Finger Gesture Tracking for Interactive Applications: A Pilot Study with Sign Languages

YILIN LIU, The Pennsylvania State University
FENGYANG JIANG, The Pennsylvania State University
MAHANTH GOWDA, The Pennsylvania State University

This paper presents *FinGTrAC*, a system that shows the feasibility of fine grained finger gesture tracking using low intrusive wearable sensor platform (smart-ring worn on the index finger and a smart-watch worn on the wrist). The key contribution is in scaling up gesture recognition to hundreds of gestures while using only a sparse wearable sensor set where prior works have been able to only detect tens of hand gestures. Such sparse sensors are convenient to wear but cannot track all fingers and hence provide under-constrained information. However application specific context can fill the gap in sparse sensing and improve the accuracy of gesture classification. Rich context exists in a number of applications such as user-interfaces, sports analytics, medical rehabilitation, sign language translation etc. This paper shows the feasibility of exploiting such context in an application of American Sign Language (ASL) translation. Noisy sensor data, variations in gesture performance across users and the inability to capture data from all fingers introduce non-trivial challenges. *FinGTrAC* exploits a number of opportunities in data preprocessing, filtering, pattern matching, context of an ASL sentence to systematically fuse the available sensory information into a Bayesian filtering framework. Culminating into the design of a Hidden Markov Model, a Viterbi decoding scheme is designed to detect finger gestures and the corresponding ASL sentences in real time. Extensive evaluation on 10 users shows a recognition accuracy of 94.2% for 100 most frequently used ASL finger gestures over different sentences. When the size of the dictionary is extended to 200 words, the accuracy is degrades gracefully to 90% thus indicating the robustness and scalability of the multi-stage optimization framework.

CCS Concepts: • **Human-centered computing → Mobile devices**; **Ubiquitous and mobile computing design and evaluation methods**.

Additional Key Words and Phrases: IoT, Wearable, Gesture, Bayesian Inference

ACM Reference Format:
Yilin Liu, Fengyang Jiang, and Mahanth Gowda. 2020. Finger Gesture Tracking for Interactive Applications: A Pilot Study with Sign Languages. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 112 (September 2020), 21 pages. https://doi.org/10.1145/3414117

1 INTRODUCTION

Five fingers of a human hand posses more than 30 degrees of freedom. Tracking these degrees of freedom to detect motion of hand gestures might ideally require multiple sensors to be placed on the hand, which can be intrusive. This paper explores the limits of feasibility of finger gesture tracking using a single sensor placed on the finger as a ring and a smartwatch worn on the wrist. While such sparse sensors may not sufficiently

Authors' addresses: Yilin Liu, yzl470@psu.edu, The Pennsylvania State University, University Park, Pennsylvania; Fengyang Jiang, fzj28@psu.edu, The Pennsylvania State University, University Park, Pennsylvania; Mahanth Gowda, mkg31@psu.edu, The Pennsylvania State University, University Park, Pennsylvania.

2474-9567/2020/9-ART112 $15.00
https://doi.org/10.1145/3414117

capture all degrees of freedom, we present a system called FinGTrAC (**Fin**ger **G**esture **Tr**acking with **A**pplication **C**ontext) that exploits application specific context to fill in for the missing sensor data. For example, basketball players maintain a specific wrist angle, careful flexing of finger joints and an optimal positioning of index and middle fingers before shooting the ball [20]. Virtual reality applications have similar prior probabilities of finger configurations. We argue that such application specific context can fill the gap in sensing, and track the main finger motion metrics of interest. While prior works[22, 32, 50, 71, 77, 85] on wearable finger gesture tracking that use non-intrusive sensors are only limited to recognition of tens of gestures, we show how a similar setup can be scaled to recognition of hundreds of gestures by exploiting application specific context. Although we do not validate any generic claim in this paper over different applications, we make our case with an example application in American Sign Language (ASL) translation. A sign language is a way of communication that uses body motion (arms, hands, fingers) and facial expressions to communicate a sentence instead of auditory speech. We show the feasibility of translating sentences composed of 100 most frequently [51] used ASL words. While the sensor data is under-constrained, we fuse them with Bayesian inferencing techniques that exploit the context information in a ASL sentence towards achieving a higher accuracy.

Finger motion tracking has a number of important applications in user-interfaces and creating accessibility for patients with motor related disorders. In augmented reality, finger motion tracking enables the richness of user interaction with augmented objects. In sports, finger motion data can be used for coaching players as well as managing injuries. In smart-health, fine-grained finger motion data analysis can reveal digital biomarkers for various motor related diseases. Finger gestures for ASL translation – the subject of this paper – can significantly help deaf and hard of hearing (DHH) in communicating with people with normal hearing ability thus motivating more than two decades of research [12, 14]. The DHH population is about 10 million [17] in US, 466 million globally and estimated to be 900 million by 2050[82].

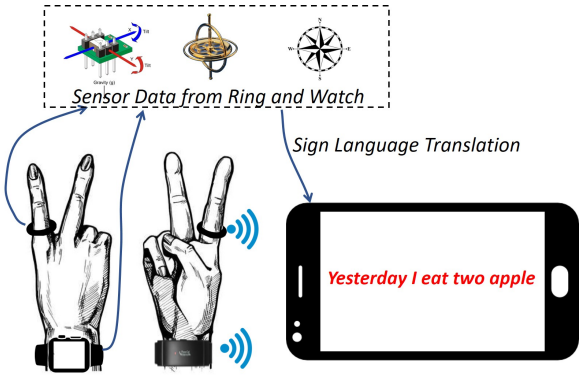


Fig. 1. ASL recognition with smart devices [40]

Prior works can be broadly classified into wearable based approaches and camera based approaches. Wearable based approaches are only limited to a few tens of gestures [22, 50, 77, 85] or use intrusive setup of sensor gloves (15-20 sensors) and Electromyography (EMG) electrodes for detecting upto hundred gestures [3, 21]. While cameras [33] track full finger motion, the accuracy depends on lighting/resolution as well as the presence of user in the camera view. Wearable based approaches in general offer more ubiquitous solution than cameras. An innovative work, SignSpeaker [25] performs translation of 103 ASL gestures with smartwatches using deep learning approaches [23] popular in speech recognition. Evaluated in Section 7, *FinGTrAC's* primary distinction lies in detecting any unseen sentence composed from a ASL dictionary, whereas SignSpeaker cannot detect unseen sentences outside the training database. Given the impracticality to train over infinitely possible sentences in a language, we believe FinGTrAC has non-trivial benefits over SignSpeaker.

In contrast to prior works, *FinGTrAC* differs in following ways: (1) *FinGTrAC* exploits application specific context in ASL for finger motion tracking and scaling gesture recognition from few tens to hundreds of gestures. (2) *FinGTrAC* uses a single sensor on one finger instead of using additional sensors on all fingers. (3) In addition to tracking hand motion during word gestures, *FinGTrAC* also tracks hand motion in between words of a sentence for higher accuracy. (4) *FinGTrAC* can detect arbitrarily new sentences without any training. (5) A minimal

training at the level of words (Section 5) by a single user is sufficient, and this model extends to any new user. Our recent work [40] presents a proof-of-concept presentation of this idea with limited evaluation. In contrast, this paper performs a thorough evaluation which includes performance characterization with different sensor placements, accuracy variation across users, comparison with state of the art approaches such as [25], potential to improvement in accuracy with human assistance etc. In addition, new techniques for finger-spelling detection are proposed and validated. The techniques proposed in the paper have been expanded with characterization of intermediate results. The raw sensor data is presented with preliminary analysis.

Fig. 1 depicts the system. A user would wear a smart ring on the index finger (index finger is the most involved finger in ASL finger gestures), and a smartwatch on the wrist. Since most people are comfortable wearing one ring and a watch in daily life, we chose this platform for the study. The Inertial Measurement Unit (IMU) sensors (accelerometers, gyroscopes, and magnetometers) on the smart ring and smart watch are used for finger tracking, and the results might be displayed on smartphone screen, or read out on speaker. A number of non-trivial challenges emerge as enumerated below. (1) The data is severely under-constrained. All fingers and both hands are involved in finger gestures, whereas *FinGTrAC* uses a sensor only on the index finger of the dominant hand. (2) The sensor data is noisy. Because of this subtle differences in ASL gestures cannot be captured accurately. (3) A huge diversity in gesturing exists across different users. *FinGTrAC* needs to maintain high accuracy and robustness despite such variations.

Towards addressing these challenges, *FinGTrAC* exploits a number of opportunities as enumerated below. (1) A ring on the index finger cannot sufficiently capture all degrees of motion of the hand to detect ASL words. Therefore, in addition to tracking finger motion during the sign gesture for a particular word, we also track hand motion in between words of a sentence. This provides a lot of contextual information. This opens up opportunities for higher accuracy gesture detection in the context of an ASL sentence than in isolation. (2) Extraction of high level motion features combined with techniques such as Dynamic Time Warping (DTW) can narrow down the search space for words while simultaneously increasing the robustness across diverse users. (3) The knowledge of the dictionary from which words are drawn can dramatically improve the accuracy of detection of finger-spelling signs. (4) Decoding words in a sentence can be modeled as a Bayesian Filtering problem (HMM) analogous to packet decoding over a noisy wireless channel. Thus, we can leverage Viterbi decoding to provide optimal sentence decoding that makes best use of the noisy sensor data.

FinGTrAC is implemented on a platform consisting of: (1) An off-the-shelf button shaped sensor (VMU931[80]) worn on the index finger like a ring. (2) SONY Smartwatch 3 SWR50 worn on the wrist. The sensor data from the watch and ring is streamed continuously to an edge device (desktop) which decodes sentences in real time (0.4 seconds even for 10 word sentences). With a study from 10 users, *FinGTrAC* achieves an accuracy of 94.2% in detecting words from 50 sentences composed from 100 most commonly used ASL words. Furthermore, scalability analysis shows that extending the dictionary size to 200 words offers a graceful degradation in accuracy.

To our best knowledge, *FinGTrAC* outperforms other wearable ASL solutions that use 15-20 sensors [3, 21]. On the other hand, in contrast to low intrusive setup that can track few tens of gestures, FinGTrAC scales recognition to hundreds of gestures. Besides what is achievable, we also discuss shortcomings. Our goal is to show the feasibility of finger motion tracking by exploiting application specific opportunities, and we use ASL as a case-study with a dictionary of 100 most frequently used ASL words. However, we note that ASL has a larger vocabulary, facial expressions, and a rich grammar. The complexity is comparable to spoken languages. Hence, full ASL translation is outside the scope of our paper, but we believe we make a significant first step. We discuss opportunities in sensing, machine learning and natural language processing towards extending this work for full ASL translation (Section 9). Considering this, our contributions are enumerated next:

(1) To the best of our knowledge, FinGTrAC is the first system that shows the limits and feasibility of using a single smart ring and a smart watch (low-intrusive) for finger motion tracking with application specific context from ASL.
(2) FinGTrAC scales gesture recognition to hundreds of gestures in contrast to tens of gestures in prior work but only requires a low fidelity training from a single user (user independent training). In the context of ASL, the recognition generalizes to unseen new sentences composed from a given dictionary of words.
(3) FinGTrAC systematically combines information from the sensors in the context of an ASL sentence into a Bayesian inferencing model for high accuracy sentence decoding.
(4) Implementation is done on user-friendly platforms of smart-ring and smart-watch for detection in real time.
(5) A study with 10 users on 50 sentences from a dictionary of 100 most popular words shows an accuracy of 94.2% .

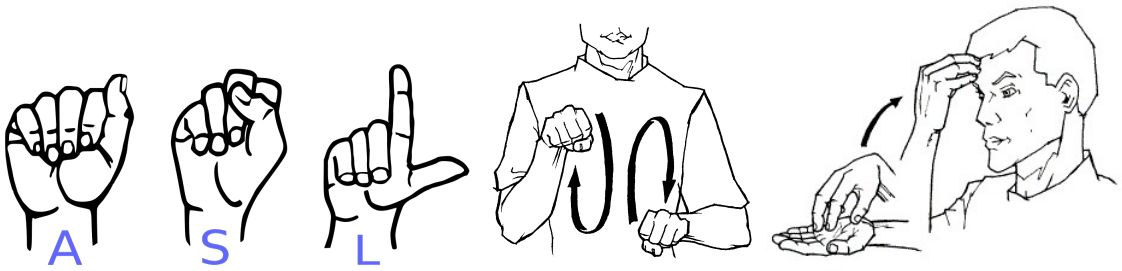
2 BACKGROUND: APPLICATION DOMAIN

We begin with a brief overview of sign languages, hand gestures, and ASL grammar.

Signing and Gestures: Sign languages use gestures instead of sound for communication. Sign languages including ASL are considered a class of natural languages with their own grammar and lexicon. Approximately, there are 137 sign languages with millions of speakers. ASL is primarily used by the in USA and parts of Canada. Majority of ASL signs involve the motion of the dominant hand including fingers. Fig. 2(a) shows the hand and finger poses for signing the letters - A, S and L. Fig. 2(b),(c) shows hand motions involved in signing “bike” and “learn”. As shown in Figure, the non-dominant hand can also be a part of ASL signs. The non-dominant hand and facial expressions are used occasionally to complement the dominant hand gestures. For example, eyebrows are raised at the end of a sentence to ask a yes/no question. ASL signs include words, 26 alphabets, and 10 digit signs [75]. The vocabulary of an ASL user is a few thousand words. Finger-spellings are used to sign words for which no direct ASL sign exist (ex: "sensor" does not have an ASL sign). After a new word is introduced by finger-spelling, a reference is created in space for the new word. Subsequent communications that reuse the same word can simply point the hands or gesture towards the spatial reference without spelling the word each time.

Table 1. ASL sentence vs English sentence

ASL Sentence	English Sentence
I Drink Water	I need to drink water
My Mother Sleep Home	My mother is sleeping at home
Ball, Boy Throw	The ball was thrown by the boy
Bathroom where?	Where is the bathroom?



(a) ASL Finger-spelling (b) ASL sign: Bike (c) ASL sign: Learn

Fig. 2. Examples of hand poses and motion in ASL

ASL Grammar: ASL grammar mainly consists of sentences in the following format: *Subject-Verb-Object* [5]. Sometimes, the word order can change when the context of the topic is established first through topicalization [59]. This results in the following format: *Object, Subject-Verb*. Table 1 shows some example ASL sentences and corresponding English translations. Facial expressions are used to denote emphasis, questions etc. Pauses indicate

end of sentences and words [24]. ASL does not have a copula [59] such as “is" to connect subject and predicate and it does not use BE verbs (am, is, are, was, were). Also, the signs do not indicate tense of a verb such as *washed* or *washing*. Therefore, the following format is often used to resolve ambiguity [37]: Time-Subject-Verb-Object. For example, WEEK-PAST I WASH MY CAR is the equivalent to saying *I washed my car last week* in English with the appropriate time and tense. Similarly, NONE/NOT is introduced at end of sentences to indicate negation. For example, TENNIS I LIKE PLAY NOT is the ASL equivalent to *I don’t like to play tennis.*

3 PLATFORM DESCRIPTION

Our ideal platform consists of a smart-ring worn on a finger and a smart-watch worn on the wrist. The ring is worn on the index finger since the index finger is more frequently involved in gestures. Smart rings that can pair with phones wirelessly to stream information as well a monitor activity are already available on the market [46, 55]. For example, the oura ring [55] is popular as a sleep tracking device and weighs between 4 and 6 grams, which is even lighter than conventional rings, and packaged in a stylish design. It is low intrusive with users finding it comfortable for wearing day and night, gym, pool etc [52], thus receiving favorable online reviews for usability [52–54]. However, most of these platforms are closed and do not provide access to raw sensor data. Therefore we use a button-shaped sensor – VMU931[80] snugly fit on the finger like a ring as shown in Fig. 3. Given that VMU931 does not support wireless streaming, we connect it with a USB cable to a Raspberry PI, which streams data to a desktop over USB WiFi.The ring (VMU931) and watch (SONY SmartWatch 3 SWR50), both generate 9 axis IMU data during ASL finger gestures - 3 axes each for Accelerometer, Magnetometer, and Gyroscope. This forms the input to *FinGTrAC*. Hand motion is tracked by the watch to obtain wrist and elbow locations using *MUSE*[67]. The ring captures subtle finger motions. Since we do not have sensors on all fingers, the sensor data does not sufficiently capture the entire ASL gesture. We now look at the raw data for special cases to discuss feasibility.

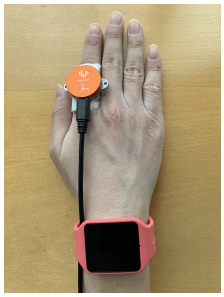


Fig. 3. ring, watch platform

4 RAW SENSOR DATA: A FIRST LOOK

Figure 4 shows the raw sensor data from watch and ring for a few interesting cases. Apart from raw IMU data from ring, and watch, *FinGTrAC* uses wrist location derived from the watch [67] for gesture detection. Tracking finger location is beyond the scope of this project (Future possibilities discussed in Section 9).

We include these examples to point out a few observations: (1) The smartwatch data for gestures of words "work" and "evening" is shown in Fig. 4(b). These two words have identical wrist locations and orientations, and their motion data look very similar to each other. Although only z-axis data of the accelerometer is shown for clarity, the data from other axes as well as the data from gyroscope and magnetometer are also similar. Thus the two words cannot be distinguished by watch data alone. However, the words have different finger motions which are accurately captured by the ring sensor data in Fig. 4(a) (The figure captures differences in finger orientation across these two gestures). Thus the ring sensor data is critical in capturing subtle finger motions considering that many word pairs in ASL (such as "Friday" and "Saturday"; "I" and "my"; finger-spellings etc) have similar wrist orientations and location. (2) For words such as "mother" and "father", the ring data is very similar (Figure 4(c)) because the words have identical index finger motion. To resolve ambiguity in such cases, *FinGTrAC* tracks wrist location differences (Figure 4(d)) across these words and incorporates the opportunity into a Bayesian filtering framework. Other challenges arise due to variation in gestures across users, noisy data etc which are addressed in Section 5. Nevertheless, a combination watch and ring data looks viable to classify most ASL gestures. (3)

Although rare, for words such as "cook" and "kitchen", both watch and ring generate very similar data (Figures 4 (e) and (f)) because both the hand poses and index finger motions are identical. *FinGTrAC* cannot handle such cases currently, but in Section 9, we argue how Natural Language Processing (NLP) [15] techniques can potentially resolve such ambiguities in the future.

5 CORE TECHNICAL MODULES

5.1 Data Segmentation

An ASL sentence is a sequence of words. Given a stream of IMU sensor data, we first segment it into alternative blocks of the following two key phases: (1) "word phases", where a user is signing word and (2) "transition phases", where the hand is moving to the new location to sign a new word. The hand pauses briefly at the beginning and end of each word [24]. We use this opportunity to segment the sensor data.

Fig. 5(a) shows the gyroscope data (magnitude) from the ring for the following sentence: "You full understand?". The data is labelled with corresponding words. "T" denotes "transition phases". Clearly, we can observe some characteristic *dips* in the measurements as pointed out in the figure. This corresponds to the micro-pauses during word transitions. Towards automatically detecting the *dips* for segmentation, we first derive a *dip template* that captures the properties of a *dip* as shown in the subplot in Fig. 5(a). Then, we perform a pattern matching with DTW[8] to search for the occurrence of the above *dip template* in the gyroscope sensor data of each sentence. This provides a robust approach for detecting transitions even when the duration and nature of the pause signature varies across users and time.

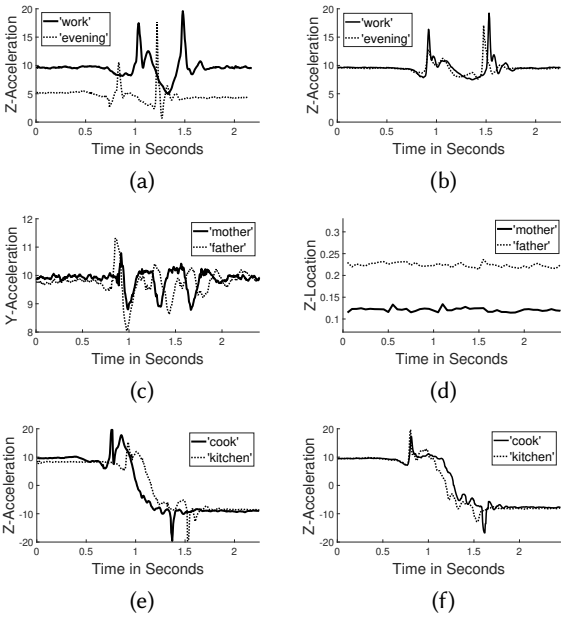


Fig. 4. Significance/limitations of ring and watch data in ASL gesture detection. The first and second columns contain ring and watch data respectively. The three rows provide a comparison of ring and watch data over three different word pairs.

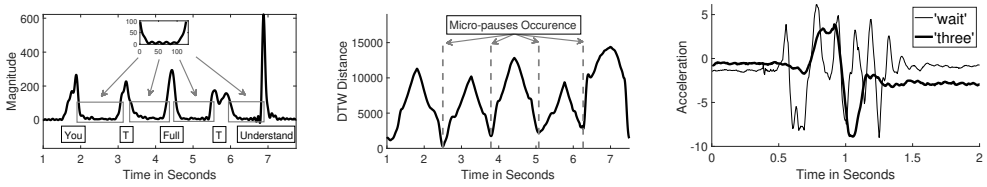


Fig. 5. (a) Micro-pauses between words can be used for segmentation (b) Segmentation based on DTW matching of the *dip* template (c) "wait" and "three" have different number of peaks in the accelerometer data

Even with words that have minimal motion of the fingers or hand during "word-phases" or "transition-phases", we observe that there is always some signal either on the watch or the ring creating the *dip* templates. Fig. 5(b) shows the DTW distance of such a pattern matching. Clearly, there are local minimas corresponding to the locations of micro-pauses which can be used to reliably segment the sensor data into "word phases" and "transition phases".

5.2 Preprocessing

We preprocess the data with below steps to improve the robustness of word detection – particularly because different users may not perform the ASL signs in an exactly identical manner.

Low Pass Filtering: Human finger motions have low frequencies below 6 Hz even with rapid motions[69]. Similarly, hand and finger motions in ASL gestures are smooth and hence dominated by low frequencies too. Therefore, we first pass the sensor data through a low pass filter with a cutoff frequency of 10 Hz. This eliminates the noise and un-intended vibrations from higher frequencies and enhances the robustness of sign detection across multiple users with minimal training.

Elimination by Periodicity: The accelerometer data (magnitude) for signs "wait" and "three "is shown in Fig. 5(c). Sign for "wait" involves a periodic motion of the hand 5 times and the data shows 5 peaks as expected. On the other hand, the periodicity pattern in "three" is different and shows only 1 peak as expected. We exploit the periodicity in narrowing down the search space. Consider matching an unknown gesture with 5 peaks with all words in the dictionary to find the best match. We first eliminate all words in the dictionary which do not have 5 peaks (such as the word "three"). This reduces the search space for the matching of the unknown word, thus improving the robustness and accuracy of further steps such as DTW. The significance of this step is quantified in the evaluation section 7.

5.3 Word Gesture Recognition and Ranking by DTW

We use Dynamic Time Warping (DTW) [8] to bootstrap word detection. Briefly, DTW is a pattern matching technique that inspects the overall shape of two signals to determine their similarity. For example, Fig. 6(a) shows the z-axis ring accelerometer data of two users gesturing the word "people". Although the overall shape is similar, parts of the motion traces happen at a faster rate or earlier for user-2 while other parts happen slower. DTW uses a dynamic programming optimization to minimally compress and stretch the two sequences relative to each other such that they provide the best overlap. Fig. 6(b) shows the two sequences after DTW optimization. DTW is known to do a good job of matching such series with similar overall shape. The residual differences between the two time series determines the similarity score between them.

Training: We first build a training database where labelled sensor data (9-axis IMU data) from the ring and watch is created for each ASL word. One user wears the prototype and signs all 100 ASL words in our dictionary. A one time training with a single user is sufficient, and no separate training is needed for each new user.

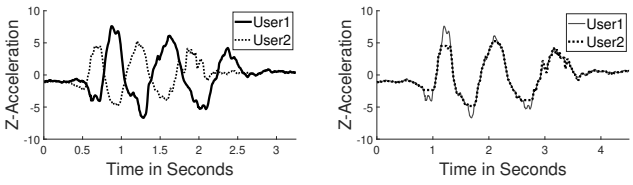


Fig. 6. (a) Accelerometer data for "people" for two users (b) Data from user-2 is compressed and stretched to match with user-1 by DTW

Word Gesture Detection: An unknown ASL gesture from a new user is matched using DTW with training data of each word in our ASL dictionary. The matching is done across all 9 dimensions of IMU data as well as the orientation estimates based on A3[86]. The word with the best match is most likely to be the unknown word. The ring data turned out to be more prominent in the word detection phase and the overall matching accuracy with DTW is 70.1%. This is because words-pairs such as "mother, father", "see, three" have similar wrist and index finger motions. Also, subtle variations in gesturing across users can cause miss-matches. To cope up with the poor accuracy of DTW matching, *FinGTrAC* considers not only the best match for an unknown word, but also the top 10 matches. The correct word is among top 10 ranks in 100% of the cases indicating promise.

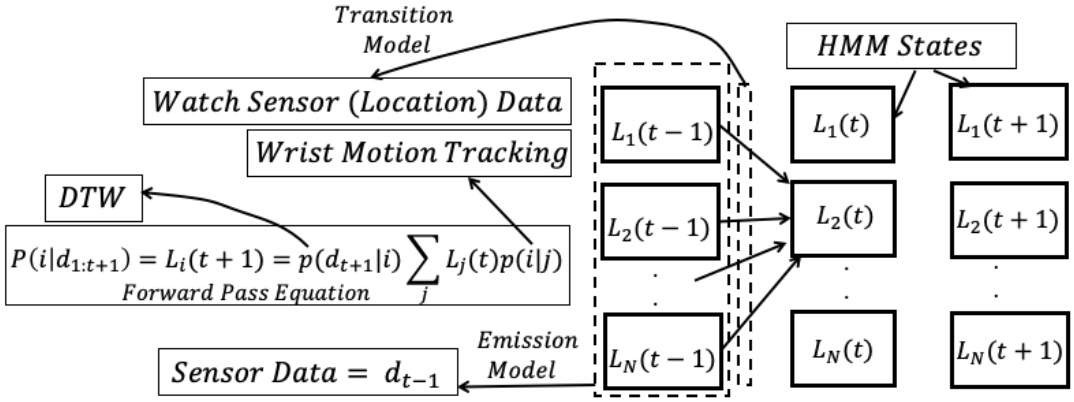


Fig. 7. HMM model for sentence detection: Raw sensor data d_t contributes to the emission probability ($p(d_{t+1}|i)$) which is computed with DTW. Wrist location data contributes to transition probability- $p(i|j)$. Likelihood $L_i(t)$ of words at all parts of the sentence is first computed with forward pass. A backward pass with Viterbi algorithm decodes the most likely sentence.

The top 10 matches for each word from DTW are further processed with a Hidden Markov Model (HMM) to improve the accuracy. The HMM, particularly, incorporates wrist location transitions between words in the context of a sentence to decode the most likely sentence. Details are elaborated next.

5.4 Sentence Decoding with HMM and Viterbi

We use results from word detection as inputs to sentence detection algorithm. While results from word recognition might be inaccurate, we incorporate transition characteristics between words towards more accurate sentence detection. The HMM architecture for sentence detection is depicted in Fig. 7. In this section, we explain the details of the HMM.

Forward Pass: An ASL sentence is a sequence of words. *FinGTrAC* models *word*s within a sentence as the hidden *states* in HMM. The HMM first computes the likelihood probability of each dictionary word occurring as the t^{th} word in the sentence during a step called *Forward Pass*. Mathematically, we denote this probability as follows: $P(i|d_{1:t}) = L_i(t)$
In other words, given all sensor data from beginning of the sentence to the current word – $d_{1:t}$, $P(i|d_{1:t})$ denotes how likely is i^{th} dictionary word to occur as the t^{th} word in the sentence. These likelihood probabilities are computed from the below recursive equation.

$$L_i(t+1) = p(d_{t+1}|i) \sum_j L_j(t)p(i|j) \tag{1}$$

There are two key probabilities that the HMM model relies on while computing the above likelihood probabilities. First, $p(d_{t+1}|i)$ denotes the **emission probability**, which indicates how likely is the word i to generate the sensor observation during the $(t+1)^{th}$ "word phase" – d_{t+1}. Secondly, $p(i|j)$ is the **transition probability** which indicates how likely is a transition from word j to i happen in a sentence given the watch location difference between word transitions. The details are elaborated below.

■ *Emission Probability*: The emission probabilities $p(d_{t+1}|i)$ are derived from DTW matching.

$$p(d_{t+1}|i) \approx DTW\_metric(d_{t+1}, d_{i,template}) \tag{2}$$

As described in Section 5.3, the 9-axis IMU data for an unknown word is matched by DTW with labelled templates for each dictionary word $d_{i,template}$. The normalized similarity metric from DTW would be used to assign the

emission probabilities. If we rank all words by similarity metrics, the correct word appears in the top 10 ranks in 100% cases. Thus we set probabilities of words beyond top 10 ranks to zero, to decrease the complexity of HMM.

■ *Transition Probability*: The transition probabilities are determined from the change in wrist location between two words as computed from the smart-watch.

We first determine wrist locations for each word in the dictionary using Kinect sensor. Kinect is only used for training data, and not a part of *FinGTrAC* system. This is a one time training overhead with a single user to determine the wrist locations for each word. Thus, Kinect locations at the start and end – $l_{st,kinect}(i)$ and $l_{ed,kinect}(i)$ for each word i in the dictionary is known with high precision. Later, during testing, a new user signs a sequence of unknown words forming a sentence. We compute wrist locations for each of these words by using techniques in MUSE [67] on the smartwatch data. Then, we update the transition probability as follows:

$$p(i|j) \propto \frac{e^{\frac{-l_d^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}},$$
$$l_d = ||(l_{st}(i) - l_{ed}(j))| - |(l_{st,kinect}(i) - l_{ed,kinect}(j))||$$

$l_{ed}(j)$ denotes the end location of word j and $l_{st}(i)$ denotes the start location of word i as determined by the smartwatch data. σ^2 denotes the standard-deviation of location errors, which is approximately $10cm$.

■ *Prior Probability*: Prior knowledge about the states in HMM is useful in improving the accuracy of decoding. For example, if DTW matching metric is similar for "I" and "have", with all other information being equal, "I" is the more likely word signed by the user since "I" is used more frequently than "have" in ASL. We use the frequencies of our dictionary words [51] $f(i)$ as prior probabilities to improve the accuracy. Since the prior probabilities do not have dependency on previous states, they are similar in nature to emission probabilities. Thus, we integrate prior probabilities into emission probabilities and update the equation 2 to below:

$$p(d_{t+1}|i) \approx f(i).DTW\_metric(d_{t+1}, d_{i,template}) \tag{3}$$

At the end of forward pass, we have likelihood probabilities for each dictionary word at each position of the sentence. We next do a backward pass using Viterbi [79] algorithm to decode the most likely sentence.

Backward Pass: Viterbi Decoding: Consider a sentence of three words and two possible decodings. First decoding, say A estimates the words to be 34^{th},61^{st},4^{th} items of dictionary while second decoding, say B estimates the words to be - 12^{th}, 41^{th}, 3^{rd}. Then, the following expressions denote the likelihood of the two decodings.

$$p(A) = p(d_1|34).p(61|34).p(d_2|61).p(4|61).p(d_3|4) \tag{4}$$
$$p(B) = p(d_1|12).p(41|12).p(d_2|41).p(3|41).p(d_3|3) \tag{5}$$

We can decide the more likely possibility among the two by comparing $p(A)$ and $p(B)$. We can extend this idea to compare all possible decodings and pick the one with highest probability. Viterbi algorithm[79] is an efficient way to do this under HMM assumption to decode the most likely sentence.

5.5 Finger-spelling Detection

English words that do not have a direct ASL sign are communicated by explicitly signing the characters forming the word. The hand and finger poses for various characters is depicted in Fig. 8. Identifying characters signed by a user is non-trivial because: (1) As seen in Fig. 8, majority of the characters (such as "ABDEFIKLRSTUVW") have identical wrist poses (positions and orientations). In cases where poses are different, the difference is too subtle to be reliably identified with variations in signing by different users. Thus, we realize that the watch data is not much useful. (2) Many characters have similar index finger poses ("GH","CO","BLX", "AISY"). Since the ring is only

worn on the index finger, the ring data may not disambiguate among such characters. Nevertheless, ring data captures more information about the signed character than the smartwatch data. Hence, we use only the ring sensor data for finger-spelling detection.

Towards finger-spelling detection, we begin by collecting training data from the ring by a single user for all the 26 characters. No new training is needed for a new user. While the ring data does not sufficiently differentiate among characters, we exploit the English dictionary for detecting the whole words even though the individual characters may not be detected reliably in isolation.

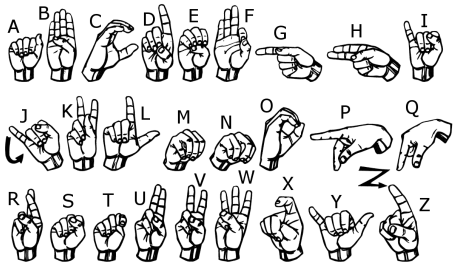


Fig. 8. Hand poses for alphabets [73]

When a new user signs an unknown word, we first segment the IMU data to obtain the sensor data for each individual character using techniques similar to the discussion in Section 5.1. Now, we consider all words in the English dictionary with exactly same number of characters as the unknown word signed by the user – we call this "search space list". Suppose the length of the word signed by the user is 6 and the first word on the "search space list" is "sensor". We measure similarity between the training data for each of the 6 characters - "s","e","n","s","o" and "r" with the data from the 1st, 2nd, 3rd, 4th, 5th and 6th segmented characters respectively from the user data for the unknown word. We compare the orientation between the training and testing data for measuring the similarity. Similarly, we derive the similarity metric with the second word on the "search space list", say "robust", and continue measuring similarity for all words in the dictionary. The dictionary word with the maximum similarity metric is decoded as the word signed by the user. We evaluate the idea with a dictionary size of 1000 words (Section 7). For working with larger dictionary sizes, the frequency of usage of the dictionary word may also be considered. If two dictionary words have similar matching metrics, the one with higher frequency could be the more likely.

6 PUTTING IT ALL TOGETHER: SYSTEM ARCHITECTURE

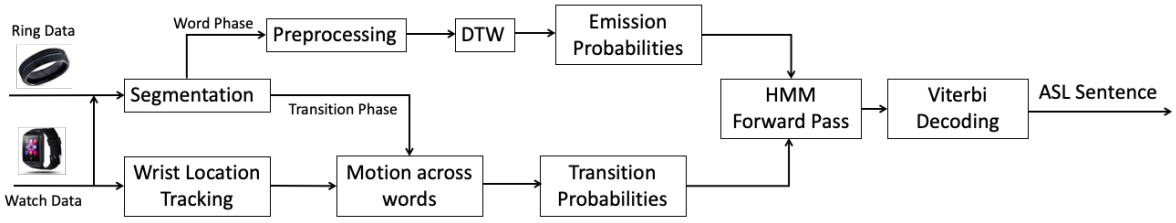


Fig. 9. *FinGTrAC* Architecture

Fig. 9 depicts the overall architecture of *FinGTrAC*. The 9-axis IMU data from the ring and the smartwatch is the input. The input data is first segmented into "word-phases" and "transition phases". Preprocessing steps together with DTW is used to determine top 10 matches for data in the word-phase. The DTW metric is used to generate the emission probabilities for the HMM. The prior probabilities of word frequencies are also incorporated into the emission probabilities. The watch data is used to track hand motion between words in the context of a sentence. This generates the transition probabilities for HMM. The emission and transition probabilities form the input to HMM which then decodes the most likely sentence using the Viterbi algorithm. The decoded sentence is read on a microphone speaker or displayed on a phone screen.

7 EVALUATION

User Study: All reported results in the paper are generated from a systematic user study campaign. Due to recruitment challenges, we could only recruit users with normal hearing ability. Thus, we conduct a study similar to guidelines followed in recently proposed wearable ASL translation works [19, 25]. We recruit users and train them extensively to perform ASL finger gestures through an online ASL tutorial. One user provides training data for 100 ASL words as discussed in Section 5 for computing the DTW similarity metric. We conduct the study with 10 test users that includes 7 males and 3 females. The volunteers are aged between 20-30, and weigh between 47kg to 96kgs. Overall, the users learnt 50 ASL sentences (3-11 words each). The sentences were composed from a dictionary of 100 most frequently used ASL words[51]. 53 of the words use both hands while 47 of the words use only the dominant hand. The 50 sentences use a total of 90 words from this dictionary. Finally, the users perform gestures of the 50 sentences ten times by wearing the ring and smartwatch platform described in Section 3. We collect 9-dimensional IMU data both from the ring as well as the smart-watch during these experiments. The data is streamed to an edge device (desktop) which decodes sentences with a latency of 0.4 seconds.

Impact of Ring Position: Although the index finger is most involved in ASL gestures and thus we place the

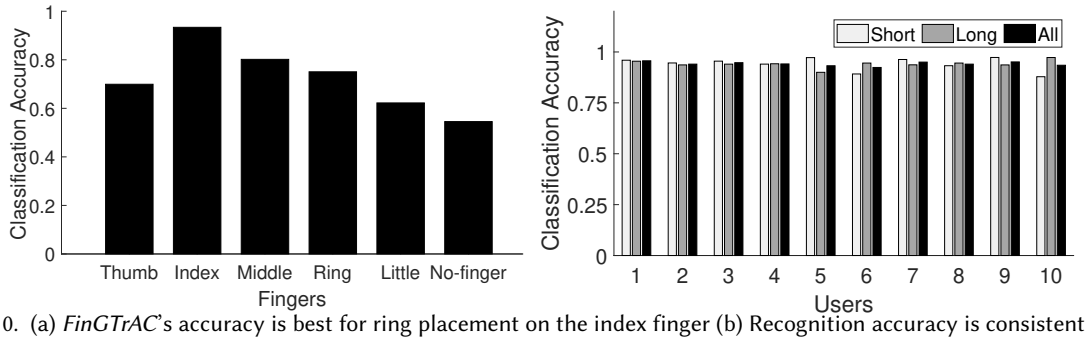


Fig. 10. (a) *FinGTrAC*'s accuracy is best for ring placement on the index finger (b) Recognition accuracy is consistent across diverse users

ring sensor on the index finger, we conduct a small scale study with just 3 users to verify how the position of the ring impacts the word classification accuracy. Fig. 10(a) shows the word classification accuracy as a function of position of the ring. The last-bar "No-finger" indicates that a ring sensor was not placed on any finger, and the classification was done only based on smartwatch data (for both DTW and HMM stages). Evidently, the index finger offers highest accuracy – upto 15% higher than the middle finger which offers second best accuracy. The accuracy is only 54% for the "No-finger" case. Thus, we conduct the study for the rest of the users by placing the ring sensor on their index finger. The next set of results presented has the ring sensor placed on the index finger.

Overall Sentence Recognition Accuracy: Fig. 11(a) shows the word error rate across different sentences. The *word error rate* of a sentence is the ratio of number of incorrectly decoded words and the total number of words. Most of the sentences are decoded correctly across users with an average word error rate of 5.86%. For cases where the words were not decoded correctly, Fig. 11(b) shows the histogram of the rank of the correct word. Evidently, the correct word is within top 5 most likely words as detected by our system. We believe this is a promising result which in combination with advances in NLP techniques can recover most of the residual errors.

Gesture Accuracy Breakup by Words: For each word gesture, the *word classification accuracy* is defined as the ratio of number of times the gesture was detected correctly to the total number of its occurrences. Fig. 11(c) shows a confusion matrix depicting the word classification accuracy. Cases of miss-classifications ("cook" and "kitchen") happen when two words have very similar index finger motion and wrist location. In other miss-classification

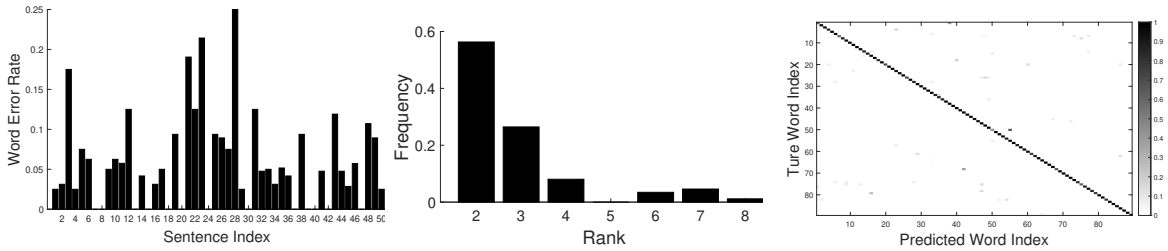


Fig. 11. *FinGTrAC*'s Performance (a) Word error rate across sentences (b) Histogram of ranks of correct words (c) Confusion matrix for word classification – [40]

examples such as "water" and "fine", the gestures are similar enough that variations in user performance can affect the accuracy. However, most words are decoded correctly with an average accuracy of 94.2% .

One hand vs both hand gestures: Our dictionary includes 47 words that use only dominant hand and the other 53 words use both hands. We did not notice any difference in accuracy between these two classes. One handed gestures have a classification accuracy of 94.47% whereas two handed gestures have an accuracy of 93.85%.

Gesture Accuracy over Users: Fig. 10(b) shows the breakup of average word classification accuracy over users for short (4 words or less), long (5 words or more), and all sentences. Evidently, *FinGTrAC* is robust over diverse users. The results are also consistent for various sentence sizes.

Comparison with Related Work: We compare with the state of the art wearable ASL recognition in SignSpeaker [25] which detects 103 ASL gestures. Thus, our dictionary sizes are similar for fair comparison. We implemented SignSpeaker's machine learning approach with Long Short Term Memory (LSTM) and Connectionist Temporal Classification (CTC).SignSpeaker trains the model with 11 users. The model is tested with 4 new users, however, the test users sign the same sentences signed by the other 11 users with no evaluation on new unseen sentences. We first reproduce this result by training the model with 9 users and testing with 1 user (10 fold cross validation [7]) for same sentences used in training. The accuracy is 97.1% which is roughly in agreement with results in SignSpeaker (Fig. 12). However, when we train the model with 49 sentences and test with a new unseen sentence (50 fold cross validation), we observe that the performance of SignSpeaker drops dramatically. While machine learning models are successful in speech recognition on unseen sentences, we believe the poor performance of machine learning based approach here is attributed to limited amount of training data in contrast to speech recognition models which have been trained with data accumulated over years. Our attempts to run the same machine learning algorithms on the ring data have also been unsuccessful in generalizing to unseen sentences for similar reasons. In comparison to SignSpeaker, FinGTrAC's accuracy for any unseen sentence is still 94.2% owing to its HMM based design that can detect any arbitrary composition of words from a given dictionary.

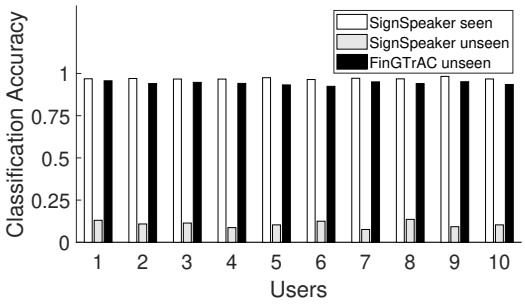


Fig. 12. In contrast to SignSpeaker's deep learning approach, *FinGTrAC*'s HMM approach generalizes well to unseen sentences

Gain by Module: Fig. 13(a) shows the improvements in accuracy delivered by various modules in *FinGTrAC*. DTW detects words with an average accuracy of 70.1%. DTW is robust to accommodate variations when the overall gesture has similarities with the standard template. However, when the user performs the gesture completely differently, DTW can fail.The preprocessing stage looks at high level features and eliminates many of the wrong

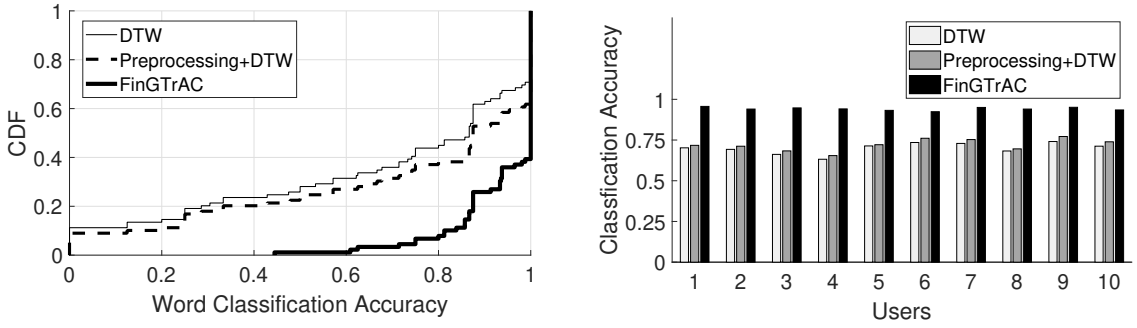


Fig. 13. Classification accuracy improves with successive stages of optimization across for all users

matches thus enhancing the accuracy to 72.2%. Although the average improvement with preprocessing looks marginal, the preprocessing steps can significantly improve the rank of correct words as discussed in Section 5.3. Further, the HMM model incorporates wrist location transitions in between words and improves the accuracy of word detection to 94.2% . Essentially, the application domain based inferencing with HMM has increased the gain fro 72.2% to 94.2% which we believe is significant. The improvement in accuracy by the technical modules is also consistent across various users (Fig. 13(b)).

Analysis of failure cases: *How bad is a failure case?* We note that most incorrectly decoded sentences have only 1-2 words that are incorrect. Even if the decoded words are incorrect, the correct word is mostly within the top 3 ranks after the HMM stage. An example is shown in the table in Fig. 14(a) where the top three ranking candidate words are shown for each of the decoded words in the sentence. Can you guess the correct sentence if we told you that only one word is wrong (i.e. correct word is not rank-1 but in top-3 ranks)? We show such examples to students in an informal setting – they were quick to identify correct sentences. Over the entire dataset, we observe that the correct word appears in top-3 ranks in 99.0% cases. As shown in Figure 14(b), if we assume that a word is classified correctly if it appears in the top-3 ranks, then the accuracy shoots up from 94.2% to 99.0%. We believe this is promising, particularly because we observe that the incorrect sentences due to miss-classifications are often incorrect grammatically and semantically. However, if the correct word is not too far-away from the incorrect word in HMM's probabilistic ranking, we believe language models can be applied to further boost the accuracy by identifying semantically and grammatically correct sentences within the search space. This will be explored in future work (elaborated in Section 9).

	w1	w2	w3	w4
rank-1	restaurant	I	drink	month
rank-2	woman	have	month	milk
rank-3	three	my	sleep	game

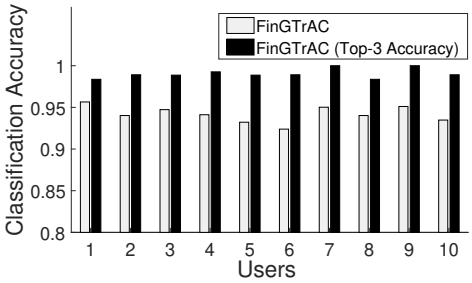


Fig. 14. (a) Table shows a decoded sentence with the top three ranks for each word position. One of the rank-1 words is incorrect. *Can you guess the correct sentence?* Answer at end of paper[1] (b) NLP techniques can boost the accuracy

Scalability Analysis: To evaluate the scalability of our system, we expand our user study by a modest amount. We augment the dictionary with 100 new words thus making the total size of our dictionary to be the top 200 popularly used ASL words [51]. The new 100 words were signed by one user using a procedure similar to

the signing of the initial 100 words. This data is entered into the training database which now contains DTW templates for 200 popularly used ASL words. In addition to the 50 sentences from our initial dataset, we add

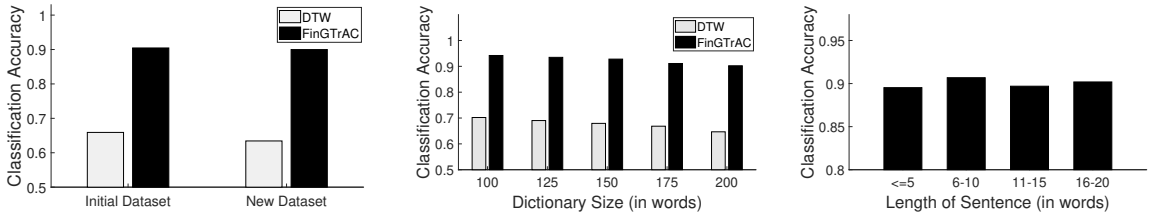


Fig. 15. (a) Accuracy of newly signed sentences is comparable to sentences from the initial dataset (dictionary size = 200) (b) FinGTrAC's accuracy degrades gracefully (c) Accuracy is consistent irrespective of the length of the sentence

20 new sentences formed from 200 words in the dictionary. The word size over sentences range from 11-20 which is signed by 3 different users. We first validate whether the *word classification accuracy* of the initial dataset of 50 sentences is similar to the new dataset of 20 sentences in Fig. 15(a). Evidently, the accuracy levels are similar which indicates that FinGTrAC's performance is independent of the words used in forming sentences. However the accuracy is lower in both cases because of the expanded dictionary size of 200 words. Fig. 15(b) expands the scalability results of FinGTrAC with respect to the number of words in the dictionary. To ensure that all dictionary sizes contain all of the words used in all sentences, the results in Fig. 15(b) only uses the original 50 sentences (in the initial dataset) which were drawn from 100 most popular words. However, we match them in a search space of varying dictionary sizes - 100, 125, 150, 175, and 200. A dictionary of size k contains k most frequently used ASL words. Fig. 15(b) shows the scalability results for both the DTW and the overall system FinGTrAC which integrates "DTW + HMM". Evidently, the *word classification accuracy* drops gracefully with more words added to the dictionary. While the DTW accuracy degrades from 70.1% to 64.7%, the overall accuracy of FinGTrAC degrades from 94.2% to 90.2%. Although the "DTW + HMM" approach itself may not be scalable for longer dictionary sizes, we believe language modeling approaches (Section 9) might help with further scalability. Fig. 15(c) shows the variation of accuracy with the length of the sentence (dictionary size = 200). We do not notice any trend in the graph because the "DTW + HMM" model in FinGTrAC is insensitive to the length of a sentence. Finally, the latency of processing varies between 0.4s to 1.1s as the length of sentences vary from 5 to 20 words, mainly because of the need to process a longer time series of data. However, we observe that the latency is insensitive to the dictionary size (over 100 to 200 words). This is because the HMM only looks at the top-10 ranked words (based on the DTW metric) regardless of the size of the dictionary (Section 5.3). On the other hand, the DTW matching stage, which is indeed affected by the size of the dictionary has negligible overhead in comparison to HMM.

Finger-spelling Detection: We consider a 1000-word dictionary of commonly used names (e.g. "John", "Adam"). We pick over 75 names from this dictionary and have users sign these names by ASL finger-spelling. We detect which of the 1000 words give the best match to determine the unknown word signed by a user. The names were decoded correctly with an overall accuracy of 96.6%. Fig. 16(a) shows the miss-classifications at the level of characters. Miss-classifications happen when two names differ in only one character (such as "Taylor" and "Naylor") and the distinguishing character has similar pose on the index finger. Fig. 16(b) shows the miss-classifications using a naive classifier that does not exploit the dictionary opportunities. By exploiting dictionary opportunities, *FinGTrAC* can reliably detect most non-ASL words.

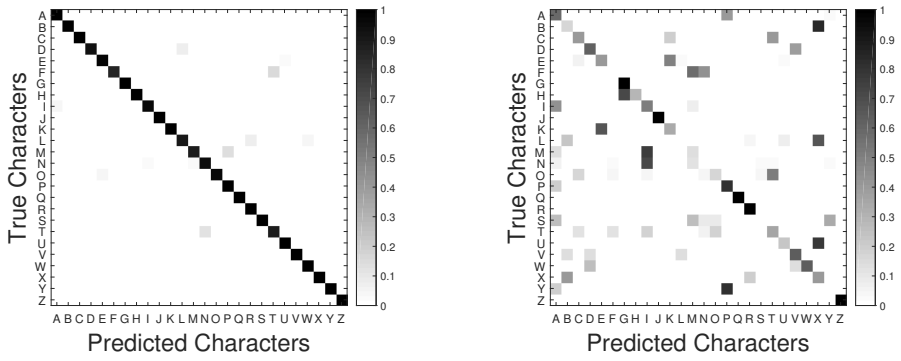


Fig. 16. Confusion matrix of character classifications (a) With dictionary (b) Without dictionary

8 RELATED WORK

Gesture and Activity Recognition: IMU, WiFi, and Acoustic signals have been extensively used for gesture recognition in a number of applications [57, 68, 78, 86]. uWave[39] uses accelerometers for user authentication and interaction with a mobile device. Specifically, they use DTW based techniques to detect 9 gestures. FingerIO [48] uses acoustic reflections for user interface (UI) applications in small devices such as smartwatches. In particular, FingerIO develops its technology based on phases of the reflected signals on orthogonal frequency-division multiplexing (OFDM) sub-carriers. FingerPing [85] uses acoustic signals for finger gesture detection. FingerPing places an acoustic-transducer and contact-microphone on the hand for transmitting acoustic signals through the body. They observe that the through-body channel varies with the finger gesture configuration using which they classify 22 different gestures. Capband [77] uses sensing with an array of capacitative sensors worn on the wrist. These sensors detect skin deformations due to finger motion thus being able to detect upto 15 hand gestures. RisQ [57] detects smoking gestures with wrist-worn IMUs using conditional random fields (CRFs). Work in [72] uses a fusion of 10 acoustic sensors (contact microphones) and IMU sensors for detecting 13 hand gestures in daily life. In contrast, *FinGTrAC* develops algorithms that are closely tied to the application for better accuracy with sparse sensors. Specifically while prior works with low intrusive sensors can only distinguish tens of gestures, FinGTrAC extends recognition to hundreds of gestures. Prior works specific to ASL are presented next.

Vision: Leap motion sensors, Kinect and computer vision research [76] can track fingers with cameras and depth imaging. Hat and neck-mounted cameras [10, 19] have been explored for ASL translation. Motion capture, depth imaging has been explored in [18, 84]. Deep learning approaches based on a hybrid of CNN and HMMs on video data is explored in [33]. 3D convolutional residual neural networks (3D- ResNets) are used for recognition of german and chinese sign languages from videos in [61]. More generally, recent papers[11, 27, 47] in computer vision can track 3D finger motion of users from videos using a combination of techniques in generative adversarial networks (GAN), convolutional neural networks (CNN), and weakly supervised learning techniques. While sophisticated computer vision techniques can easily track fingers and thus enable translation of larger ASL dictionary sizes, cameras do not provide ubiquitous tracking and need clear view and lighting conditions. In contrast, our goal is to explore the limits and feasibility of sensing with wearable devices.

WiFi: WiFi based hand/finger gesture detection has been explored [35, 42, 65] that use wireless channel and doppler shift measurements for ASL recognition. WiFinger [35] uses channel state information (CSI) of wireless reflections with discrete wavelet transforms (DWT), and DTW for classifying upto 9 finger gestures. Work in [42] uses directional antennas and collects WiFi signal strength measurements of reflections from the body. Using pattern matching techniques, the work shows the feasibility of detecting 25 finger gestures. Work in [65] uses

WiFI CSI together with feature extraction and SVM techniques to distinguish upto 5 gestures. Wisee [62] uses doppler shifts of wireless reflections to classify 9 different gestures. SignFi [41] is an innovative work that uses wireless channel measurements from WiFi APs for ASL recognition over a larger dictionary. We admit that WiFi based detection is completely passive without on-body sensors. However, we believe wearable approaches can complement WiFi based techniques in an ubiquitous and adhoc setting (ex: outdoors) with no availability of WiFi.

Wearable Sensors including Gloves: An extensive review of glove systems is provided in [3]. The survey indicates that works between 2007 to 2017 use a combination of one handed and two handed gloves with accelerometers, flex sensors and EMG sensors to detect upto 100 ASL words. Flex sensor technology [63, 64] is popularly used for detecting finger bends. These sensors are typically made of resistive carbon elements whose resistance depends upon the bend radius. Works [60, 66, 70] use flex sensors on each of the fingers to be able to distinguish upto 40 sign language gestures. Optical resistance based finger bend technologies have also been adopted [31, 81]. By combining a light emitting diode with a light dependent resistance, the optical technology detects the finger bend by observing the variation in resistance due to variation in light intensity resulting from finger bending. Work in [74] uses 5 optical flex sensors for detecting 25 words in Malaysian sign language using probabilistic classification models. Similarly work [30] uses a glove made of optical flex sensors in detecting 25 words and basic postures in Korean Sign Language. Tactile sensor technology [34] detects whether a finger is curled or straight by using a robust polymer thick film device. The resistance of the material decreases as pressure on the finger increases thus sensing the state of the finger. Work in [66] detects upto 8 ASL gestures using such tactile sensors. Hall Effect Magnetic Sensors [13] are used for ASL recognition over 9 ASL words using logistic regression algorithms. The core idea is to place a strong magnet on the palm, and place small magnetic sensors on finger tips. As the finger bends, it comes closer to the palm and hence closer to the strong magnetic field. By sensing this, the bending of fingers can be detected for use in gesture classification. Works with Electromyography (EMG) sensors [36, 83] typically use tens of electrodes to detect muscle activity for ASL word recognition using support vector machines, nearest neighbor, naive bayes, and other classification algorithms. In addition, a number of glove based systems which combine the above sensors with IMUs have been proposed with some of them being commercially available on the market. 5DT Data Glove uses[1] flex sensors on all fingers to measure bending angles of fingers. Using this, works [29] have shown detection of upto 26 ASL alphabets. Work in [56] uses commercially available CyberGlove [16] as well as IMU sensors to detect ASL words from a dictionary of 107 words. In contrast to above works that use 5-20 of embedded sensors on hands, *FinGTrAC* uses a low intrusive platform of smartwatches and rings to detect ASL gestures with a similar accuracy without compromising on the size of the dictionary. Closest to our work is a recently proposed innovative system called SignSpeaker [25] that detects sentences composed of 103 ASL words using smartwatch. While SignSpeaker is less intrusive than *FinGTrAC* given it only requires a smartwatch, we believe *FinGTrAC* differs from SignSpeaker in an important way. FinGTrAC can detect unseen sentences whereas SignSpeaker cannot detect unseen sentences (Evaluated in Section 7) with the evaluation being only done with seen sentences. Thus, we believe the non-deep learning approach in FinGTrAC generalizes well in comparison to SignSpeaker where training needs to be done for all possible sentences which entails exponentially larger training overhead in the size of the dictionary. While deep learning with RNNs/LSTMs [23] is popular in speech recognition, the success is attributed to large training datasets accumulated over years. We discuss potential methods to integrate deep learning into our problem domain in Section 9. Work [9] uses 6 rings for ASL translation but lacks in any detail. AuraRing [58], a recent work, tracks the index finger precisely using a magnetic wristband and ring. We believe *FinGTrAC* can integrate such new hardware innovations into the model for improving the accuracy in the DTW stage.

9 DISCUSSION AND FUTURE WORK

Implement-ability in Real Time: *FinGTrAC* involves three computation modules. (1) DTW for word classification/ranking (2) HMM and viterbi decoder for sentence decoding (3) Wrist location tracking. The first two only involves low weight computation. Unlike wireless communications where the packet sizes are long, the size of a sentence is only a few words making viterbi algorithm tractable. The third module is perhaps the most complex one since it involves particle filters. Currently, we can run the particle filter on a desktop machine/edge device in real time. Given our bandwidth requirement is minimal 50kb/s, FinGTrAC operates in real time with a sentence decoding latency of 0.4 seconds, and a battery energy consumption of 150mW with WiFi streaming [26].

Other IoT Applications: Domain specific modeling of ASL helps improve the accuracy of gesture tracking. Similarly, we will also consider finger motion tracking for other applications by exploiting application-specific context. Analysis of finger motion data of tennis, baseball, basketball players can provide valuable information for coaching and injury management, as well as provides bio-markers of motor diseases [2].

Limitations and Future Work: We believe this is a significant first step, yet full ASL translation is a complex problem. In particular, the small dictionary size is a limitation since 200 ASL words do not offer sufficient coverage for practical usage in real world. There are also other limitations including not handling facial expressions and inability to exploit the rich ASL grammar. Therefore, we will explore the following approaches to address these limitations in the future.

(1) Non deep learning techniques such as Natural Language Processing (NLP) and Semantic Context to augment dictionary: FinGTrAC benefits from NLP. For example, a sentence *"You can hear me?"* was wrongly decoded as a grammatically incorrect sentence: *"You can more me?"* In fact, the miss-classified word ("hear") was the second most likely word. Thus, we believe that a deeper integration of NLP techniques can improve accuracy to substantially expand the dictionary size of *FinGTrAC*. Language modeling techniques[44] from speech recognition will also be considered. The semantic context of a sentence offers rich information to correct miss-classifications. A sentence - *"Sleep my bed don't_mind"* was classified as *"Think my bed don't_mind"*. Clearly, the first sentence is more meaningful than the second. We plan to incorporate context information not only within sentences but also across sentences using similarity metrics from word2vec [43], WordNet[45], context from FrameNet [6] etc. We will first attempt to directly incorporate such NLP constraints not only within sentences but also across sentences to augment the dictionary size as much as possible. However, as an alternative we will also explore deep learning approaches elaborated next.

(2) Deep Learning to Enhance Dictionary: Unlike speech processing or image recognition, no large dataset is available for tracking finger motion with IMU. Thus, FinGTrAC uses HMM to minimize training overhead and validate the feasibility. RNNs [44] are very popular in speech processing and language translations. With enough training data accumulated over years, they can learn complex functions as well as intelligently use old observations in decision making using LSTMs [23]. To bootstrap training data generation, we will exploit finger motion from videos [47] of ASL tutorials and news. We believe this will substantially increase the size of *FinGTrAC*'s dictionary.

(3) Facial Expressions: Main expressions of interest include lowering/raising eyebrows; anger, happiness, sadness; jaw-motion; winking; nodding; head-tilt etc. We will explore ear-worn sensors (IMU and biological sensors) to track such expressions [4, 28]. Ear-worn sensors can be used naturally as earphones or ornaments – the design would be non-intrusive. EMG (facial muscles), EEG (brain signals) and EOG (eye signals) can be integrated easily into earphones [49] which provide opportunity for sensing facial expressions. While prior works detect isolated expressions, we plan to integrate the sensing deeply into language and machine learning in the future study.

(4) User study with expert ASL users: Lack of testing with fluent ASL users is a limitation of our system. However, we believe that the FinGTrAC's results are promising enough to develop on our ideas above for enhancing dictionary

size and incorporating facial expressions such that the system is feasible for full fledged ASL translation. We plan to validate such an end-to-end ASL translation system with expert ASL users.
(5) Finger Motion Tracking and Additional Sensors: Towards pushing the limits of accuracy, *FinGTrAC* will explore free form tracking of finger joint locations. While this is non-trivial with under-constrained information, the anatomical constraints of fingers open up opportunities [38]. Modeling the constraints dramatically decreases the search space for finger poses thus making it feasible to track. We will also measure the trade-offs in intrusiveness and accuracy in using additional sensors or a second ring. We also consider through body acoustic vibrations from ring to smartwatch [85] or integrating a single EMG electrode into wrist-watch for 3D finger motion tracking while still keeping the sensor footprint small.

10 CONCLUSION

This paper shows the feasibility of finger gesture classification by using low-weight wearable sensors such as a smart-ring and smart-watch which are becoming popular in recent times. For such platforms, the importance of application context in scaling gesture recognition from tens to hundreds of gestures has been demonstrated through a case-study with ASL translation. Previous approaches which use motion tracking cameras cannot offer ubiquitous tracking, while wearable solutions that require sensor-gloves, or EMG sensors are too cumbersome. In contrast, *FinGTrAC* shows feasibility with a ubiquitous and easy-to-use platform with minimal training. In building the system, *FinGTrAC* uses a probabilistic framework incorporating the noisy and under-constrained motion sensor data, as well as contextual information between ASL words to decode the most likely sentence. A systematic user study with 10 users shows a word recognition accuracy of 94.2% over a dictionary of 100 most frequently used ASL words. Despite progress, this is only the first step. We believe the results from this work offer significant promise in extending this work. Opportunities in sensing, machine learning, and natural language processing can be exploited to extend this work towards enabling a realistic solution with low intrusive wearables for seamless ASL translation.

ACKNOWLEDGMENTS

This research was partially supported by a NSF grant CNS-1909479.

REFERENCES

[1] 5DT 2019. 5DT Data Glove Ultra - 5DT. https://5dt.com/5dt-data-glove-ultra/.
[2] Rocco Agostino, Antonio Currà, Morena Giovannelli, Nicola Modugno, Mario Manfredi, and Alfredo Berardelli. 2003. Impairment of individual finger movements in Parkinson's disease. *Movement disorders* (2003).
[3] Mohamed Aktham Ahmed, Bilal Bahaa Zaidan, Aws Alaa Zaidan, Mahmood Maher Salih, and Muhammad Modi bin Lakulu. 2018. A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017. *Sensors* 18, 7 (2018), 2208.
[4] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2017. CanalSense: face-related movement recognition system based on sensing air pressure in ear canals. In *ACM UIST*.
[5] Benjamin J Bahan. 1997. Non-manual realization of agreement in American Sign Language. (1997).
[6] Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*.
[7] Yoshua Bengio and Yves Grandvalet. 2004. No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research* 5, Sep (2004), 1089–1105.
[8] Donald J Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series.. In *KDD workshop*.
[9] Bracelets and Rings 2013. Bracelets and rings translate sign language. https://www.cnet.com/news/bracelet-and-rings-translate-sign-language/.
[10] Helene Brashear et al. 2003. Using multiple sensors for mobile sign language recognition. Georgia Institute of Technology.
[11] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. 2018. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 666–682.
[12] Ming Jin Cheok, Zaid Omar, and Mohamed Hisham Jaward. 2019. A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics* 10, 1 (2019), 131–153.

[13] Tushar Chouhan, Ankit Panse, Anvesh Kumar Voona, and SM Sameer. 2014. Smart glove with gesture recognition ability for the hearing and speech impaired. In *2014 IEEE Global Humanitarian Technology Conference-South Asia Satellite (GHTC-SAS)*. IEEE, 105–110.
[14] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7784–7793.
[15] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research* 12, Aug (2011), 2493–2537.
[16] Cyber Glove 2017. Cyber Glove III - Cyber Glove Systems LL. http://www.cyberglovesystems.com/cyberglove-iii/.
[17] Deaf Statistics 2011. How many deaf people are there in United States. https://research.gallaudet.edu/Demographics/deaf-US.php.
[18] Facial Expressions 2017. Facial expressions in American Sign Language. https://www.newscientist.com/article/2133451-automatic-sign-language-translators-turn-signing-into-text/.
[19] Biyi Fang et al. 2017. DeepASL: Enabling Ubiquitous and Non-Intrusive Word and Sentence-Level Sign Language Translation. In *ACM SenSys*.
[20] Finger Placement 2011. Finger Placement When Shooting a Basketball. https://www.sportsrec.com/476507-finger-placement-when-shooting-a-basketball.html.
[21] Jakub Gałka, Mariusz Mąsior, Mateusz Zaborski, and Katarzyna Barczewska. 2016. Inertial motion sensing glove for sign language gesture acquisition and recognition. *IEEE Sensors Journal* 16, 16 (2016), 6310–6316.
[22] Marcus Georgi, Christoph Amma, and Tanja Schultz. 2015. Recognizing Hand and Finger Gestures with IMU based Motion and EMG based Muscle Activity Sensing.. In *Biosignals*. 99–108.
[23] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2016. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems* 28, 10 (2016), 2222–2232.
[24] François Grosjean and Harlan Lane. 1977. Pauses and syntax in American sign language. *Cognition* 5, 2 (1977), 101–117.
[25] Jiahui Hou et al. 2019. SignSpeaker: A Real-time, High-Precision SmartWatch-based Sign Language Translator. *MobiCom* (2019).
[26] Junxian Huang, Feng Qian, Alexandre Gerber, Z Morley Mao, Subhabrata Sen, and Oliver Spatscheck. 2012. A close examination of performance and power characteristics of 4G LTE networks. In *ACM MobiSys*.
[27] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. 2018. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 118–134.
[28] Iravantchi et al. 2019. Interferi: Gesture Sensing using On-Body Acoustic Interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 276.
[29] Kazuma Iwasako, Masato Soga, and Hirokazu Taki. 2014. Development of finger motion skill learning support system based on data gloves. *Procedia Computer Science* 35 (2014), 1307–1314.
[30] Eunseok Jeong, Jaehong Lee, and DaeEun Kim. 2011. Finger-gesture recognition glove using velostat (ICCAS 2011). In *2011 11th International Conference on Control, Automation and Systems*. IEEE, 206–210.
[31] Jong-Sung Kim, Won Jang, and Zeungnam Bien. 1996. A dynamic gesture recognition system for the Korean sign language (KSL). *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 26, 2 (1996), 354–359.
[32] Minwoo Kim, Jaechan Cho, Seongjoo Lee, and Yunho Jung. 2019. IMU Sensor-Based Hand Gesture Recognition for Human-Machine Interfaces. *Sensors* 19, 18 (2019), 3827.
[33] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. 2018. Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. *International Journal of Computer Vision* (2018).
[34] Hyung-Kew Lee, Jaehoon Chung, Sun-Il Chang, and Euisik Yoon. 2008. Normal and shear force measurement using a flexible polymer tactile sensor with embedded multiple capacitors. *Journal of Microelectromechanical Systems* 17, 4 (2008), 934–942.
[35] Hong Li, Wei Yang, Jianxin Wang, Yang Xu, and Liusheng Huang. 2016. WiFinger: talk to your smart devices with finger-grained gesture. In *ACM UbiComp*.
[36] Yun Li, Xiang Chen, Xu Zhang, Kongqiao Wang, and Z Jane Wang. 2012. A sign-component-based framework for Chinese sign language recognition using accelerometer and sEMG data. *IEEE transactions on biomedical engineering* 59, 10 (2012), 2695–2704.
[37] Lifeprint 2017. ASL Grammar. https://www.lifeprint.com/asl101/pages-layout/grammar.htm.
[38] John Lin, Ying Wu, and Thomas S Huang. 2000. Modeling the constraints of human hand motion. In *Proceedings workshop on human motion*. IEEE.
[39] Jiayang Liu, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. 2009. uWave: Accelerometer-based personalized gesture recognition and its applications. *Pervasive and Mobile Computing* (2009).
[40] Yilin Liu, Fengyang Jiang, and Mahanth Gowda. 2020. Application Informed Motion Signal Processing for Finger Motion Tracking using Wearable Sensors. In *IEEE ICASSP*.
[41] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. 2018. Signfi: Sign language recognition using wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 23.
[42] Pedro Melgarejo, Xinyu Zhang, Parameswaran Ramanathan, and David Chu. 2014. Leveraging directional antenna capabilities for fine-grained gesture recognition. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*.

ACM, 541–551.
[43] Tomas Mikolov, Kai Chen, Gregory S Corrado, and Jeffrey A Dean. 2015. Computing numeric representations of words in a high-dimensional space. US Patent 9,037,464.
[44] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.
[45] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.
[46] Motiv Ring 2020. Motiv Ring | 24/7 Smart Ring | Fitness + Sleep Tracking | Online Security. https://mymotiv.com/.
[47] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2018. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 49–59.
[48] Rajalakshmi Nandakumar et al. 2016. Fingerio: Using active sonar for fine-grained finger tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1515–1525.
[49] Anh Nguyen et al. 2016. A lightweight, inexpensive in-ear sensing system for automatic whole-night sleep stage monitoring. In *ACM SenSys*.
[50] Viet Nguyen, Siddharth Rupavatharam, Luyang Liu, Richard Howard, and Marco Gruteser. 2019. HandSense: capacitive coupling-based dynamic, micro finger gesture recognition. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 285–297.
[51] OSF 2020. OSF|SignData.csv. https://osf.io/ua4mw/.
[52] Oura Ring 2018. Oura Ring Review. https://www.wareable.com/health-and-wellbeing/oura-ring-2018-review-6628.
[53] Oura Ring 2019. Oura Ring - What we learned about the sleep tracking ring. https://www.cnbc.com/2019/12/20/oura-ring-review---what-we-learned-about-the-sleep-tracking-ring.html.
[54] Oura Ring 2019. Oura Ring review - The early adopter catches the worm. https://www.androidauthority.com/oura-ring-2-review-933935/.
[55] Oura Ring 2020. Oura Ring: The most accurate sleep and activity tracker. https://ouraring.com/.
[56] Cemil Oz and Ming C Leu. 2011. American Sign Language word recognition with a sensory glove using artificial neural networks. *Engineering Applications of Artificial Intelligence* 24, 7 (2011), 1204–1213.
[57] Abhinav Parate et al. 2014. Risq: Recognizing smoking gestures with inertial sensors on a wristband. In *ACM MobiSys*.
[58] Farshid Salemi Parizi, Eric Whitmire, and Shwetak Patel. 2019. AuraRing: Precise Electromagnetic Finger Tracking. *ACM IMWUT* (2019).
[59] Deborah Chen Pichler. 2002. Word order variation and acquisition in American Sign Language. (2002).
[60] Nikhita Praveen, Naveen Karanth, and MS Megha. 2014. Sign language interpreter using a smart glove. In *2014 International Conference on Advances in Electronics Computers and Communications*. IEEE, 1–5.
[61] Junfu Pu, Wengang Zhou, and Houqiang Li. 2019. Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4165–4174.
[62] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. 2013. Whole-home gesture recognition using wireless signals. In *Proceedings of the 19th annual international conference on Mobile computing & networking*. ACM, 27–38.
[63] Giovanni Saggio. 2012. Mechanical model of flex sensors used to sense finger movements. *Sensors and Actuators A: Physical* 185 (2012), 53–58.
[64] Giovanni Saggio, Francesco Riillo, Laura Sbernini, and Lucia Rita Quitadamo. 2015. Resistive flex sensors: a survey. *Smart Materials and Structures* 25, 1 (2015), 013001.
[65] Jiacheng Shang and Jie Wu. 2017. A robust sign language recognition system with multiple wi-fi devices. In *Proceedings of the Workshop on Mobility in the Evolving Internet Architecture*. ACM, 19–24.
[66] Deepak Sharma, Deepak Verma, and Poras Khetarpal. 2015. LabVIEW based Sign Language Trainer cum portable display unit for the speech impaired. In *2015 Annual IEEE India Conference (INDICON)*. IEEE, 1–6.
[67] Sheng Shen, Mahanth Gowda, and Romit Roy Choudhury. 2018. Closing the Gaps in Inertial Motion Tracking. In *ACM MobiCom*.
[68] Michael Sherman, Gradeigh Clark, Yulong Yang, Shridatt Sugrim, Arttu Modig, Janne Lindqvist, Antti Oulasvirta, and Teemu Roos. 2014. User-generated free-form gestures for authentication: Security and memorability. In *ACM MobiSys*.
[69] Mohammad Sharif Shourijeh, Reza Sharif Razavian, and John McPhee. 2017. Estimation of Maximum Finger Tapping Frequency Using Musculoskeletal Dynamic Simulations. *Journal of Computational and Nonlinear Dynamics* 12, 5 (2017), 051009.
[70] Ahmad Zaki Shukor, Muhammad Fahmi Miskon, Muhammad Herman Jamaluddin, Fariz bin Ali, Mohd Fareed Asyraf, Mohd Bazli bin Bahar, et al. 2015. A new data glove approach for Malaysian sign language detection. *Procedia Computer Science* 76 (2015), 60–67.
[71] Nabeel Siddiqui and Rosa HM Chan. 2020. Multimodal hand gesture recognition using single IMU and acoustic measurements at wrist. *PloS one* 15, 1 (2020), e0227039.
[72] Nabeel Siddiqui and Rosa HM Chan. 2020. Multimodal hand gesture recognition using single IMU and acoustic measurements at wrist. *Plos one* 15, 1 (2020), e0227039.

[73] Signlanguage Characters [n. d.]. Drawing hands in sign language poses. http://albaneze.weebly.com/step-3---drawing-hands-in-sign-language-poses.html.
[74] Tan Tian Swee, AK Ariff, Sh-Hussain Salleh, Siew Kean Seng, and Leong Seng Huat. 2007. Wireless data gloves Malay sign language recognition system. In *2007 6th International Conference on Information, Communications & Signal Processing*. IEEE, 1–4.
[75] Richard A Tennant and Marianne Gluszak Brown. 1998. *The American sign language handshape dictionary.* Gallaudet University Press.
[76] Carlo Tomasi, Slav Petrov, and Arvind Sastry. 2003. 3d tracking= classification+ interpolation. In *null*. IEEE, 1441.
[77] Hoang Truong et al. 2018. CapBand: Battery-free Successive Capacitance Sensing Wristband for Hand Gesture Recognition. In *ACM SenSys*.
[78] Yu-Chih Tung and Kang G Shin. 2015. Echotag: Accurate infrastructure-free indoor location tagging with smartphones. In *ACM MobiCom*.
[79] Andrew Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory* 13, 2 (1967), 260–269.
[80] VMU931 2018. New Rugged and Compact IMU. https://variense.com/product/vmu931/.
[81] Lefan Wang, Turgut Meydan, and Paul Williams. 2017. Design and evaluation of a 3-D printed optical sensor for monitoring finger flexion. *IEEE sensors journal* 17, 6 (2017), 1937–1944.
[82] WHO 2020. World Health Organization: Deafness and Hearing Loss. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.
[83] Jian Wu, Lu Sun, and Roozbeh Jafari. 2016. A Wearable System for Recognizing American Sign Language in Real-Time Using IMU and Surface EMG Sensors. *IEEE J. Biomedical and Health Informatics* (2016).
[84] Yuancheng Ye, Yingli Tian, Matt Huenerfauth, and Jingya Liu. 2018. Recognizing American Sign Language Gestures from within Continuous Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2064–2073.
[85] Cheng Zhang et al. 2018. FingerPing: Recognizing Fine-grained Hand Poses using Active Acoustic On-body Sensing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM.
[86] Pengfei Zhou, Mo Li, and Guobin Shen. 2014. Use it free: Instantly knowing your phone attitude. In *ACM MobiCOm*.

[1]Restaurant I drink milk.