Research Article
Journal of Research and Review in Science, 58-65
Volume 5, December 2018
DOI: 10.36108/jrrslasu/8102/50(0170)

ORIGINAL RESEARCH

CONVERSION OF SIGN LANGUAGE TO TEXT AND SPEECH USING MACHINE LEARNING TECHNIQUES

Victoria A. Adewale [1], Dr. Adejoke O. Olamiti [2]

[1]*Crawford University, Faith-City, Igbesa, Ogun State, Nigeria,* [2] *University of Ibadan, Ibadan, Nigeria*

[2]*Department of Neuroimaging Sciences, Center for Clinical Brain Sciences, University of Edinburgh, Edinburgh UK*

Correspondence
Victoria A. Adewale, *Crawford University, Faith-City, Igbesa, Ogun State, Nigeria.*

Email: bimpsyade@gmail.com

Abstract:

Introduction: Communication with the hearing impaired (deaf/mute) people is a great challenge in our society today; this can be attributed to the fact that their means of communication (Sign Language or hand gestures at a local level) requires an interpreter at every instance. Conversion of images to text as well as speech can be of great benefit to the non-hearing impaired and hearing impaired people (the deaf/mute) from circadian interaction with images. To effectively achieve this, a sign language (ASL – American Sign Language) image to text as well as speech conversion was aimed at in this research.

Aims: To convert ASL signed hand gestures into text as well as speech using unsupervised feature learning to eliminate communication barrier with the hearing impaired and as well provide teaching aid for sign language.

Materials and Method: The techniques of image segmentation and feature detection played a crucial role in implementing this system. We formulate the interaction between image segmentation and object recognition in the framework of FAST and SURF algorithms. The system goes through various phases such as data capturing using KINECT sensor, image segmentation, feature detection and extraction from ROI, supervised and unsupervised classification of images with K-Nearest Neighbour (KNN)-algorithms and text-to-speech (TTS) conversion. The combination FAST and SURF with a KNN of 10 also showed that unsupervised learning classification could determine the best matched feature from the existing database. In turn, the best match was converted to text as well as speech.

Results: The introduced system achieved a 78% accuracy of unsupervised feature learning.

Conclusion: The success of this work can be attributed to the effective classification that has improved the unsupervised feature learning of different images. The pre-determination of the ROI of each image using SURF and FAST, has demonstrated the ability of the proposed algorithm to limit image modelling to relevant region within the image.

To Keywords: Image and Speech processing; Text-to-Speech (TTS); Unsupervised Learning; FAST and SURF algorithms.

All co-authors agreed to have their names listed as authors.

1. INTRODUCTION

Communication has been defined as an act of conveying intended meanings from one entity or group to another through the use of mutually understood signs and semiotic rules. It plays a vital role in the existence and continuity of human. For an individual to progress in life and coexist with other individuals there is the need for effective communication. Effective communication is an essential skill that enables us to understand and connect with people around us. It allows us to build respect and trust, resolve differences and maintain sustainable development in our environment where problem solving, caring and creative ideas can thrive. Poor communication skills are the largest contributor to conflict in relationships. The indicators of poor communication include inattentiveness, arguments, vilification, and language barrier between the communicators. All of these factors do not only affect the physically fit people but also the physically challenged. Research has shown that over nine (9) billion people at intervals, all over the world are physically challenged in terms of communication; blind, deaf or mute [1]. Investigating the barrier of communication between the hearing-impaired and the hearing person has led to the need of providing a means of bridging this communication gap.

Extant literatures capture some of the dynamics of solving the problems facing effective communication, although not without observed missing links.

Academic and industrial researchers have recently been focusing on analyzing images of people and there has been a surge interest in recognizing human gestures. A research on scene segmentation of images was carried out using deep learning techniques; the classification yielded 53.8% accuracy [2]. In relation to conversion of sign language, there is the need to explore other image classification techniques to enhance accurate classification.

Also, a novel method for unsupervised learning of human action categories was presented by Juan Carlos to automatically learn the probability distributions of the spatial-temporal words and the intermediate topics corresponding to human action categories. This was achieved by using latent topic models such as the probabilistic Latent Semantic Analysis (pLSA) model and Latent Dirichlet Allocation (LDA)[3]. This study was aimed at recognizing general human actions which can be said to be ambiguous; a more specific human action identification approach using unsupervised learning will yield a better result.

Furthermore, an approach to convert signed ASL alphabets using unsupervised learning feature (Gaussian model) has also been used to learn set of features similar to edges using an autoencoder; a softmax classifier was used to classify the learned features. Results showed that the higher the training set the higher the level of accuracy-(1200 training set produced 95.62% accuracy and 6000 training set resulted in 98.20%) and the disparity in the training and test error [4]. Further studies on how to capture and convert ASL words would be a plus to this study.

2. METHODOLOGY

The aim of the study as earlier stated was to provide an unsupervised learning feature of signed hand gestures while the system returns corresponding output as text and speech. The following were the measurable methods employed in actualising the aim:

1. Segmentation of captured signed gestures of ASL as inputs
2. Feature extraction of the segmented images
3. UFL and classification of several images
4. *Text* and *Speech* synthesis of classified images

Figure 1 gives an overview of the system

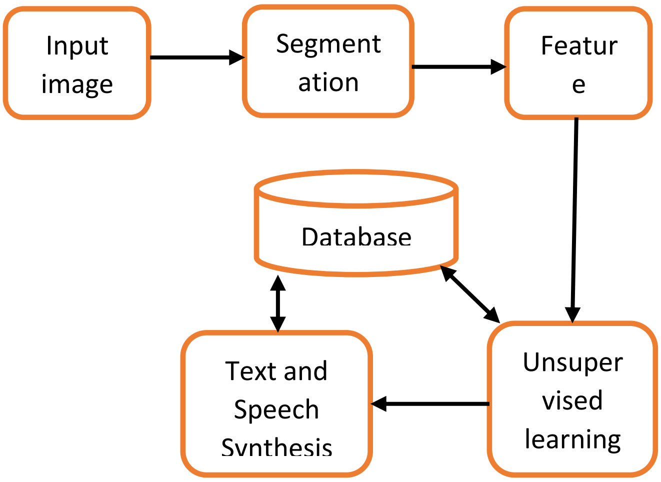


Fig 1: System Overview

2.1 Segmentation of captured signed gestures of ASL as inputs

The aim of segmentation was to convert images into more meaningful and easy to analyse portions. Segmentation does the job of partitioning an image into multiple segments which help to locate the objects and boundaries (curves, arcs, lines, etc.) in an image in binary form. The set of images captured from the Kinect sensor using the Image Acquisition Tool in MATLAB would be selected and fed into the Image Segmenter in MATLAB which is then converted to grayscale image. The threshold of the images is then obtained by converting grayscale images into binary images to determine the high level contrast of the images. Such images can then be cropped or resized. The segmentation process is represented as shown in fig. 2.

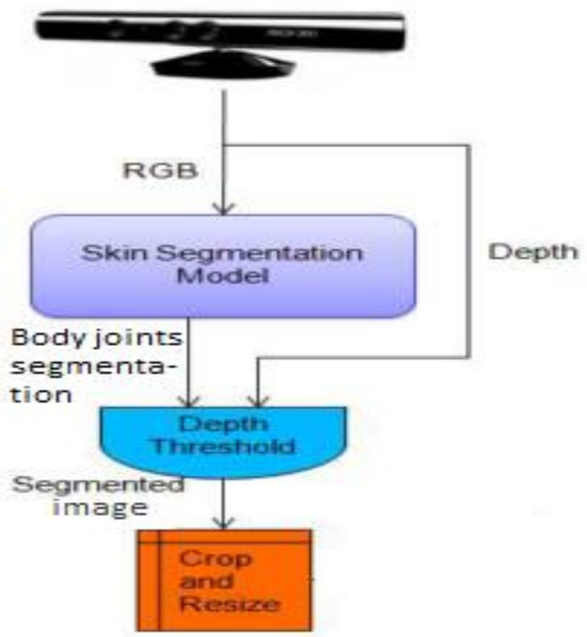


Fig 2: Image Segmentation

2.2 Feature extraction of the segmented images

Sign Language usually involves movement in the upper part of the body; head, shoulder, hands and elbows coordinates are retained while other parts are discarded [5]. To satisfy this need, key points corresponding to high-contrast locations such as object edges and corners were used. These features are intended to be non-redundant, informative and relevant for the intended use. Extracting ROI from images has been very much challenging as it is the base for further image analysis, interpretation and classification. A rectangular ROI whose outline consists of four segments joining the four corner points is used to make computational statistics feasible [6].

The vertices of an ROI outline may be positioned anywhere with respect to the array of image pixels, so the same Rectangular ROI superimposed on the pixel array may appear.

2.3 UFL and classification of several images

The identification of interest points present within the space of an image is important in the determination of the image's ROI, therefore the method being proposed in this paper maximizes the number of interest points detected within a sample image through the use of the combination of FAST corner detector and SURF detector.

2.3.1 Fast and Surf Points For K-Nearest Neighbour UFL

If FAST corner points and SURF key points are respectively represented by the sets F= {f_1,f_2,f_3,…,f_L } and S= {s_1,s_2,s_3,…,s_L}, then the combination of these two algorithms can be represented by a set A (i.e. $A = F \cup S$). The two key criteria which distinguish keypoints belonging to an ROI from those that do not belong to the desired region are location and description. These combination take its root from K-Nearest Neighbour (KNN) algorithm in equation (1) used by [7] for classification of description of each point into either foreground or background therefore, it requires training samples.

$$\frac{\sum_{j=1}^{L}|f_i - f_j|}{n(F)} < \frac{\sum_{j=1}^{L}|s_i - s_j|}{n(S)}$$

equation (1)

A new set of extracted data will be fed into the system for training in order to learn the set of unsupervised features. In the implementation of the KNN, each feature point to be categorised is allocated the highest occurring label from the closest 10 neighbours (Medium KNN), thus the points labeled as the foreground are grouped together to form the desired Region.

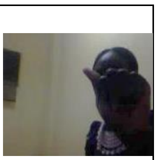
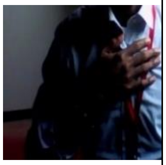
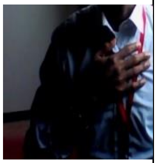
2.4 Text and Speech synthesis of classified images

Text-to-Speech (TTS) refers to the ability of computers to read text aloud. A TTS Engine converts written text to a phonemic representation, and then converts the phonemic representation to waveforms that can be output as sound. Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware. After the successful classification of these features, the important task is to generate appropriate text and speech output for every input image using MATLAB Speech Synthesizer.

3. RESULTS AND DISCUSSION

Sample images of different ASL signs were collected using the Kinect sensor using the image acquisition toolbox on MATLAB. About five hundred (500) data samples (with each sign count five and ten (5-10)) were collected as the training data. The reason for this is to make the algorithm very robust for images of the same database in order to reduce the rate of misclassification. Examples of the images collected is shown in Table 1

Table 1: Coloured images for training

			
Imageset A1	ImagesetA2	ImagesetA3	ImagesetA4
			
Imageset B1	ImagesetB2	ImagesetB3	ImagesetB4

3.1 SEGMENTATION OF IMAGES

Batch segmentation for all training samples was carried out to convert the coloured images into binary form with MATLAB Image Segmenter and Batch Processor toolbox. Basically only the set of binarised images are useful in feature detection and extraction. The segmented form of the images in Table 1 is represented in Table 2.

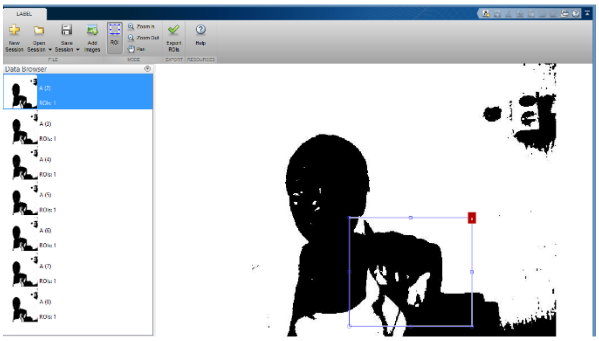
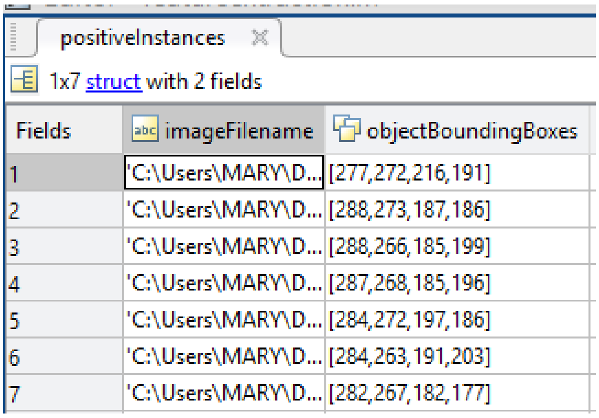


Fig 3: ROI Labelling of ASL Images



positiveInstances

1x7 struct with 2 fields

Fields	imageFilename	objectBoundingBoxes
1	'C:\Users\MARY\D...	[277,272,216,191]
2	'C:\Users\MARY\D...	[288,273,187,186]
3	'C:\Users\MARY\D...	[288,266,185,199]
4	'C:\Users\MARY\D...	[287,268,185,196]
5	'C:\Users\MARY\D...	[284,272,197,186]
6	'C:\Users\MARY\D...	[284,263,191,203]
7	'C:\Users\MARY\D...	[282,267,182,177]

Fig 4: Object Bounding Box of Labelled ASL images

The bounding boxes in Figure 4 for each image alongside its path name is generated and stored in a .mat file for further processing.

Table 2: Segmented ImageSet of Coloured Images

 SegmentedImageA1	 SegmentedImageA2	 SegmentedImageA3	 SegmentedImageA4
 SegmentedImageB1	 SegmentedImageB2	 SegmentedImageB3	 SegmentedImageB4

3.2 FEATURE EXTRACTION

Once segmentation process has been successfully carried out, the next thing is to load the image database. A for loop is used to read an entire folder of images and store them in MATLAB's memory for labeling. The image training labeler function of MATLAB in Figure 3 is employed to do this and then ROI generated by equation 1 was used for selection of the training set.

3.3 UNSUPERVISED CLASSIFICATION OF ASL IMAGES

In the implementation of Fast and Surf points for KNN mentioned in equation (1), each feature point to be classified is allocated the highest occurring label from the closest 10 neighbours, thus the points labeled as the foreground are grouped together to form the desired region.

3.3.1 STAGES FOR UFL AND CLASSIFICATION

1. PREPARE COLLECTION OF IMAGES TO SEARCH

Read the set of reference images each containing a different object. Multiple views of the same object are included in the collection shown in Figure 5 in order to capture hidden or occluded areas.

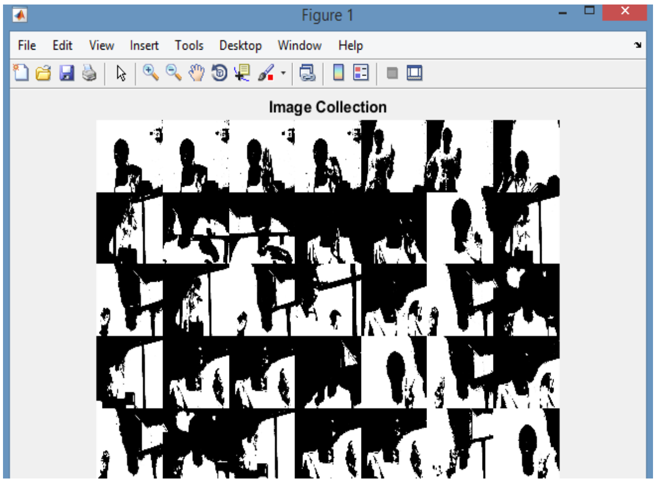


Fig 5: Image Collection of different ASL

2. DETECT FEATURE POINTS IN IMAGE COLLECTION

Detect and display feature points in first image as shown in Figure 6. Use of local features serves two purposes. It makes the search process more robust to changes in scale and orientation and reduces the amount of data that needs to be stored and analysed. Then Detect features in the entire image collection.

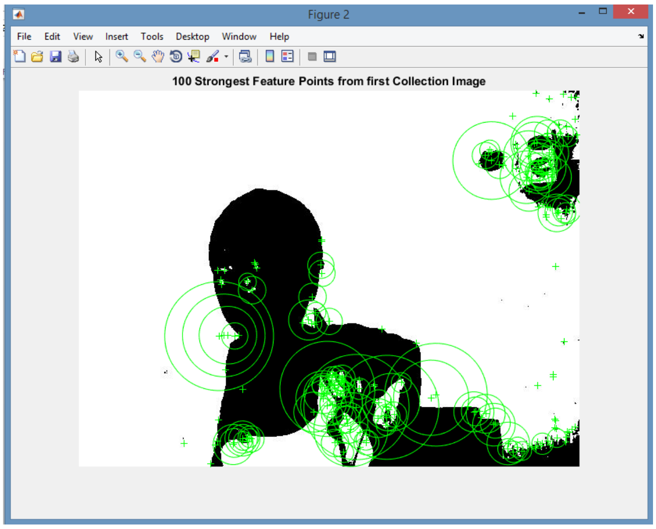


Figure 2
File Edit View Insert Tools Desktop Window Help
100 Strongest Feature Points from first Collection Image

Fig 6: First Image in Collection

3. BUILD FEATURE DATASET

All of the features from each image are combined into a matrix. The matrix was then used to initialize a KDTreeSearcher object from the Statistics Toolbox. This object allows for fast searching for nearest neighbours of high-dimensional data. In this case, a nearest neighbour of FAST and SURF descriptor are used as view of the same point.

4. CHOOSE QUERY IMAGE

An entirely new set of images as shown in Figure 7 outside the trained images are supplied. In other words, it is an imageset that is not part of the training set.

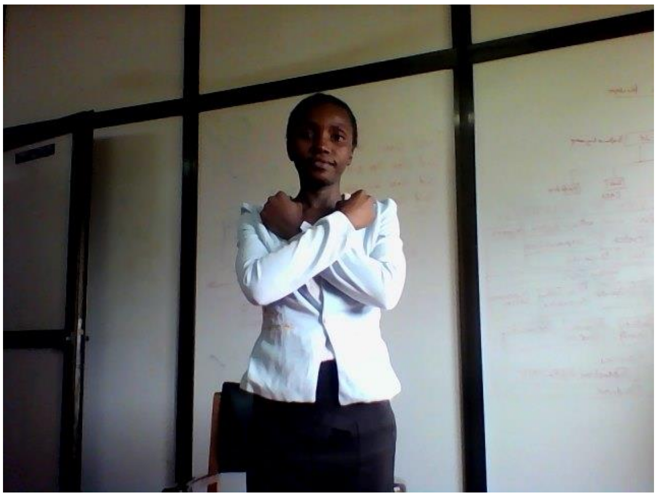


Fig 7: New Image to be classified

5. DETECT FEATURE POINTS IN QUERY IMAGE

The query image is converted into grayscale and threshold to obtain a segmented image; then the ROI of the image is captured before the features are extracted using fast corner points and surf keypoints for effective and efficient recognition. The features detected from Figure 8a are represented in Figure 8b.

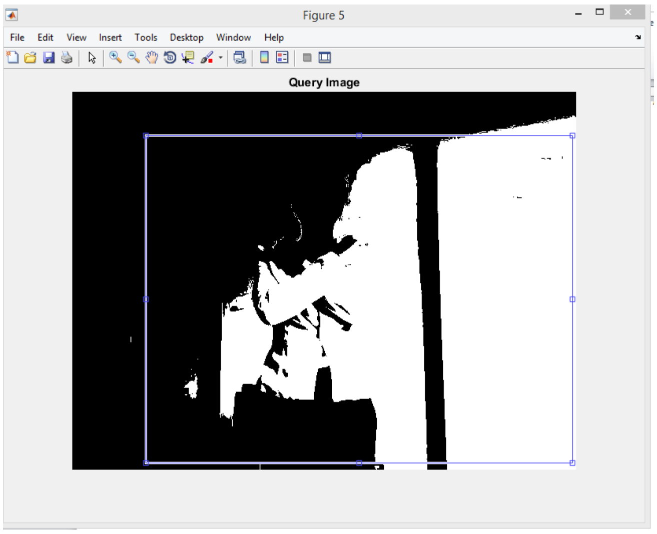


Fig 8a: ROI Selection

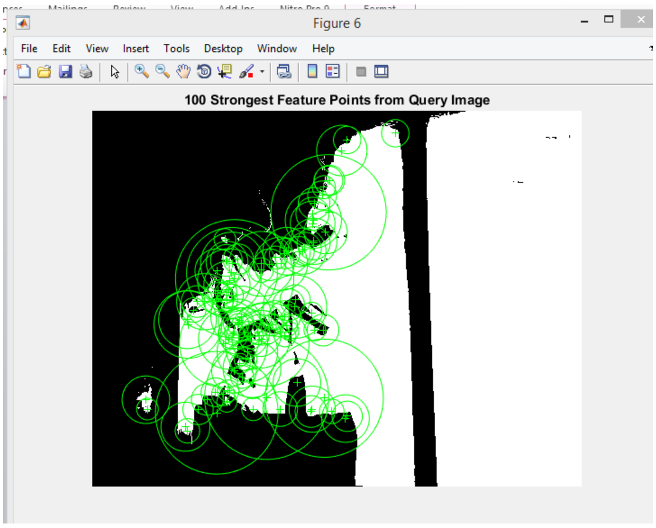


Figure 6
File Edit View Insert Tools Desktop Window Help
100 Strongest Feature Points from Query Image

Fig 8b: Detected Features

6. SEARCH IMAGE COLLECTION FOR THE QUERY IMAGE

For all of the features in the query image, ten nearest neighbours in the dataset were considered to compute the distance to each neighbour. The KNN-search function returns the nearest neighbours, even if none of the features are a close match. To throw away those bad matches, we will use a ratio of the ten closest neighbour distances. The histogram function was used to count the number of features that matched from each image. Each pair of indices, in the indexIntervals of Figure 9 constitutes an index interval that corresponds to an image.

The strength with which each image in the collection matches the query image can be viewed in image collection shown in Figure 10. Size of each image in Figure 11 is proportional to the proximity of matching features. It was observed that some other images are still considered as either a strong or weak match. These are outliers that will be eliminated in the next step.

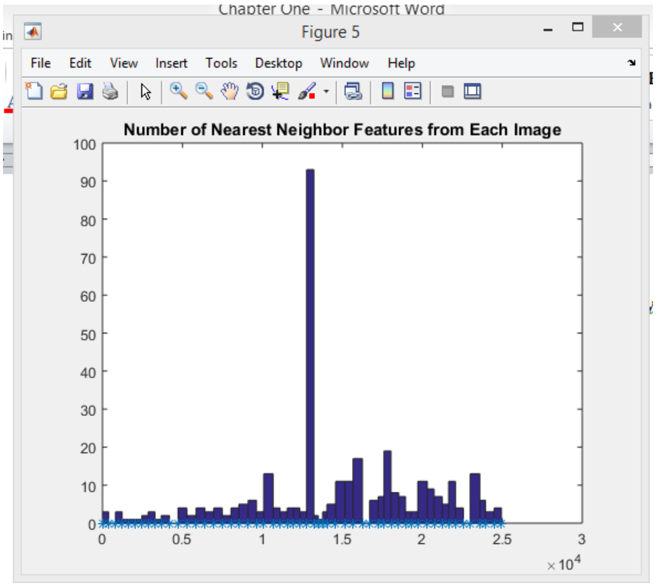


Fig 9: Histogram of matched images

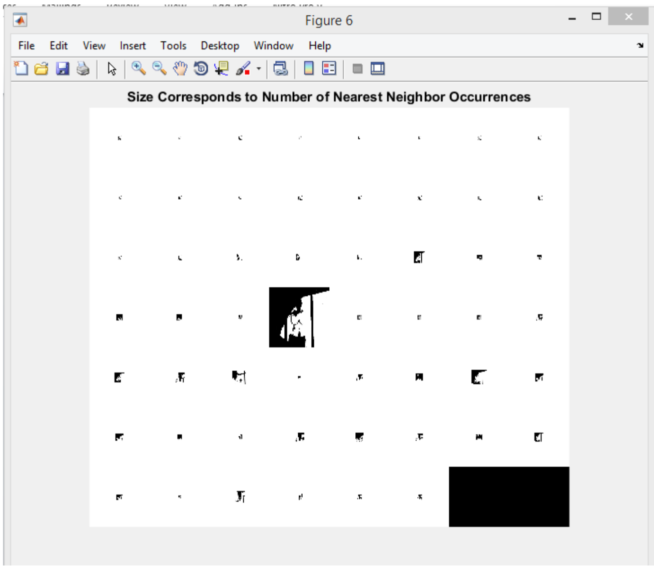


Figure 6
File Edit View Insert Tools Desktop Window Help
Size Corresponds to Number of Nearest Neighbor Occurrences

Fig 10: Set of Matched Images

7. ELIMINATE OUTLIERS USING DISTANCE TESTS

To prevent false matches, it is important to remove those nearest neighbour matches that are far from their query feature. The poorly matched features can be detected by comparing the distances of the first and second nearest neighbour. If the distances are similar, as calculated by their ratio, the match is rejected as shown in Figure 11. Additionally, matches that are far apart were ignored. These processes were repeated for other new set of images for unsupervised learning and classification.

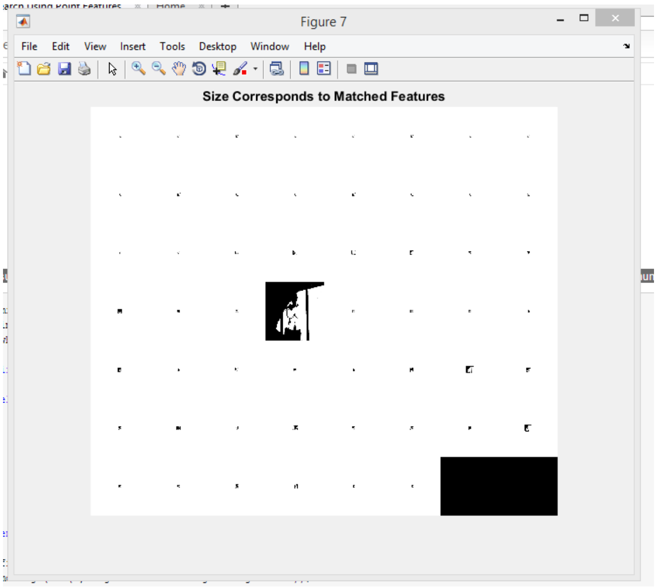
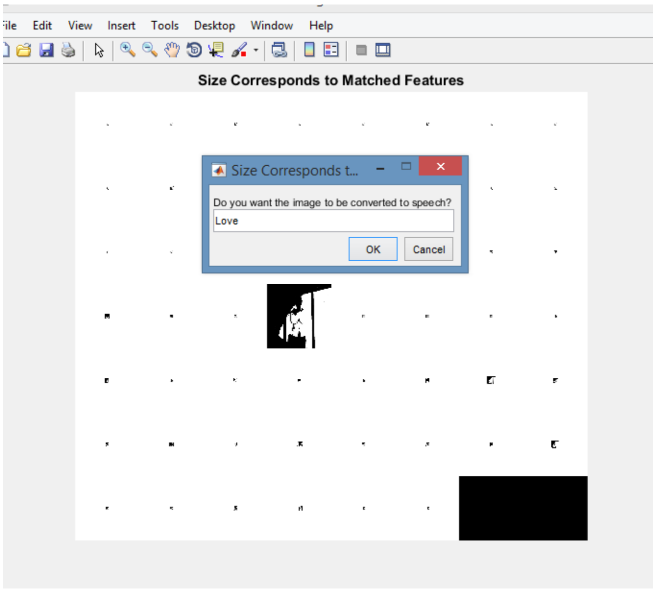


Figure 7
File Edit View Insert Tools Desktop Window Help
Size Corresponds to Matched Features

Fig 11: Best Matched Feature

3.4 TEXT-TO-SPEECH SYNTHESIS

The Text-To-Speech Synthesizer function of MATLAB was used to convert the string of the filename of the best matched feature in the collection to speech in Figure 12.


File Edit View Insert Tools Desktop Window Help
Size Corresponds to Matched Features
Size Corresponds t...
OK
Cancel

Do you want the image to be converted to speech?

Love

Fig 12: Text-To-Speech

3.4.1 RESULT OF SUPERVISED AND UNSUPERVISED CLASSIFICATION

Table 3 shows the result of correctly classified images:

Table 3: Result of Classification

Image Samples	Number of image Samples per sign	Supervised Feature Learning (Classification)	Unsupervised Feature Learning (Classification)
A	10	**1**	1
B	10	**1**	1
C	10	**1**	1
D	9	**1**	0
E	10	**1**	0
F	10	**1**	1
G	10	**1**	1
H	9	**1**	1
I	10	**1**	1
K	9	**1**	1
L	10	**1**	1
M	10	**0**	0
N	10	**1**	0
O	10	**1**	1
P	10	**1**	1
Q	10	**1**	0
R	10	**1**	1
S	10	**1**	1
T	10	**1**	1
U	10	**1**	1
V	7	**1**	1
W	8	**1**	1
X	10	**1**	0
Y	10	**1**	0
Love	10	**1**	1
Master	9	**0**	1
Father	5	**1**	1
Mother	9	**1**	1
You	9	**1**	1
Me	8	**1**	1
Your	9	**1**	1
Start	10	**1**	1
End	10	**1**	1
Man	10	**1**	1
Come	10	**1**	0
To	10	**1**	1
Meat	10	**1**	1
Want	9	**1**	1
Church	10	**1**	1
Name	8	**1**	1
God	8	**1**	1
Food/Eat	10	**1**	1
What	10	**1**	1
Are	10	**1**	1
My/Mine	10	**1**	1
Water	10	**1**	1
House	10	**1**	0
Have	10	**1**	1

The final result in Figure 13 shows 92% correct classification using supervised learning and 78% correct classification using unsupervised learning.

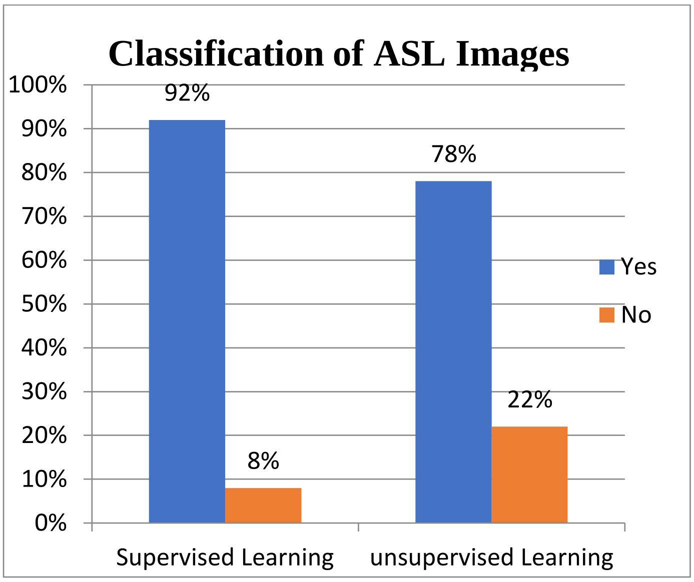


Fig 13: Classification of ASL Images

3.5 DISCUSSION

The study findings show that American Sign Language (ASL) is commonly used in Nigeria by the hearing impaired hence; five hundred (500) ASL images were collected as training set. From the collection of images, a database of forty-nine (49) different signs was used. Having subjected the set of images to batch segmentation, features of each signs were detected and extracted from specific bounding-box of Region of Interest (ROI) to aid supervised learning. The combination FAST and SURF with a KNN of 10 also showed that unsupervised learning classification could determine the best matched feature from the existing database. In turn, the best match was converted to text as well as speech. The introduced system achieved a 92% accuracy of supervised feature learning and 78% of unsupervised feature learning.

REFERENCES

[1] V. Padmanabhan and M. Sornalatha, “Hand gesture recognition and voice conversion system for dumb people,” *Int. J. Sci. Eng. Res.*, vol. 5, no. 5, 2014.

[2] C. Chen, J. Chen, and A. Ryan, “Scene Segmentation of 3D Kinect Images with Recursive Neural Networks,” 2011. [Online]. Available: http://cs.nyu.edu/. [Accessed: 14-Mar-2017].

[3] J. C. Niebles, H. Wang, L. Fei-Fei, J. C. Niebles, H. Wang, and L. Fei-Fei, “Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words,” *Int J Comput Vis*, 2008.

[4] P. A. Ajavon, “An Overview of Deaf Education in Nigeria,” vol. 109, no. 1, pp. 5–10, 2006.

[5] D. Mart, “Sign Language Translator using Microsoft Kinect XBOX 360 TM.”

[6] Xinapse, “Region of Interest (ROI) Algorithms,” 2018. .

[7] A. Li, W. Jiang, W. Yuan, D. Dai, S. Zhang, and Z. Wei, “An Improved FAST + SURF Fast Matching Algorithm,” *Procedia - Procedia Comput. Sci.*, vol. 107, no. Icict, pp. 306–312, 2017.