SCEPTRE: a Pervasive, Non-Invasive, and Programmable Gesture Recognition Technology

Prajwal Paudyal, Ayan Banerjee, and Sandeep K.S. Gupta
IMPACT Lab, `http://impact.asu.edu`
Arizona State University, Tempe, Arizona
{Prajwal.Paudyal, abanerj3, sandeep.gupta}@asu.edu

ABSTRACT

Communication and collaboration between deaf people and hearing people is hindered by lack of a common language. Although there has been a lot of research in this domain, there is room for work towards a system that is ubiquitous, non-invasive, works in real-time and can be trained interactively by the user. Such a system will be powerful enough to translate gestures performed in real-time, while also being flexible enough to be fully personalized to be used as a platform for gesture based HCI. We propose SCEPTRE which utilizes two non-invasive wrist-worn devices to decipher gesture-based communication. The system uses a multi-tiered template based comparison system for classification on input data from accelerometer, gyroscope and electromyography (EMG) sensors. This work demonstrates that the system is very easily trained using just one to three training instances each for twenty randomly chosen signs from the American Sign Language(ASL) dictionary and also for user-generated custom gestures. The system is able to achieve an accuracy of 97.72 % for ASL gestures.

Author Keywords

Sign language processing; gesture-based interfaces; assistive technology; wearable and pervasive computing.

ACM Classification Keywords

K.4.2 Social Issues: Assistive technologies for persons with disabilities; H.5.2 User Interfaces: User-centered design

INTRODUCTION

Non-verbal communication is a big part of day-to-day interactions. Body movements can be a powerful medium for non-verbal communication, which is done most effectively through gestures [6]. However the human computer interfaces today are dominated by text based inputs and are increasingly moving towards voice based control [9]. Although speech is a very natural way to communicate with other people and computers, it can be inappropriate in certain circumstances that require silence, or impossible in the case of deaf people [26]. Expressibility is also lost because a lot more could be communicated if machines were trained to recognize gestures on top of the traditional User Interface(UI) elements like text input and speech. These factors have inevitably nudged us towards gesture based communication.

There are several challenges that need to be addressed and solved before we can fully implement such methods. The first one is the lack of a common protocol for gesture based communication, and the second one is the lack of a framework that can successfully translate such communication gestures to meaningful information in real-time. Another important aspect of designing gesture based communication methods is that they need to be pervasive and non-invasive. A survey of ASL users that was taken as part of this research to better assess such requirements, which is summarized in Table 2. The participants of this survey were university students between ages of 19-38, who had completed or were working towards an ASL course at ASU. Everybody in the survey was comfortable conveying simple ideas using ASL.

The system that is proposed as part of this work, SCEPTRE, fulfills these requirements. The entire system is comprised of an Android smartphone or a Bluetooth enabled computer and one to two commercial grade Myo devices that are worn on the wrist to monitor accelerometer, electromyogram(EMG) and orientation. Data is aggregated and preprocessed in the smartphones and is either processed locally or sent to a server. This flexible approach makes it very pervasive as the users can walk around anywhere with the ready-to-use system [4],[2]. Furthermore the devices can be worn underneath a shirt inconspicuously. In the future smart-watches equipped with EMG sensors can be used instead. The main challenge of this research was to develop a method to store, retrieve and most importantly match gestures effectively, conclusively and in real-time.

There are various known methods for gesture recognition and time-series analysis, such as Support Vector Machines, Linear Correlation, Hidden Markov Models, and Dynamic Time Warping [24], [14], [7], [10]. Two classes of methods are usually used for time-series analysis and data-matching: One is the learning based, in which models are developed and stored and the other one is the template based, in which the new time-series data is compared to a saved template. Although learning based models were tried, the template based models were preferred to allow user-driven training with the least number of repetitions and to make sure the system is usable

IUI 2016, March 7–10, 2016, Sonoma, CA, USA.
Copyright © 2016 ACM 978-1-4503-4137-0/16/03 ...$15.00.
http://dx.doi.org/10.1145/2856767.2856794

in real-time. A very compelling advantage of utilizing a user-trained approach is that the gesture space can start small and can gradually grow as it is being used. This will also give the users an ability to discover a pattern of expression, train their own system and use that to communicate seamlessly even with users of other systems or languages. Due to the flexibility of the system, a user who is well versed in a particular gesture based system such as ASL, can choose to train the system according to the specifications of that language. She can then add custom gestures or shortcuts to add to that system if desired.

To demonstrate this, the system was first trained by randomly selecting ten control-oriented signs that are intuitive to use such as drawing a square or a letter in the English alphabet in the air with one or both hands. Other more intuitive gestures such as 'twisting one hand' or 'waving right' etc. were also used. The recognition rate on these 'custom' gestures can be kept very high since the system is designed in such a way that while using non-language based signs, the system is able to guide a user to select only those signs that are sufficiently 'different' from the signs that the system is already trained in; this is achieved in a similar way as if a test gesture was being recognized. The system accepts a new gesture only if recognizing the gesture causes no loss in overall accuracy. This is called the 'guided' mode in the system.

To demonstrate the extendibility of the system, it was then trained for 20 ASL signs by turning the 'guided mode' off. Experiments were done both when the system was trained in these signs by the user, and when the system was used with training data only from the other 'pool' of users. Due to the vast difference in EMG signals between users, the latter experiments did not yield very favorable results. However, recognition rates of 97.72% was achieved when all three of EMG, accelerometer and orientation sensors were used and the database consisted of three instances of the gesture trained by the same user. Using any two of the three sensors, the highest recognition accuracy of 95.45% was achieved when using EMG sensors paired with orientation sensors. With only one of the sensors the highest accuracy that was achieved was 93% for orientation sensor. Data was collected from 10 healthy adults between the ages of 22 and 35, each of whom performed a total of 20 ASL gestures.

Usage

The system is envisioned to be used in two primary use cases:
1. User-to-user communication where at least one user uses a sign language: In this scenario, the user wishes to communicate with a person using a gesture based communication medium, but the other person cannot understand that mode of communication. The first user, then begins the translation mode and performs a gesture. The hub receives the data, processes it and sends the recognized output to the user via a text message or audio message as seen in Figure 2 or converted to animated models as facilitated by systems like that described by Marshall et al. [31].

2. User-to-computer Interactions: In this scenario, the user of system uses ASL or some form of custom gestures to communicate with a computer system. The gestures transmitted can be converted to a regular language for communication purposes just as one would use an audio input to form sentences, or they can be used to trigger specific actions according to pre-configured preferences. For example a user may use a certain gesture to signal she wants to control a certain device. The smartphone hub can then trigger a connection to that device, say a garage door, and the user can give a specific gesture command to open the door, and then another one to disconnect as seen in Figure 1.

RELATED WORK

Most existing work on sign language recognition utilize image/video based recognition techniques while some other utilize data gloves [21], [19], [12], [17], [29]. Although a high accuracy is achieved in the aforementioned data-glove based approaches, one major drawback is that the system is wired and invasive as it interferes with day-to-day activities of the user. According to Fang et al. [12], commercially existing data-gloves are also too expensive for large scale deployment. The system that we propose, utilizes Bluetooth enabled sensors that appear in wearables such as smart-watches, and are thus pervasive and non-invasive. They can also be easily worn underneath a shirt to make them less conspicuous.

The image and video based recognition systems [33, 23, 16, 18, 5] use either annotated fingers, color bands etc, or just plain raw images and video from users to translate gesture based communication. Some studies also utilized multiple cameras [37] to enhance the accuracy by some margin. However, image and video based systems are dependent on presence of cameras in the right positions and are affected by background colors. Another drawback to these systems is that image and video processing is computationally expensive which makes the recognition system's response time slower. Also, heavy dependence on Hidden Markov Model (HMM) based methods makes training the system cumbersome and data intensive. Unlike these approaches, using portable wristwatch like sensors to detect gestures is far less invasive and a template based recognition approach does not require many training data-sets. The computational complexity also decreases as there is no need to deal with data-intensive images and videos, this allows the system to be real-time as well as environment independent.

Li et al. [20] has done a study on combining accelerometer and EMG sensors to recognize sub-word level gestures for Chinese Sign Language and Chen et al. [8, 39] show that combination of multiple sensors helps to increase the recognition accuracy. Following this path, we add the EMG measurements from eight built-in pods in the Myo device to get information that can be leveraged to detect subtle finger movements and configurations which are essential for detecting certain signs and distinguishing them from others. The orientation sensors help to distinguish between signs that have a similar movement and muscle patterns but different rotational primitives. Accelerometer sensors detect movement and position of the hands. The combination of all these sensors into one commercial grade, wireless sensor has given the ability to deploy gesture and sign-language recognition abilities in a far more practical and user-friendly way. Bluetooth technol-

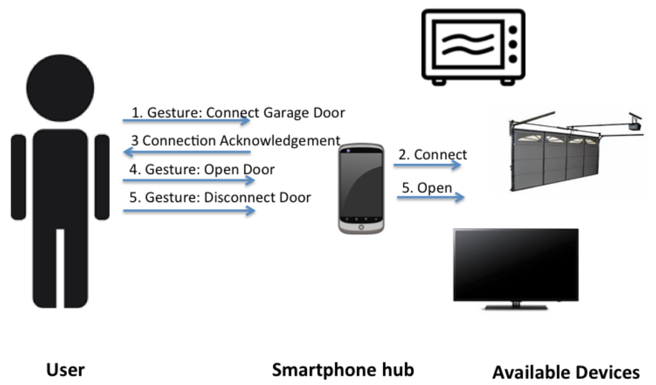
1. Gesture: Connect Garage Door
3 Connection Acknowledgement
4. Gesture: Open Door
5. Gesture: Disconnect Door
2. Connect
5. Open
User
Smartphone hub
Available Devices

Figure 1. Usage for HCI/ Control system.

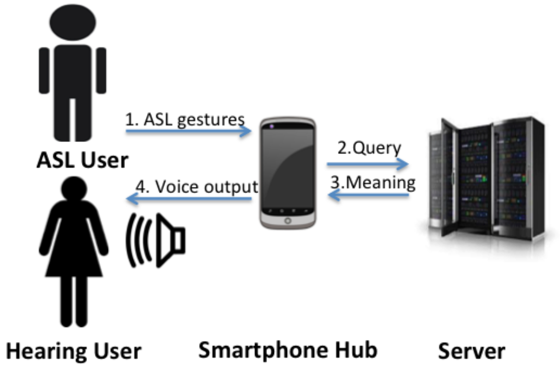

Figure 2. Usage for Person-to-Person Communication.

Table 1. Summary of existing Gesture Recognition tools.

Existing Gesture Recognition Algorithms	Mobile	User Extendable	Requires Video	Real Time	User Independent	Invasive	Language Independent
[21]	×	×	×	✓	×	✓	×
[19]	NA	×	×	×	✓	✓	×
[12]	×	✓	×	NA	✓	✓	×
[33]	×	×	✓	✓	×	✓	×
[8]	×	×	×	✓	×	✓	×
Sceptre(Proposed)	✓	✓	×	✓	×	×	✓

ogy for transfer of data makes the system wireless and easy to use. The overall accuracy is also increases due to the use of all three types of signals.

Thus by coupling pervasive technologies with simple algorithms we achieve high accuracy rates and great mobility. Starner et al. propose a solution that might be user and language independent as possibilities, however an implementation is not provided [33]. Moving this system towards complete user-independence is part of future work. The visual nature of sign language communication is however not limited to just hand gestures; expressions and body gait also play important parts in the overall communication. However analyzing facial expressions and body gaits will be part of future work.

PRELIMINARIES

American Sign Language Specifics

Signs in ASL are composed of a number of components: the shape the hands assumes, the orientation of the hands, the movement they perform and their location relative to other parts of the body. One or both hands may be used to perform a sign, and facial expressions and body tilts also contribute to the meaning. Changing any one of these may change the meaning of a sign [15], [3], [35], [11]. This is what gives sign-language the immense power of expressibility, however this also means that multi-modal information has to be processed for sign-language recognition.

Orientation is the degree of rotation of a hand when signing. This can appear in a sign with or without movement. The gyroscope sensors provide rotation data in three axes, which is utilized in recognizing orientation. Data from EMG sensors is utilized in detecting muscle tensions to distinguish handshapes, which is another very important feature of the signs.

Location, or *tab*, refers to specific places that the hands occupy as they are used to form words. ASL is known to use 12 locations excluding the hands [36]. They are concentrated around different parts of the body. The use of a passive hand and different shapes of the passive hand is also considered valid locations. Movement or sig, refers to some form of hand action to form words. ASL is known to use about twenty movements, which include lateral, twist, flex, opening or closing [34]. A combination of accelerometer and gyroscope data is instrumental in distinguishing signs based on movement and signing space.

Sign language is a very visual medium of communication, the system that this paper proposes however tries to capture this visual information, in a non-visual manner to ensure pervasiveness. Thus, multi-modal signals are considered to get as close of an estimate as possible.

Problem Description

The primary task of the system is to match gestures. Since multi-modal signals are being considered, they are compared individually to the signals stored in the database and the results are combined. The form of the input data is described in detail Section Methodology, but in short it consists of an array of time-series data. The gesture database has pre-processed iterations of the same gesture along with other gesture data. The problem that the system is solving is comparing this test gesture data to each of the existing sample data using DTW

Algorithm 1 Problem Description.

```
procedure MATCH GESTURES(test)
    for (i in 1 to database_size) do
        $template \leftarrow next\_template$
        for (j in 1 to number_of_sensors) do
            $distance[j] \leftarrow dist(test[j], template[j])$
        end for
        Overall_dist = combine(distance)
    end for
    Output = min(Overall_dist)
end procedure
```

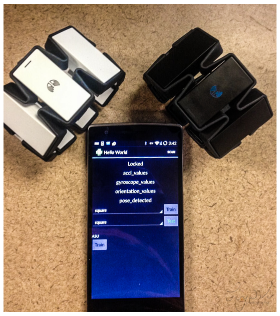

Figure 3. Myo Devices and Android Smartphone with the Sceptre App showing.

and Energy based techniques as appropriate and then combining the results at the end to give a distance score. Then, at the very end, the task is to recognize the gesture based on the least distance. Algorithm 1 provides a high level overview of this approach.

REQUIREMENTS

The need/demand for sign language recognition software for accessibility has been previously established, but accessing if there is a need or demand for systems that can utilize an established natural language as a human computer interface has not been studied (to our knowledge). As part of this research a survey was taken where 13 users of ASL expressed their opinions on these topics that included their willingness to use a gesture based system, the perceived practicality of using ASL for HCI, the acceptable time constraints for user to user and user to computer communications etc. The results of the survey are summarized in Table 2.

Extendability

With more and more systems with voice command interfaces, it only seems plausible that a system that understands sign language will be desirable. To the question 2 in the survey, more than 75% of the ASL users agreed that they were very likely to use ASL for HCI, and the rest would be somewhat likely to use such a system. More than 75% people also responded that it was very important that the system be easily trainable, and another 15% thought it was somewhat important. This follows from the knowledge [15] that ASL (and sign language in general) varies a lot over geographical, racial and cultural bounds. Thus, the ability to add unrecognized signs or new signs easily is very important. When ASL users were asked if they were likely to use the system if it could be trained using 3 training instances, more than 76% said very likely and the rest responded they were somewhat likely to spend the time training. This requirement puts the constraint on the system that it has to be relatively easy to train. This system can be trained using only three instances of training per sign. The relationship between training instances and accuracy is summarized in SectionResults and Evaluation.

Time Constraints

Communication is a real-time activity and the need for fast response times in both person-to-person communication as well as HCI is well established. A timing delay of 0.1 s is considered 'instantaneous' while a delay of 0.2 s to 1 s is noticeable to the user, but generally acceptable. If a system takes more than 1 s, then some indication needs to be given to user about the 'processing' aspect [22]. The survey question regarding timing backs this up: while more than 76% of ASL users said they were very likely and an additional 15% said they were somewhat likely to use the system if the response time was under 3 s, this number falls to 53% and 0% very likely responses when the response times were 3-7 s, and more than 7 s respectively. This establishes a strict constraint for recognition time for the system. To meet these standards the system has to be light-weight and should not rely on computationally expensive methods. Also, if the processing is delayed due to unpredictable issues like client network sluggishness, a proper indication has to be given to the user of the wait.

Non-Invasiveness

Historically there has been a certain stigma associated with signing. It was only in 2013 that the White House published a statement in support of using sign language as an official medium of instruction for children, although it has been known to be a developmentally important for deaf children for years [1]. Given this stigma, and due to other more obvious reasons like usability, an interpretation system should be as non-invasive as possible. The survey results confirm this, as more than 90% thought that it is important or very important for the system to be non-invasive.

Pervasiveness

The other disadvantage of using more invasive technology like data-gloves or multiple-cameras is that it makes the system either tethered to a certain lab based environment settings or it requires bulky equipment that discourages its use in an everyday setting [21, 19, 12]. This makes the use of smartwatch like wireless equipment that can work on the streets as well as it works in a lab more desirable.

To make sure that the proposed system meets these requirements some performance metrics are decided upon which are discussed in detail in Subsection Performance Metrics and subsequently the system is tested w.r.t to these metrics as discussed in Subsection Design of experiments.

SYSTEM ARCHITECTURE

The system consists of two Myo Devices connected via Bluetooth to an Android device which acts as a hub as shown in Figures 3 and 4. For the EMG part a Bluetooth enabled computer is used to obtain the data since the Myo framework

Table 2. Summary of ASL users' survey on a continuum from very unlikely (0) to very likely (5).

Question	1	2	3	4	5
Interest in trying solutions for gesture based HCI	0	0	1	3	9
Interest in using ASL for gesture based HCI	0	0	3	4	6
Likeliness to use a system that requires 3 trainings instances	0	0	3	1	9
Importance of non-invasivenss	0	1	3	2	7
Likeliness to use if wait time 0-3 s	0	1	2	4	6
Likeliness to use if wait time 3-7 s	1	1	4	5	2
Likeliness to use if wait time 7-10 s	4	4	5	0	0
Likeliness to try a prototype	0	1	3	4	5
Likeliness of the use of this system by a native speaker for communication	0	1	1	5	6

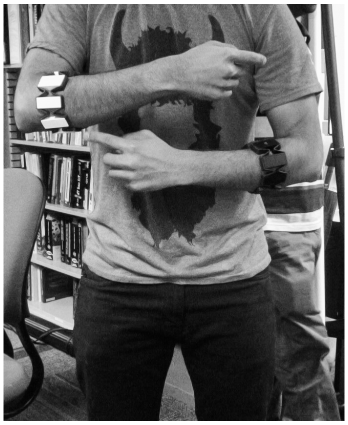

Figure 4. Deployment: A user with a wrist-band devices on each hand performing a gesture.

does not yet facilitate streaming EMG data to smart phones. For tests that do not use EMG data, Android device collects and does some pre-processing on the data of the gesture performed and stores this as time-series data. All this is sent to the server for processing. Three to four instances for each gesture data are collected, processed and stored. A high level overview of the pre-processing that is done to the data before it is stored in the database is summarized in Figure 5.

At test time when a gesture is performed, the system processes the gesture-data in a similar way as described above. In the Architecture Figure 6, this is encapsulated in the 'Preprocessing' block. After this is done, the system compares this data with the other gesture data stored in the database using a specialized comparison method. The comparison method comprises of testing accelerometer data, orientation data and EMG data by different methods and later combining the results before calculating a 'nearness' score for each pair. After this step a ranking algorithm is employed in the server to compute a list of stored gestures that are closest to the test gesture and best one is returned. Details on these techniques is given in Section Methodology. The system using the same methods of comparison is able to help the user be selective in the signs that the system is trained on, this is called the 'guided mode' as explained in Subsection Training. A smart ranking architecture is used when the size of the database grows to ensure that the results come back in acceptable time to be a real-time application. This is discussed in more details in Subsection Ranking and Optimization. The system also periodically uploads all recognized gesture data and other meta-data like user specific accuracy, new signs trained etc. to the server. This information can be used to make future predictive methods more efficient which is not part of this work.

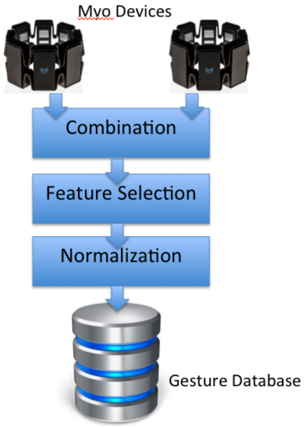

Figure 5. Pre-processing: Data is collected from the Myo devices, processed and then stored in Database.

Training

Training the system requires two Myo devices, a hub for data collection, and server for processing. The hub and the server may be one and the same. A typical training sessions consists of the following steps: 1. The user selects either the 'guided mode' or 'ASL' mode. (in the smartphone interface) 2. She selects a sign from a dropdown list, or creates a new sign. 3. She performs the sign. 4. The system annotates the input data with the name of the sign and stores the data. 5. If guided mode was selected, the system performs a scan of the sign database specific to the user and determines if there are any clashes. If not, the user is asked to repeat the sign two more times after which the sign is ready to use. If however there was a clash, then the user is suggested to choose another sign instead. 6. If the 'ASL' mode was selected, the system does not give such feedback, and simply asks the user to train the system two more times. 7. The gesture is ready to use.

METHODOLOGY

The overall methodology of gesture recognition consists of the following steps as shown in Figure 6.

Pre-processing

Data is collected from two Myo devices while the gesture is being performed (Figure 5). As soon as the end is detected or

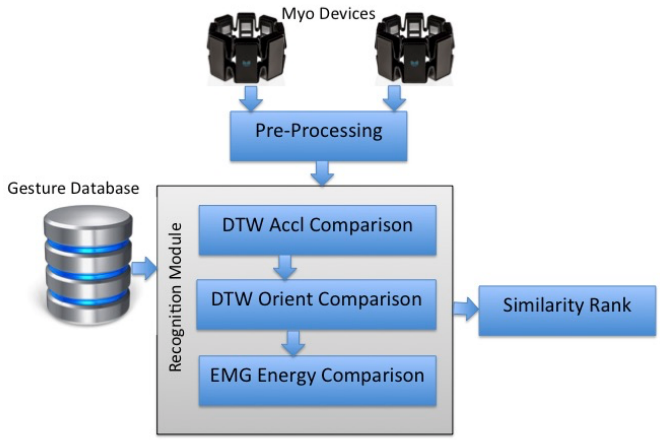

Figure 6. Gesture Comparison and Ranking: New data is collected, processed and compared with existing gesture data to get a ranked-list of similarities.

signaled, the pre-processing begins. The data collected from two hands is aggregated into one data-table then stored into a file as an array of time-series data. At 50 Hz. of sampling rate a 5 s. gesture data will consist of: 6 accelerometer Vectors of length 250 each, 6 gyroscope Vectors of Length 250 each, 6 orientation Vectors of Length 250 each, 16 EMG Vectors, Length 250 each. We combine all this for simplicity into a 34 × 250 matrix. Each time-series is transformed to make sure the initial value is 0, by subtracting this value from all values in the time-series. This will help prevent errors when the user performs the sign in a different starting positions. Normalization is then done by representing all values as floats between 0 and 1 by min-max method.

Orientation values are received in the form of three time-series in terms of unit quaternions. The pitch, yaw and roll values are obtained from the quaternion values w, x, y and z by using the following equations:

$$roll = tan^{-1}(\frac{2(wx+yz)}{-x^2y^2}) \tag{1}$$
$$pitch = sin^{-1}(max(-1, min(1, 2(wy - zx))))$$
$$yaw = tan^{-1}(\frac{2(wz+xy)}{-y^2z^2}).$$

EMG Energy

After correctly identifying the location of each of the individual pods of the two Myo devices, data is stored and shuffled in such a way that the final stored data is aligned from EMG pod-1 to EMG pod-8. This gives flexibility to the end-user as she no longer has to remember 'how' to put on the devices, and can thus randomly put either Myo device on either hand and expect good results. This is a great leap towards the pervasiveness and ease of use of this technology.

While we found that the DTW based testing worked very good for accelerometer and orientation comparisons, the EMG comparison numbers were not satisfactory. This is because, EMG activations seem to be much more stochastic and a 'shape' based comparison did not yield good results. The shapes formed by performing the same gesture for two different people are drastically different as seen in Figure 7, but they also differ in shape significantly enough between two repetitions by the same person to not give a small DTW distance. However, it was found by experimentation that gestures tend to activate the same EMG pods when repeated. Thus an energy based comparison was tried:

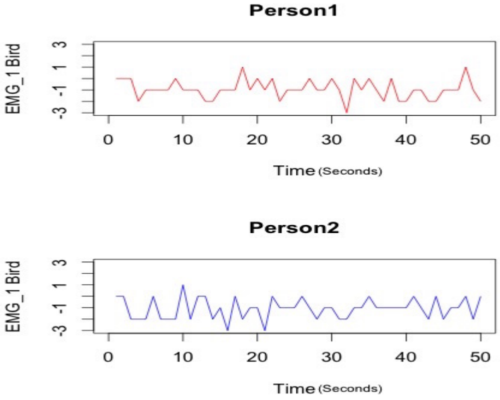

Figure 7. Variation in EMG signals between two users.

EMG energy 'E' on each pod is calculated as the sum of squares of x[n] , the value of the time-series at point 'n' .

$$E = sum(x[n]^2). \tag{2}$$

Dynamic Time Warping

We used four different approaches for comparing accelerometer and orientation data: a. Euclidian distance b. Regression c. Principal Component Analysis (PCA) and d. DTW. Euclidian distance simply compares the two time-series using mean-squared error. Regression analysis fits a model to the time-series and uses this model to compare best fit for the test gesture. Given a set of features from the time-series for a gesture, PCA derives the optimal set of features for comparison. According to [25], DTW is a well-known technique to find an optimal alignment between two given (time-dependent) sequences under certain restrictions. It was found that DTW based methods gave best results. DTW based approached proved ideal in our use-case since it could gracefully handle the situations when the sign was performed slower or faster than the stored training data, or performed with a small lag. Traditionally, DTW has been used extensively for speech recognition, and it is finding increasing use in the fields of gesture recognition as well [30], specially when combined with Hidden Markov Models [38]. SCEPTRE takes a simpler approach by randomly re-sampling the training and test datasets based on the least number of points in either one, then doing a DTW based distance analysis on them. First a DTW analysis of Accelerometer Data is run and a ranked list of 'probable' signs is passed on for DTW based analysis of orientation Data which in turn creates an even shorter ranked list to be processed by the final EMG algorithm.

On another approach, the normalized distances from each of the outputs is taken and the sum of squares of the final out-

Algorithm 2 Overall Gesture recognition method.

```
procedure GestureRecognition
    $t \leftarrow$ elapsed time from start of the experiment
    $test\_gesture \leftarrow$ get gesture data from the user
    $S_G \leftarrow$ query list of recognized gestures from database
    $normalized\_accl \leftarrow accel\_values - first\_accl\_value$
    $normalized\_orient \leftarrow orientation - first\_orientation$
    $normalized\_emg \leftarrow emg\_values - first\_emg\_value$
    for each trained gesture $T_G \in S_G$ do
        $dtw\_accl \leftarrow dtw(training\_accl, normalized\_accl)$
        $dtw\_orient \leftarrow dtw(training\_orient, normalized\_orient)$
        $T_G^E \leftarrow$ calculated energy of each pod for the training gesture
        $emg\_energy\_diff \leftarrow T_G^E$ - test_emg_energy
        $scal\_lim \leftarrow \{0, 1\}$
        $scaled\_emg\_diff \leftarrow$ scale(emg_energy_diff,scal_lim)
        $scaled\_accl\_dist \leftarrow$ scale(dtw_accl, scal_lim)
        $scaled\_orient\_dist \leftarrow$ scale(dtw_orient, scal_lim)
        $combined\_nearness \leftarrow$ compute nearness from Eq. 3
    end for
    Sort $T_G$ with respect to combined nearness.
    $recognized\_gesture \leftarrow$ top $T_G$ in the sorted list
end procedure
```

put is taken as an indication of 'closeness' of a test sign to a training sign. This is the approach that was chosen due to its computational simplicity and speed.

Combination

The overall 'nearness' of two signs is computed to be the total distance between those signs which is obtained by adding up the scaled distances for accelerometer, orientation and EMG as discussed above. An extra step of scaling the distance values between (0,1) is performed so as to give equal weight to each of the features. Also, since we have 8 EMG pods and only 3 each of accelerometer and orientation sensors, we use the following formula for the combination. The formula is for combining accelerometer sum of distances and EMG sum of distances. Similar techniques were applied for the other combinations. An algorithmic summary can be found in Equation 3.

$$dist = (8cs(sc\_accl\_comb) + 3cs(sc\_emg\_comb))/24. \quad (3)$$

where $cs()$ is a function that returns the sum of columns, sc_accl_comb is a data frame that holds the combined accelerometer DTW distances for both hands for all trained signs, and sc_emg_comb is a data frame that holds the combined EMG energy distances for both hands for all trained signs. The overall algorithm that is employed for recognition is summarized in Algorithm 2.

Ranking and Optimization

Timing constraints due to the real-time nature of the application requires us to optimize the recognition algorithm to be efficient, specially as the gesture space increases. As the number of gestures in the database increases to beyond 60, as we can see in Figure 8, the recognition time for identifying one gesture goes beyond the .5 s mark, which we have chosen to represent a real-time timing constraint. Thus, to deal with this situation, we modify our comparison algorithm to first compare to one stored instance of each gesture, then choose the top 'n' number of gestures which when compared to 'k' of each, still allowing the time-constraint to be fulfilled. We then, proceed with the normal gesture comparison routine on only these gesture instances and thus keep the recognition time within our constraints. All the variables for this method viz. the 'n', 'k' are calculated dynamically by what is allowed by the timing constraint 'tc', thus making this approach fluid and adaptable to more vigorous time constraints if required. The overall optimization algorithm is summarized in Algorithm 3.

Algorithm 3 Optimization to meet time constraints.

```
$databasesize \leftarrow sizeofgesturedatabaseforuser$
$time\_to\_compute \leftarrow estimatedtimefrompreviousiterations$
if $time\_to\_compute < time\_contraint$ then
    Compute Similarity Score with one template per gesture
    $top\_gestures \leftarrow list(top\_gestures, time\_to\_compute, time\_constrant)$
    Utilize Algorithm 2
    END
end if
```

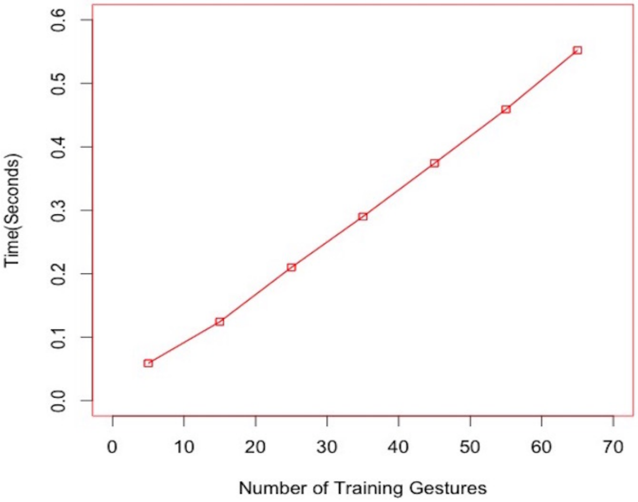

Figure 8. Database Size vs. Recognition Time.

RESULTS AND EVALUATION

Design of experiments

Experiments were designed to test the application with scalability in mind. The system begins on an "empty slate" and the gestures are trained interactively. Since a template based comparison is being utilized, we store the entire data set plus some features. After all the data is collected, the system is ready to be trained for another gesture. A sample application screen can be found in Figure 3. It is advised to provide 3-4 training instances of the same gesture to improve accuracy of recognition but that is not required. We test the system with 1, 2, 3, and 4 instances of each gesture to determine the correlation between number of instances saved and the recognition accuracy rate. That is all that is needed at the training phase. During testing phase we evaluate our system based on time it takes for recognition and on combination of features that produce the best results. Then we evaluate how recognition time increases with the increase in the number of gestures.

The way gestures are performed by the same person are generally similar but they have a tendency to vary slightly over time. Although, considering signals generated by a portable device that the user can put on and off by herself, provides the system a lot of portability and pervasiveness, it comes with a drawback that the signals we receive might not always be the same. To account for this experiments were performed that allowed users to put on the devices themselves, they were allowed to face any direction, and perform the gesture as long as they stayed in the Bluetooth range of the hub. Also, samples were collected over multiple sessions to account for variability in signing over time. The test subjects were 10 University students between the ages of 20 and 32 who performed the gestures in multiple sessions while also varying direction they were facing and if they were standing up/sitting down. The subjects viewed a video recording of the gestures being performed by ASL instructors before performing them with the system.

Prototyping and Choice of Gestures

20 ASL signs were chosen to prototype the system. They were carefully chosen such that a. A good mix of the various ASL components as discussed in Section Preliminaries and b. To include signs that are very close in some components but different in others. The choice of signs and the break-down of components for 10 of the 20 chosen signs is summarized in Table 3. The other 10 signs that were chosen are hurt, horse, pizza, large, happy, home, home, mom, goodnight, wash. Detailed analysis of system performance is give in Section Results and Evaluation.

Performance Metrics

The system is evaluated on each of the requirements discussed in Section Requirements. Pervasiveness of the system is justified since it is wireless and can work in Bluetooth range of the hub device which can either do the processing itself or offload it to a cloud server. Invasiveness is a harder metric to evaluate, as this seems to be more subjective. However, with wearable devices like smartwatches, wristbands such as FitBit, Myo becoming increasingly compact, wireless and even stylish [13], the proposed system is much less invasive then any of the alternatives. (see Related Work). Accuracy is undoubtedly one of the fundamental performance metrics for any recognition system. This basic function of the system is to facilitate real-life conversations, thus the system has to be able to function in real-time. Another aspect of the system is that it is extendable, thus the ease in scalability to a larger number of signs is very important. With the increase in the gesture space, the recognition time will also increase, the system should be able to scale up with a reasonable recognition time. Although there is a slight dip in recognition rate, the potential for scalability showcases the system's flexibility. Thus the experiments focus on these three metrics to evaluate the system:

1. Recognition time
2. Extensibility
3. Recognition rate (Accuracy)

Device Requirements

The device requirements for the experiments are:

Sensors: a. EMG, accelerometer and orientation (or Gyroscope) sensors that go between the wrist and elbow of a user on both hands. The desired frequency of each channel is 50 Hz. b. Wireless interface to transmit raw data from each of these sensors. c. A smartphone that acts as a hub to receive, pre-process, communicate the data to server, and later to display the results. d. A server that stores previous data in an individualized database per user, compares gestures, finds a match, and sends information back to the smartphone.

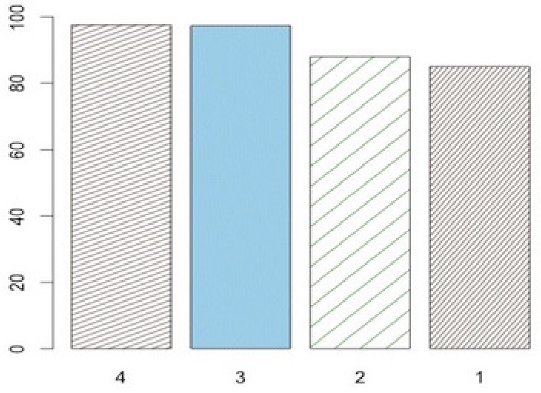

Figure 9. Training Data vs. Accuracy.

All experiments were performed with the setup described in Section System Architecture with devices with the following configurations:

Two Myo Sensors: Medical Grade Stainless Steel EMG sensors, Highly sensitive nine-axis IMU containing three-axis gyroscope, three-axis accelerometer, three-axis magnetometer. Expandable between 7.5 - 13 inches, wieghing 93 grams with a thickness of 0.45 inches. Communication channel: Bluetooth. Processor:ARM Cortex M4 Processor Hub Specifications(Android): OS: Android OS, v5.0 Chipset: Qualcomm MSM8974 Snapdragon 800 CPU: Quad-core 2.3 GHz Krait 400 Memory: 2 GB RAM Mac Specifications(Server) Processor:3.06GHz Intel Core i3 processor with 4MB level 3 cache; supports Hyper-Threading Momory: 8 GB DDR

Results

It was determined through experiments that three is a good choice for the number of training instances for each gesture. This formed a good compromise between usability and results as shown in Figure 9. Thus, each gesture was performed three times at varying times in the day. Then all data was aggregated, and annotated according to the gesture performed. The testing comprised of taking one dataset and comparing it with all other data sets and then estimating the correct gesture. With the 'guided mode' turned on, a recognition rate of 100% was achieved as expected. Then 20 randomly selected ASL Signs were used. The best accuracy was obtained by a tiered-combination method of all three features. The relative accuracy of other methods can be seen in Figure 11. This helps confirm that the combination of all three features produces the highest results.

Figure 8 shows how the recognition time increases with the increase in the number of gestures that are stored in the system. With 65 gestures in the database, a recognition time of

Table 3. Component breakdown of 10 of the chosen signs.

Number	Name of Sign	Location(Tab)	Movement	Orientation	Number of Hands
1	Blue	Around the Face	Wrist-Twist	Away from signer	1
2	Cat	Around the face	Outward and closing	Towards center	2
3	Cost	Trunk	Wrist-Flex	One towards signer one towards center	2
4	Dollar	Trunk	Two Wrist-Flexes	One towards signer one towards center	2
5	Gold	Side of the head	Away from center, change handshape twist	Towards the signer	1
6	Orange	Mouth	Open and close twice	Towards user	1
7	Bird	Mouth	Pinch with Two fingers	Away from signer	1
8	Shirt	Shoulder	Twice Wrist Flex	Facing down	2
9	Large	Trunk	Both Hands Away from Center	Facing away from signer	2
10	Please	Trunk/Chest	Form a circle	Both facing towards signer	1

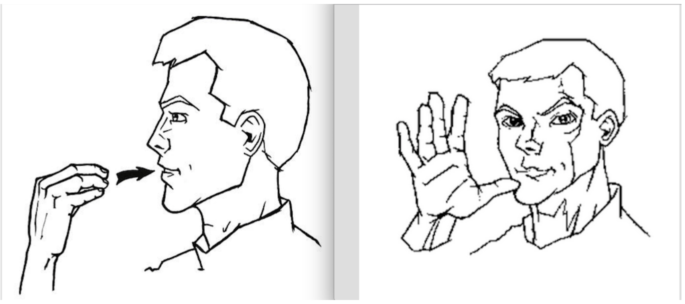

Figure 10. Gesture for ’Mom’ Vs. ’Eat’.

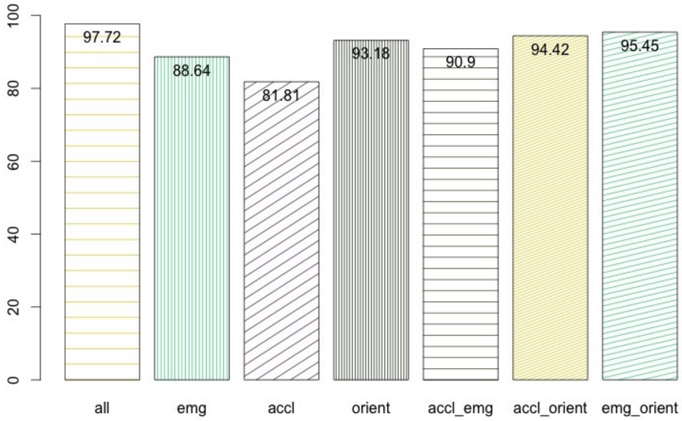

Figure 11. Features vs. Accuracy.

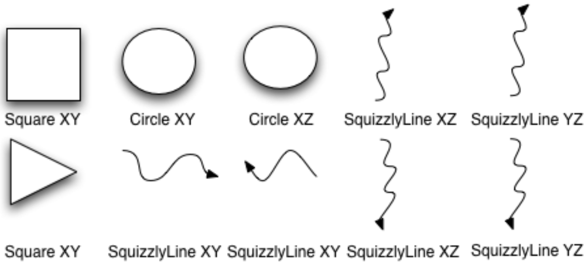

Figure 12. User Trained Gestures.

0.552 s was achieved. (processed on a 3.06 GHz Intel Core i3 8 GB RAM Mac). This fast response time means that the system qualifies for use for real-time communications. This time can be further improved upon by utilizing a more powerful server and optimizing the data queries, which will be part of future work.

Figure 13 shows the performance of the system based on recognition results for a combination of the features. The gestures tested were ’day’, ’enjoy’, ’eat’, ’happy’, ’home’, ’hot’, ’large’, ’mom’, ’pizza’, ’shirt’, ’still’, ’wash’ and ’wave’. The different sub-figures show that although comparing solely based on accelerometer performs good for some gestures like ’happy’, it is ineffective in distinguishing between others like ’day’. This can be explained because the nature of performing this gesture is very similar to other gestures in overall hand movements, but are different when it comes to finger configurations. Thus the performance gets better as shown in Figure 13c when EMG sensors are brought into the equation. Like we discussed earlier, the least number of recognition errors overall occur when all three of the sensor data are fused as seen in Figure 13d. This gives an insight on where the system is error-prone and thus can be optimized. Server level Optimization based on such input will be part of future work. We envision a server that continuously monitors success-failure data and becomes better by implementing machine learning algorithms to give weights to the different features.

These results can be understood better in context with the breakdown information given for each gesture in Table 3. For instance, Figure 10 shows screenshots of a person performing the gesture ’mom’ and another one performing ’eat’. These two gestures are very similar in ’Location’ and ’Movement’ but are distinctive when it comes to ’HandShape’ and ’Orientation of Hand’. Thus, while the system isn’t able to distinguish between these two gestures based on Accelerometer information only as seen in Figure 13a, the system does well when EMG information is included as seen in Figure 13c.

The guided mode as discussed in Usage was trained using 3 iterations each of 10 custom signs shown in Figure 12. An

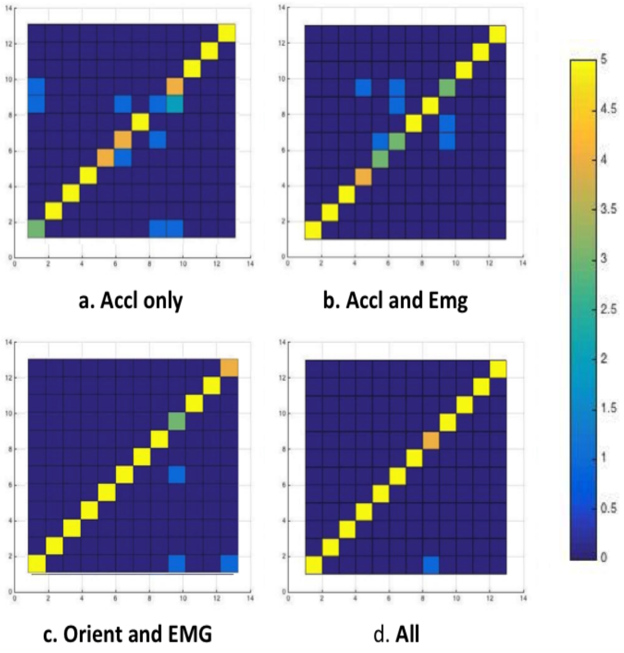

Figure 13. Gesture Recognition vs. Features used.

accuracy of 100% is achieved for this mode as expected, because clashes are detected at training time and avoided.

The results in Figure 11 show an accuracy of 97.72% for 20 ASL signs when all accelerometer, orientation and EMG are used. The database consisted of signs from 10 different people, and each person performed each sign 3 times. This table also lists the accuracy of other variations in the combination of these three input signals. The accuracy varied when individual databases were separated, however it stayed in the (97-100)% bracket when all signals were used on a dataset consisting of template gestures from the test user.

DISCUSSION

Continuous Processing: The gestures that were recognized in this test were isolated in nature. A start and stop button, or a timer was utilized to gather data. This was done to test the algorithms in isolation first. However, when the system is actually deployed to the public, a continuous monitoring algorithm has to be utilized. This can be done by using windowing methods and optimizing based on best results. However, we expect that the system will require a lot more training.

Dictionary Size: Another thing that can be improved upon is the dictionary size and the support for Signed Alphabet. Signed alphabet give an ASL user the ability to use English Language Alphabet to visually spell out a word. This is an important aspect of sign language communication since not all english words or ideas have sign language counterparts. The dictionary size used is currently 20 ASL words and 10 user invented gestures. The video based sign language dictionary website signasl.org currently has over 30,000 videos for sign languages in use, so there is a lot of room to grow into.

Framework Limitations: Currently due to limitations of the Myo framework for Android, final tests involving EMG data signals had to be done using laptop computers. Two Mac computers were used to gather the data simultaneously via Bluetooth channels and then combine them and store them to a cloud based data storage system. A separate script was triggered to process the stored data at test time. With the anticipated release of new framework for Myo Devices, this computation can be done at the mobile phone level and data can be sent to the server only at designated intervals.

User Independence: EMG data fluctuates a lot between people. This is apparent from Figure 7. More research can be done on data gathering and feature selection of the EMG data pods to come up with a 'unifying' framework that works for everyone. This will be a definitive direction in attaining user-independence while using EMG signals.

Search Algorithm: The search algorithm that is implemented can be improved upon to decrease the recognition time. Hierarchal searching can be done, in which training gestures are clustered according to the energies in each time-series data, and only those gestures are compared which have comparable energies.

SCEPTRE in collaboration with other HCI techniques such as brain-driven control [27, 28, 32] can revolutionize the way people interact with computers. In addition to the uses of HCI or Sign Language Communication, this technology can also be extended to uses in the domains of activity recognition, food intake estimation or even physiotherapy.

CONCLUSIONS

DTW and energy based comparison methods are fast and highly effective. Template based comparison techniques have a great advantage of lossless information storing. Comparing accelerometer and orientation data between gestures is best done by using Dynamic Time Warping methods which muscle movement (EMG) data is best compared by comparing total energies in each pods. The structure present in Sign Languages like ASL can be divided into components which explain the success of certain signals in recognizing them. The overall accuracy of the system is increased by a smart combination of these signals without compromising on speed, recognition rate, or usability. The sensors that are used are non-invasive and versatile thus allowing people to effectively utilize them in day-to-day situations without drawing much attention. The future direction of this research is in incorporating continuous sign processing, user independence in the system and increasing the dictionary size.

ACKNOWLEDGEMENTS

This project was partly funded by NIH NIBIB (Grant #EB019202) and NSF IIS (Grant #1116385). We thank Julie Stylinski, Lecturer at ASU for her invaluable ASL domain knowledge and for facilitating the survey. We thank all the survey participants as well as volunteers that helped to train/test the system. We thank Subhasish Das of IMPACT lab, ASU for his technical contribution to get the project going.

REFERENCES

1. Daphne Bavelier Aaron J. Newman. A critical period for right hemisphere recruitment in american sign language processing. In *Multimodal Signals: Cognitive and Algorithmic Issues*, pages 76–80, 2002.
2. Frank Adelstein, Sandeep KS Gupta, Golden Richard, and Loren Schwiebert. *Fundamentals of mobile and pervasive computing*, volume 1. McGraw-Hill New York, 2005.
3. B.J. Bahan. *Non-manual Realization of Agreement in American Sign Language*. UMI Dissertation Services. UMI, 1996.
4. Ayan Banerjee and Sandeep KS Gupta. Analysis of smart mobile applications for healthcare under dynamic context changes. *Mobile Computing, IEEE Transactions on*, 14(5):904–919, 2015.
5. Sara Bilal, Rini Akmeliawati, Amir A Shafie, and Momoh Jimoh E Salami. Hidden markov model for human to computer interaction: a study on human hand gesture recognition. *Artificial Intelligence Review*, 40(4):495–516, 2013.
6. Baptiste Caramiaux, Marco Donnarumma, and Atau Tanaka. Understanding gesture expressivity through muscle sensing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 21(6):31, 2015.
7. Feng-Sheng Chen, Chih-Ming Fu, and Chung-Lin Huang. Hand gesture recognition using a real-time tracking method and hidden markov models. *Image and vision computing*, 21(8):745–758, 2003.
8. Xiang Chen, Xu Zhang, Zhang-Yan Zhao, Ji-Hai Yang, Vuokko Lantz, and Kong-Qiao Wang. Hand gesture recognition research based on surface emg sensors and 2d-accelerometers. In *Wearable Computers, 2007 11th IEEE International Symposium on*, pages 11–14. IEEE, 2007.
9. Philip R Cohen and Sharon L Oviatt. The role of voice input for human-machine communication. *proceedings of the National Academy of Sciences*, 92(22):9921–9927, 1995.
10. Andrea Corradini. Dynamic time warping for off-line recognition of a small gesture vocabulary. In *Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 2001. Proceedings. IEEE ICCV Workshop on*, pages 82–89. IEEE, 2001.
11. E. Costello. *Random House Webster's American Sign Language Dictionary*. Random House Reference, 2008.
12. Gaolin Fang, Wen Gao, and Debin Zhao. Large vocabulary sign language recognition based on fuzzy decision trees. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 34(3):305–314, 2004.
13. Sandeep KS Gupta, Tridib Mukherjee, and Krishna Kumar Venkatasubramanian. *Body area networks: Safety, security, and sustainability*. Cambridge University Press, 2013.
14. Deng-Yuan Huang, Wu-Chih Hu, and Sung-Hsiang Chang. Vision-based hand gesture recognition using pca+ gabor filters and svm. In *Intelligent Information Hiding and Multimedia Signal Processing, 2009. IIH-MSP'09. Fifth International Conference on*, pages 1–4. IEEE, 2009.
15. T. Johnston and A. Schembri. *Australian Sign Language (Auslan): An introduction to sign language linguistics*. Cambridge University Press, 2007.
16. Daniel Kelly, Jane Reilly Delannoy, John Mc Donald, and Charles Markham. A framework for continuous multimodal sign language recognition. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 351–358. ACM, 2009.
17. Kyung-Won Kim, Mi-So Lee, Bo-Ram Soon, Mun-Ho Ryu, and Je-Nam Kim. Recognition of sign language with an inertial sensor-based data glove. *Technology and Health Care*, 24(s1):S223–S230, 2015.
18. 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*, 141:108–125, 2015.
19. T Kuroda, Y Tabata, A Goto, H Ikuta, M Murakami, et al. Consumer price data-glove for sign language recognition. In *Proc. of 5th Intl Conf. Disability, Virtual Reality Assoc. Tech., Oxford, UK*, pages 253–258, 2004.
20. Yun Li, Xiang Chen, Jianxun Tian, Xu Zhang, Kongqiao Wang, and Jihai Yang. Automatic recognition of sign language subwords based on portable accelerometer and emg sensors. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ICMI-MLMI '10, pages 17:1–17:7, New York, NY, USA, 2010. ACM.
21. Rung-Huei Liang and Ming Ouhyoung. A real-time continuous gesture recognition system for sign language. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 558–567. IEEE, 1998.
22. Robert B Miller. Response time in man-computer conversational transactions. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, pages 267–277. ACM, 1968.
23. Sushmita Mitra and Tinku Acharya. Gesture recognition: A survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 37(3):311–324, 2007.
24. Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

25. Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

26. Yuji Nagashima. Present stage and issues of sign linguistic engineering. *HCI, 2001*, 2001.

27. Koosha Sadeghi Oskooyee, Ayan Banerjee, and Sandeep KS Gupta. Neuro movie theatre: A real-time internet-of-people based mobile application. 2015.

28. Madhurima Pore, Koosha Sadeghi, Vinaya Chakati, Ayan Banerjee, and Sandeep KS Gupta. Enabling real-time collaborative brain-mobile interactive applications on volunteer mobile devices. In *Proceedings of the 2nd International Workshop on Hot Topics in Wireless*, pages 46–50. ACM, 2015.

29. Nikhita Praveen, Naveen Karanth, and MS Megha. Sign language interpreter using a smart glove. In *Advances in Electronics, Computers and Communications (ICAECC), 2014 International Conference on*, pages 1–5. IEEE, 2014.

30. Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

31. Éva Sáfár and Ian Marshall. The architecture of an english-text-to-sign-languages translation system. In *Recent Advances in Natural Language Processing (RANLP)*, pages 223–228. Tzigov Chark Bulgaria, 2001.

32. Javad Sohankar, Koosha Sadeghi, Ayan Banerjee, and Sandeep KS Gupta. E-bias: A pervasive eeg-based identification and authentication system. In *Proceedings of the 11th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, pages 165–172. ACM, 2015.

33. Thad Starner, Joshua Weaver, and Alex Pentland. Real-time american sign language recognition using desk and wearable computer based video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(12):1371–1375, 1998.

34. W.C. Stokoe, D.C. Casterline, and C.G. Croneberg. *A Dictionary of American Sign Language on Linguistic Principles*. Linstok Press, 1976.

35. W.C. Stokoe, ERIC Clearinghouse for Linguistics, and Center for Applied Linguistics. *The Study of Sign Language*. ERIC Clearinghouse for Linguistics, Center for Applied Linguistics, 1970.

36. R.A. Tennant and M.G. Brown. *The American Sign Language Handshape Dictionary*. Clerc Books, 1998.

37. Christian Vogler and Dimitris Metaxas. Asl recognition based on a coupling between hmms and 3d motion analysis. In *Computer Vision, 1998. Sixth International Conference on*, pages 363–369. IEEE, 1998.

38. Andrew D Wilson and Aaron F Bobick. Parametric hidden markov models for gesture recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(9):884–900, 1999.

39. Xu Zhang, Xiang Chen, Yun Li, Vuokko Lantz, Kongqiao Wang, and Jihai Yang. A framework for hand gesture recognition based on accelerometer and emg sensors. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 41(6):1064–1076, 2011.