Presented at the DLSU Research Congress 2015
De La Salle University, Manila, Philippines
March 2-4, 2015

Sign-language Recognition through Gesture & Movement Analysis (SIGMA)

Ian Lim, Joshua Lu, Claudine Ng, Thomas Ong and Clement Ong*
College of Computer Studies – De La Salle University Manila, Philippines
**clem.ong@delasalle.ph*

Abstract: Sign language is used as the means of communication by the speech-impaired. It is a language which uses hand gestures and movements to convey meaning. Several studies have attempted to use computers to recognize hand gestures using either a data glove, which uses several sensors to determine hand pose and attitude, or a vision based system, which uses a camera to determine the hand position and gestures made. Both approaches have their inherent strengths and limitations. This paper introduces a system that combines a prototype data glove with computer vision in order to translate Filipino Sign Language for medical purposes. Ten words with similarities in gestures were selected resulting in an accuracy of 80%.

Keywords: Automated sign language recognition; flex sensor; inertial measurement unit; gesture recognition

1. INTRODUCTION

Based on Right diagnosis' [11] extrapolated statistics there are approximately 862,416 speech impaired persons in the Philippines. Most prominently, sign language comes to the aid of the deaf and speech impaired. Filipino Sign Language is currently being used by 54% sign language users in the Philippines [6]. FSL is the ordered and rule-governed visual communication which has risen naturally and embodies the cultural identity of the Filipino community of signers [6].

Sign language translation is needed everywhere. In education, social services, and most importantly health care services. In the Philippines however, most health care workers do not understand sign language. They provide daily treatment and care to many people, with some who are speech impaired. It should be noted that most common errors in the field of medicine is because of miscommunication [7]. When the clients and health care providers do not share a common language, a qualified sign language interpreter can facilitate communication. They are necessary in situations wherein the important information needs to be exchanged such as taking a patient's medical history, giving diagnosis, performing medical procedures, explaining treatment planning, and during emergency situations [8]. However, there is a lack of sign language interpreters in the field of medicine.

Using computers to recognize sign language and convert this to text is not new; many successful systems have been developed, although only a handful for FSL. All systems use either an instrumented glove to read hand pose and motion (gestures), or use computer vision to do the same [10] [12] [13].

Computer vision approaches tend to require significant amounts of computing power, while glove-based approaches provide direct reading but are inherently unable to provide hand versus body position information due to the relative measurement nature of the glove sensors.

Sign-language Recognition through Gesture and Movement Analysis (SIGMA) combines a data glove with image processing. The guiding principle of SIGMA is to use the inherent direct reading

capabilities of an instrumented glove with "just enough" video processing to provide sparse hand position information, relative to the signer's body, over time. The end-goal is to produce a system that will be within the computational capabilities of a smartphone, working without additional requirements from the Cloud.

2. SIGMA SYSTEM

SIGMA aims to break the barriers of communication between the speech-impaired and those who are not especially in the medical field wherein proper communication is important. To achieve this goal, a prototype data glove was developed to capture hand posture and attitude information, and a vision system is used, and limited to, detecting the hand position with respect to the body.

2.1. Data Glove

The data glove is composed of an Arduino platform, four LEDs, and two types of sensors - the flex sensor and the IMU. Nine flex sensors (two flex sensors per finger and one on the thumb) and an IMU was used to capture hand pose and attitude information. The Arduino is responsible for timing the sensor when and from which sensor to capture data. This is also where the Analog-to-Digital conversion takes place with the use of an analog multiplexer circuit. After conversion, the data is processed and sent to the PC for training or recognition.

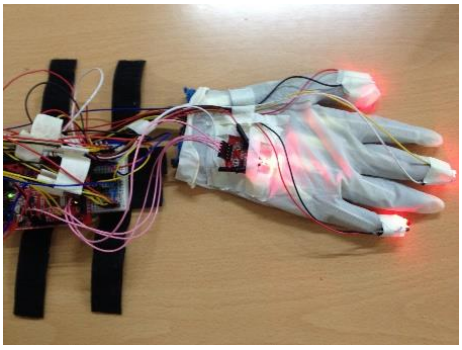


Figure 1 – Data Glove

2.2. Flex Sensor

Flex sensors utilized are passive resistive units, as shown in Figure 3. The sensor changes resistance proportionately to the amount of bend. As the sensor is flexed, the resistance varies from 25K - 125K ohms. The sensor is 5.5cm, which ensures that it reads the bending of only one joint. The flex sensors were used in a voltage divider so that it generates voltage values rather than resistance values.

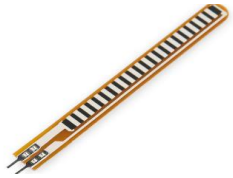


Figure 2 – Flex Sensor

2.3. Inertial Measurement Unit

Inertial Measurement Units (IMUs) is a self-contained system that measures linear and angular motion usually with a triad of gyroscopes and triad of accelerometers [3]. The accelerometer generates three analog signals describing the accelerations along each of its axes produced by, and acting on the device. The gyroscope also outputs three analog signals that describe the angular rate about each of the sensor axes.

Figure 3 - IMU

2.4. Data Acquisition

Data acquisition is accomplished through an Arduino equivalent; a single-board microcontroller, intended to make building interactive objects or environments more accessible [2]. The Arduino is responsible for the collection of data from the different sensors connected to it. It is used to manage the sensor data accurately by using proper timing to decide which of the sensor data will be converted into digital value. After data acquisition, the microcontroller sends all the data to the computer through serial communication.

2.5. Vision System

The vision system uses a webcam configured to capture at a rate of 8 frames per second with a

resolution of 640x480 in order to detect the hand position with respect to the body. The color representation used to determine the components were selected based on certain conditions. The color red was chosen because it is the most visible among all other color choices. It is also the color of the selected LEDs that would be integrated with the data glove to determine the position of the hand of the user. On the other hand, the color blue is chosen because it is the only color aside from red that can be easily determined by the image processing. The process begins by acquiring the coordinates of the face of the signer. This is then followed with the vision system that gets an RGB frame from the video.

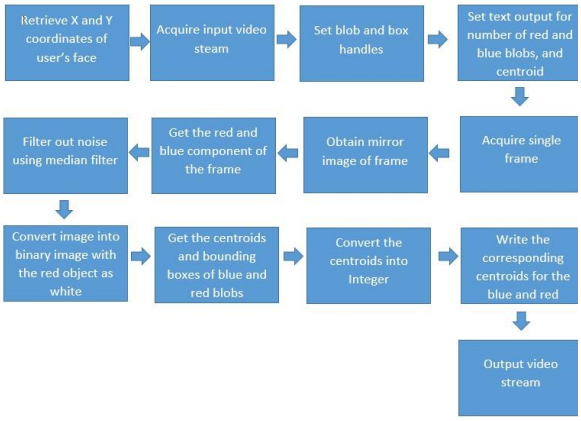


Figure 4 – Vision System Block Diagram

After obtaining the RGB frame, the system then extracts the red channel matrix and subtracts it with the gray image of the RGB frame.

The result is then converted into a corresponding Binary Image using the proper threshold value, which is determined through trial and error.

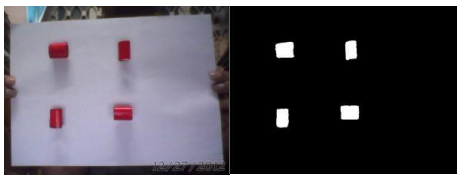


Figure 5 – Color Image and Resulting Binary Image

The process then gets the blue component of the image and filters out the noise through a median filter. This is followed by getting the centroids and bounding boxes of the blobs for the blue and red components. The different centroids of the red and blue components and the coordinates of the face of the user were used to compute for the Euclidean distance. The computed Euclidean distance is then stored into a text file which will be used for the preprocessing module.

3.METHODOLOGY

For the initial data, the flex sensors were calibrated in order for it to give a more accurate and reliable output. After calibration, the data glove started capturing multiple values of finger flexion and hand movement. Each gesture was performed ten times and was captured using two separate software. While the data glove was capturing, the vision system used MATLAB's functionalities to connect and capture live video feed through the webcam. The vision system starts capturing when a user clicks on the signer's face. This will also determine its X and Y coordinates, which is a baseline for computing the distance between the hand and the face.

The initial experimentation made use of six gestures to establish optimal data gathering technique. These six gestures were specifically chosen because they represent similarities with the positioning of the hand, delivery of the gesture, and have different levels of difficulty and length in performing the gesture. The gesture 'Wednesday' is static, similar to the other gestures such as the 24 letters of the alphabet, the days of the week, and the months of the year. The gesture 'Doctor' shows similarities in the position of the user's hand in comparison to how the user performs the gestures 'Injection' and 'Allergy', wherein all three gestures have the right hand placed over the left hand. There are also instances wherein the movements and positioning of the two hands of the user are performed similarly such as in the case of the gesture 'Drug Test'.

After capturing, data were manually segmented to determine the starting and the ending of the gestures. Since two separate software were used to capture, another software is used to combine the data together to form a feature vector sequence. The feature vector sequence were then labeled into:

- Allergy
- Doctor
- Drug Test

- Headache
- Injection
- Wednesday

After the analysis of the data of the initial experimentation, another experiment is conducted. This experimentation made use of ten gestures to establish the number of feature vectors needed for the Hidden Markov model. These ten gestures were chosen because of having similar motion and hand form. The ten gestures are:

- After
- Allergy
- Always
- Before
- Blood Pressure
- Cold
- Cough
- Doctor
- Everyday
- Drug Test

The gesture of these ten words can be seen in the Section 8, Appendix.

4. RESULTS AND DISCUSSION

An analysis of the data was conducted by calculating the variance of each feature per gesture. A small variance indicates that the data tend to be very close to the mean and hence to each other, while a high variance indicates that the data are very spread out around the mean and from each other [6]. However, for gestures, a high variance should be the result since gesture data changes, which will result to the values being spread out. This is the reason why calculating for the variance of the variances is used in order to visualize the inconsistency of data.

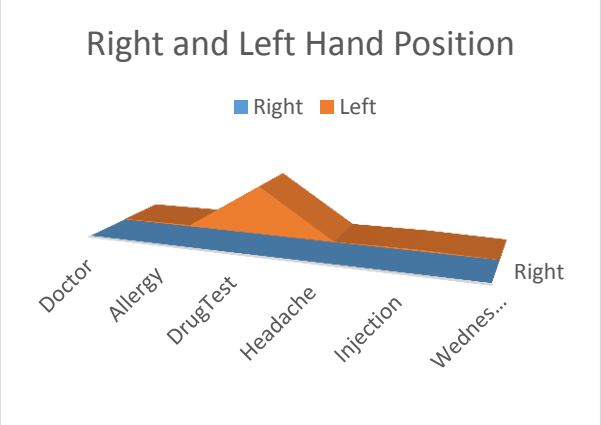


Figure 6 – Detected Image Position Variance for Right and Left Hand

As seen in Figure 6, the left hand has a high variance compared to the right hand. The reason for this is because the vision system that detects the left hand does not depict the color blue consistently, thus, it cannot track the left hand properly.

The calibration process is the establishment of individual flex sensor values and ranges. In order to calibrate the flex sensors, the signer is asked to open their fist for three secods and close their fist for another three seconds to let the system capture the high and low points of each analog input. As seen in Figure 7, the flex sensor in the thumb has the highest variance. This is due to the fact that during the calibration process, wherein the signer clinches his/her fist, the thumb bends slightly compared to the other fingers. This causes the flex sensor values to be sensitive to bending during the data gathering process.

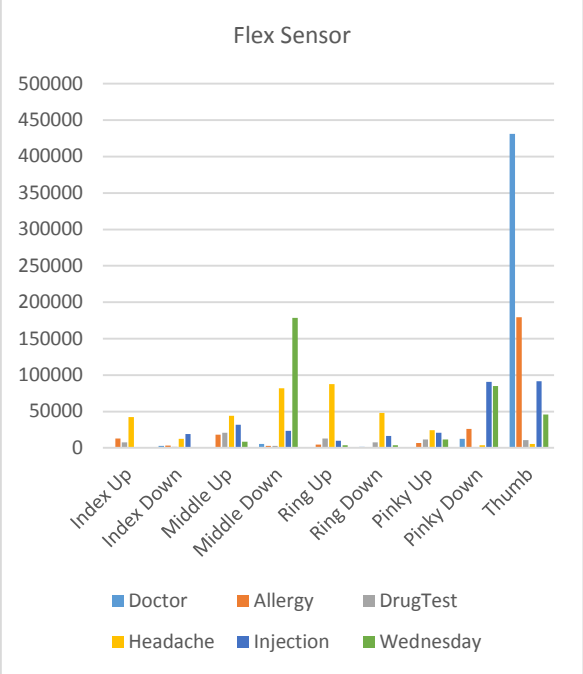


Figure 7 - Variance of Flex sensors

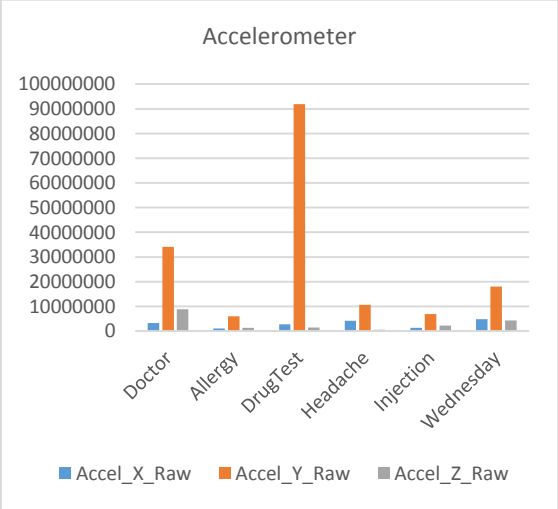


Figure 8 – Variance of Accelerometer

As seen in Figure 8, the gesture 'Drug Test' has the highest variance value in the y-axis. This is because of the constant movement of the gesture wherein the hands move from left to right, and maintained at a vertical position for a certain amount of time as seen below, thus, creating different accelerometer values unlike for the gesture 'Injection', wherein the right hand only moves to one position in the axis which results a low variance, compared to doctor, allergy, headache, and Wednesday. Figures 9 and 10 show the hand positions / motions associated with "Drug test" and "Injection", respectively.

Figure 9 – Drug Test

Figure 10 – Injection

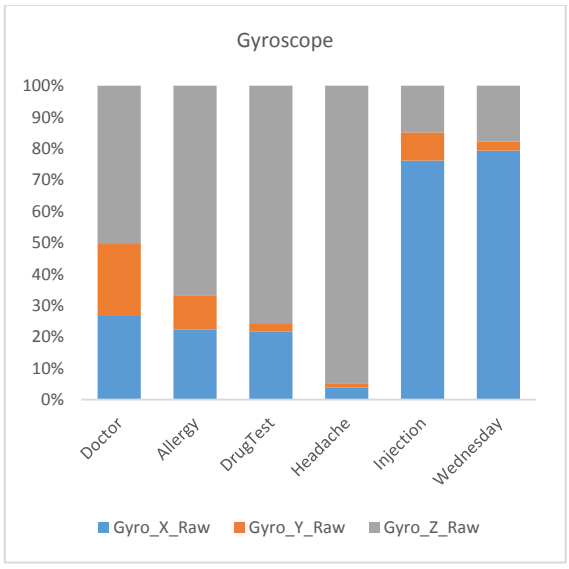


Figure 11- Variance of Gyroscope

As seen in Figure 11, the gyroscope has a high variance on X and Z axes. Since gyroscopes measure rotation and discriminate against linear movement, the gyroscope output has lower noise than accelerometers. On the other hand, gyros inherently have drift which build over time, which can explain the large variances recorded.

Tables 1-6 show confusion matrixes of the data set by using cross validation of overlapping and block data. It can be seen that the overlapping processes 6 through 9 have a close percentage of accuracy while the 5-block and 7-block process without overlap have accuracies of 80 and 61.29%, respectively.

Table 1 – Confusion Matrix using average of features with overlapping window size of 6 feature vectors

```
---------------------- Overall Results -----------------------
SENT: %Correct=76.32 [H=29, S=9, N=38]
WORD: %Corr=76.32, Acc=76.32 [H=29, D=0, S=9, I=0, N=38]
---------------------- Confusion Matrix ----------------------
             A  A  A  B  B  C  C  D  D  E
             f  l  l  e  l  o  o  o  r  v
             t  l  w  f  o  l  u  c  u  e
             e  e  a  o  o  d  g  t  g  r
             r  r  y  r  d     h  o  T  y  Del [ %c / %e]
After        0  0  0  1  0  0  0  0  0  0    0 [ 0.0/2.6]
Allergy      0  3  0  0  0  0  1  0  0  0    0 [75.0/2.6]
Always       0  0  3  0  0  0  0  0  3  0    0 [50.0/7.9]
Before       1  0  0  2  0  0  0  0  0  0    0 [66.7/2.6]
BloodPressure 0 0  0  0  3  1  0  0  0  0    0 [75.0/2.6]
Cold         0  0  0  0  0  3  0  0  0  0    0
Cough        0  0  0  0  0  0  4  0  1  0    0 [80.0/2.6]
Doctor       0  0  0  0  0  0  0  1  0  0    0
DrugTest     0  0  0  0  0  0  0  0  5  0    0
Everyday     0  0  0  0  0  0  0  0  1  5    0 [83.3/2.6]
Ins          0  0  0  0  0  0  0  0  0  0
============================================================
```

Table 2 – Confusion Matrix using average of features with overlapping window size of 7 feature vectors

```
---------------------- Overall Results -----------------
SENT: %Correct=70.59 [H=24, S=10, N=34]
WORD: %Corr=70.59, Acc=70.59 [H=24, D=0, S=10, I=0, N=34]
---------------------- Confusion Matrix ----------------
      A  A  A  B  B  C  D  D
      f  l  l  e  l  o  o  r
      t  l  w  f  o  l  c  u
      e  e  a  o  o  d  t  g
      r  r  y  r  d     o  T  Del [ %c / %e]
Afte  2  0  0  0  0  0  0  0    0
Alle  0  2  0  0  2  0  0  0    0 [50.0/5.9]
Alwa  0  0  5  0  0  0  0  0    0
Befo  0  0  0  3  0  0  0  0    0
Bloo  0  0  0  0  1  0  0  0    0
Cold  0  0  1  0  1  4  0  1    0 [57.1/8.8]
Doct  0  0  0  0  0  0  2  0    0
Drug  0  0  0  0  0  0  0  5    0
Ever  0  0  0  0  0  0  0  5    0 [ 0.0/14.7]
Ins   0  0  0  0  0  0  0  0
============================================================
```

Table 3 – Confusion Matrix using average of features with overlapping window size of 8 feature vectors

```
---------------------- Overall Results ----------------------
SENT: %Correct=75.00 [H=30, S=10, N=40]
WORD: %Corr=75.00, Acc=75.00 [H=30, D=0, S=10, I=0, N=40]
---------------------- Confusion Matrix ----------------------
      A  A  A  B  B  C  C  D  D  E
      f  l  l  e  l  o  o  o  r  v
      t  l  w  f  o  l  u  c  u  e
      e  e  a  o  o  d  g  t  g  r
      r  r  y  r  d     h  o  T  y  Del [ %c / %e]
Afte  2  0  0  1  0  0  0  0  0  0    0 [66.7/2.5]
Alle  0  2  0  0  1  0  0  0  0  0    0 [66.7/2.5]
Alwa  0  0  4  0  0  0  0  0  0  0    0
Befo  0  0  0  4  0  0  0  0  0  0    0
Bloo  0  0  0  0  3  0  0  0  0  0    0
Cold  0  0  3  0  2  1  0  0  0  0    0 [16.7/12.5]
Coug  0  0  1  0  0  0  2  0  0  0    0 [66.7/2.5]
Doct  0  0  0  0  0  0  0  5  0  0    0
Drug  0  0  1  0  0  0  1  0  4  0    0 [66.7/5.0]
Ever  0  0  0  0  0  0  0  0  0  3    0
Ins   0  0  0  0  0  0  0  0  0  0
============================================================
```

Table 4 – Confusion Matrix using average of features with overlapping window size of 9 feature vectors

```
---------------------- Overall Results -----------------
SENT: %Correct=72.41 [H=21, S=8, N=29]
WORD: %Corr=72.41, Acc=72.41 [H=21, D=0, S=8, I=0, N=29]
---------------------- Confusion Matrix ----------------
      A  A  B  B  C  C  D  D  E
      f  l  e  l  o  o  o  r  v
      t  w  f  o  l  u  c  u  e
      e  a  o  o  d  g  t  g  r
      r  y  r  d     h  o  T  y  Del [ %c / %e]
Afte  3  0  1  0  0  0  0  0  1    0 [60.0/6.9]
Alwa  0  3  0  0  0  0  0  0  0    0
Befo  0  0  1  0  1  0  0  0  0    0 [50.0/3.4]
Bloo  0  0  0  5  0  0  0  0  0    0
Cold  0  0  0  0  1  0  0  1  0    0 [50.0/3.4]
Coug  0  2  0  0  0  2  0  1  0    0 [40.0/10.3]
Doct  0  0  0  0  0  0  1  0  0    0
Drug  0  0  0  0  0  0  0  3  0    0
Ever  0  0  0  0  0  0  0  1  2    0 [66.7/3.4]
Ins   0  0  0  0  0  0  0  0  0
============================================================
```

Table 5 – Confusion Matrix using average of features with window size of 5 feature vectors

```
---------------------- Overall Results -----------------
SENT: %Correct=80.00 [H=28, S=7, N=35]
WORD: %Corr=80.00, Acc=80.00 [H=28, D=0, S=7, I=0, N=35]
---------------------- Confusion Matrix ----------------
      A  A  B  B  C  C  D  D  E
      l  l  e  l  o  o  o  r  v
      l  w  f  o  l  u  c  u  e
      e  a  o  o  d  g  t  g  r
      r  y  r  d     h  o  T  y  Del [ %c / %e]
Alle  2  1  0  0  0  0  1  0  0    0 [50.0/5.7]
Alwa  0  3  0  0  0  0  0  1  0    0 [75.0/2.9]
Befo  0  0  5  0  0  0  0  0  0    0
Bloo  0  2  0  3  0  0  0  0  0    0 [60.0/5.7]
Cold  0  0  0  0  4  0  0  0  0    0
Coug  0  1  0  0  0  1  0  0  0    0 [50.0/2.9]
Doct  0  0  0  0  0  0  4  0  1    0 [80.0/2.9]
Drug  0  0  0  0  0  0  0  3  0    0
Ever  0  0  0  0  0  0  0  0  3    0
Ins   0  0  0  0  0  0  0  0  0
============================================================
```

Table 6 – Confusion Matrix using average of features with window size of 7 feature vectors

```
---------------------- Overall Results ----------------------
SENT: %Correct=61.29 [H=19, S=12, N=31]
WORD: %Corr=61.29, Acc=61.29 [H=19, D=0, S=12, I=0, N=31]
---------------------- Confusion Matrix ---------------------
      A   A   A   B   B   C   C   D   D   E
      f   l   l   e   l   o   o   o   r   v
      t   l   w   f   o   l   u   c   u   e
      e   e   a   o   o   d   g   t   g   r
      r   r   y   r   d       h   o   T   y   Del [ %c / %e]
Afte  1   0   0   0   0   0   0   0   0   1     0 [50.0/3.2]
Alle  0   0   0   0   0   0   0   1   0   0     0 [ 0.0/3.2]
Alwa  0   0   1   0   0   0   0   0   2   0     0 [33.3/6.5]
Befo  0   0   0   3   0   0   0   1   0   0     0 [75.0/3.2]
Bloo  0   0   1   0   2   0   0   0   0   0     0 [66.7/3.2]
Cold  0   0   0   0   0   3   0   0   1   0     0 [75.0/3.2]
Coug  0   1   3   0   0   0   3   0   0   0     0 [42.9/12.9]
Doct  0   0   0   0   0   0   0   2   0   0     0
Drug  0   0   0   0   0   0   0   0   3   0     0
Ever  0   0   0   0   0   0   0   0   1   1     0 [50.0/3.2]
Ins   0   0   0   0   0   0   0   0   0   0
=============================================================
```

It appears that overlapping to smooth transitions is unnecessary and detrimental. The higher performance of the smaller block size for which the averaging is done indicates sensor values are changing significantly within five sampling intervals; more robust (i.e. predictive) techniques will be needed to reduce the data representation required while maintaining recognition accuracy.

Looking at

Table **5**, majority of the gestures resulted in 75% and above recognition rate. The gesture with the highest recognition rate is "Doctor". It is an example of an accurately recognized gesture wherein it was correctly recognized 80% of the time.

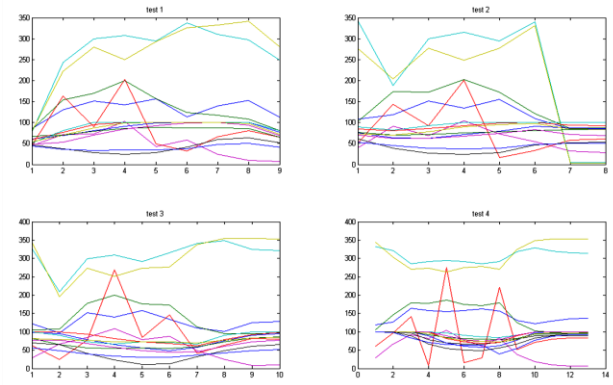


Figure 12 – Plotted data of the gesture "Doctor"

As seen in Figure 12, the data of test 1, 3, and 4 are similar to each other, while test 2 was incorrectly recognized by the HMM as "Everyday". The reason for this gesture to be recognized incorrectly is caused by performing incorrectly the gesture wherein the performed gesture is "significantly" different from the ideal gesture or the trained gesture set, and length of time it was performed. It can also be noted that the misrecognition of these gestures may be due to the user's performance of an incomplete gesture or when the signer stops and drops their hands in the middle of the gesture.

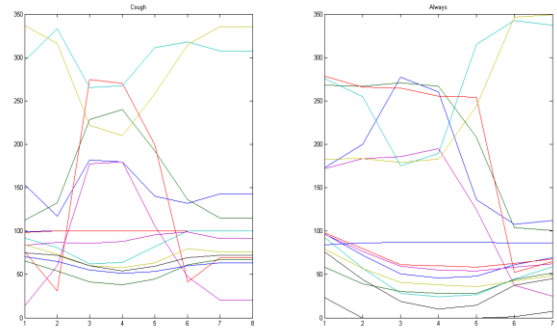


Figure 13 – Cough vs Always

Figure 14 – Signing for Cough (Left) vs. Always (Right); note gesture similarity.

The gesture "Cough" was recognized only 50% of the time. It is being recognized as "Always" since they have strongly similar motion and hand form, as seen above where the right hand almost has the same coordinates and has the same "closed-fist" pose.

5. CONCLUSIONS

SIGMA combines a data glove housing a 6-DOF IMU and nine flex sensors with image processing. Information is combined to derive more accurate hand pose over time, favoring direct reading to reduce computational cost. The system currently achieves an 80% recognition rate on a 10-word FSL vocabulary composed of medical terms, a number of which have strongly similar gestures.

6. FUTURE WORK

The overall system is expected to be enhanced and modified in several areas in order to cope with the full vocabulary target of 23 medical terms, the alphabet, numbers zero to 10 and the days of the week. For the data glove, flex sensor anchoring needs to be improved and a shorter, larger resistance-delta sensor needs to be utilized for the thumb. To address gyroscope drift on the IMU, sensor fusion will be implemented. More robust modeling, coupled with other Machine Learning (ML) techniques are expected to reduce data representation requirements while still increasing vocabulary size and recognition accuracy.

7. REFERENCES

[1] A. Bose, "How to Detect and Track Red Objects in Live Video in MATLAB," Arindam's Blog, 10 November 2013. [Online]. Available: http://arindambose.com/blog/?p=72. [Accessed 222 October 2014].

[2] "Arduino," 22 October 2014. [Online]. Available: http://www.arduino.cc/.

[3] "IMU Inertial Measurement Unit," Xens, [Online]. Available: https://www.xsens.com/tags/imu/. [Accessed 22 October 2014].

[4] L. Martinez, "An Introduction to Filipino Sign Language," Philippine Deaf Resource Center, Manila, 2004.

[5] L. Martinez , "Primer on Filipino Sign Language," 1 December 2012. [Online]. Available: http://opinion.inquirer.net/41909/primer-on-filipino-sign-language. [Accessed 22 October 2014].

[6] M. Abuan, "Calls made for a national language for the deaf," The Carillon, 2009. [Online]. Available: http://archive.is/mrGi4. [Accessed 5 February 2014].

[7] N. Aoki, K. Uda, T. Kiuchi and T. Fukui, "Impact of miscommunication in medical dispute cases in Japan," *International Journal for Quality in Health Care,* vol. 20, no. 5, pp. 358-362, 2008.

[8] RID, "INTERPRETING IN HEALTH CARE SETTINGS," [Online]. Available: http://www.rid.org/UserFiles/File/pdfs /Standard_Practice_Papers/Drafts_June_2006 /Health_Care_Settings_SPP.pdf. [Accessed 20 February 2014].

[9] S. Kalla, "Statistical Variance," Explorable, 15 March 2009. [Online]. Available: https://explorable.com/statistical-variance. [Accessed 22 October 2014].

[10] S. Krishna, S. Lee, P. Wang and J. Lang, "Sign Language Traslation," 2012.

[11] "Statistics by Country for Speech impairment," [Online]. Available: http://www.rightdiagnosis.com/s /speech_impairment/stats-country.htm. [Accessed 28 January 2014].

[12] T. Zimmerman, J. Lanier, C. Blanchard, S. Bryson and Y. Harvil, "A Hand Gesture Interface Device," in *Conference on Human Factors in Computing Systems and Graphic Interface*, Redwood City, 1987.

[13] V. Aguios, C. Mariano, E. Mendoza and J. Orense, "A Portable Letter Sign Language Translator," Manila, 2007.

8. APPENDIX – WORDS TESTED

1. After

2. Allergy

3. Always

4. Before

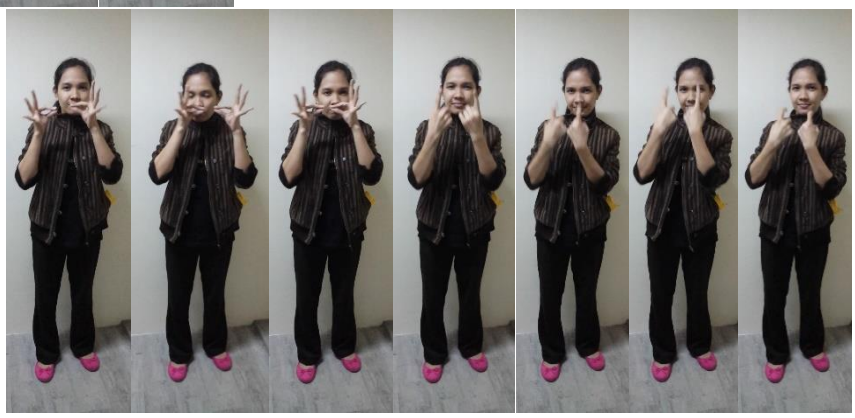
5. Blood Pressure

6. Cold

7. Cough

8. Doctor

9. Everyday



10. Drug Test