Robust Visual Lips Feature Extraction Method for Improved Visual Speech Recognition System

Mahmuod.H. Mahmmed
Dept. of Electrical Engineering
University of Technology,
Baghdad, Iraq.

Thamir.R. Saeed
Asst. Prof
Dept. of Electrical Engineering
University of Technology
Baghdad, Iraq.

Wissam. H. Ali
Dept. of Electrical Engineering
University of Technology,
Baghdad, Iraq.
wisamas1976@yahoo.com

Received on: 07/06/2016
Accepted on: 29/09/2016

Abstract-*Recently, automatic lips reading ALR acquired a significant interest among many researchers due to its adoption in many applications. One such application is in speech recognition system in noisy environment, where visual cue that contain some integral information added to the audio signal, as well as the way that person merges audio-visual stimulus to identify utterance**.** The unsolved part of this problem is the utterance classification using only the visual cues without the availability of acoustic signal of the talker's speech. By taking into considerations a set of frames from recorded video for a person uttering a word; a robust image processing technique is used to isolate the lips region, then suitable features are extracted that represent the mouth shape variation during speech. These features are used by the classification stage to identify the uttered word. This paper is solve this problem by introducing a new segmentation technique to isolate the lips region together with a set of visual features base on the extracted lips boundary which able to perform lips reading with significant result. A special laboratory is designed to collect the utterance of twenty six English letters from a multiple speakers which are adopted in this paper (UOTEletters corpus). Moreover; two type of classifier (using Numeral Virtual generalization (NVG) RAM and K nearest neighborhood KNN) where adopted to identify the talker's utterance. The recognition performance for the input visual utterance when using NVG RAM is 94.679%, which is utilized for the first time in this work. While; 92.628% when KNN is utilize.*

Keywords- *visual speech, feature extraction, AV letters recognition, classification.*

How to cite this article: M.H. Mahmmed, T.R. Saeed, and W.H. Ali, "Robust Visual Lips Feature Extraction Method for Improved Visual Speech Recognition System," *Engineering & Technology Journal*, Vol. 36, Part A, No. 2, pp. 136-145, 2018.

1. Introduction

Great availability of pattern recognition systems and multimedia devices motivate researchers to be interested in visual speech recognition system design. One important reason behind building visual speech recognition system is the need for quiet communication between human and machine that the operation of the human still has a weakness in this communication process. [1] Furthermore, the increase in the development and cheapness of digital computer and FPGA (field programmable gate array) technology promises low cost hardware, hence it is simple to implement this system [2].

Providing a computational solution to visual speech recognition can be divided into three main tasks, or stages: Firstly, the mouth region is automatically or manually detected for a given set of images (lips segmentation). Secondly, suitable visual features from the given mouth images that represent the utterance of the speakers are detected. Finally, these features are used as input to recognition network in order to identify the speaker utterance.

Here a review some important approaches introduced by the researchers recently in order to solve the problem of utterance classification visually with and without the presence of acoustic signal which proposed in this paper.

Harry Mcgurk [3] gave the first flash on the importance of VSR system. They made a simple experiment by collecting a mix of peoples to be trained and tested on a different utterance The result indicated a 99% correct accuracy response for auditory only; wherein auditory – visual resulted in a percentage error of about 52%, due to the confusion between soundtrack and the lips movement. In [4] an improved version of automatic lip reading is developed using vector quantization for the extracted features from the lips region to reduce the segmented error. A

DOI: https://doi.org/10.30684/etj.36.2A.4
2412-0758/University of Technology-Iraq, Baghdad, Iraq

dynamic time wrapping is used for the classification parts. The proposed system was tested on multi talker pronouncing the alphabetic letters. Results from combined acoustic and visual speech recognition showed an improvement in performance reaching 66% for visual only.An algorithm for segmenting the speaker's lip and features extraction is proposed in [5]. The algorithm is based on converting the image from RGB region to HI (hue intensity) region then using Markov random field modeling to determine red hue current region. The final stage is to extract the region of interest ROI to get the geometric feature. 95% correct segmented lips images are obtained that verified the success of the algorithm. While, a method of integrating both types of lips features information, (color information and Edge information) is presented by Zhang [6]. The proposed system consists of multi stages; first using hue transformation to enhance the color space, second using the notable red hue as an indicator to locate the position of the lips, finally contract the inner and outer boundary for the lips region and calculating the edge point. The result shows that the correct lips feature can be extracted successfully. S.L. Wang et al. [7] introduced a new lips feature extraction based on geometric and based model. A 16-point lip model is used to represent the lip contour. Using FCMS (fuzzy clustering method incorporating shape function) to get the probability map and a region area can be calculated**.** Experimental results show that the proposed research satisfies correct results for 5,000 lip images of over 20 people. In [8] a new approach for automatic lips point feature on a talker face is proposed. The extracted visual information is then classified in order to recognize the utterance of some French words. Experiments revealed that the system recognition is 73%. Muzaffer Doğan [9] developed an application using Microsoft MS Kinect camera to recognize Turkish color names to be used in the education of hearing impaired children. The proposed model consists of two stages. The first predefines lips point using MS kinetic face by the SDK program. The second extracts feature from these points by calculating the corner angle between them. K-nearest Neighbors KNN classifier is used to classify the word with Manhattan and Euclidian distance. As a result, the isolated words are classified with a success rate of 78.22%. In this paper a new technique for improving visual speech recognition are proposed including the three stages mentioned. These techniques was tested for a twenty six alphabet English words from thirty speakers pronunciations, however it can be extended to work for different sets of words and other languages as well. The next sections in this paper are outlined as follows. Section 2 describes new approach to isolate the moth region from face area then segmented the lips boundary. Section 3 explains one type of lips feature extraction (geometric model). Section 4 describes the two proposed classifiers used to identify the speakers' utterances. Section 5 illustrates the proposed visual speech recognition system. Section six the simulation results for testing the classification rate of the input visual speech for the proposed VSR systems. Section 7 outlines major conclusions of this work.

2. Lips Segmentation

The proposed visual speech recognition system is organized into three modules: 1) lips segmentation from mouth image; 2) visual lips feature extraction (model base); 3) utterance classification. Before lips segmentation process start a preprocessing operation on the input must be applied, .The first operation is selecting the best sets of frame from the input video frames that represent the pronounced word. The second one is to detect the face area from the input image then face part detection process is applied to isolate the moth region therefore, to identify this operation a face parts detection method based on Viola – Jons algorithm VJA is considered in this paper [10]. VJA is one of the first real time face detection algorithm [11]. Furthermore, it is a machine approach for face part detection, which is capable of processing image extremely quickly and achieving high detection rate as shown in Figure 1. Once the mouth is isolated, a new method is proposed depending on skin color segmentation to isolate the lip boundary correctly. Furthermore, it removes any additional unwanted clusters area within the lips image, which is important for extracting good lips movements' features. At the beginning, the effective mouth images for the utterance video are extracted and converted to gray scale based on word segmentation process and VJA algorithms as shown in Figure 2.

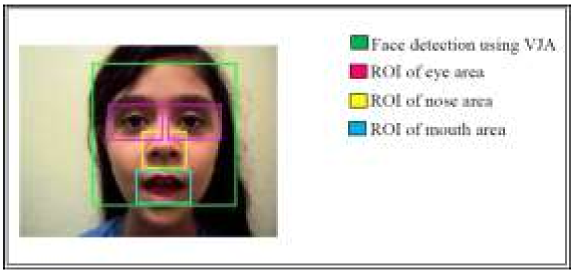
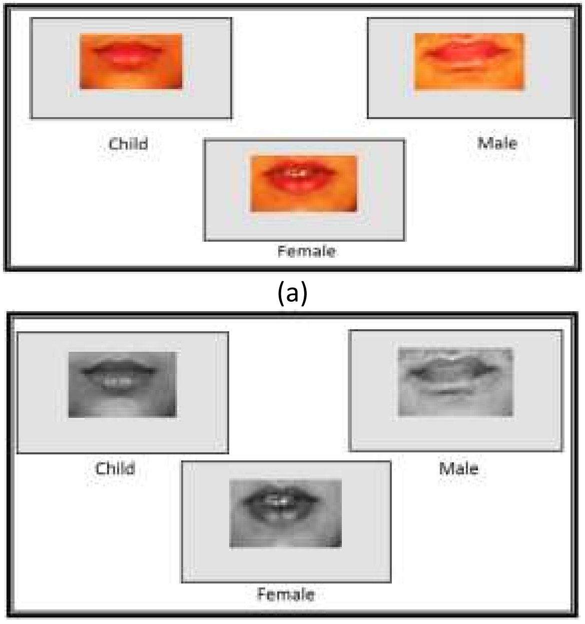


Figure 1: Region of interest (ROI) to detect face parts.



(a)

(b)

Figure 2: Effective selective mouth images for three subjects. (a) Color image, (b) Intensity image

The main aim of the proposed segmentation algorithm (PSA) is to extract the lips boundary from each selected frame according to the following steps:

1. Crop the mouth region window from the binary image of the selected frames, based on VJA as shown in Figure 3. The binary image of the mouth will be represented as a $(N \times M)$ matrix of zeroes and ones.

2. Remove the surrounding boundary in order to obtain the edges of the mouth using the following operations.

a. Find the number of ones in each row of the binary image matrix.

b. Remove the rows that have the number of one's less than 10% of N (experimentally selected).

c. Finding the number of ones in each column of the binary image matrix.

d. Remove the columns that have the number of one's less than 10% of M. Figure 4 shows the output image of this step.

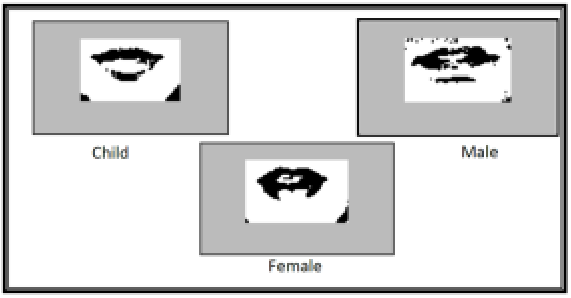


Figure 3: Effective selective frames binary output from stage 1

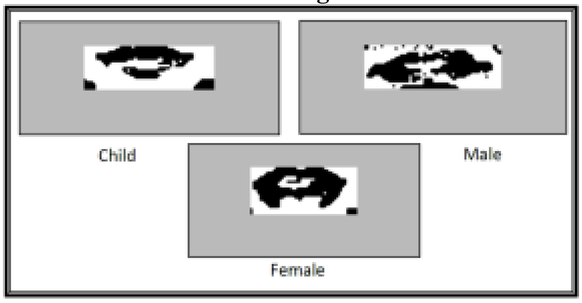


Figure 4: Effective selective frames binary output from stage 2

3. Trace the objects that do not belong to the mouth and remove them according to the following steps:

a. All exterior bounded objects in binary image matrix.

b. Remove all objects that have number of pixels less than $1/25$ of the number of image pixels

After this step is done, the binary image appears as illustrated in Figure 5.

4. Extract the final mouth boundary by removing all rows and columns that contain zero pixels only as demonstrated in Figure 6.

3. Visual Feature Extraction (Geometric Base Model)

In general, visual lips features representation approaches are divided into two parts; pixel-based and model based. The model-based features are adopted for the proposed VSR system to extract one types of visual lips features. A method is proposed for extracting the geometric features depending on specific pixel labeled automatically on the lip's boundary.

Geometric or Points of Interest (POI) represent one of the visual speech features in this work. After extracting the exact lip boundary, a proposed method will be used to extract the geometric features by setting age lip pixels. Eight pixels are considered to calculate these features, as shown in Figure 7, according to the following steps:

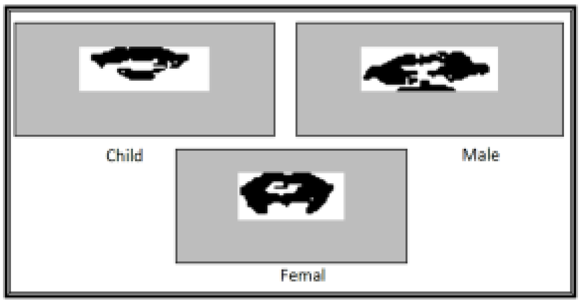


Figure 5: Effective selective frames after boundary operation

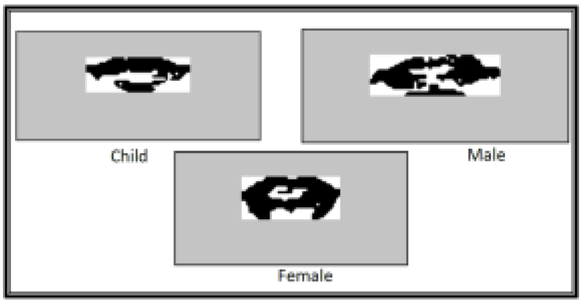


Figure 6: Final frames for correct lips boundary.

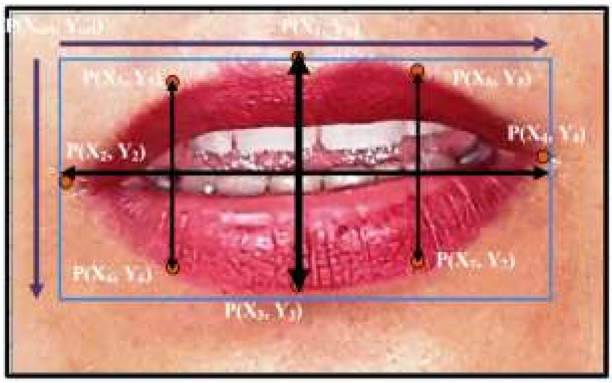


Figure 7: Eight pixels on the lips edges

1. Determine the exact lips boundary images and assign the size as $N \times M$, where N and M is the number of rows and columns of binary mouth image respectively.

2. Localize the eight pixels of interest on the lips boundary by calculating the Cartesian coordinate of these pixels as follows. (Assuming the left upper corner of the image is the original point).

a. For P$_1$, X_1 and Y_2 are computed as:

$$X_1 = \mathbb{Z}\left(\frac{N}{2}\right), \qquad Y_1 = 1 \tag{1}$$

b. For P$_2$, x2 and y2 are computed as:

$X_2 = 1$

$$Y_2 = Index\left(\mathbb{Z}\left(\frac{length\ of\ vector\ Index}{2}\right)\right) \tag{2}$$

Where Index is the vector of the locations of X_2 vector that is equal logic 1.

c. For P3, X_3 and Y_3 are computed as:

$$X_3 = \mathbb{Z}\left(\frac{N}{2}\right), \qquad Y_3 = M \tag{3}$$

d. For P$_4$, X_4 and Y_4 are computed as:

$X_4 = N$

$$Y_4 = Index\left(\mathbb{Z}\left(\frac{length\ of\ vector\ Index}{2}\right)\right) \tag{4}$$

a. For P$_5$, X_5 and $Y_5$are computed as:

$$X_5 = \mathbb{Z}\left(\frac{X_1 + X_2}{2}\right) \tag{5}$$

$Y_5 =$
$first\ location\ of\ X_5\ column\ vector\ that\ is\ equal\ to\ 1$

b. For P$_{6,}$ X_6 and Y_6 are computed as:

$$X_6 = \mathbb{Z}\left(\frac{X_1 + X_2}{2}\right) \tag{6}$$

$Y_6 = last\ location\ of\ X_6\ column\ vector\ that\ is\ equal\ to\ 1$

c. For P$_{7,}$ X_7 and $Y_7$are computed as:

$$X_7 = \mathbb{Z}\left(\frac{X_3 + X_4}{2}\right) \tag{7}$$

$Y_7 = first\ location\ of\ X_6\ column\ vector\ that\ is\ equal\ to\ 1$

d. For P$_{8,}$ and Y$_8$ are computed as:

$$X_8 = \mathbb{Z}\left(\frac{X_3 + X_4}{2}\right) \tag{8}$$

$Y_8 = first\ location\ X_8\ column\ vector\ that\ equal\ 1$

3. After computing the required edges points for the lips image, four features (height, width, left height and right height) will be extracted as shown in fig. (7). The following equations will be used to compute these features:

$$F1 = Y_3 - Y_1\ ,where\ F1\ is\ height\ of\ the\ lips \tag{9}$$
$$F2 = X_4 - X_2\ ,where\ F2\ is\ width\ of\ the\ lips \tag{10}$$
$$F3 = Y_6 - Y_5\ ,where\ F3\ is\ the\ left\ height\ of\ the\ lips \tag{11}$$
$$F4 = Y_7 - Y_8\ ,where\ F4\ is\ the\ rifgt\ height\ of\ the\ lips \tag{12}$$

4. Proposed Classifiers

The last stage of the proposed VSR system is the classification part. For the geometric features (visual cues as input signals), two types of decision classifier are suggested in order to get highest probability of correct discrimination. All of these classifiers are based on traditional classification algorithms, which include Numeral Virtual Generalization Random access memory (NVG-RAM) and K- Nearest Neighborhoods (KNN).

I. NVG-RAM

Virtual Generalizing Random Access Memory Weightless Neural Networks (VG-RAM WNN) is an efficient learning mechanism device that provides simple implementation and fast training and testing [12]. (VG-RAM WNN) is a RAM based neural network that just need memory capacity which is enough to store the data with regard to the training set in the nodes of these networks. This memory stores the input-output pairs appearing during training phase, instead of only the output. In the test phase, the memory of

VG-RAM WNN nodes is searched associatively by comparing the input presented to the network with all inputs in the input-output pairs trained. The output of each VG-RAM WNN neuron is taken by the distance function implemented by VG-RAM WNN nodes called Hamming distance [13].

$$d^{HD}(x,y) = \sum_{i=0}^{m-1} [Z_{x,i} \neq Z_{y,i}] \qquad (13)$$

Where;$d^{HD}(x,y)$ is the hamming distance between the objects x and y, i is the index of respective element reading Z out of the total number of variables m. In other word HD produce the number of Dissimilarity between the variables paired by i. The generalization of VG-RAM has been ensured through storage of all the address patterns of a node and searching for the nearest class to the unknown pattern. Figure 8 represents the circuit diagram for VG_RAM WNN [14, 15].

Although, VG-RAM is an important technique in the field of pattern recognition, but it suffers from drawback. The weakness of this method, that it can handle only the binary patterns. Therefore, an efficient numeral virtual generalization random access memory NVG-RAM classifier is suggested to overcome this issue [16].

The N-VGRAM uses the Manhattan distance [17] instead of Hamming distance as in ordinary VG-RAM in order to determine the matching of the numeral input-output pairs in recall phase. Rectilinear distance (Manhattan distance), considered by Hermann Minkowski, is a form of geometry in which the usual distance function of metric or Euclidian geometry is replaced by a new metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates. To determine the Manhattan distance MD between two vectors $p \; and \; q$ and, each vector consists of a set of numeric elements that can be determined by taking sum of the Manhattan distance for these elements [18]:

$$MD(\boldsymbol{p},\boldsymbol{g}) = \sum_{i=1}^{n} |p_i - q_i| \qquad (14)$$

Where $\boldsymbol{p} = (p_1, p_2, p_3, \dots p_n)$,
$\boldsymbol{g} = (g_1, g_1, g_3, \dots \dots g_n)$ and $\boldsymbol{n}$ is the number of elements in each vector.

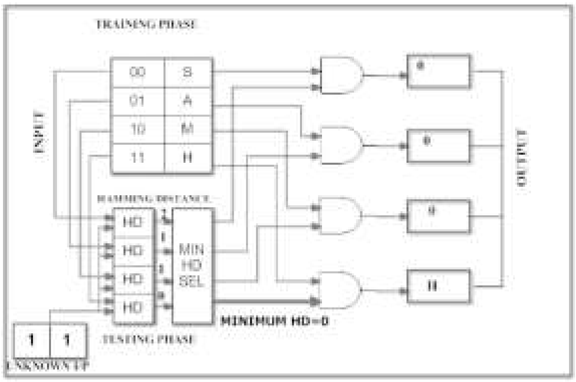


Figure 8: Circuit diagram for VG-RAM

Moreover, the size of memory used to store the input-output pairs in NVG-RAM is determined by the number of training sets because all training pairs are stored in this memory. So that, as well as the training set increases, the RAM capacity increases too, also the speed of this method limits the number of comparisons because it depends on the number of training sets. Therefore, this approach adopts a minimization technique to reduce the size of memory used by minimizing the number of learning sets and this leads to increase in the network's speed. Although, number of training set is reduced, the NVG-RAM maintains a high level of the performance accuracy.

II. KNN

K- Nearest Neighbors algorithm (KNN) is a non-parametric method used for classification and pattern recognition. The simplicity and the good accuracy in this algorithm is the one which makes it distinctive in relation to the rest of the algorithms. For a working knowledge of this algorithm, let's consider a two class problem. The similar sample most likely will have the same class assignment. That means the KNN will classify each similar sample to one class. For an unknown test sample, the KNN decision will depend on closest distance between the tested sample and each nearest sample per class, then the new sample will be assigned to one of these classes as shown in Figure 9 [19].

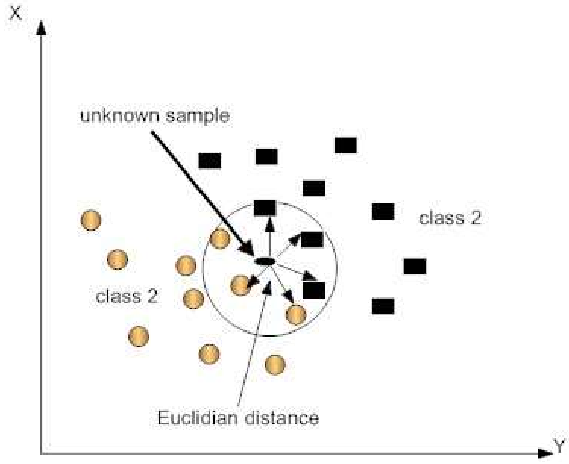


Figure 9: Classification example of KNN with two classes

Here k=5 distance. The Euclidian distance can be calculated as follows:

$$ED_{x,y} = \sqrt{\sum_{i=1}^{i}(x_i - y_i)^2} \qquad (15)$$

Nearest neighbor rule has some problems in classifying an unknown pattern. If there is m number of samples per pattern, then to ensure the nearest neighbor a computed m distances must be found from the test pattern to each of the sample points. Also, it is important to store all these m sample points. This leads to an increase of the mathematical operation as well as storage complexity of the algorithm. As the number of features increases, more number of training data samples is required; hence it increases the storage and computational complexities [20].

5. Proposed Visual Speech Recognition System

As mentioned in the previous section ASR system consists of two parts: acoustic and visual speech recognizer. This work has adopted only the visual part, accordingly a visual speech recognition system VSR is proposed as shown in Figure 10. This system consists of three stages: the first stage is subdivided into two sub stages. The first sub stage is the data corpus for the speakers' utterances and the second is the process of face partition detection and segmentation as explained in the previous sections.

The second stage is the process of visual features extractions from the speaker's lips movement. One types of features are suggested, to meet as much as possible, the required information of visual cues for the utterance. The final stage is the classifications process. In this stage the extracted visual features from the utterance lip images will be used to identify the thirty speakers' utterances

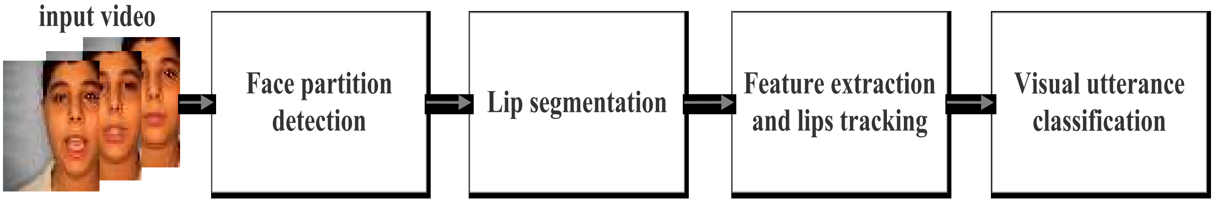

input video
Face partition detection
Lip segmentation
Feature extraction and lips tracking
Visual utterance classification

Figure 10: Bock diagram of the proposed VSR system.

In the proposed VSR system I, the visual speech classifier is a Numeral VG-RAM (NVG-RAM) neural network. The NVG-RAM algorithm is considered in this work due to the simplest arithmetic operations used in this approach, so it can be readily built by FPGA. Moreover, this classifier can handle a decimal number instead of binary number used in traditional VG-RAM.

This type of classifier is characterized by; it can learn in one shoot. The learning of proposed system I is just storing the input-output pairs in RAM. MATLAB code is written for verify the proposed system. In this program, a two dimensional matrix was utilized instead of RAM. The implementation of NVG-RAM in FPGA required reducing the number of training set of the visual speech geometric features, because a limited number of slices exist in the hardware kit. The recall phase of this network depends on the minimum Manhattan distance between the unknown pattern and all pairs stored in the RAM. The Manhattan distance between the unknown pattern and all stored pairs are evaluated to find the smallest distance. The index of the minimum distance is assigned to the number of the class in output field. The desired class is fetched to network output.

Twenty-six words of the visual speech alphabet letters are needed to be classified by the NVG_RAM network with visual geometric features as a features classifier. The training set is prepared for twenty values of visual feature for each word from the thirty speakers utterances, thus the number of vectors in training set is 780 vectors for all the twenty-six alphabet characters. Each vector contains twenty one fields; the first twenty fields are used to store the magnitude of the width F1, height F2, left height F3, and right height F4 geometric features (four fields for each feature), while the last field is used to hold the class number. Therefore, the RAM of this network consists of 780 locations; each location consists of twenty-one fields. Each field has number of cells depending on the number of bit in equivalent binary number of the features. When unknown visual speech utterance is received, the geometric features F1,F2,F3, and F4 are calculated for this visual utterance After that, these features are passed to the NVG-RAM classifier. The classifier finds the nearest pairs in training sets to extract the feature, based on Manhattan distance. The identifier can recognize type of the visual spoken word according to the minimum distance.

The K - Nearest Neighbors KNN discriminator is a popular method for audio-video speech recognition system; this is due to its simplicity and power in acquiring successful classification rate on visual speech application. The key idea behind this classifier, for the training process, is that similar sets of extracted geometric features vectors from the speaker's utterance belong to similar class (word). The unknown input utterances to the classifier are discriminated based on the distance to the nearest neighbors. During the recognition process, each isolated word (letter) belongs to the closest neighbor, based on the Manhattan or Euclidian distance. Furthermore, the input data in KNN classifier does not need further training when a new word is added to the training data set.

In the proposed VSR system II, the VSR system is built based on KNN classifier. VSR KNN classifier will be used to classify twenty-six alphabet letters from thirty speaker utterances. The training sets of VSR KNN consist of seven hundred and eighty visual feature vectors (as explained in the proposed VSR system I) belonging to twenty six word uttered by the speakers. The training phase of the KNN classifiers will assign each visual features vector set to it as own class (utterance letter). When unknown spoken word by the speaker is received, the geometric features F1, F2, F3, and F4 are calculated for this visual speech. After that, these features are passed to the KNN classifier with k value 1. The classifier will compute the Euclidian distance between each feature vector at the training sets with the extracted feature vector from the unknown input word. Through the minimum distance, the identifier can recognize type of the visual spoken word.

6. Simulation Result

The performance evaluations of the two proposed visual speech classifiers (NVG-RAM and KNN) are explored in this section using the module base features, which are more suitable to represent the variation of lip region during speech. Moreover, less mathematical computation is required to extract these features. Once the eight points of interest surrounding the lip boundary are obtained, the geometric based features (width, height, left height, and right height) can be derived and their application in recognizing utterances can be checked. A person's utterance dependent discriminator for the English alphabet letters is used in each of the proposed six classifiers of the VSR system.

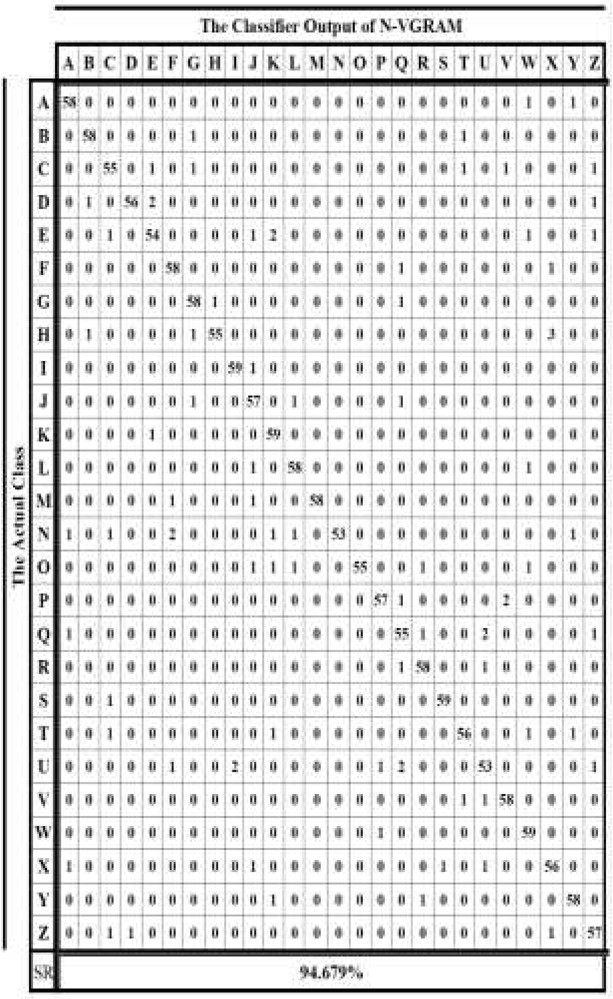
To evaluate the performance of the proposed VSR systems, two sets from thirty persons' utterances videos were adopted. Each set consists of the utterance videos from thirty talkers with different genders and ages, where, every speaker utters the alphabet English letters from 'A' to 'Z'. All the simulation results, including automatic extraction of the visual features from the input speakers' videos, training and testing of the classifiers performance of the VSR system, are performed using MATLAB2014a program.

The proposed lips geometric features produced high performance when used as inputs to NV-GRAM classifiers in order to recognize twenty six speakers' utterances as shown in Table (1). As demonstrated in this Table, all the twenty – six letters have a success rate greater than 88.33%. The best four classified letters are 'I', 'K', 'S', and 'W' with probability of correct recognition greater than 98% for the two tested sets. The worst recognition rate of this classifier was found in the identification of letters 'N' and 'M' that is 88.33%. This is due to the similarity between letters 'N' and 'M', as well as, letters 'U' and 'Q', where the variation of the lip movements are mostly the same during the pronunciations of these letters. The remaining eighteen letters are classified with recognition rate greater than 90%, depending on the way that the speakers utter the specific letter.

Figure 11 shows the performance of the VSR NVG-RAM classifier for the thirty speakers individually. Eight speakers of the two tested sets have full utterance identification, while the success rate of eleven persons is greater than 95%. The probability of correct classification of the other speakers ranged from 78.84% to

94.23%, but one drawback is in speaker index fifteen, where the success rate is 63.46% because the speaker moves his head widely during speech making the extracted features inaccurate. The overall success rate of this classifier is 94.68% when identifying *UOTEletters* corpora

Table 1: Confusion matrix of VSR-NVGRAM classifiers



	The Classifier Output of N-VGRAM																									
The Actual Class	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
B	0	58	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
C	0	0	55	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1
D	0	1	0	56	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
E	0	0	1	0	54	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
F	0	0	0	0	0	58	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
G	0	0	0	0	0	0	58	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
H	0	1	0	0	0	0	1	55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
I	0	0	0	0	0	0	0	0	59	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	1	0	0	57	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
K	0	0	0	0	1	0	0	0	0	0	59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	1	0	58	0	0	0	0	0	0	0	0	0	0	1	0	0	0
M	0	0	0	0	0	1	0	0	0	1	0	0	58	0	0	0	0	0	0	0	0	0	0	0	0	0
N	1	0	1	0	0	2	0	0	0	0	1	1	0	53	0	0	0	0	0	0	0	0	0	0	1	0
O	0	0	0	0	0	0	0	0	0	1	1	1	0	0	55	0	0	1	0	0	0	0	1	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	57	1	0	0	0	0	2	0	0	0	0
Q	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	55	1	0	0	2	0	0	0	0	1
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	58	0	0	1	0	0	0	0	0
S	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	59	0	0	0	0	0	0	0
T	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	56	0	0	1	0	1	0
U	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	1	2	0	0	0	53	0	0	0	0	1
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	58	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	59	0	0	0
X	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	56	0	0
Y	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	58	0
Z	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	57
SR	94.679%																									

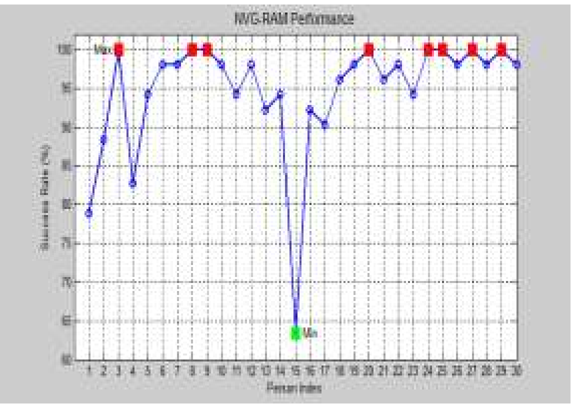


Figure 11: speaker utterance identification using VSR-NVGRAM classifier

The overall success rate of the utterance identification, using the VSR –KNN classifier, is 92,628% as illustrated by the confusion matrix in Table 2. The best performance observed from this table is for letter 'I'. The success rate of this letter is 98.35% for all testing sets.

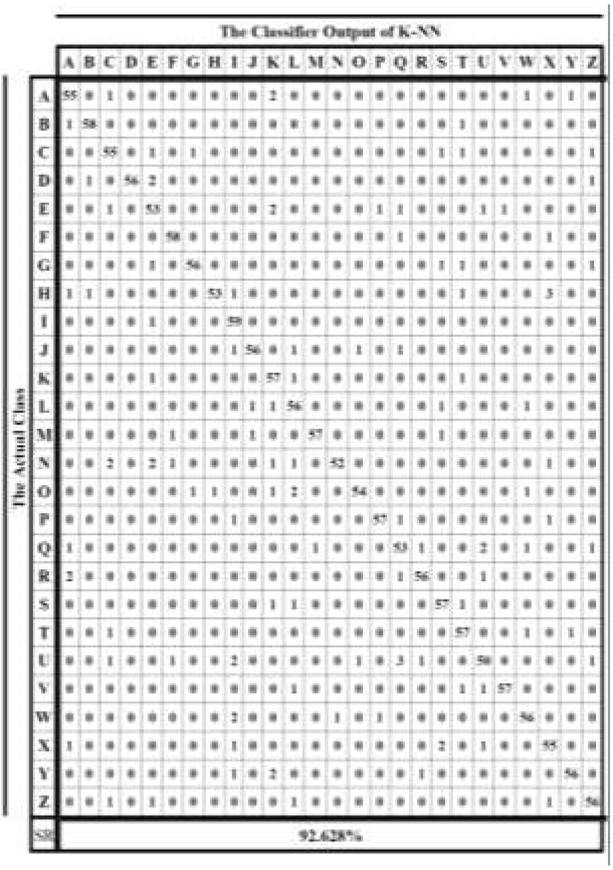
The probabilities of correct recognition for (five) letters varied from 83.3% to 88.3%, because the height feature value of these letters are closed, hence they will be classified as different letters. The lowest P_{CC} found in this classifier is for letter 'U', because of the confusion of this letter with letter 'Q'. The probabilities of correct discrimination for the remaining letters are above 90%, depending on the overlapping between them.

The total probability of classification based on speaker index using VSR-KNN is the same as that for letter recognition as shown in Figure 12. From this figure, the success rate for four speakers is 100%, while twelve speakers have a recognition rate above 95%. The PCC for thirteen speakers are found over the range from 78 to 94.23. As in VS-NVGRAM, the minimum recognition rate is 59.61 and it's for person fifteen.

7. Conclusion

Recently, automatic speech recognition systems sparked off a great deal of interest. The ASR system is consists in the main of two parts, acoustic and visual speech identification. In this paper, the focus is on VSR system only. The proposed VSR systems classify efficiently twenty-six English alphabet letters of uttered by thirty speakers. The identification processes depend only on the visual information in mouth shape variation for those talkers. A new proposed segmentation method is utilized to isolate the lips boundary from the mouth images. The result is quite impressive as compared with previous approaches, where the lip region is bounded accurately without any additional pixel noise. The success rate, when applying this algorithm on the utterance images from thirty is 98.58%.Using new approaches for visual feature extraction, one types of visual (geometric) features are extracted from the lip images that represent the speakers' utterances. The geometric feature utilizes a set of pixels points surrounding the lips area to calculate the height, width, left height, and right height of the mouth boundary. Based on the visual geometric features, the simulation result shows that, the proposed VSR NVG RAM classifiers have a success rate about 94,649% while, 92.62% for VSR KNN

Table 2: Confusion matrix of VSR-NVGRAM classifiers



		The Classifier Output of K-NN																									
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
The Actual Class	A	55	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
	B	1	58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	C	0	0	55	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1
	D	0	1	0	56	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	E	0	0	1	0	53	0	0	0	0	0	2	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0
	F	0	0	0	0	0	58	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
	G	0	0	0	0	1	0	56	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1
	H	1	1	0	0	0	0	0	53	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3	0	0
	I	0	0	0	0	1	0	0	0	59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	J	0	0	0	0	0	0	0	0	1	56	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0
	K	0	0	0	0	1	0	0	0	0	0	57	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	L	0	0	0	0	0	0	0	0	0	1	1	56	0	0	0	0	0	0	1	0	0	0	1	0	0	0
	M	0	0	0	0	0	1	0	0	0	1	0	0	57	0	0	0	0	0	1	0	0	0	0	0	0	0
	N	0	0	2	0	2	1	0	0	0	0	1	1	0	52	0	0	0	0	0	0	0	0	0	1	0	0
	O	0	0	0	0	0	0	1	1	0	0	1	2	0	0	54	0	0	0	0	0	0	0	1	0	0	0
	P	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	57	1	0	0	0	0	0	0	1	0	0
	Q	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	53	1	0	0	2	0	1	0	0	1
	R	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	56	0	0	1	0	0	0	0	0
	S	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	57	1	0	0	0	0	0	0
	T	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	57	0	0	1	0	1	0
	U	0	0	1	0	0	1	0	0	2	0	0	0	0	0	1	0	3	1	0	0	50	0	0	0	0	1
	V	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	57	0	0	0	0
	W	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	1	0	0	0	0	0	0	56	0	0	0
	X	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0	1	0	0	55	0	0
	Y	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	56	0
	Z	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	56
	SR	92.628%																									

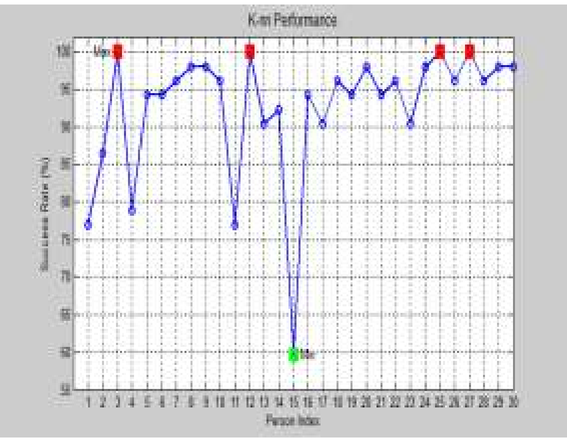


Figure 12: speaker utterance identification using VSR-KNN classifier

References

[1] B. Singh, V. Rani and N. Mahajan, "Preprocessing In ASR for Computer Machine Interaction Computer Science and Software Engineering," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 3, pp. 397-399, 2012.

[2] M. C. Herbordt, Y. Gu, T. VanCourt, J. Model, B. Sukhwani and M. Chiu, "Computing Models for FPGA-Based Accelerators," Computing in Science & Engineering, vol. 10, no. 6, pp. 35-45, 2011.

[3] H. McGuRk and J. MacDonald, "Hearing lips and seeing voice," Nature, vol. 274, pp. 747-748, 1976.

[4] E. Petajan and N. M. Brooke, "An Improved Automatic Lipreading System to Enhance Speech Recognition," in CHI '88 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, 1988.

[5] M. Likvin and F. Luthon, "Lip features automatic extraction," in Image Processing, ICIP 98. Proceedings. International Conference, Chicago, 1998.

[6] X. Zhang and R. M. Mersereau, "Lip feature extraction towards an," in Image Processing, Proceedings International Conference on (Volume:3), Vancouver, 2000.

[7] S. L. Wang, W. H. Lau and S. H. Leung, "A new real-time lip contour extraction algorithm," in Acoustics, Speech, and Signal Processing, Proceedings. (ICASSP '03).IEEE International, Hong Kong, 2003.

[8] S. Werda, W. Mahdi and A. B. Hamadou, "Lip Localization and Viseme Classification for Visual Speech Recognition," International Journal of Computing & Information Sciences, vol. 5, no. 1, pp. 62 - 75, 2007.

[9] A. Yargıç and M. Doğan, "A Lip Reading Application on MS Kinect Camera," in Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium, Albena, 2015.

[10] A. El Maghraby, M. Abdalla, O. Enany and M. Y. El, "Detect and Analyze Face Parts Information using Viola- Jones and Geometric Approaches," International Journal of Computer Applications, vol. 101, no. 3, pp. 23-28, 2014.

[11] V. Vezhnevets, V. Sazonov and A. Andreeva, "A Survey on Pixel-Based Skin Color Detection Techniques," in Proceeding. Graphicon, 2003.

[12] A. F. De Souza, C. Badue, F. Pedroni, S. S. Dias, H. Oliveira and S. F. de Souza, "VG-RAM Weightless Neural Networks for Face Recognition," in Face Recognition, InTech, 2010, pp. 172-185.

[13] A. F. De Souza, F. D. Freitas and A. G. C. de Almeida, "High performance prediction of stock returns with VG-RAM weightless neural networks," in High Performance Computational Finance (WHPCF), IEEE Workshop, New Orleans, 2010.

[14] A. F. De Souza1, F. Pedroni1 and E. Oliveira, "Automated Free Text Classification of Economic Activities using VG-RAM Weightless Neural Networks," in Seventh International Conference on Intelligent Systems Design and Applications,IEEE, Brazil, 2007.

[15] A. F. De Souza and C. Badue, "Improving

VG-RAM WNN Multi-label Text Categorization via Label Correlation," in Eighth International Conference on Intelligent Systems Design and Applications, Brazil, 2008.

[16] I. A. Hashim, J. W. Abdul Sadah and T. R. Saeed, "Efficient Numeral VG-RAM Pattern Recognition Using Manhattan Distance Calculation and Minimization Algorithm," Kasmera Journal, vol. 43, no. 2, pp. 111-122, 2015.

[17] P. Grabusts, "The choice of metrics for clustering algorithms," in Proceedings of the 8th International Scientific and Practical Conference, Latvia, 2011.

[18] A. Prakash, D. M. Urs, L. S, P. H. S and M. A. Anusuya, "Comparitive Study of Various Distance Measures for Isolated Speech Recognition Application," International Journal Of Engineering Sciences & Research Technology, vol. 4, no. 6, pp. 961-972, 2015.

[19] A. K. Abdul Hassan and M. S. Kadhm, "An Efficient Image Thresholding Method for Arabic Handwriting Recognition System," Eng. &Tech.Journal, vol. 34, no. 1, pp. 26-34, 2016.

[20] B. El Kessab, C. Daoui, B. Bouikhalene and R. Salouan, "Fingerprint Recognition Using Gabor Filter with Neural Network," Eng. & Tech. Journal, vol. 32, no. 2, pp. 393-354, 2014.

Author(s) biography

Dr. Wissam H. Ali, PhD degree in electrical engineering, pattern recognition, department of Electrical Engineering, University of Technology

.