	
Translation (See Section 3.1.3)	Cross-Language model	
System **Performance**	Retrain core classifiers (See Sections 3.1 & 4)	Evaluation dataset matched to use case(s) **and**
		Train datasets leveraging any combination of the following: • Dataset(s) matched to use case(s) • Dataset(s) partially matched to use case(s) • Synthetic data (See Section 4.4)
	Leverage 3D sensor signal	• Modest computational cost • Retrain core classifiers (see above)

Table 5

Speech to Sign

Capability	Technical Approach	Dependencies/Requirements
Translation (See Section 3.1.3)	See ASL to English translation	
User Interface	Text	
	Avatar	Technology exists, but its availability is unclear. (See Section 3.2)

Table 6

Mobility-Specific Issues

Capability	Technical Approach	Dependencies/Requirements
Signal Acquisition (See Section 3.3.1)	Leverage smartphone-based sensors (newer versions have 2D/3D cameras)	
Compute (See Section 3.3.2)	Leverage smartphone-resident AI/GPU chip	
	Leverage cloud solution	Communication channel bandwidth, security, and latency
	Retrain classifiers to optimize computational performance	See performance under ASL to English recognition and translation

6. REFERENCES

1. Oliver Adams, Graham Neubig, Trevor Cohn, Steven Bird, Quoc Truong Do, and Satoshi Nakamura, Learning a Lexicon and Translation Model from Phoneme Lattices, Proceedings of the Conference on Empirical methods in Natural Language Processing, pg. 2377-2382, Austin Texas, pp. 1-5 November 2016.

2. Simon Alexanderson and Jonas Beskow, Towards Fully Automated Motion Capture of Signs – Development and Evaluation of a Key Word Signing Avatar, ACM Transactions on Accessible Computing (TACESS) – Special Issue on Speech and Language Processing for AT, Volume 7 Issue 2, July 2016.

3. Antreas Antoniou, Amos Storkey, Harrison Edwards, Data Augmentation, Generative Adversarial Networks, arXiv:1711.04340, November 2017.

4. Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali, The American Sign Language Lexicon Video Dataset, Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR) Conference, Workshop for Human Communicative Behavior Analysis, June 2008.

5. C. Fabian Benitez-Quiroz, Kadir Gökgöz, Ronnie B. Wilbur, Aleix M. Martinez, Discriminant Features and Temporal Structure of Nonmanuals in American Sign Language, PLoS ONE 9(2), January 2014.

6. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

7. L.C. Barczak, N.H. Reyes, M. Abastillas, A. Piccio, and T. Susnjak; A New 2D Static Hand Gesture Colour Image Dataset for ASL Gestures; Res. Lett. Inf. Math Sci., 2011, Vol. 15, pp. 12-20.

8. Britta Bauer and Karl-Friedrich Kraiss, Toward an Automatic Sign Language Recognition System Using Subunits, International Gesture Workshop (GW), Gesture and Sign Language in Human-Computer Interaction, May 2002.

9. S. Bhowmick, S. Kumar, and A. Kumar, "Hand Gesture Recognition of English Alphabets using Artificial Neural Network," IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS) 2015.

10. J. Albert Bickford and Kathy Fraychineaud, "Mouth Morphemes in ASL: A Closer Look," Theoretical Issues in Sign Language Research Conference, Florianopolis, Brazil, December 2006.

11. Patrick Buehler, Mark Everingham, and Andrew Zisserman, “Learning Sign Language by Watching TV (Using Weakly Aligned Subtitles),” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

12. D. Stein, J. Bungeroth, H. Ney, “Morpho-Syntax Based Statistical Methods for Sign Language Translation,” In: 11th EAMT, Oslo, Norway, pp. 169–177 (2006).

13. Xiujuan Chai, Guang Li, Yushun Lin, Zhihao Xu, Yili Tang, Xilin Chen, Ming Zhou, “Sign Language Recognition and Translation with Kinect,” The 10th IEEE International Conference on Automatic Face and Gesture Recognition (FG2013), Apr. 22-26, 2013, Shanghai, China.

14. Xiujuan Chai, Guang Li, Xilin Chen, Ming Zhou, Guobin Wu, Hanjing Li, “VisualComm: A Tool to Support Communication between Deaf and Hearing Persons with the Kinect,” ASSETS’13: The 15th International ACM SIGACCESS Conference on Computers and Accessibility Proceedings.

15. Xiujuan Chai, Hanjie Wang, Fang Yin and Xilin Chen, “Communication tool for the hard of hearings,” The Sixth International Conference on Affective Computing and Intelligent Interaction (ACII2015), 2015, Xi'an, China.

16. W. Chan, N. Jaitly, Q.V. Le, and O. Vinyals, “Listen, attend and spell,” arXiv preprint arXiv:1508.01211, 2015.

17. Caitlin S. Channer, “Coarticulation in American Sign Language, Linguistics PhD Disseration,” University of New Mexico, Dec. 1, 2012.

18. Xilin Chen, Hanjing Li, Tim Pan, Stewart Tansley, Ming Zhou, “Kinect Sign Language Translator Expands Communication Possibilities,” Microsoft Research, 2013.

19. Xilin Chen, “Sign Language Recognition and Translation Based on Kinect,” Institute of Computing Technology, Chinese Academy of Sciences, Microsoft Research Asia Faculty Summit 2012.

20. Christopher Conly, Zhong Zhang, and Vassilis Athitsos, “An Integrated RGB-D System for Looking Up the Meaning of Signs,” Pervasive Technologies Related to Assistive Environments (PETRA), July 2015.

21. Helen Cooper and Richard Bowden, “Learning Signs from Subtitles: A Weakly Supervised Approach to Sign Language Recognition,” In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, pp. 2568 - 2574, 2009.

22. Helen Cooper, B Holt, & Richard Bowden, “Sign language recognition,” In Visual Analysis of Humans (pp. 539-562). Springer London, 2011.

23. Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, Richard Bowden, “Sign Language Recognition using Sub-Units,” Journal of Machine Learning Research,13(Jul):2205–2231, 2012.

24. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

25. Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M. Rush, “What You Get Is What You See: A Visual Markup Decompile,” arXiv preprint arXiv:1609.04938, 2016.

26. Liya Ding and Alei M. Martinez, “Modelling and Recognition of the Linguistic Components in American Sign Language, Image and Vision Computing,” 27(12):1826-1844, November 2009.

27. Cao Dong, Ming C. Leu, and Zhaozheng Yin, “American Sign Language Alphabet Recognition Using Microsoft Kinect,” IEEE Computer Vision and Pattern Recognition (CVPR) HANDS Workshop 2015.

28. Philippe Dreuw, Carol Neidle, Vassilis Athitsos, Stan Sclaroff, and Hermann Nye, “Benchmark Databases for Video-Based Automatic Sign Language Recognition,” International Conference on Language Resources and Evaluation (LREC) Workshop, Morocco 2008.

29. Philippe Dreuw, Jens Forster, Hermann Ney, “Tracking Benchmark Databases for Video-Based Sign Language Recognition,” Trends and Topics in Computer Vision European Conference on Computer Vision (ECCV), pages 286-297, 2010.

30. J. Forster, O. Koller, C. Oberdorfer, Y. Gweth, H. Ney, “Improving Continuous Sign Language Recognition: Speech Recognition Techniques and System Design,” 4th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT) 2013.

31. Brandon Garcia and Sigberto Alarcon Viesca, “Real-time American Sign Language Recognition with Convolutional Neural Networks,” Stanford University CS231n Convolutional Neural Networks for Visual Recognition Report 214, Spring 2016.

32. Srujana Gattupalli, “Sign Gesture Spotting in American Sign Language using Dynamic Space Time Warping,” Master's Thesis, University of Texas at Arlington, May 2013.

33. Isabelle Guyon, Vassilis Athitsos, Pat Jangyodsuk, and Hugo Jair Escalante, “The ChaLearn Gesture Dataset, Machine Vision and Applications,” Volume 25, Issue 8, pp 1929–1951, November 2014.

34. Mary Harper, “The Automatic Speech Recognition in Reverberant Environments (ASpIRE) Challenge,” 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ, 2015, pp 547-554.

35. Matt Huenerfauth and Pengfei Lu, “Eliciting Spatial Reference for a Motion-Capture Corpus of American Sign Language Discourse,” 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, 2010.

36. Matt Huenerfauth, Pengfei Lu, Hernisa Kacorri. 2015, “Synthesizing and Evaluating Animations of American Sign Language Verbs Modeled from Motion-Capture Data, “Proceedings of the 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), INTERSPEECH 2015, Dresden, Germany.

37. Pat Jangyodsuk, Christopher Conly, and Vassilis Athitsos, “Sign Language Recognition using Dynamic Time Warping and Hand Shape Distance Based on Histogram of Oriented Gradient Features,” Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments (PETRA), Article 50, 2014.

38. B. H. Juang and Lawrence R. Rabiner, “Automatic Speech Recognition — A Brief History of the Technology Development,” in Elsevier Encyclopedia of Language and Linguistics, Second Edition, Elsevier, 2005.

39. Aradhana Kar and Pinaki Sankar Chatterjee, “An Approach for Minimizing the Time Taken by Video Processing for Translating Sign Language to Simple Sentences in English,” International Conference on Computational Intelligence and Networks, 2015.

40. Purushottam Kar, Achla M. Raina, Amitabha Mukerjee, “Spoken and Sign Languages: A Cross Modal Study,” 28th All India Conference of Linguists, Banaras Hindu University, 2006.

41. K. Kawahigasi, Y. Shirai, J. Miura and N. Shimada “Automatic Synthesis of training Data for Sign Language Recognition Using HMM,” PROC.ICCHP, pp.623-626, 2006.

42. Jonathan Keane, Dianne Brentari, and Jason Riggle, “Coarticulation in Fingerspelling,” In Stefan Keine and Sjayne Sloggett (eds.), Proceedings from the North East Conference Linguistic Society (NELS), 42, vol. 1, 261-272, 2012.

43. C. Keskin, F. Kirac, Y. E. Kara, and L. Akarun, “Real Time Hand Pose Estimation using Depth Sensors. 2011 Computer Vision Workshops, pp. 1228-1234, 2011.

44. Taehwan Kim, Jonathan Keane, Weiran Wang, Hao Tang, Jason Riggle, Gregory Shakhnarovich, Diane Brentari, and Karen Livescu, “Lexicon–Free Fingerspelling Recognition from Video: Data, Models, and Signer Adaptation,” Computer Speech & Language, Elsevier, Vol. 46, Nov. 2017.

45. T. Kim, J. Keane, W. Wang, H. Tang, J. Riggle, G. Shakhnarovich, D. Brentari, and K. Livescu, “Lexicon-Free Fingerspelling Recognition from Video: Data, Models, and Signer Adaptation,” Computer Speech and Language (to appear).

46. Diederik P Kingma and Max Welling, “Auto-Encoding Variational Bayes,” International Conference on Learning Representations (ICLR), 2014.

47. Oscar Koller, J. Forster, and H. Ney, “Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers,” Computer Vision and Image Understanding, volume 141, pages 108-125, December 2015.

48. Oscar Koller (Koller 2016a), 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,” IEEE Computer Vision and Pattern Recognition (CVPR) Conference, Las Vegas, 2016.

49. Oscar Koller (Koller 2016b), Hermann Ney, and Richard Bowden, “Automatic Alignment of HamNoSys Subunits for Continuous Sign Language Recognition,” International Conference on Language Resources and Evaluation (LREC), Workshop, 2016.

50. W.W. Kong, Surenda Rananath, “Automatic Hand Trajectory Segmentation and Phoneme Transcription for Sign Language,” Proceedings Inter. Conf. Automatic Face and Gesture Rec, 2008, 1-6.

51. W.W. Kong and Surenda Rananath, “Towards subject independent continuous sign language recognition: a segment and merge approach,” Pattern Recognition, 47(2014), 1294-1308.

52. Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” Advances in Neural Information Processing Systems (NIPS), 2012.

53. A. Kuznetsova, L.L. Taixe, and B. Rosenhahn, “Real-time sign language recognition using a consumer depth camera,” Computer Vision Workshops, pp. 83-90, 2013.

54. H. Lane, R. Hoffmeister, and B. Bahan, “A Journey into the Deaf-World,” DawnSign Press, San Diego, CA, 1996.

55. Aaron Lawson, M. Linderman, M. Leonard, A. Stauffer, B. Pokines, and M. Carlin, “Perturbation and Pitch Normalization as Enhancements to Speaker Recognition,” 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, 2009, pp 4533-4536.

56. Tuan Anh Le, Atilim Gunes Baydin, Robert Zinkov, Frank Wood, “Using Synthetic Data to Train Neural Networks Is Model-Based Reasoning,” In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 9–14 May 2017; pp. 3514–3521.

57. Yushun Lin, Xiujuan Chai, Yu Zhou, and Xilin Chen, “Curve Matching from the View of Manifold for Sign Language Recognition,” The 12th Asian Conference on Computer Vision (ACCV 2014), Workshop on Feature and Similarity Learning for Computer Vision (FSLCV14), 2014, Singapore.

58. Jingjing Liu, Bo Liu, Shaoting Zhang, Fei Yang, Peng Yang, Dimitris N. Metaxas, Carol Neidle, "Non-manual Grammatical Marker Recognition Based on Multi-scale Spatio-temporal analysis of Head Pose and Facial Expression," Image and Vision Computing, Elsevier, 32 (2014) pp. 671-681.

59. Pengfei Lu, Matt Huenerfauth. 2011, "Data-Driven Synthesis of Spatially Inflected Verbs for American Sign Language Animation," ACM Transactions on Accessible Computing. Volume 4 Issue 1, November 2011.

60. Pengfei Lu and Matt Huenerfauth, "CUNY American Sign Language Motion-Capture Corpus: First Release," Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, The 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.

61. Pengfei Lu and Matt Huenerfauth, "Collecting and Evaluating the CUNY ASL Corpus for Research on American Sign Language Animation," Computer Speech & Language, Vol. 28, Issue 3, pages 812-831, May 2014.

62. Pengfei Lu and Matt Huenerfauth, "Learning a Vector-Based Model of American Sign Language Inflecting Verbs from Motion-Capture Data," Proceedings of the 3rd Workshop on Speech and Language Processing for Assistive Technologies (SPLAT), Montreal, June 2012.

63. M.-T. Luong and C.D.Manning, "Stanford neural machine translation systems for spoken language domains," in Proceedings of the International Workshop on Spoken Language Translation, 2015.

64. Aleix M. Martinez, Ronnie B. Wilbur, Robin Shay, and Avi Kak, "Purdue RVL-SLLL ASL Database for Automatic Recognition of American Sign Language," Proceedings of IEEE International Conference on Multimodel Interfaces, 2002.

65. Lucas Matney, "Ava Gives the Deaf and Hard-of-Hearing a More Present Voice in Group Conversations," TechCrunch, November 21, 2016.

66. J. McDonald, R. Wolfe, J. Schnepp, J. Hochgesang, D. G. Jamrozik, M. Stumbo, L Berke, M. Bialek, F. Thomas, "An automated technique for real-time production of lifelike animations of American Sign Language," Universal Access in the Information Society 14(4):2016. pp. 551-566.

67. Rachel Metz, "How Armbands Can Translate Sign Language," MIT Technology Review, February 2016.

68. Ross Mitchell, Travas Young, Travas; Bellamie Bachleda, and Michael Karchmer (2006). "How Many People Use ASL in the United States?: Why Estimates Need Updating" (PDF). Sign Language Studies. Gallaudet University Press. 6 (3).

69. Sunita Nayak, “Representation and learning for sign language recognition,” PhD Dissertation, University of South Florida, 2008.

70. Carol Neidle and Christian Vogler, “A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface (DAI),” Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012, Istanbul, Turkey.

71. Carol Neidle, Ashwin Thangali, and Stan Sclaroff, “Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus,” 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012, Istanbul, Turkey, May 27, 2012.

72. Carol Neidle, Judy Kegl, Dawn MacLaughlin, Benjamin Bahan, Robert G. Lee, *The Syntax of American Sign Language, Functional Categories and Hierarchical Structure*, The MIT Press, Cambridge MA, 2000.

73. Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout, “Multi-Scale Deep Learning for Gesture Detection and Localization,” In ECCV ChaLearn Workshop on Looking at People, 2014.

74. Tan Dat Nguyen and Surendra Ranganath, “Facial Expressions in American Sign Language: Tracking and Recognition, Pattern Recognition 45” (2012), Elsevier, pp 1877-1891.

75. Xunbo Ni, Gangyi Ding, Xunran Ni, Xunchao Ni,Qiankun Jing, JianDong Ma, Peng Li, and Tianyu Huang, “Signer-Independent Sign Language Recognition Based on Manifold and Discriminative Training,” ICICA 2013, Part I, CCIS 391, pp. 263–272, 2013.

76. Cham Leang and Patricia O’Neill-Brown, “CIA’s Accessibility Technologies: Present and Future,” 33rd Annual Pacific Rim (PacRim) International Conference on Disability and Diversity, Honolulu, Oahu, Hawaii, October 9, 2017.

77. Markus Oberweger, Paul Wohlhart, and Vincent Lepetit, “Hands Deep in Deep Learning for Hand Pose Estimation,” In Proceedings of 20th Computer Vision Winter Workshop (CVWW), 2015.

78. Arika Okrent, “Why Great Sign Language Interpreters Are So Animated,” The Atlantic, November 2, 2012.

79. Eng-Jon Ong and Richard Bowden, “Learning Sequential Patterns for Lipreading,” Proceedings of the British Machine Vision Conference (BMVC), pages 55.1-55.10. BMVA Press, September 2011.

80. Eng-Jon Ong, Helen Cooper, Nicolas Pugeault, Richard Bowden, “Sign language Recognition using Sequential Pattern Trees,” IEEE Computer Vision and Pattern Recognition Conference (CVPR), pp. 2200-2207, 2012.

81. Achraf Othman, Zouhour Tmar, Mohamed Jemni, “Toward Developing a Very Big Sign Language Parallel Corpus,” International Conference on Computers for Handicapped Persons (ICCHP), pages 192-199, 2012.

82. Achraf Othman and Mohamed Jemni, “English-ASL Gloss Parallel Corpus 2012: ASLG-PC12,” 5th Workshop on the Representation and Processing of Sign Languages, International Conference on Language Resources and Evaluation, 2012.

83. Achraf Othman and Mohamed Jemni, “A Novel Approach for Translating English Statements to American Sign Language Gloss,” International Conference on Computers for Handicapped Persons (ICCHP), 2014.

84. David S. Pallett, “A Look at NIST’s Benchmark ASR Test: Past, Present, and Future,” Proc. IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop 2003, pp. 483-488, Virgin Islands, Dec. 2003.

85. Prajwal Paudyal, Ayan Banerjee, and Sandeep K.S. Gupta, “SCEPTRE: A Pervasive, Non-Invasive, and Programmable Gesture Recognition Technology,” Proc. Of the 21st Intl. Conf. on Intelligent User Interfaces (IUI), NY, NY, March 2016.

86. David M. Perlmutter, “What is Sign Language?, Linguistic Society of America,” https://www.linguisticsociety.org/files/Sign_Language.pdf.

87. Tomas Pfister, J Charles, A Zisserman, “Large-Scale Learning of Sign Language by Watching TV (Using Co-occurences),” British Machine Vision Conference, 2013.

88. Tomas Pfister, J Charles, A Zisserman, “Domain-adaptive Discriminative One-shot Learning of Gestures,” European Conference on Computer Vision, 2014.

89. Tomas Pfister, “Advancing Human Pose and Gesture Recognition,” DPhil Thesis, University of Oxford, April 2015.

90. Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen, “Sign Language Recognition Using Convolutional Neural Networks,” ECCV 2014 Workshops, Part I, LNCS 8925, pp. 572–578, 2015.

91. Lionel Pigou, Mieke Van Herreweghe and Joni Dambre, “Sign classification in sign language Corpora with deep neural networks,” International Conference on Language Resources and Evaluation (LREC), Workshop, Proceedings. pg. 175-178, 2016.

92. Lionel Pigou, Mieke Van Herreweghe and Joni Dambre, “Gesture and Sign Language Recognition with Temporal Residual Networks,” Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR) Conference, 2017.

93. Nicolas Pugeault and Richard Bowden, “Spelling It Out: Real-Time ASL Fingerspelling Recognition,” 1st IEEE Workshop on Consumer Depth Cameras for Computer Vision, International Conference on Computer Vision (ICCV), Barcelona Spain, 2011.

94. Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra, “Stochastic Backpropagation and Approximate Inference in Deep Generative Models,” International Conference on Machine Learning (ICML), 2014.

95. Ashok K Sahoo, Gouri Sankar Mishra and Kiran Kumar Ravulakollu, “Sign Language Recognition: State of the Art (REVIEW ARTICLE),” Asian Research Publishing Network (ARPN) Journal of Engineering and Applied Sciences, Vol 9, No 2, February 2014.

96. Pushkar Shukla, Abhisha Garg, Kshitij Sharma, and Ankush Mittal, “A DTW and Fourier Descriptor based approach for Indian Sign Language Recognition,” 3rd International Conference on Image Information Processing, 2015.

97. SignAll, “SignAll Reveals the Future of Deaf Accessibility at CES 2018!,” https://www.businesswire.com/news/home/20180104006011/en/SignAll-Reveals-Future-Deaf-Accessibility-CES-2018%21, BusinessWire, 2018.

98. Matt Simon, “The Remarkable Tech Bringing the Deaf and Hearing Worlds Together,” WIRED Magazine, June 2016.

99. K. Simonyan, A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv Technical Report, 2014.

100. d'Armond L. Speers, “Representation and learning for sign language recognition,” PhD Dissertation, Georgetown University, 2002.

101. T. Starner and A. Pentland, “Real-time American Sign Language recognition using desk and wearable computer based video,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(12):1371–1375, 1998.

102. Thad Starner, Joshua Weaver, and Alex Pentland, “A Wearable Computer-based American Sign Language Recognizer,” Personal Technologies, Volume 1 Issue 4, pages 241–250, 1997.

103. W Stokoe, “Sign Language Structure: An Outline of the Visual Communication System of the American Deaf, Studies in Linguistics: Occasional Papers” Linstok Press, Silver Spring, MD, 1960.

104. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going Deeper with Convolutions,” arXiv:1409.4842 [cs],, Sept. 2014.

105. Rachael Tatman, “Beyond the Hands: Hurdles in Automatic Sign Language Recognition,” Presentation to HLT Group at MIT Lincoln Laboratory, 27 June 2017.

106. A Thangali, J Nash, S Sclaroff, C Neidle, “Exploiting Phonological Constraints for Handshape Inference in ASL Video,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011.

107. Clayton Valli (Editor), Peggy Swartzel Lott (Illustrator), Daniel Renner (Illustrator), Rob Hills (Illustrator), *The Gallaudet Dictionary of American Sign Language*, Gallaudet University Press, 2006.

108. Clayton Valli, Ceil Lucas, Kristin J. Mulrooney, Miako N. P. Rankin, “Linguistics of American Sign Language An Introduction,” 5th Edition, Gallaudet University Press, Washington, D.C., 2011.

109. Haijing Wang, Alexandra Stefan, Sajjad Moradi, Vassilis Athitsos, Carol Neidle, and Farhad Kamangar, “A System for Large Vocabulary Sign Search,” Workshop on Sign, Gesture and Activity (SGA), September 2010.

110. Hanjie Wang, Jingjing Fu, Yan Lu, Xilin Chen, Shipeng Li, “Depth Sensor Assisted Real-time Gesture Recognition for Interactive Presentation,” Journal of Visual Communication and Image Representation. vol.24, no.8, pp.1458-1468, 2013.

111. R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, “Sequence-to-sequence models can directly transcribe foreign speech,” arXiv preprint arXiv:1703.08581, 2017.

112. Ronnie Wilbur and Avinash C. Kak, “Purdue RVL-SLLL American Sign Language Database, School of Electrical and Computer Engineering,” Technical Report TR-06-12, Purdue University, September 2006.

113. Poline Yanovich, Carol Neidle, and Dimitris Metaxas, “Detection of Major ASL Sign Types in Continuous Signing for ASL Recognition,” Language Resources and Evaluation Conference (LREC), Portorož, Slovenia. May 25-27, 2016.

114. Fang Yin, Xiujuan Chai and Xilin Chen, “Iterative Reference Driven Metric Learning for Signer Independent Isolated Sign Language Recognition,” 14th European Conference on Computer Vision (ECCV2016), 2016, Amsterdam, Netherlands.

115. Fang Yin, Xiujuan Chai, Yu Zhou and Xilin Chen, “Semantics Constrained Dictionary Learning for Signer-independent Sign Language Recognition,” The 22nd IEEE International Conference on Image Processing (ICIP2015), 2015, Quebec City, Canada.

116. Fang Yin, Xiujuan Chai, Yu Zhou and Xilin Chen, “Weakly Supervised Metric Learning towards Signer Adaptation for Sign Language Recognition,” British Machine Vision Conference (BMVC2015), 2015, Swansea, UK.

117. Hanjie Wang, Xiujuan Chai, Yu Zhou and Xilin Chen, “Fast Sign Language Recognition Benefited From Low Rank Approximation,” The 11th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015), 2015, Slovenia.

118. Hanjie Wang, Xiujuan Chai, Xiaopeng Hong, Guoying Zhao and Xilin Chen, “Isolated Sign Language Recognition with Grassmann Covariance Matrices,” ACM Transactions on Accessible Computing, 2016, 8(4):14.

119. Y. Zhang, W. Chan, and N. Jaitly, “Very deep convolutional networks for end-to-end speech recognition,” arXiv preprint arXiv:1610.03022, 2016.

120. Xingyi Zhou, Quingfu Wan, Wei Zhang, Xiangyang Xue, and Yichen Wei, “Model-based Deep Hand Pose Estimation,” Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI), pg. 2421-2427, NY, NY, 2016.

121. Workshop Proceedings, 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, Language Resources and Evaluation Conference (LREC), Portoroz, Solovenia, 28 May 2016.

This page intentionally left blank.

7. ONLINE RESOURCES AND DICTIONARIES

1. Signing Math & Science, https://signsci.terc.edu/.
2. Rochester Institute of Technology (RIT) National Technical Institute for the Deaf (NTID) ASL Video Dictionary and Inflection Guide, https://www.rit.edu/ntid/dictionary/.
3. ASL Sign Language Dictionary, http://www.handspeak.com/word/.
4. ASL Sign Language Translation, http://www.handspeak.com/translate/.
5. Start ASL American Sign Language Dictionary, https://www.start-american-sign-language.com/american-sign-language-dictionary.html.
6. Microsoft Cognitive Services - Emotion API, https://azure.microsoft.com/en-us/services/cognitive-services/emotion/.
7. Karen Nakamura, About ASL: Deaf Resource Library, http://www.deaflibrary.org/asl.html.
8. DPAN.TV, The Sign Language Channel: https://dpan.tv/.