International Journal of Computer Applications Technology and Research
Volume 5– Issue 3, 141 - 146, 2016, ISSN:- 2319–8656

Holistic Approach for Arabic Word Recognition

Talaat M. Wahbi
College of Computer Science and Information Technology
Sudan University of Science and Technology
Khartoum, Sudan

Mohamed E. M. Musa
College of Computer Science and Information Technology
Sudan University of Science and Technology
Khartoum, Sudan

Abstract. Optical Character Recognition (OCR) is one of the important branches. One segmenting words into character is one of the most challenging steps on OCR. As the results of advances in machine speeds and memory sizes as well as the availability of large training dataset, researchers currently study Holistic Approach "recognition of a word without segmentation". This paper describes a method to recognize off-line handwritten Arabic names. The classification approach is based on Hidden Markov models.. For each Arabic word many HMM models with different number of states have been trained. The experiments result are encouraging, it also show that best number of state for each word need careful selection and considerations.
Keywords: pattern recognition; HMM; Holistic approach; offline recognition: Arabic word recognition

1. INTRODUCTION

One of the important branches of pattern recognition is Optical character recognition (OCR). OCR concerns depend on the stages of pattern recognition system. Firstly, it addresses text types like: digits, letters, and words. The second stage is segmentation to letters or strokes in case of using word or continues to the next stage without segmentation (holistic approach). In this approach a word is treated and identified as entity. The third stage is preprocessing which detect the main problems in scanning or writing like skew or noise. Feature extraction stage in OCR is a main issue to classify recognition to online in case of using pen moving direction, pen press,….etc beside the image, or offline in case of image features. The final stage is to use classifier to evaluate the previous stages.

All these stages make different challenges to detect the best parameters for each stage, and this increase when we use a cursive handwritten. For example Arabic words have many letter's shapes, dots, in addition to letter's overlapping. All these difficulties in Arabic language itself let Arabic being late in progress that has been achieved in the field of handwritten word recognition. Another challenge is the lack of special task handwritten datasets.

Arabic names used today have so much repetition, such as names of the prophets (Muhammad, Ibrahim... etc.) and names of the Caliphs and compound names whose first element is Abd (slave of God) (Abdullah, Abdul Rahman... etc.), and there are many examples of repetitive names (such as Adil, Awad ... etc.), together with a few common names. Therefore the idea of designing a system that uses the Holistic Approach to quickly recognize the common names and resort to the use of the Analytical Approach to recognize the names that are not common (Figure 1), is worthy of consideration for the probability of designing an effective system to recognize the names. This paper examines the effectiveness of the first part of this system which is the use of probabilistic neural networks in the inclusive Recognition of the most common Arabic names.

The rest of the paper is organized as follows: Section 2 sketches some related studies in HWR using HMMs. Section 3 briefly introduces HMMs. Section 4 describes in general SUST names dataset. Section 5 illustrates and outlines the results achieved by the experiments performed. Finally, a conclusion is drawn with future work outlooks in section 6.

2. RELATED STUDIES

The application of HMMs to Arabic OCR was first attempted by Amin and Mari[1]. Subsequently Khorsheed and Clocksin [2] present a technique for the offline recognition of cursive Arabic script based on an HMM. AlKhateeb et al. design a word-based off-line recognition system using Hidden Markov Models (HMMs). They extract several structural features and a group of intensity features using a sliding window. Experiments were carried out using the IFN/ENIT database which contains 32,492 handwritten Arabic words [3].Volker Märgner eta al presents the IfN's Offline Handwritten Arabic Word Recognition System. The system uses Hidden Markov Models (HMM) for word recognition, and is based on character recognition without explicit segmentation [4]. Somaya Alma'adeed et al. present a complete scheme for unconstrained Arabic handwritten word recognition based on a multiple hidden Markov models (HMM) [5]. Ramy Al-Hajj and Chafic Mokbel present results of a language independent handwritten recognition baseline system developed to recognize cursive handwritten words. The system is based on a stochastic Hidden Markov Model [6].

3. HIDDEN MARKOV MODEL (HMM)

A hidden Markov model is a stochastic finite state machine, specified by a tuple (S;A;π) where

S is a discrete set of hidden states with cardinality N,

π is the probability distribution for the initial state

$$\pi(i) = P(s_i) \qquad s_i \in \mathcal{S}$$

A is the state transition matrix with probabilities:

$$a_{ij} = P(s_j \mid s_i) \qquad s_i, s_j \in \mathcal{S}$$

Where the state transition coefficients satisfy

$$\sum_{s_j \in \mathcal{S}} a_{ij} = 1\,,\ s_i \in \mathcal{S}$$

The states themselves are not observable. The information accessible consists of symbols from the alphabet of observations O = (o1, ... ,oT) where T is the number of samples in the observed sequence. For every state an output distribution is given as

$$b_i(k) = P(o_t = k \mid s_i) \qquad k \in \mathcal{O},\ s_i \in \mathcal{S}$$

Thus, the set of HMM parameters θ consists of the initial state distribution, the state transition probabilities and the output probabilities. MMs can be used for classification and pattern recognition by solving the following problems:

The Evaluation Problem: Given the model with parameters θ, calculate the probability for an observation sequence O, Let O=(o1,… ,oT) denote the observation sequence and S=(s1,… ,sT) ; a state sequence. The probability P(O|θ) can be obtained by Forward Algorithm.

The Decoding Problem: Find the optimal state sequence for an observation sequence argmaxSϵST P(S|O,θ). This can be done by the Viterbi algorithm [7].

The Learning Problem: Given an observation sequence O and the HMM parameters, find the parameters Ө which maximize P(O|θ) i.e. Ө argmaxθ P(O|θ). This question corresponds to training an HMM. The state sequence is not observable. Therefore, the problem can be viewed as a missing-data problem, which can be solved via an EM-type algorithm. In the case of HMM training [7], this is the Baum-Welch algorithm. A tutorial on HMM models, the estimation problems mentioned above, and their applications to modeling a recognition system can be found in [7].

4. SUST NAMES DATASET

Arabic males names data set was used to detect the efficiency of the suggested comparison, it's a new dataset publish by SUST ALT group, it contain about 40,000 sample for 40 common males and females name in Sudan (this statistical depend on a previous dataset from the same group), figure (1) below shows the form used in the data collection process.

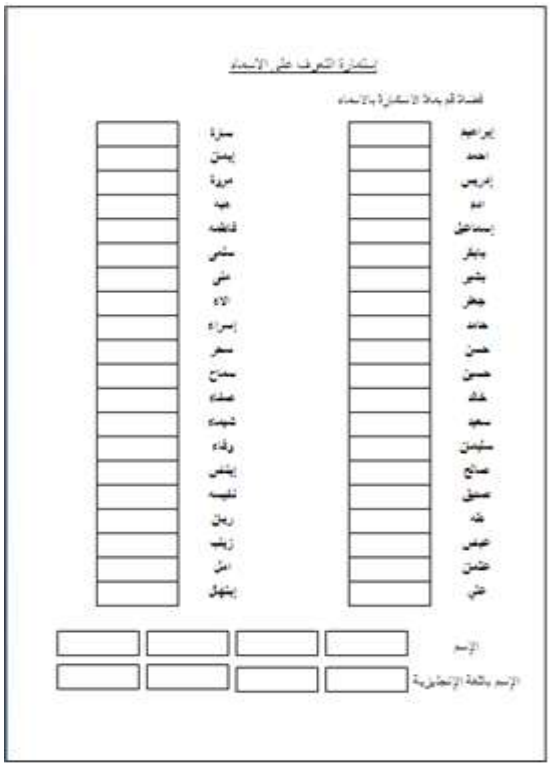


Figure 1: SUST names dataset

5. EXPERIMENT AND RESULTS

The main Recognition system stages are: Preprocessing, framing, features extraction, vector quantization, classification. We choose males names from SUST dataset with 100 samples per class for training and 50 samples per class for testing, figure (2) show all processes in details as follows.

5.1. Preprocessing

This stage contain many sub stages:

- Noises remove.
- Cropping and binarization: extract just handwritten word image and get the binary image.
- Resizing: to cope image dimensions difference in size, forming all images size to be 60X140.

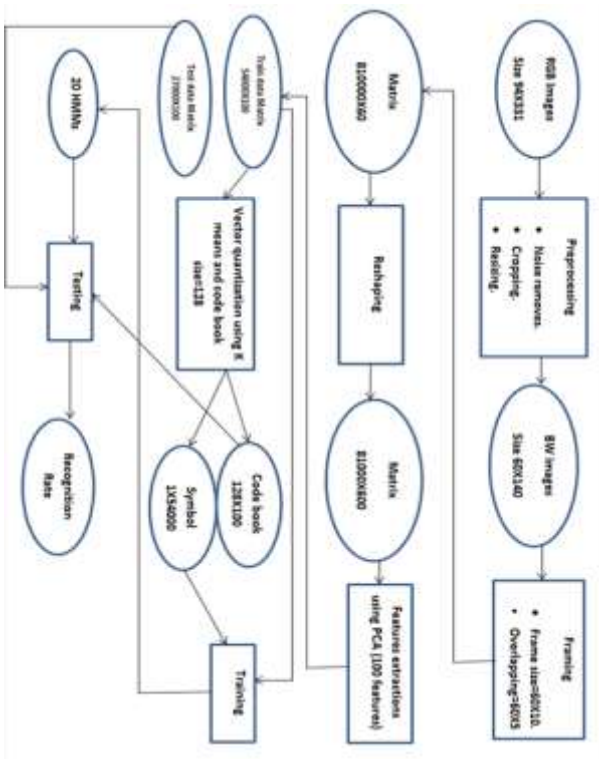


Figure (2): show general view of recognition system

5.2. Framing

Any vector split into frames by window size equal to half frame and frame size equal to 60X10 pixels. The choosing of this size depends on experiments.

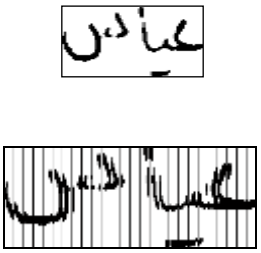


Figure (3): Upper image show the image after resized and the lower show image after framing.

5.3 Reshaping

Each frame reshape in to vector 1X600

5.4 Features extraction

Three testing done to choose the best features number using PCA, 100 features founded to been the best one.

5.5 Vector quantization

Codebook generated using k means clustering algorithm with 128 clusters, this number is the dominated in many researches [8, 9].

5.6 Classification

Model Discriminant HMM (MD-HMM) is used. The main goal of classification is to address states number effect in the HMM recognition system. 20 HMMs trained using Viterbi, states number set is {3,4,5,6,7} and their recognition rates shown in table (1).

Table (1): train and test recognition rates for states set

states	train	test
3	78.2%	52.1%
4	84.2%	54.3%
5	88.3%	56.6%
6	91.35%	59%
7	92.4%	63%

6. TEST CONFUSION MATRICES

6.1 In case 3 states

Table (2): three states model confusion matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	27	1	3	0	3	0	0	1	2	0	4	0	0	1	2	1	1	0	1	3
2	2	23	1	3	8	0	2	0	1	0	1	0	0	0	2	3	4	0	0	0
3	5	0	26	1	0	9	1	0	1	0	1	0	0	2	1	1	0	0	2	0
4	1	3	2	17	10	5	1	2	1	1	1	0	0	1	3	0	1	0	0	1
5	0	3	0	3	19	0	0	7	1	3	2	2	0	0	1	0	9	0	0	0
6	3	0	20	0	0	19	3	0	2	0	0	0	0	0	0	0	2	0	0	1
7	1	0	6	2	1	5	23	0	0	0	8	0	0	3	0	0	1	0	0	0
8	2	0	1	0	4	0	0	31	4	1	3	0	0	0	1	0	2	0	1	0
9	1	1	1	0	0	0	0	0	41	0	0	1	0	1	1	0	0	2	0	1
10	2	3	1	0	7	0	2	0	0	22	4	0	1	0	1	0	1	3	0	3
11	8	0	1	0	2	0	3	6	0	0	29	0	0	0	0	0	0	0	0	1
12	0	0	0	1	1	0	0	1	1	2	2	29	0	5	2	2	0	2	1	1
13	2	0	5	0	1	3	2	1	2	0	0	0	32	1	1	0	0	0	0	0
14	5	0	4	1	0	0	4	1	0	0	0	1	0	25	0	6	0	1	2	0
15	6	0	1	2	5	1	1	3	1	0	2	0	1	1	22	1	2	1	0	0
16	2	2	0	1	0	2	4	0	0	0	3	0	0	2	0	27	5	0	1	1
17	0	4	0	3	6	0	0	0	1	1	4	0	0	0	0	4	26	0	1	0
18	2	0	2	0	1	0	0	0	0	3	0	0	0	2	0	3	0	27	6	4
19	4	0	1	0	0	0	0	1	0	0	3	0	0	0	0	3	0	9	28	1
20	6	0	0	0	0	1	0	2	1	0	6	1	0	1	0	2	0	1	1	28

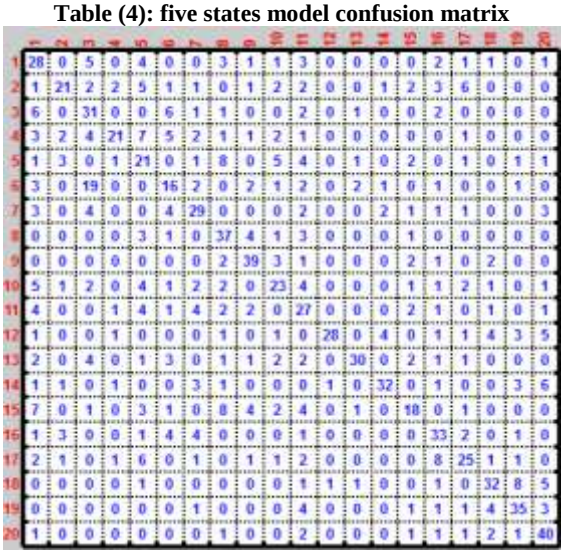
6.2 In case 4 states

Table (3): four states model confusion matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	24	0	3	0	4	0	0	0	0	0	2	0	1	0	4	3	2	1	2	4
2	2	27	2	2	6	0	1	0	1	1	4	0	0	0	1	2	1	0	0	0
3	6	0	26	0	0	6	2	0	3	0	1	0	2	2	1	0	0	1	0	0
4	2	2	5	19	6	4	4	2	1	1	1	0	0	0	2	0	1	0	0	0
5	2	2	0	0	19	0	0	8	0	3	5	0	1	0	2	0	6	0	1	1
6	3	0	13	0	0	20	7	0	1	3	1	0	0	0	1	0	1	0	0	0
7	3	0	5	1	0	3	30	0	0	0	4	0	0	1	0	0	0	0	0	3
8	0	0	0	1	2	0	1	35	0	3	1	0	0	0	4	1	0	2	0	0
9	1	0	0	0	0	0	1	5	36	1	1	1	0	0	1	0	1	1	0	1
10	4	1	2	0	5	0	1	1	2	26	3	0	0	0	2	0	1	1	0	1
11	10	0	1	0	0	1	2	6	3	1	22	0	0	0	1	0	1	0	0	2
12	0	0	0	0	0	0	0	3	1	1	1	29	1	3	0	1	0	5	2	3
13	1	0	3	0	1	3	0	0	1	2	1	0	35	0	2	0	0	1	0	0
14	2	1	1	1	0	0	4	1	0	0	0	0	1	29	0	4	0	0	4	2
15	6	0	1	1	3	1	2	7	3	1	2	0	0	0	19	0	2	1	1	0
16	3	1	3	0	0	0	1	0	0	0	1	0	0	2	2	32	4	0	0	1
17	3	1	0	2	10	0	1	0	0	1	1	0	0	0	2	5	23	0	1	0
18	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0	4	0	30	6	6
19	1	0	1	0	0	0	1	3	1	1	3	1	0	0	3	3	0	2	27	3
20	2	0	1	0	0	0	0	2	0	1	4	0	0	0	2	3	0	0	0	35

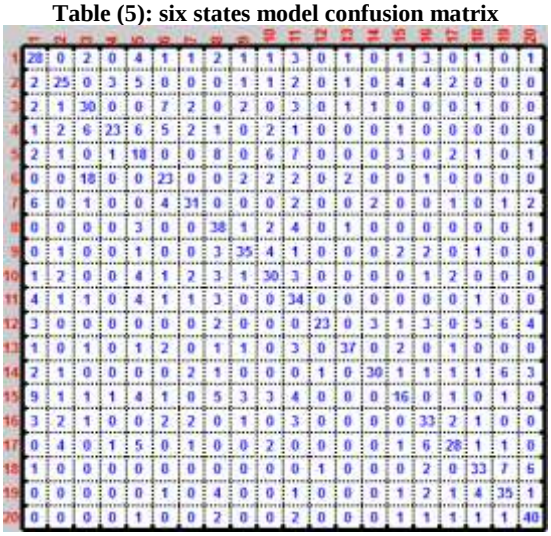
6.3 In case 5 states

Table (4): five states model confusion matrix



6.4 In case 6 states

Table (5): six states model confusion matrix



6.5 In case 7 states

Table (6): seven states model confusion matrix

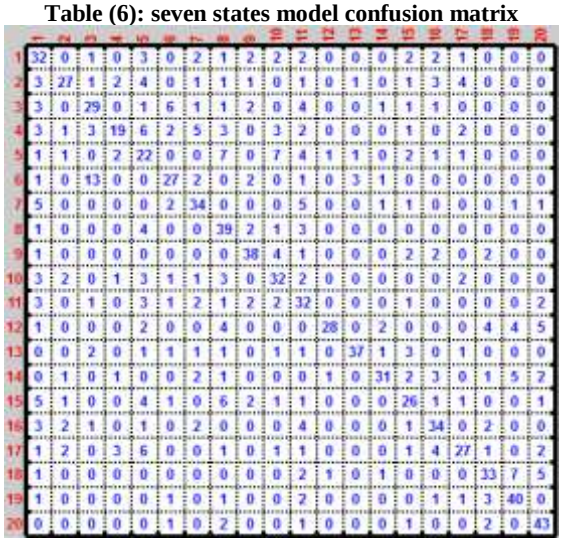


Table (7): explain names and theirs letters numbers

Name/class		Letters no.
1. Abass	عباس	4
2. Hesn	حسن	3
3. Khalid	خالد	4
4. Osman	عثمان	5
5. Saiad	سعيد	4
6. Salih	صالح	4
7. Sedig	صديق	4
8. Soliman	سليمان	6
9. Taha	طه	2
10. Adm	آدم	3
11. Ahmed	أحمد	4
12. Ali	علي	3
13. Babekir	بابكر	5
14. Bsheer	بشير	4
15. Ebraheem	إبراهيم	7
16. Edrees	إدريس	5
17. Esmaeail	إسماعيل	7
18. Gafr	جعفر	4
19. Hamed	حامد	4
20. Hesain	حسين	4

The following issues are observed from the above confusion matrices:

- **States increasing and number of letters**

The Table(8) displays the best number of states for different classes.

Table (8): explain classes and state/states according to best recognition rates

Name/class	States	Name/class	states
1. Abass	7	11. Ahmed	6
2. Hesn	4,7	12. Ali	3,4
3. Khalid	5	13. Babekir	6,7
4. Osman	6	14. Bsheer	5
5. Saiad	7	15. Ebraheem	7
6. Salih	7	16. Edrees	7
7. Sedig	7	17. Esmaeail	6
8. Soliman	7	18. Gafr	6,7
9. Taha	3	19. Hamed	7
10. Adm	7	20. Hesain	7

The above table display clearly the effect of numbers of letters in recognition, especially in short names like Ali and Taha (short names has best recognition rates with small number of states). The rest of names (medium and high) have higher recognition rates with higher number of states.

From Table (9) we note that five names which have highest confusion with other names are (greater than 30 samples): Abass, Khalid, Saiad, Soliman, and Ahmed. All these names have 4 letters except (Soliman). Also these name have high recognition rates with one class except (Saiad).

- **The relation between states and error rates**

Table (10) shows the names which six or more confused samples

The major confused classes are: Saiad with Soliman, salih with Khalid, and Gafr with Hamed and not vice versa. Overlapping play the main role in this confusion as shown in Figure (5).

Table (9): The relation between increasing the number of states and numbers of samples confused with others classes

Name/ Class	3 states model	4 states model	5 states model	6 states model	7 states model
1. Abass	52	52	41	37	36
2. Hesn	17	8	11	16	10
3. Khalid	49	42	41	31	22
4. Osman	17	8	7	6	9
5. Saiad	49	37	39	38	38
6. Salih	26	18	27	25	16
7. Sedig	23	28	22	13	19
8. Soliman	25	38	31	35	33
9. Taha	18	17	17	13	13
10. Adm	11	20	21	32	22
11. Ahmed	44	37	40	41	37
12. Ali	5	3	2	2	3
13. Babekir	2	6	6	6	5
14. Bsheer	20	8	8	6	7
15. Ebraheem	15	30	15	18	19
16. Edrees	26	26	26	26	18
17. Esmaeail	28	20	19	14	13
18. Gafr	19	15	16	18	15
19. Hamed	16	17	19	23	17
20. Hesain	17	27	26	19	18

Table (10): explain classes and state/states according to best recognition rates

Name/ class	3 states model	4 states model	5 states model	6 states model	7 states model
1. Abass					
2. Hesn	Saiad				
3. Khalid	Salih			Salih	
4. Osman	Khalid		Saiad		
5. Saiad	Soliman, Esmaeail	Soliman	Soliman	Soliman, Ahmed	Soliman, Adm
6. Salih	Khalid	Khalid, Sedig	Khalid	Khalid	Khalid
7. Sedig	Ahmed				
8. Soliman					
9. Taha					
10. Adm	Saiad				
11. Ahmed	Abass	Abass			
12. Ali					
13. Babekir					
14. Bsheer					
15. Ebraheem		Soliman	Gafr	Abass	
16. Edrees					
17. Esmaeail		Saiad	Edrees		
18. Gafr			Hamed	Hamed	Hamed
19. Hamed	Gafr				
20. Hesain					

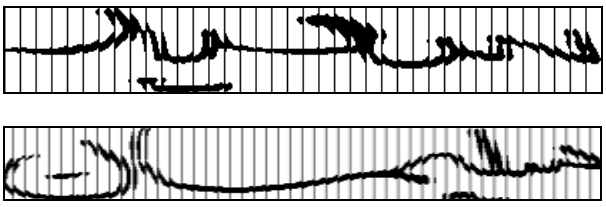


Figure (4): Show overlapping images for (Saiad), and (Soliman).

7. CONCLUSION AND FUTURE WORK

This paper discusses the effects of number of states in a HMM for handwritten Arabic names recognition. According to the results many improvements may achieved. For instance, different set of features may give better results such as Chain code and wavelet code and wavelet. Also adding a post processing component may boost the recognition rate. Another important issue to put in consideration is using multiple states for different words.

8. REFERENCES

[1] Amin and J. Mari, "Machine recognition and correction of printed Arabic text," IEEE Trans. on Systems, Man, and Cybernetics, vol. 19, no. 5, 1989, pp.1300-1306.

[2] M. Khorsheed and W. Clocksin, "Structural Features Of Cursive Arabic Script", The 10th British Machine Vision Conference, University of Nottingham, Nottingham-UK, September-1999.

[3] J.AlKhateeba,J.Rend, J.Jiangb, and H.Al-Muhtaseb Offline handwritten Arabic cursive text recognition using Hidden Markov Models and re-ranking, Pattern Recognition Letters,Volume 32 Issue 8, June, 2011 .

[4] V. Maegner, H.El Abed, M.Pechwitz,"Offline Handwritten ArabicWord Recognition Using HMM -a Character Based Approach without Explicit Segmentation" .

[5] S. Almaadeed, C. Higgens, and D. Elliman, "A New Preprocessing System for the Recognition of Off-lineHandwritten Arabic Words", IEEE International Symposium on Signal Processing and Information Technology, December, 2001.

[6] R.Al-Hajj, C.Mokbel," HMM-Based Arabic handwritten cursive reconition system".

[7] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition," Proc. of the IEEE, vol. 77, n. 2, pp 257-285, Feb. 1989.

[8] K.C Jung,S.M Yoon,H.J Kim**,** Continuous HMM applied to quantization of on-line Korean character paces,Pattern Recognition Letters, Volume 21, Issue 4, April 2000, Pages 303-310.

[9] Kenichi Maruyama, Makoto Kobayashi, Yasuaki Nakano, HirobumiYamada. Cursive Handwritten Word Recognition