APPLICATION INFORMED MOTION SIGNAL PROCESSING FOR FINGER MOTION TRACKING USING WEARABLE SENSORS

Yilin Liu, Fengyang Jiang, Mahanth Gowda

Pennsylvania State University

ABSTRACT

Finger motion tracking has a number of applications in user-interfaces, sports analytics, medical rehabilitation and sign language translation. This paper presents a system called *FinGTrAC* 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). 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. This paper shows the feasibility of exploiting such context in an application of American Sign Language (ASL) translation. Non-trivial challenges arise due to noisy sensor data, variations in gesture performance across users and the inability to capture data from all fingers. *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 detection accuracy of 94.2% for 100 most frequently used ASL finger gestures over different sentences.

Index Terms— Internet of Things, Wearable, Gesture, Bayesian Inference

1. INTRODUCTION

This paper presents a system called FinGTrAC (**Fin**ger **G**esture **Tr**acking with **A**pplication **C**ontext) that explores the limits and feasibility of fine-grained finger gesture tracking with a single ring worn on a finger and a smartwatch worn on the wrist. The five fingers posses more than 30 degrees of freedom which is in-feasible to track with a sensor on only one finger. However, we note that application specific opportunities provide sufficient 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 [1]. 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. We make our case with an 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 [2] 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.

Prior works on finger motion tracking are limited to a few gestures [3, 4] or use intrusive setup of sensor gloves (15-20 sensors) and Electromyography (EMG) electrodes [5, 6, 7] to track 100 ASL gestures. A combination of algorithms from simple regression models to artificial neural networks[8] have been used in these works, however context information of hand motion in between words has not been modeled thus requiring tens of sensors. While cameras [9, 10, 11, 12] can track full finger motion using deep learning approaches, they benefit from availability of large datasets for training and validation. In contrast, our work explores the limits and feasibility of bayesian inferencing models when extensive training data is unavailable or difficult to generate. A recent work [13] performs ASL translation of 103 words with smartwatches but the training and testing has been done with same sentences. Thus, it would be impractical to train for all possible sentences in any language. In contrast to above, *FinGTrAC* offers application specific context to track finger motion gestures using a low intrusive platform with minimal training overhead. To the best of our knowledge *FinGTrAC* is the first work that uses only two wearable sensors to track 100 ASL gestures with minimal training overhead. Results demonstrate that a single user training suffices to accommodate diverse new users. New training is not necessary for new users.

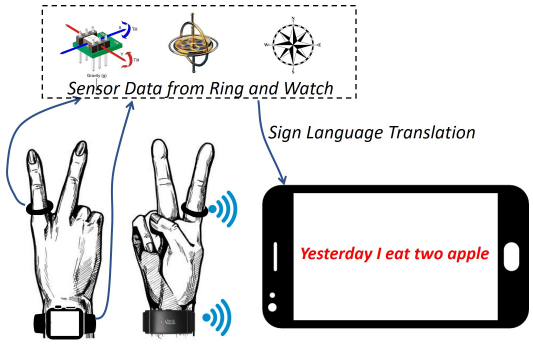

Fig. 1: ASL detection with smart devices

A *FinGTrAC* user needs to 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 as shown in Fig. 1. The Inertial Measurement Unit (IMU) sensors (accelerometers, gyroscopes, and magnetometers) on these devices are used for finger tracking. This is a non-trivial problem with a number of challenges. (1) All fingers, and both hands are involved in finger gestures, whereas *FinGTrAC* uses a sensor only on the index finger. Thus, the data is under-constrained. (2) The sensor data is noisy and the difference between ASL gestures can be subtle. (3) *FinGTrAC* needs to maintain high accuracy and robustness despite variations in the way different users may perform the same gesture.

FinGTrAC exploits a number of opportunities to deal with under-constrained sensor data and reducing the search space for finger gestures. (1) Fingers have a high degree of freedom and the ring (on index finger) cannot capture all of them sufficiently enough to identify a word. Therefore, we also track hand motion in between words of a sentence. Particularly, the motion of the watch (on wrist) in between words of a sentence 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) Analogous to decoding a packet over a noisy channel in wireless networking, decoding words in a sentence can be modeled as a Bayesian Filtering problem (HMM). Thus, we leverage Viterbi decoding to provide optimal sentence decoding from noisy sensor data.

2. BACKGROUND: APPLICATION DOMAIN

ASL uses gestures instead of speech for communication. Majority of ASL signs involve motion of the dominant hand including fingers. The non-dominant hand including facial expressions are used occasionally to complement the dominant hand gestures. ASL grammar mainly consists of sentences in the following format: *Subject-Verb-Object* [14]. Sometimes, the word order can change when the context of the topic is established first through topicalization [15]. This results in the following format: *Object, Subject-Verb*. Facial expressions are used to denote emphasis, questions etc. Pauses indicate end of sentences and words [16]. The signs do not indicate tense of a verb such as *washed* or *washing*. Therefore, the following format is often used to resolve ambiguity [17]: 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.

Fig. 2: Smart ring, watch

3. PLATFORM DESCRIPTION

Smart rings that can pair with phones wirelessly to stream information and monitor activity are available on the market [18]. However, most of these platforms are closed and do not provide access to raw sensor data. Thus, our platform consists of a button shaped sensor – VMU931[19] snugly fit on the finger like a ring as shown in Fig. 2 and a smartwatch. 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.

4. CORE TECHNICAL MODULES

4.1. Data Segmentation and Preprocessing

ASL sentences include brief pauses between words [16]. Let us denote the sensor data during signing of a word as "Word Phase" and the data when the hand is moving from one word to another word as "Transition Phase". Sentences can be decomposed into above two phases by detecting change points in the observed sensor data. We first determine a finger print of the change points for sample sentences. For a new sentence, we find occurrences of this pattern using a DTW[20] based pattern matching to decompose a sentence into words. We also perform other pre-processing step such as low pass filtering (cut off freq $10Hz$) to eliminate high frequency noisy as well as eliminate unlikely matches for a particular word by looking at features such as number of peaks in the data.

4.2. Word Gesture Recognition and Ranking by DTW

We use Dynamic Time Warping (DTW) [20] to bootstrap detection of ASL words. For example, Fig. 3(a) shows the z-axis ring accelerometer data of two users gesturing the word "people". Although the overall shape is similar, parts of the

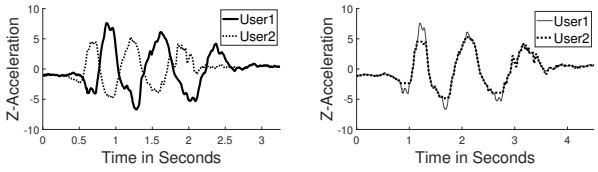

Fig. 3: (a) Accelerometer data for "people" for two users (b) Data from user-2 is compressed and stretched to match with user-1 by DTW

motion traces happen at a faster rate or earlier for user-2 while other parts happen slower. DTW minimally compresses and stretches the two sequences relative to each other such that they provide the best overlap. Fig. 3(b) shows the two sequences after DTW optimization. The residual differences between the two series determines their similarity score.

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 for *FinGTrAC*, and no separate training is needed for new users.

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[21]. 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 the way users perform gestures 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. Our data indicates that the correct word is among top 10 ranks in 100% of the cases indicating promise. The top 10 matches for each word from DTW are further processed with a Hidden Markov Model (HMM) to improve the accuracy of word detection. 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.

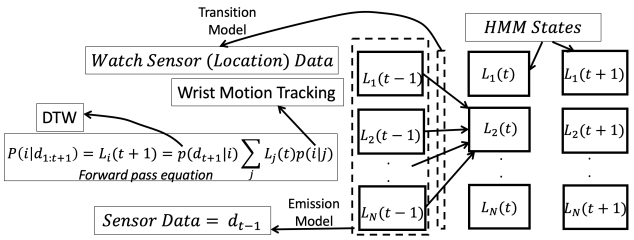

Fig. 4: HMM model relating sensor data to model parameters

4.3. Sentence Decoding with HMM and Viterbi

While results from word recognition with DTW might be inaccurate, we incorporate transition characteristics between words towards more accurate sentence detection. The HMM architecture is depicted in Fig. 4, details elaborated below.

Forward Pass: *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 4.2, 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 the similarity metric, the correct word appears in the top 10 ranks in 100% cases. Thus we set probabilities of words beyond top 10 ranks to be zero, to decreasing the computational 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 [22]. 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 [23] on the smartwatch data. Then, we update the transition probability as follows:

$$p(i|j) \propto \frac{e^{\frac{-l_d^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}},$$
$$l_d = ||(l_{st}(i) - l_{ed}(j))| - |(l_{st,kinect}(i) - l_{ed,kinect}(j))|| \tag{3}$$

$l_{ed}(j)$ 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

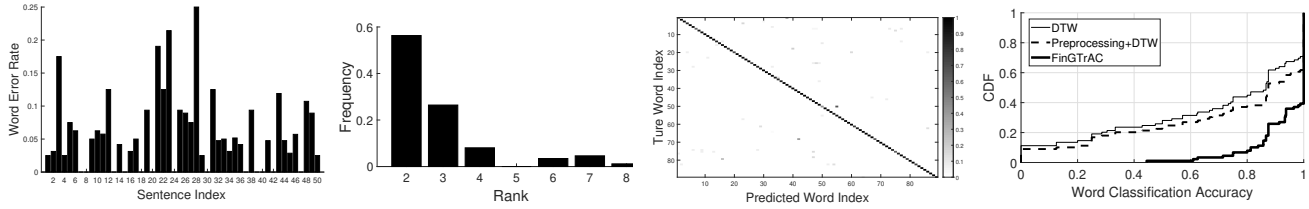

Fig. 5: *FinGTrAC*'s Performance (a) Word error rate across sentences (b) Histogram of ranks of correct words (c) Confusion matrix for word classification (d) Classification accuracy improves with successive stages of optimization across for all users

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 [2] $f(i)$ as prior probabilities to improve the accuracy. We integrate prior probabilities into emission probabilities and update equation 2 as below:

$$p(d_{t+1}|i) \approx f(i).DTW\_metric(d_{t+1}, d_{i,template}) \quad (4)$$

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 [24] algorithm to decode the most likely sentence.

5. 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 other wearable ASL translation works [13, 11], and only recruit users with normal hearing ability and train them extensively to perform ASL finger gestures through a 3 hour long online ASL tutorial. Here, we conduct the study with 10 users that include 7 males and 3 females. The users are aged between 20 to 30 years. Overall, the users signed 50 ASL sentences (3-11 words each) that included 90 words from our 100 word ASL dictionary. The sentences were composed from a dictionary of 100 most frequently used ASL words[2]. 53 of the words use both hands while 47 of the words use only the dominant hand. 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, which is processed further with our algorithms.

Overall Sentence Recognition Accuracy: Fig. 5(a) shows the word error rate across 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. 5(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 can combine with natural language processing (NLP) to recover 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. 5(c) shows a confusion matrix depicting the word classification accuracy. Missclassifications can occur when two gestures are similar. However, most words are decoded correctly by *FinGTrAC* with an average accuracy of 94.2% . The accuracy figures are similar for both one hand and two-handed gestures.

Gain by Module: Fig. 5(d) shows the improvements in accuracy delivered by various modules in *FinGTrAC*. DTW detects words with an average accuracy of 70.1%. 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 from 72.2% to 94.2% which we believe is significant.

6. CONCLUSION

FinGTrAC shows the feasibility of finger gesture classification from low-weight wearable sensors such as a smart-ring and smart-watch with minimal training. The importance of application context in boosting accuracy has been demonstrated by a case-study of sign language translation. *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. RNNs[25] and LSTMs[26] which have revolutionized speech processing are applicable for ASL translation too. Lack of extensive training dataset might be a bottleneck, however we plan to exploit transfer-learning approaches to train wearable sensor data from videos of online ASL tutorials. The finger motion captured in such videos can be used for training wearable motion sensors. Natural language processing and language modeling techniques [25] can be integrated into RNNs thus boosting the accuracy. Facial expressions in ASL can be detected using a combination EMG (facial muscles), EEG (brian signals), and EOG (eye signals) with an embedded earphone sensors [27]. Given plenty of such opportunities, we believe our preliminary study offers enough promise to pursue research in this direction towards extending ASL translation to large dictionary sizes/dialects.

7. REFERENCES

[1] "Finger placement when shooting a basketball," https://www.sportsrec.com/476507-finger-placement-when-shooting-a-basketball.html.

[2] "Osf—signdata.csv," https://osf.io/ua4mw/.

[3] Marcus Georgi, Christoph Amma, and Tanja Schultz, "Recognizing hand and finger gestures with imu based motion and emg based muscle activity sensing.," in *Biosignals*, 2015, pp. 99–108.

[4] Cheng Zhang et al., "Fingerping: Recognizing fine-grained hand poses using active acoustic on-body sensing," in *ACM CHI*, 2018, p. 437.

[5] The Duy Bui and Long Thang Nguyen, "Recognizing postures in vietnamese sign language with mems accelerometers," *IEEE sensors journal*, vol. 7, no. 5, pp. 707–712, 2007.

[6] Boon Giin Lee and Su Min Lee, "Smart wearable hand device for sign language interpretation system with sensors fusion," *IEEE Sensors Journal*, vol. 18, no. 3, pp. 1224–1232, 2018.

[7] Mohamed Aktham Ahmed, Bilal Bahaa Zaidan, Aws Alaa Zaidan, Mahmood Maher Salih, and Muhammad Modi bin Lakulu, "A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017," *Sensors*, vol. 18, no. 7, pp. 2208, 2018.

[8] Cemil Oz and Ming C Leu, "American sign language word recognition with a sensory glove using artificial neural networks," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 7, pp. 1204–1213, 2011.

[9] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden, "Deep sign: Enabling robust statistical continuous sign language recognition via hybrid cnn-hmms," *International Journal of Computer Vision*, vol. 126, no. 12, pp. 1311–1325, 2018.

[10] Oscar Koller, Jens Forster, and Hermann Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, 2015.

[11] Biyi Fang, Jillian Co, and Mi Zhang, "Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation," in *ACM Sensys*, 2017.

[12] Will Zou, Shenghuo Zhu, Kai Yu, and Andrew Y Ng, "Deep learning of invariant features via simulated fixations in video," in *Advances in neural information processing systems*, 2012, pp. 3203–3211.

[13] Jiahui Hou et al., "Signspeaker: A real-time, high-precision smartwatch-based sign language translator," *MobiCom*, 2019.

[14] Benjamin J Bahan, "Non-manual realization of agreement in american sign language.," 1997.

[15] Deborah Chen Pichler, "Word order variation and acquisition in american sign language.," 2002.

[16] François Grosjean and Harlan Lane, "Pauses and syntax in american sign language," *Cognition*, vol. 5, no. 2, pp. 101–117, 1977.

[17] "Asl grammar," https://www.lifeprint.com/asl101/pages-layout/grammar.htm .

[18] "Motiv ring — 24/7 smart ring — fitness + sleep tracking — online security," https://mymotiv.com/.

[19] "New rugged and compact imu," https://variense.com/product/vmu931/.

[20] Donald J Berndt and James Clifford, "Using dynamic time warping to find patterns in time series.," in *KDD workshop*. Seattle, WA, 1994, vol. 10, pp. 359–370.

[21] Pengfei Zhou, Mo Li, and Guobin Shen, "Use it free: Instantly knowing your phone attitude," in *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 2014, pp. 605–616.

[22] "Microsoft kinect2.0," https://developer.microsoft.com/en-us/windows/kinect.

[23] Sheng Shen, Mahanth Gowda, and Romit Roy Choudhury, "Closing the gaps in inertial motion tracking," in *ACM MobiCom*, 2018.

[24] Andrew Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[25] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur, "Recurrent neural network based language model," in *Eleventh annual conference of the international speech communication association*, 2010.

[26] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[27] Anh Nguyen et al., "A lightweight and inexpensive in-ear sensing system for automatic whole-night sleep stage monitoring," in *ACM Sensys*, 2016.