ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 5, November 2014

Multiple Classifiers Combination Applied to OCR of Tifinagh Alphabets

Brahim SABIR, Yassine KHAZRI, Ahmed JADIR, Bouzekri TOURI, Mohamed MOUSSETAD
Faculté des Sciences Ben M'Sik, Ecole Supérieure de Technologie Essaouira
Faculté des Sciences Ben M'Sik, Faculté des Sciences Ben M'Sik

Abstract*— Optical character Recognition (OCR) is a technic that converts scanned or printed text images into editable text for further processing. Many OCR solutions have been proposed and used in commercial systems. However not much can be found about OCRs for the Tifinagh Alphabets. This paper has been an attempt towards the development of an OCR for Tifinagh Alphabets. The proposed Optical Tifinagh Alphabets Recognition algorithm includes all traditional steps of an OCR tool: normalization of scanned image, binarization, segmentation, feature extraction and a decision tree based on neural networks to match read Tifinagh alphabets with trained pattern. In addition to that and in order to resolve the high rates of confusion issue; a confusion matrix has been identified, and for classification step, multiple classifiers were implemented for the identified set of confused alphabets to select the recognized ones. Based on various parameters, the proposed algorithm has been developed using Matlab, and compared with a commercial OCR tool, mainly insight software of Cognex. The presented method has the benefit that the accuracy of recognition is much higher when compared to a neural network, and a commercial OCR that performs the same operation. In addition to that, the use of the multiple classifiers for alphabets with high rates of confusion enhances the time consumption.

***Index Terms*— Artificial Neural Network, Cognex, Insight software, Multiple Classifiers, OCR, Tifinagh Alphabets.**

I. INTRODUCTION

The Amazigh (Berber) language is spoken in North Africa mainly for oral communication and has been introduced in mass media and in the educational system in Morocco [4], and it is "an official language of Morocco, as common to all Moroccans without exception heritage."(Moroccan Constitution 2011). The establishment of The "Royal Institute of the Amazigh Culture" (IRCAM) carried out a major action to standardize the Amazigh language. The Tifinagh is the writing system of the Amazigh language; and The Amazigh alphabets, called "Tifinagh-IRCAM", adopted by the Royal Institute of the Amazigh Culture, was officially recognized like belonging to the basic multilingual planned by the International Organization of Standardization (ISO). The figure.1 represents the repertoire of Tifinagh which is recognized and used in Morocco; the number of the alphabetical phonetic entities is 33. [9] In contrast to Latin and Arab, the Amazigh alphabet is never cursive which facilitates the operation of segmentation.

Fig.1.Tifinagh Alphabets - IRCAM

II. RELATED WORKS

The OCR can contribute tremendously to the advancement of documents identifications, e-books producing, questionnaires processing, exam papers processing and many other applications. The OCR is characterized mainly by computational cost and accuracy rate. The methods popularly used for OCR are: template matching, feature matching and structural analysis, which are with not high accuracies rates. The used classification methods are[1]:

- Statistical Methods in OCR: Statistical classifiers like k-Nearest-Neighbor (KNN), which store and compare all training samples.
- Artificial neural networks(ANNs) in OCR: is a classification based on learning from examples in addition to SVM(Support vector machines)
- SVMs in OCR (kernel methods support vector machines): (SVMs) have higher computational complexity and memory requirements than ANNs, and higher classification accuracy.

A multiple classifier is a combination of all methods. Some efforts have been reported in literature for Amazigh characters: based on the Hough transformation [9], the statistical and geometrical approaches [9], [10], artificial neural network [11], [12], [13], Hidden Markov Models, syntactical method based on the finite-state machines [13], and dynamic programming. The recognition rate for printed characters is high than 99% and for handwritten characters is high than: 96% [13]. The performance of an OCR is related to: sample data, preprocessing technique, feature representation, classifier structure and learning algorithm. The training time of statistical classifiers is linear with the number of classes, and the training time is generally proportional to the square of number of samples. SVMs have

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 5, November 2014

been demon-strated superior classification accuracies to neural classifiers in many experiments [11]. Neural classifiers consume less storage and computation than SVMs[12].

III. SCOPE OF THIS WORK

The aim of this work is to classify and recognize an input image containing Tifinagh characters using three step approaches. In the first step the Tifinagh text image is segmented and in the second step it classifies and recognizes these characters using Neural Network. The third step will be the resolution of high confusion rates by multiple classifiers approach, and the proposed method is illustrated as:

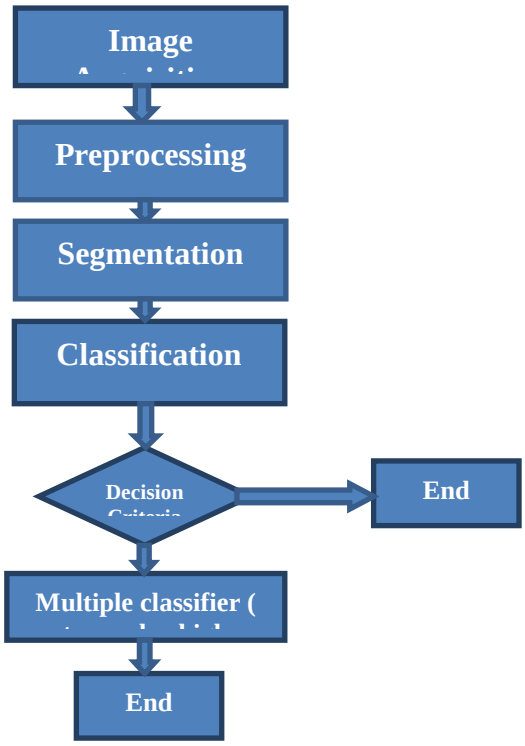


Fig.2.Proposed method for Tifinagh character recognition

A. Preprocessing

The input image will be preprocessed in order to perform noise cleaning.

Convert to gray scale, binary image, remove salt and paper noise. A Color, a gray level, or a binary image of Tifinagh will be binarized based on:

For each point (x,y), X is random variable [0,254]:

If I(x,y)<X : level of the selected point = 1

If I(x,y)>X : level of the selected point = 0

B. Segmentation

The segmentation algorithm will extract each part as separate characters. (The non-cursive characteristic of the Tifinagh Alphabets facilitates the preprocessing step). The segmentation step will be achieved in order to construct the input matrix (35*35).

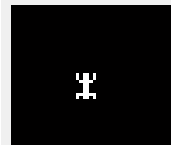
Image with noise	Image without noise	Segment
		

Fig.3. Segmentation step

The alphabets: all the 33 alphabets (sorted based on IRCAM alphabets) have 1 glyph except the alphabets(4,10,21) which have two glyphs and the alphabet 7 which has 3 glyphs.

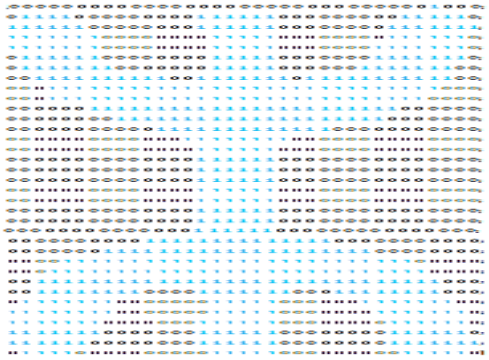
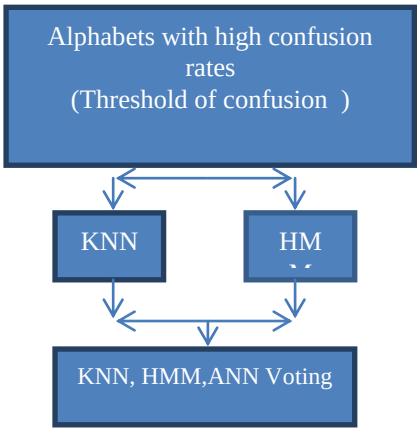


Fig.4. the alphabet YAZ (alphabet 32)

C. Classification



Our goal is to create a neural network that takes a picture of a Tifinagh character as an input and tells us which letter of the alphabet it represents. The neural network classifier will be used initially to classify the inputted tifinagh alphabets. For the alphabets with high rates of confusion a multiple classifier will be implemented.

D. The Back propagation Algorithm

The algorithm has as input the segmented alphabets (matrix 35*35) and outputs (vector column 33 alphabets). The training begins with random weights, and the goal is to adjust

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 5, November 2014

them so that the error will be minimal.
A: Activation function, X input and W the weight:

$$A_j(\bar{x},\bar{w}) = \sum_{i=0}^{n} x_i w_{ji}$$

Sigmoidal output (O) function:

$$O_j(\bar{x},\bar{w}) = \frac{1}{1+e^{A_j(\bar{x},\bar{w})}}$$

The goal of the training process is to obtain a desired output when certain Inputs (Matrix of Tifinagh Alphabet) are given. Since the error is the difference between the actual and the desired output, the error depends on the weights, and we need to adjust the weights in order to minimize the error E :

$$E_j(\bar{x},\bar{w},d) = \left(O_j(\bar{x},\bar{w}) - d_j\right)^2$$

The error of the network will be the sum of the errors of all the neurons in the output layer:

$$E(\bar{x},\bar{w},\bar{d}) = \sum_j \left(O_j(\bar{x},\bar{w}) - d_j\right)^2$$

$$E(\bar{x},\bar{w},\bar{d}) = \sum_j \left(O_j(\bar{x},\bar{w}) - d_j\right)^2$$

E. Multiple classifier

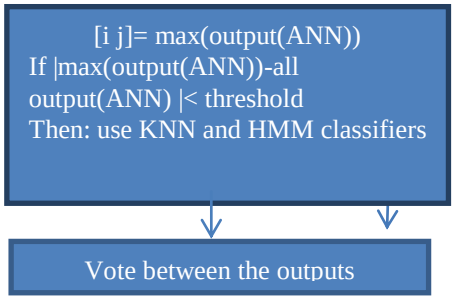


Fig.5. Multiple classifier step.

KNN

SampleS[i,j]=input image: matrix 35*35
Training [i,j]= is the references alphabets matrix alphabets 35*35
Group G[i]= [1…35]
class = knnclassify (sample, training, group)
How many times we have row i of sample is closest to row i of training:
For i=1 to 35
  For j=1 to 35
    If |S[i,j]- T[i,j]|<=ξ
  Then count=count+1

IV. PATTERN RECOGNITION WITH HMM

Based on training set on OCR-Cognex and training sets resulted from ANN:
V = {v1. . . v33} set of the vertices, and E = {(v1, v2), (v2, v5), (v5, v5), (v5, v4), (v5, v4)}….for the edges.V and E are finite sets.By a labeling of the vertices of the graph G = (V,E). The weights correspond to confusion rate (the results of Commercial tool cognex):

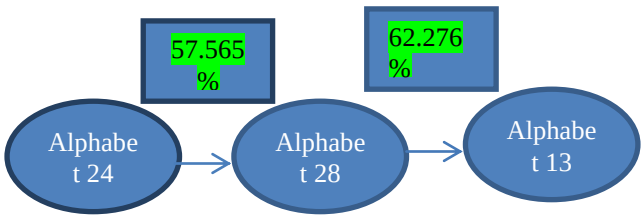


Fig.6. Example of Confusion probability based on commercial tool of .Cognex

	Alphabet1	Alphabet2
Alphabet1	A11	A12
Alphabet2	A21	A22

Alphabet2 has the values = [X1 X2 …..X33] generated from ANN.
Alphabet1 =[Y1 Y2 …Y33]generated from ANN.
E=[matrix (2 vectors : the output vector of ANN)
Transition matrix done from Cognex results.

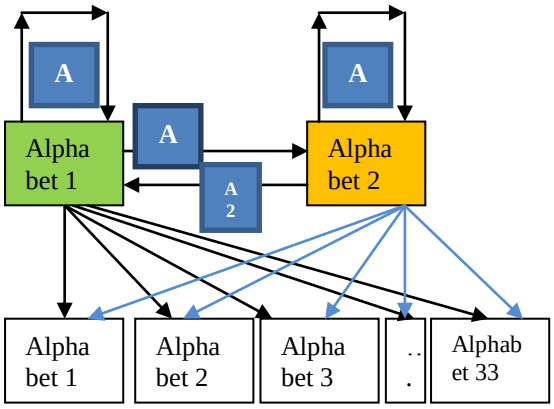


Fig.7.HMM presentation

The output of ANN is a 33 vector.
For each state: [1 2 3 4 5 6 7 8 9 10 11 …….33] : the ouput will be a probability sequence.
A sequence of states and emissions from a Markov model, will be generated.

V. OCR OF COGNEX CAMERA

A. Commercial tool Cognex Camera

Alphabets are {1,2…,33}, according to order of IRCAM Cognex OCR is one of the proprietary Optical Character Recognition tool, which also provides a good amount of accuracy, so we tried to perform OCR on the same set of images (Tifinagh Alphabets) to observe the result produced by OCR of Cognex. The processing time of Cognex OCR is

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 5, November 2014

also evaluated.

B. Overview of OCR Maw on Cognex[14]

Used for alphabets with one glyph, The OCR Max function performs Optical Character Recognition through a process of segmentation and classification. This is done by comparing the images of the segmented characters to the trained characters in the font.

C. TrainPatMaxPattern[14]

Used for alphabets with more than one glyph, A PatMax® pattern is a collection of geometric features where each feature is a point on the boundary between two regions of dissimilar pixel values. Train Pat Max Pattern trains a pattern, and then Find Pat Max Patterns is used to locate one or more instances of that pattern in an image.

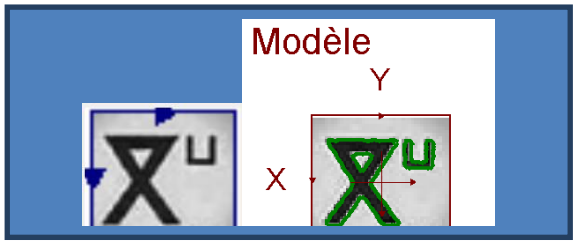


Fig.8.Alphabet with two glyphs.

VI. EXPERIMENTAL RESULTS

A. Cognex camera algorithm

Tests were done using images from IRCAM Web site, and Insight-Explorer 4.8.1 software was used to implement the job of Cognex camera.

Alphabets	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1						66%									
2		2%										25%			
3	8%		94%	20%			27%	46%			65%		76%		
9									57%	26%					
13															
14			94%												
15									57%						
16															
17															60%
18															
19														84%	
22															
25															
27															
28															
29					94%										
32															

Alphabets	17	18	19	21	22	23	24	25	26	27	28	29	30	31	32	33
1																
2																
3															50%	
9																
13											62%					
14																
15			84%													
16	61%	15%												66%		
17																
18				10%												
19												84%				
22								93%		78%			52%			
25					93%											
27						67%			49%							
28							57%									
29																
32																94%

Table.1.Confusion results of Cognex

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 5, November 2014

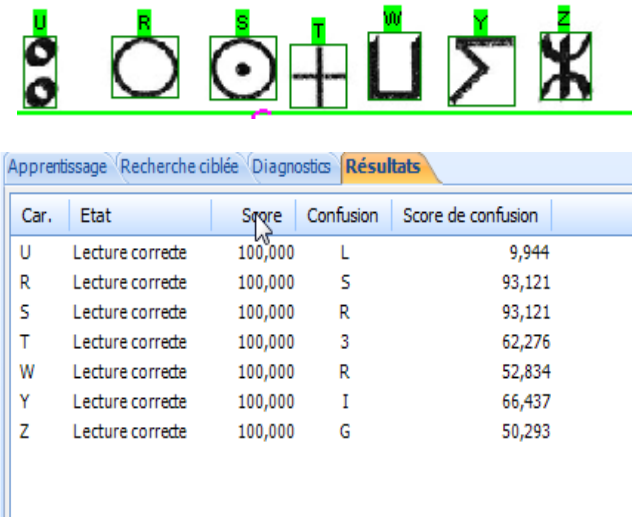


Fig.9. Some Results of recognition rates of Cognex

B. Proposed method

*a. ANN with a input matrix (7*5)*

To illustrate the accuracy of the proposed Tifinagh OCR algorithm, the performance was tested using images from IRCAM Web site. MATLAB 7.10.0 R2010a Version /64-bit is used to implement the Proposed Tifinagh OCR algorithm on an Dell Intel ® Core ™ CPU M640@2.8GHZ, RAM 4Go, machine running a 64-bit operating, system -MS Windows 7. For A matrix input 7*5:

Input Alphabet	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Recognized	1	12	3	3	5	16	17	18	10	10	11	12	3	3	5	16	17
Recognition rate	86%	50%	25%	25%	87%	49%	34%	49%	49%	49%	84%	50%	25%	25%	83%	49%	34%

Input Alphabet	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Recognized	18	19	17	21	22	23	24	25	26	27	28	29	30	31	32	33
Recognition rate	49%	81%	34%	80%	79%	85%	87%	80%	85%	85%	85%	83%	87%	86%	82%	82%

Table.2. Matrix input (7*5): Accuracy: 72.7% with an average confusion rate:36%.

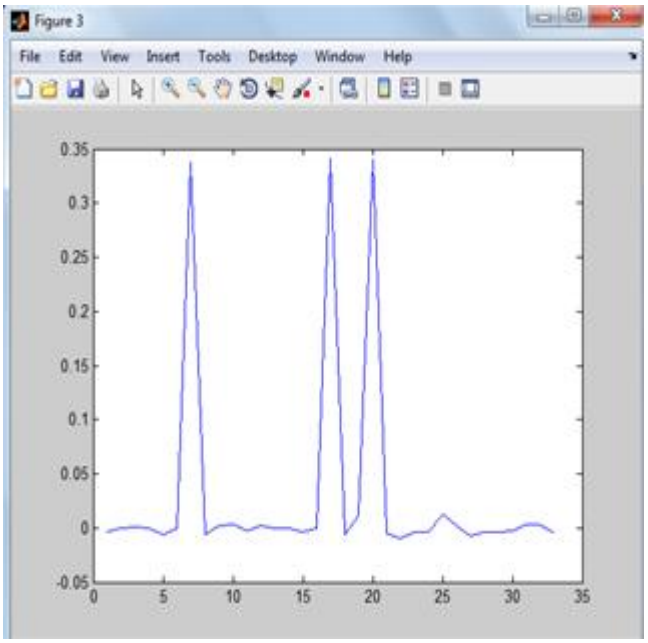


Fig.10. Confused alphabets 17, 20 and 7, the input alphabet is 20, the output one is 17.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 5, November 2014

*B. ANN with a input matrix (35*35)*

Alphabet	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Recognized	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Recognition rate	94%	93%	92%	92%	94%	93%	93%	94%	92%	92%	94%	94%	94%	93%	93%	94%	94%

Alphabet	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Recognized	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Recognition rate	93%	93%	82%	93%	92%	94%	94%	93%	94%	94%	92%	94%	94%	94%	0.93	94%

Table.3. Matrix input (35*35): Accuracy: 100% with an average confusion rate: 7%.

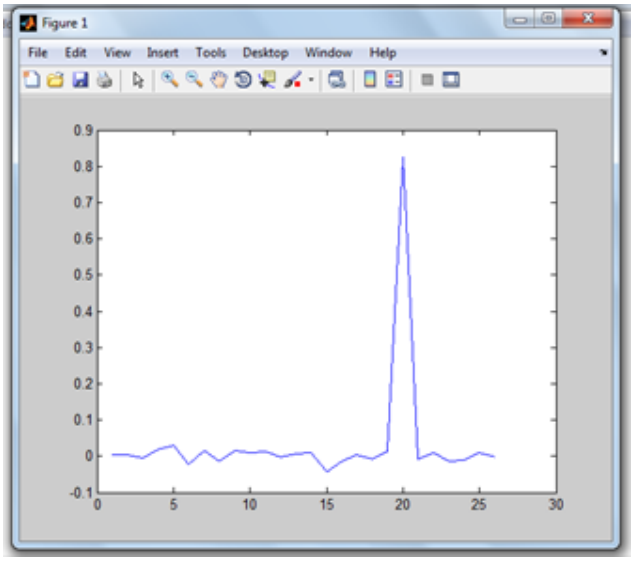


Fig.11. Alphabet 20 (was confused within 7*5 input matrix), and within 35*35 matrix input the confusion disappeared.

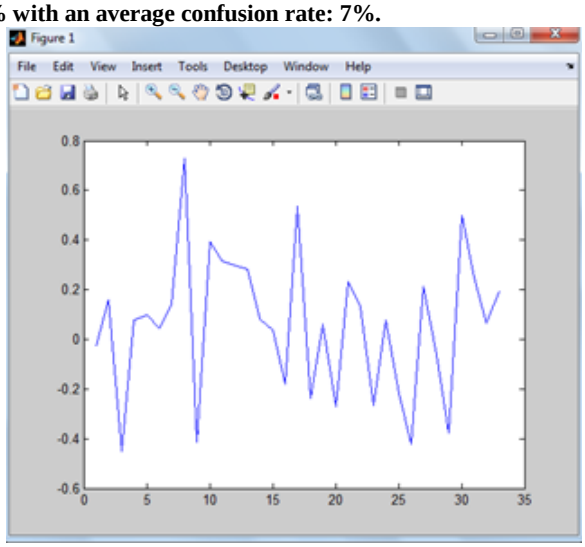


Fig.12. High rates of confusion (hand writing alphabet 6, result is 8 instead of 6) : high confusion rates with : 6→8,17 and 30

C. Use of multiple classifier for high confusion rates

Using hand writing alphabets, the confusion rates increase, as shown in the example below (read of the alphabet 6).
For HMM implementation: the vector will be : (0 for the other alphabets)

Alphabet	Output of ANN	Probability
8	0,7286	15%
17	0,535	11%
10	0,3906	8%
11	0,3161	6%
12	0,2997	6%
31	0,2482	5%
7	0,1409	3%
30	0,5008	10%
13	0,282	6%
33	0,1975	4%
2	0,1592	3%
22	0,1319	3%
5	0,0962	2%
14	0,0821	2%
24	0,0766	2%
6	0,0462	1%
15	0,0378	1%
21	0,2336	5%
32	0,0641	1%
19	0,0599	1%
4	0,0787	2%
27	0,2107	4%

Table.4.Use of HMM for alphabets with high confusion rates

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 5, November 2014

VI. CONCLUSION

In this project, we have, an OCR for Tifinagh alphabets. The OCR system has been quite successful in the recognition of input images. The experimental result shows us that with the test image (matrix of 35*35) with same font type the accuracy rate is 99 -100%. A comparison study of Cognex is carried out with proposed algorithm. Searches and comparisons are time consuming, however the accuracy of proposed method is high compared to other OCRs. The execution time for example for a 33 alphabets image of size 2500X2500 took around 2 seconds.

Method	Recognition rates-same font type(%)	Error rates(%)	Time consumption (33 alphabets)	average confusion rate
Cognex OCR	93.94	6.06	1 second	57%
Proposed algorithm (ANN with 7*5 matrix input)	72.7%	27.3%	1 second	36%
Proposed algorithm (ANN with 35*35 matrix input)	100%	0%	2 seconds	7%

Table.5.Recognition rates, Error rates and time consumption.

APPENDIX

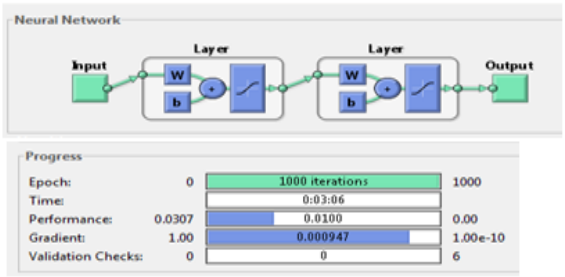


Fig.13. ANN training for 7*5 input matrix (Time: 3min 06 sec).

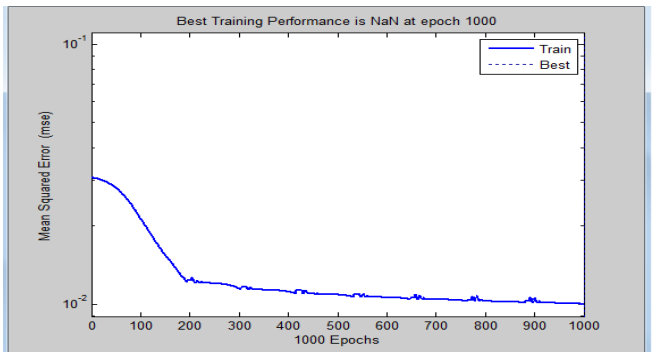


Fig.14.Training performance of matrix(7*5)

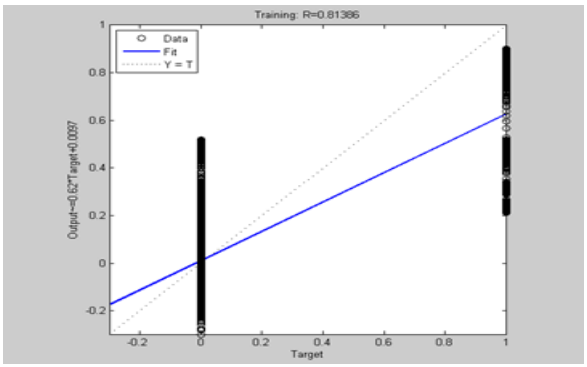


Fig.15. Regression of neural network in training stage (matrix input 7*5)

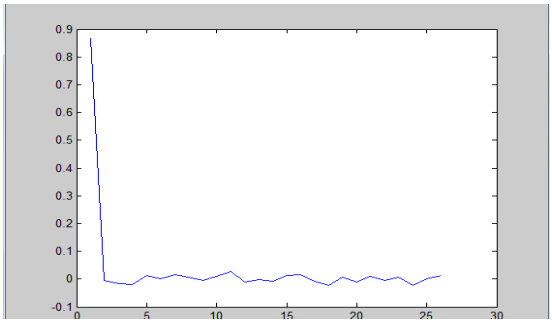


Fig.16. Non confused alphabet 2, confusion rate 0.7865, input matrix (7*5)

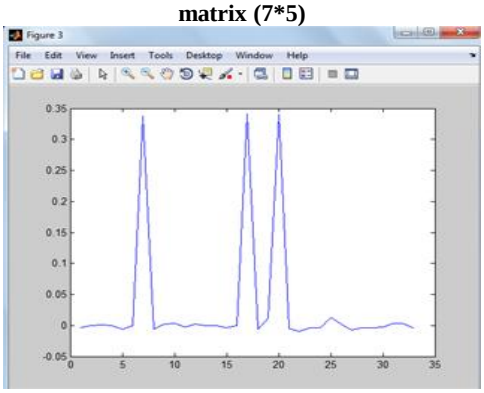


Fig.17. Confused alphabets 17, 20 and 7), the input alphabet is 20.

Fig.18.Alphabet 5 (YAD) segmented (matrix 35*35)

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 5, November 2014

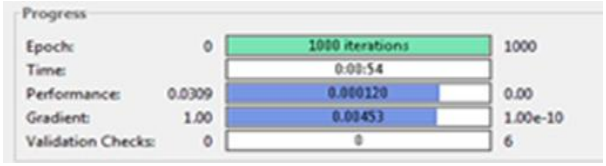


Fig.19.Training stage ANN with an input matrix 35*35

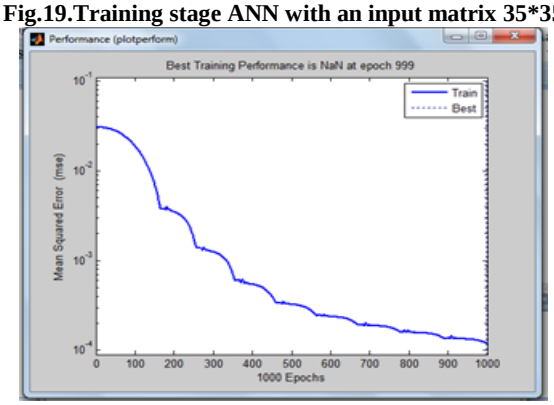


Fig.20.Training performance (matrix input 35*35)

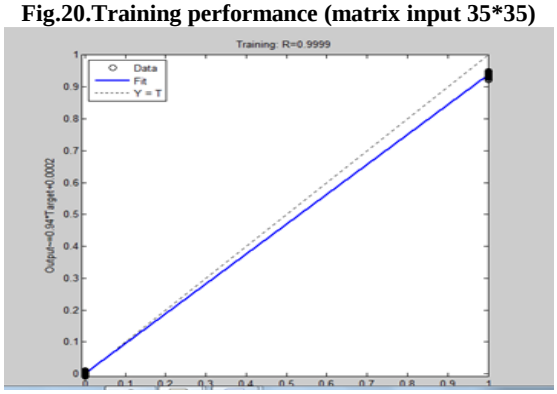


Fig.21.Regeression on ANN (input matrix 35*35)

REFERENCES

[1] Suruchi G. Dedgaonkar, Anjali A. Chandavale, Ashok M. Sapkal, "Survey of Methods for Character Recognition", International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 5, May 2012.

[2] Claus Bahlmann, Bernard Haasdonk, and Hans Burkhardt.On-line handwriting recognition with support vector machines. a kernel approach. In Proc. of the 8th IWFHR, pages 49.54, 2002.

[3] A. AL-Shatnawi, S. AL-Salaimeh, F. AL-Zawaideh and O. Khairuddin, "Offline Arabic Text Recognition-An Overview", World of Computer Science and Information Technology Journal (WCSIT), 1(2011), 184-192.

[4] Youssef EsSaady, Ali Rachidi, Mostafa El Yassa, DrissMammassIRF-SIC Laboratory, University Ibn Zohr, Agadir,Morocco, essaady2110@yahoo.fr, rachidi.ali@menara.ma, melyass@gmail.com, mammass@univ-ibnzohr.ac.ma,"Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character", International Journal of Advanced Science and Technology Vol. 33, August, 2011.

[5] Ahmad T. Al-Taani ahmadta@yu.edu.jo,Department of Computer Sciences Yarmouk University, Irbid, Jordan Saeed Al-Haj shaj@cs.nmsu.edu, Department of Computer Science, New Mexico State University, New Mexico, USA.

[6] "Recognition of On-line Arabic Handwritten Characters, Using Structural Features", JOURNAL OF PATTERN RECOGNITION RESEARCH 1 (2010) 23-37, Received January 15, 2010. Accepted July 2, 2010.

[7] O. Bencharef, M. Fakir, N. Idrissi, B. Bouikhalen et B.Minaoui, « Application de la géométrie riemannienne à la reconnaissance des caractères Tifinagh», Agadir-Morocco, 06-07 Mai 2011.

[8] R. El Ayachi, M. Fakir, B. Bouikhalene, S.Safi «OFFLINE PRINTED AMAZIGH SCRIPTS RECOGNITION». JATIT, vol. 20, No. 2, 2010.

[9] O. Bencharef,M. Oujaoura, B. Minaoui, M. Fakir and R. ElAyachi, « Recognition of Isolated Printed Tifinagh Characters»

[10] Y. Ait Ouguengay, M. Taalabi, "Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe: Phase d'apprentissage", Systèmes intelligents-Théories et applications, Paris: Europia, cop. 2009 (impr. au Maroc), ISBN-102909-285553, 2009.

[11] Y.ESSAADY, A.RACHIDI, M.ELYASSA, D.Mammass.2011. AMHCD: A database for Amazigh Handwritten character Recognition Research, International Journal of Computer Applications (0975 – 8887), Volume 27– No.4, August 2011.

[12] Cognex user guide (www.cognex.com).