Angle-of-Arrival-Based Gesture Recognition Using Ultrasonic Multi-Frequency Signals

Hui Chen, Tarig Ballal, Mohamed Saad and Tareq Y. Al-Naffouri
Computer, Electrical and Mathematical Sciences & Engineering
King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900
Email: {hui.chen; tarig.ahmed; mohamed.saad; tareq.alnaffouri}@kaust.edu.sa

***Abstract*—Hand gestures are tools for conveying information, expressing emotion, interacting with electronic devices or even serving disabled people as a second language. A gesture can be recognized by capturing the movement of the hand, in real time, and classifying the collected data. Several commercial products such as Microsoft Kinect, Leap Motion Sensor, Synertial Gloves and HTC Vive have been released and new solutions have been proposed by researchers to handle this task. These systems are mainly based on optical measurements, inertial measurements, ultrasound signals and radio signals. This paper proposes an ultrasonic-based gesture recognition system using AOA (Angle of Arrival) information of ultrasonic signals emitted from a wearable ultrasound transducer. The 2-D angles of the moving hand are estimated using multi-frequency signals captured by a fixed receiver array. A simple redundant dictionary matching classifier is designed to recognize gestures representing the numbers from '0' to '9' and compared with a neural network classifier. Average classification accuracies of 95.5% and 94.4% are obtained, respectively, using the two classification methods.**

I. Introduction

Gestures have played an essential role in conveying information in human society for centuries. Novel gesture recognition systems are emerging as people are trying to explore new ways to interact with machines efficiently. In-air gestures, rather than setting commands on electronic devices by touching the screen or the touchpad, has their unique advantages [1]. Users can perform gestures naturally away from the screen and more complex gestures can be integrated. In general, vision, inertial sensors, radio signals and ultrasonic signals are four main types of technology to handle this task.

Vision based methods mainly rely on video streams captured by cameras and analyze the data frames. A Finger-Earth Movers Distance (FEMD) algorithm [2] combined with commercial Microsoft Kinect Sensor is proposed to perform the recognition task. As an alternative, IR cameras combined with LEDs can track the movement of the fingers and hence interpret the information, which works as an air-writing recognition system [1]. However, the camera-based techniques require high computational power and are sensitive to surrounding illumination conditions. Inertial sensors [3] provide another solution for motion tracking, but the sensors need a subsystem for transmitting the collected data to a central processor to analyze in real time, which makes the wearable system bulky. Wi-Fi signals [4] have been exploited recently by wisely leveraging the ubiquitous radio signals. Efforts have also been put in employing ultrasound signals by utilizing Doppler effect such as in Dolphin [5] and Wavesound [6] systems. It is inspiring that these systems are constructed with off-the-shelf components, but the lack of accuracy and limitation of gesture types restrict the applications of these systems.

In this paper, we present a novel system for hand gesture recognition. In the proposed system, we track the movement of a hand based on the AOA information of the received ultrasound signal. A handheld ultrasonic transmitter that can be triggered to send multi-frequency signals is adopted in the system. After detecting the signals, an ultrasonic receiver array extracts horizontal and vertical angle information to represent the real-time location of the transmitter. To classify angle observation into gestures, machine learning methods can be used. There have been a variety of machine learning approaches such as SVM (Support Vector Machine) [7], HMM (Hidden Markov Machine) [1], NN (Neural Networks) [8] to classify gestures. Other methods like FSM (Finite State Machine) and particle filters are also mentioned in [9]. However, these methods require high computational complexity and machine learning models need a large database of training data. To handle this problem in a more efficient way, we propose a redundant-dictionary-matching classifier to simplify the classification process and the results are compared with neural network based classification.

This paper is organized as follows. Section II presents the proposed gesture recognition system which includes AOA estimation, outlier rejection, redundant-dictionary based and neural-network based classification. Section III describes the conditions under which the experiments were carried out. In section IV, we analyze the performance of the system and evaluate the results. Section V concludes the whole work with suggestions of future directions.

II. The Proposed Gesture Recognition System

The proposed system is based on the following concept: The 3-D location of a target can be represented by a horizontal angle $\alpha \in [0°, 180°]$, a vertical angle $\beta \in [0°, 180°]$ and a distance r observed from the center of a receiver array. Fig.1 shows the horizontal angle α_3 and the vertical angle β_3 corresponding to hand location 3. For convenience, we will use $\theta_x = 90° - \alpha$ and $\theta_y = 90° - \beta$ to represent horizontal and vertical angles in the interval $[-90°, 90°]$. A gesture can

This work is supported by the KAUST-MIT-TUD consortium under grant OSR-2015-Sensors-2700.

ISBN 978-0-9928626-7-1 © EURASIP 2017

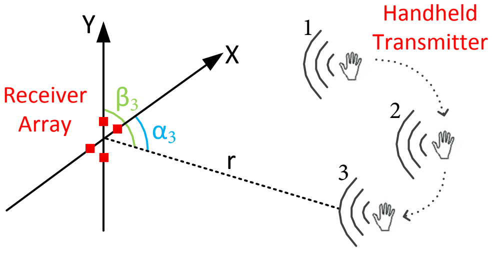


Fig. 1. Gesture Recognition Using Angle-of-Arrival

be represented as the variation of the 3-D location of the hand with time, i.e., $[\theta_x(t), \theta_y(t), r(t)]$. It is obvious that each of θ_x, θ_y and r changes in a unique way for each unique gesture. Using all the three components is expected to deliver better results compared to using only one or two components. However, since calculating the distance r requires stringent synchronization between the transmitter and the receivers, which adds to system complexity, the proposed system utilizes only 2-D AOA information (θ_x, θ_y) to detect and classify hand gestures.

The proposed system is made up of three main processes carried out in order: AOA estimation, outlier rejection and gesture classification. Each of these processes will be described in the following subsections.

A. AOA Estimation

The transmitted signal consists of multiple frequencies and the estimate of the phase difference $\hat{\psi}_{x,i} \in (-\pi, \pi]$ at the i^{th} carrier frequency f_i observed between sensor u and sensor v can be estimated as the angle of the CPS (Cross Power Spectrum) of the two signals:

$$\hat{\psi}_{x,i} = \text{ang}(Y_u(f_i) \cdot Y_v^*(f_i)) = \hat{\phi}_{x,i} - 2\pi N_{x,i}, \qquad (1)$$

where Y_u and Y_v are the DFT (Discrete Fourier Transform) of the received signals at sensor u and sensor v (respectively), $(\cdot)^*$ indicates the complex conjugate operation, $\hat{\phi}_{x,i}$ is the actual phase difference and $N_{x,i}$ is an integer. In the sequel, we will discuss the estimation of the horizontal angle. The vertical angle can be obtained similarly from a perpendicular pair of sensors.

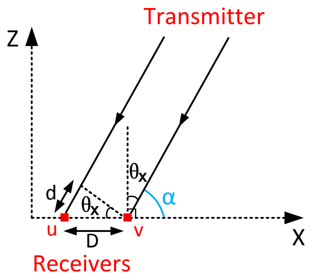


Fig. 2. Far-field Model

In a far-field scenario as shown in Fig.2, the relationship between the phase difference and the horizontal angle θ_x can be expressed as:

$$\sin(\hat{\theta}_x) = \frac{d}{D} = \frac{\hat{\phi}_{x,i}\, c}{2\pi f_i\, D}, \qquad (2)$$

where d is the range difference between the transmitter and the two receivers, c is the speed of propagation and D is the distance between the two sensors. Equation (2) can be used to calculate the AoA. However, that requires knowledge of $\hat{\phi}_{x,i}$, while we only observe $\hat{\psi}_{x,i}$ as in (1). Unless the sensor baseline D is restricted to be less than half of the wavelength of the received frequency, the integer $N_{x,i}$ is not guaranteed to be zero. Therefore, a mechanism to recover the latter integer is essential for AoA estimation using phase observations. In [10] a method was developed to recover the integer ambiguity parameters for multi-frequency signals. In this paper, we draw on the idea in [10] to develop an AoA estimator without explicitly calculating the ambiguity integers.

Using the same condition on the relationship between a frequency pair in [10], the following grid search method can be used to estimate the AoA. Namely, we search the whole range $[-90°, 90°]$ for the angle that matches the observations best. Let θ be a hypothesized emitter location. The corresponding observed phase can be calculated, based on (1) and (2), as

$$\tilde{\psi}_{x,i}(\theta) = \text{wrap}(\tilde{\phi}_{x,i}(\theta)) = \text{wrap}\left(\frac{2\pi f_i D \sin(\theta)}{c}\right), \qquad (3)$$

where $wrap$ performs the phase wrapping operation in (1). After performing (3) for all available frequencies, and over the entire range $[-90°, 90°]$ (using a suitable step), the final AoA estimate can be obtained as

$$\hat{\theta}_x = \text{arg}\ \min_{\theta} \sum_{<i>} (|\hat{\psi}_{x,i} - \tilde{\psi}_{x,i}(\theta)|), \qquad (4)$$

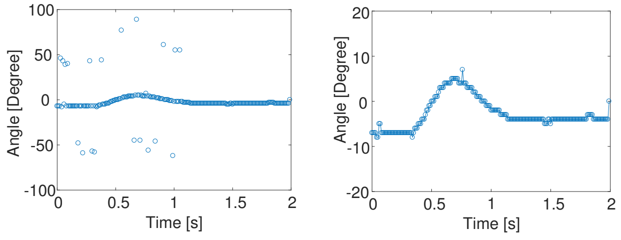
where the summation is carried over all the available frequencies (minimum two frequencies are required [10]).

B. Outlier Rejection

Due to angle estimation errors, outliers may occur in the angle information. Given that the velocity of the moving object is limited, any jumps in angle measurement between two points that exceeds a pre-defined threshold can be treated as an outlier. Here, a simple outlier rejection procedure is adopted which detects the outliers by thresholding the derivative of $\hat{\theta}_x(t)$ (same applies for $\hat{\theta}_y(t)$) and replacing the outliers with the closest neighbor values. Fig.3 shows an example of the impact of outlier rejection performed on real data.

C. Redundant-Dictionary Based Classification

1) Gesture Template: In our system, a gesture is represented by the change in horizontal and vertical angle with time. When the gesture is an alphanumeric character (as the case for all the gestures considered in this paper), for instance, the resultant pattern of angle variation depends on the specific way that gesture is performed. For example, writing a character from left to right results in a different angle sequence from



(a) Before Rejection (b) After Rejection

Fig. 3. Outlier Rejection

that result from a right-to-left writing of the same letter. In this work, we consider a set of gestures with a pre-defined movement pattern. Fig.4 shows templates of these patterns. We consider two classification approaches: a redundant-dictionary based approach and a neural-network based approach. In this subsection we will explain the first approach. The second approach will be discussed in the following section.

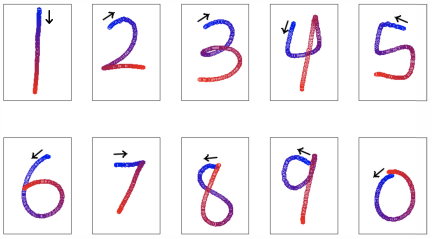


Fig. 4. Templates of the Gestures

2) Template Dictionary: As mentioned earlier, each gesture is represented by a sequence of angle measurements $[\hat{\theta}_x(t), \hat{\theta}_y(t)]$. Suppose that we have K measurements for each of the horizontal and vertical angles taken at time instants $t = 1, 2, \cdots, K$. For each gesture, an idealistic pair of sequences $[\theta_x(t), \theta_y(t)]$ can be generated assuming certain starting and end points, in addition to gesture timing, etc. A template dictionary can be created by combining the two sequences of each gesture in a column vector to form a matrix $\boldsymbol{A}_t$ with size $2K \times M$, where M is the number of different gestures. The inner-products of the dictionary columns can be obtained as

$$\boldsymbol{B} = \boldsymbol{A}_t^T \boldsymbol{A}_t, \tag{5}$$

where $\boldsymbol{B}$ is a $M \times M$ matrix. In the proposed system, we consider 10 gestures, i.e., $M = 10$, and each angle sequence consists of $K = 200$ measurements (given that angles are measured each 10 ms, this is equivalent to a total of 2 sec). The inner-product matrix indicates how easily or difficultly a gesture is distinguishable from the others. Fig.5 depicts the normalized inner products of the columns of a template dictionary for the 10 gestures. We can see from Fig.5 that some gesture pairs exhibit relatively high cross correlation, and hence it is likely that these gestures be confused with each other.

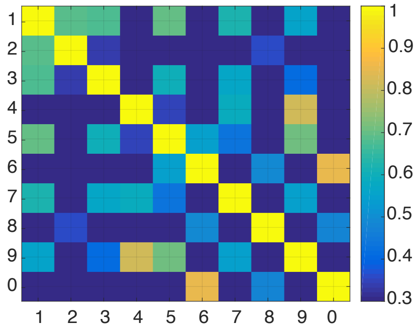


Fig. 5. Inner-product Matrix

3) Redundant Dictionary: Even when the users follow the same movement pattern, the extension of the gesture in both time and space differs from one user to another. To increase the robustness of the recognition method, we generate different multiple templates for each gesture and add them to extend the template dictionary dictionary. These added templates represent variations of the original template with different movement speed, timing, starting and ending points. The goal here is to make the dictionary redundant to account for some of the inherent uncertainty in the gestures. Fig.6 shows an example of the redundant dictionary templates. The original template is a concatenation of idealistic horizontal and vertical angle sequences. The extended template represent similar information, but with a shorter time duration and the timing is generally different. Namely, the gesture is delayed by 20 samples and is 80 samples shorter in duration. Other extended templates may have even shorter duration but exhibits a similar angle variation pattern from the point the movement starts to the point where it stops. In this paper, we delay and compress each template by multiples of 20 samples to extend the dictionary into a 400×210 redundant dictionary matrix $\boldsymbol{A}_r$. Further, we adjust each column of the dictionary to have zero mean and unit second norm.

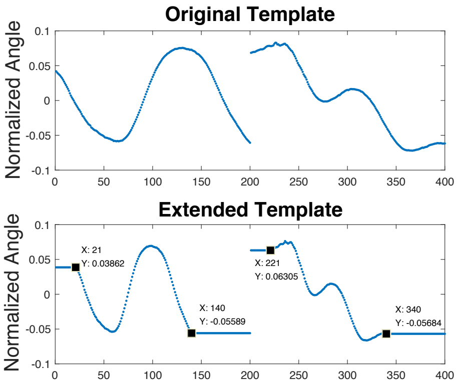
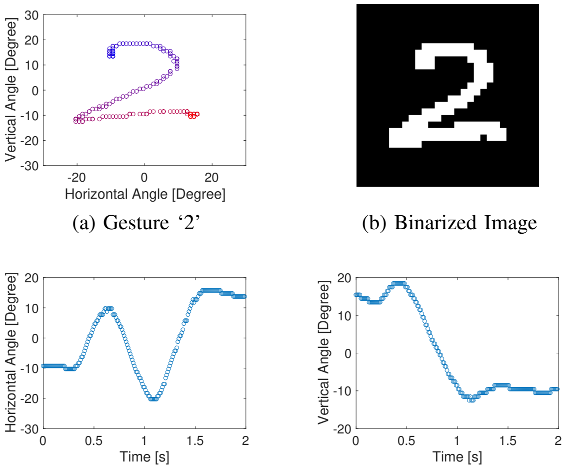


Fig. 6. One Example of the Redundant Dictionary

4) Classification: Usually, one gesture lasts 1-2 seconds, and thus, each gesture can be represented using at most 200 horizontal data and 200 vertical data. In cases where the signal is received for less than 2 seconds, zeros are padded to generate a 400×1 vector $\boldsymbol{g}$ of the concatenated angle observations. To carry out the classification task, a simple matrix-vector multiplication operation is performed. Namely,

$$\boldsymbol{r} = \boldsymbol{A}_r^T \boldsymbol{g}. \qquad (6)$$

Then the location of the peak of $\boldsymbol{r}$ is used as an indicator of the gesture type. In other words, the gesture with a template that has the maximum inner product with the angle observation vector is declared as the detected gesture.



(a) Gesture ‘2’

(b) Binarized Image

(c) Horizontal Angle

(d) Vertical Angle

Fig. 7. Gesture Reconstruction

D. Neural-Network Based Classification

1) Training Data: Three models M_{angle}, M_{std} and M_{cons} are built based on different training data sets. The model M_{angle} is trained using normalized angle data to compare with the dictionary-based classifier from the previous subsection, whereas the other two models are trained using images. The model M_{std} is trained using MNIST [11] database, while M_{cons} is trained using reconstructed gesture images. A stacked autoencoder model from Matlab Neural Network Toolbox [12] with two hidden layers, as well as a softmax layer, is implemented and used with each of the three models.

2) Image Reconstruction: A gesture images is reconstructed placing the horizontal and vertical angle measurements on the two axes of a 2-D coordinate system and marking the points in the 2-D plain where a pair of horizontal and vertical angles occurs. An example is shown in Fig.7 (a) where the gesture starts from blue to red. A 28×28 pixels binary version of the image is shown in Fig.7 (b), while AoA measurements are plotted in Fig.7 (c) and Fig.7 (d). Fig.7 not only validates the concept of AOA-based gesture recognition but also suggests using the reconstructed images in the gesture classification process.

Classification results obtained using the three models from this subsection, and using the dictionary-based approach from the previous subsection, will be presented in the following section.

III. Experiments

When the transmitter is triggered, a series of pulses will be sent from the ultrasound transducer. Each pulse lasts for 1.5 ms and is repeated 100 times per second. The transmitted signal consists of 3 Hanning-windowed sinusoidals of frequencies of $20kHz$, $21.5kHz$ and $23kHz$. A receiver array of four elements arranged as two orthogonal pairs collect the signals at a sampling rate of 192 kHz. Fig.8 shows an example of the received signal.

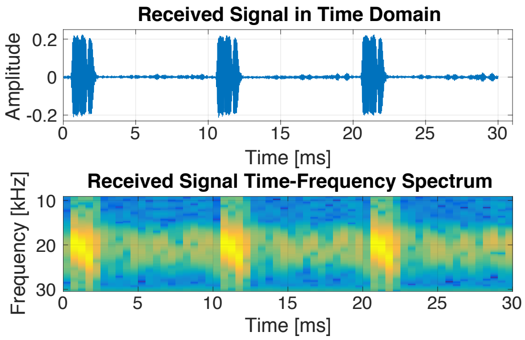


Fig. 8. An Example of Received Signal

To evaluate the performance of the proposed system, 10 volunteers were asked to perform gestures according to the following instructions:

- Try to write each number based on the template;
- The duration of each gesture should be between 1 and 2 seconds;
- Sit around 1 metre in front of the receiver array;
- The movement of the hands should be within a square of 80 cm by 80 cm centered around the receiver array;
- Repeat each gesture 10 times with a break after each gesture.

All the experiments were carried out in a typical office room with a temperature around $24°C$ and a set of total 1000 gestures was acquired.

After removing outliers, each gesture is converted to a 28×28 binary image. From the 100 gestures of each volunteer, 50 were picked up for the training set and the remaining 50 were left for testing. The model M_{std} was trained using binary MNIST data with a threshold of 0.15 while M_{angle}, M_{cons} were trained with the same gesture set but with two different data format (i.e., angle measurements and 28×28 images). To make a fair comparison, the three neural network models used the same testing data.

IV. Results and Discussions

A. AOA Estimation

Before dealing with gestures, tests were conducted to evaluate AOA estimation accuracy. Since the array is symmetric, it suffices to present results for the horizontal angle. The transmitter was placed 1.5 meters away from the receiver array and 7 angles from 0° to 75° with a step of 15° were tested by changing the location of the transmitter. Each angle was measured 200 times and the results are summarized in Table I.

ISBN 978-0-9928626-7-1 © EURASIP 2017

TABLE I
ANGLE OF ARRIVAL ACCURACY TEST

	0°	15°	30°	45°	60°	75°
RMSE	1.966	1.925	1.691	2.520	0.604	1.102
Bias	-1.955	-1.895	-1.620	-2.470	-0.145	0.155
STD	0.207	0.339	0.487	0.500	0.588	1.094

B. Classification Results

One typical classification result of number ‘8’ using the redundant-dictionary approach is shown in Fig.9, which plots the correlation $\boldsymbol{r} = \boldsymbol{A}_r^T \boldsymbol{g}$. The large values in the interval $(148, 168)$ indicate that this gesture representation $\boldsymbol{g}$ is more correlated with the templates representing the gesture ‘8’, which is the true gesture.

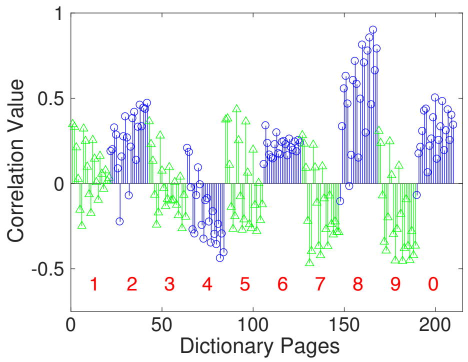


Fig. 9. An Example of Dictionary Matching for Gesture ‘8’

As an example, the confusion matrix of the redundant dictionary classifier, which gives an overall accuracy of 95.5%, is shown in Table II. It can be seen that the highest error rate occurs with the number ‘4’. A summary of the results for the 4 classifiers that were tested are shown in Table III. From the table, we conclude that the difference between gestures and hand-written digit database causes unfavorable performance in the M_{std} model, which uses standard images of the numbers 0–9 for training.

TABLE II
CONFUSION MATRIX OF DICTIONARY-BASED CLASSIFIER

Actual	Classified Gestures (95.5% of 1000 gestures)									
Gesture	‘1’	‘2’	‘3’	‘4’	‘5’	‘6’	‘7’	‘8’	‘9’	‘0’
‘1’	96						3	1		
‘2’	4	91					5			
‘3’	5	1	91				3			
‘4’				81		5			9	5
‘5’	3				95			1	1	
‘6’						97				3
‘7’	9						91			
‘8’							1	99		
‘9’					1		1		97	1
‘0’						1				99

TABLE III
NEURAL NETWORK CLASSIFICATION RESULTS

Classifier	Dictionary	M_{angle}	M_{std}	M_{cons}
Accuracy	95.5%	94.4%	66.5%	91.1%

V. CONCLUSION

This work is an attempt to detect gestures using AOA information (of a handheld ultrasonic transmitter) collected via a receiver array. In order to track the gesture route, 2-D angle of arrival estimation is carried out. An outlier rejection algorithm is used to smooth the data, which is then classified by a redundant dictionary with an accuracy of 95.5%. This dictionary-based classifier requires less computational resources and no training but still shows comparable performance to that of a neural network classifier.

Occasional erroneous classification of some gestures such as the confusion between ‘4’ and ‘9’, and between ‘1’ and ‘7’ requires further efforts to improve the templates and refine the classification part. And the search-based method for AOA estimation can also be revisited by judiciously confining the search to the appropriate angle interval based on prior information about the gesture and/or by leveraging the previous angle measurements.

For future work, this system can be extended to include more complex gestures such as letters to fulfill the needs of on-air writing. Also, image-based classification and machine learning methods can be enhanced and evaluated thoroughly. Furthermore, range information might also be utilized to provide depth information and improve the identifiability of different gestures.

REFERENCES

[1] M. Chen, G. AlRegib, and B.-H. Juang, “Air-writing recognition-part i: Modeling and recognition of characters, words, and connecting motions,” *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 3, pp. 403–413, 2016.

[2] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, “Robust part-based hand gesture recognition using kinect sensor,” *IEEE transactions on multimedia*, vol. 15, no. 5, pp. 1110–1120, 2013.

[3] C. Amma, M. Georgi, and T. Schultz, “Airwriting: a wearable handwriting recognition system,” *Personal and ubiquitous computing*, vol. 18, no. 1, pp. 191–203, 2014.

[4] B. Kellogg, V. Talla, and S. Gollakota, “Bringing gesture recognition to all devices.” in *NSDI*, vol. 14, 2014, pp. 303–316.

[5] Y. Qifan, T. Hao, Z. Xuebing, L. Yin, and Z. Sanfeng, “Dolphin: Ultrasonic-based gesture recognition on smartphone platform,” in *Computational Science and Engineering (CSE), 2014 IEEE 17th International Conference on*. IEEE, 2014, pp. 1461–1468.

[6] S. Gupta, D. Morris, S. Patel, and D. Tan, “Soundwave: using the doppler effect to sense gestures,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1911–1914.

[7] K. K. Biswas and S. K. Basu, “Gesture recognition using microsoft kinect®,” in *Automation, Robotics and Applications (ICARA), 2011 5th International Conference on*. IEEE, 2011, pp. 100–103.

[8] E. Stergiopoulou and N. Papamarkos, “Hand gesture recognition using a neural network shape fitting technique,” *Engineering Applications of Artificial Intelligence*, vol. 22, no. 8, pp. 1141–1158, 2009.

[9] S. Mitra and T. Acharya, “Gesture recognition: A survey,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, 2007.

[10] T. Ballal and C. J. Bleakley, “Doa estimation for a multi-frequency signal using widely-spaced sensors,” in *Signal Processing Conference, 2010 18th European*. IEEE, 2010, pp. 691–695.

[11] Y. LeCun, C. Cortes, and C. J. Burges, “The mnist database of handwritten digits,” 1998.

[12] H. Demuth and M. Beale, “Matlab neural network toolbox users guide version 6. the mathworks inc,” 2009.