ISSN 2229-5518

Handwritten Amazigh Character Recognition System for Image Obtained by Camera Phone

[1]Youssef Rachidi, [2]Zouhir Mahani
[1] Laboratory IRF-SIC, faculty of sciences Ibn Zohr University, Agadir, Morocco
[2] Laboratory of Engineering Sciences & Management of Energy, Ibn Zohr University, Agadir, Morocco
rachidi.8691@gmail.com, z.mahani@uiz.ac.ma

Abstract— In this paper; we introduce a system of automatic recognition of Amazigh characters based on the Random Forest Method in non-constrictive pictures that are stemmed from the terminals Mobile phone. After doing some pretreatments on the picture, the text is segmented into lines and then into characters. In the stage of characteristics extraction, we are representing the input data into the vector of primitives of the zoning types, of diagonal, horizontal, Gabor filters and of the Zernike moment. These characteristics are linked to pixels' densities and they are extracted on binary pictures. In the classification stage, we examine four classification methods with two different classifiers types namely the Support vector machines (SVM) and the Random Forest method. After some checking tests, the system of learning and recognition which is based on the Random Forest has shown a good performance on a basis of 100 models of pictures.

Index Terms— Handwritten Amazigh Character Recognition, Tifinagh, mobile phone, Zoning, Gabor Filters, Zernike Moments, SVM, Random Forest.

---

1 INTRODUCTION

The automatic recognition of handwritten or printed characters Amazigh a subject of research and experimentation. The problem is not yet solved despite the fact that results have reached fairly high rates in some applications [1]. Some attempts have been done to improve the current situation [1]. In this context, we have employed a recognition system of handwritten Amazigh characters extracted from a picture taken by camera phone [2]. Indeed, in the primitives' extraction stage, our approach is based on primitives of the Zoning types [3], of Diagonal [4], of Horizontal, of Gabor Filters and of the Zernike's moment [5] [6] [7] [8]. These primitives will supply a Random Forest in the learning and recognizing phases. On a database of handwritten Amazigh, segmented and isolated characters acquired by camera phone, obtained an encouraging results on the majority of this characters. The limit of this adopted approach is that it is not operational on some extracted characteristics of Zernike's moments [9]. The database contains 100 samples of 33 classes, collected from 5 different writers. As a result the database consists of 3300 samples. For classification stage we have used two classifiers: the Support Vector Machine (SVM) and the Random Forest and for each classifier we employed a set of different features extraction methods

Habitually, the phases form the structures of handwriting recognition system are: Pre-processing, Segmentation, Feature extraction, Classification and Post-processing [2].

In this paper, our objective is mainly interested in the development of handwriting Amazigh character recognition system and Improvement of the Recognition Rate by Random Forest, in which the images are obtained by camera phone.

The paper is organized as follows. In section 2, the proposed the pre-processing and gives descriptions of the methods that we used throughout the OCR process, which includes the following stages: Binarization, Noise removing, skew detection and correction and Segmentation. The Database of Characters Amazigh (Tifinagh) in the section 3. The feature extraction procedure adopted in the system is detailed in the section 4. Section 5 describes the classification and recognition using support vector machines (SVM), and Random Forest. Section 6 presents the experimental results and comparative analysis. Finally, the paper is concluded in section 7.

2 PRE-PROCESSING

The procedure of preprocessing which refines the scanned input image includes several steps: Binarization, for transforming gray-scale images in to black and white images, noises removal, and skew correction performed to align the input paper document with the coordinate system of the scanner and segmentation into isolated characters [1].

2.1 Binarization and Noise Remo- oval

We used the Otsu method for binarization [10] this method of thresholding is performed as a preprocessing step to remove

ISSN 2229-5518

the background noise from the picture prior to extraction of characters and recognition of text. Fig.1 (a) shows a sample input handwritten character image and Fig.1(b) shows the binarized image after the thresholding step using Otsu method.

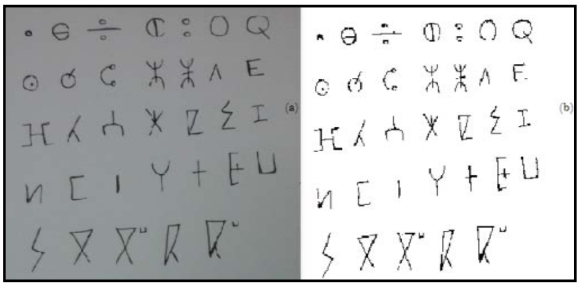


Fig1. (a) Example of an input image, (b) Thresholded image with Otsu method.

Noise which is in the images is one of the big difficulties in optical character recognition process. The aim of this part is to remove and eliminate this obstacle; there are several methods that allow us to overcome this problem. In this work we decided to use the morphology operations to detect and delete small areas of less than 30 pixels [2].

2.2 Skew detection and correction

Skew correction methods are used to align the paper document with the coordinate system of the scanner. Main approaches for skew detection include line correlation [11], projection profiles [12], Hough transform [13], etc. For this purpose two steps are applied. First, the skew angle is estimated. Second, the input image is rotated by the estimated skew angle. In this paper, we use the Hough transform to estimate a skew angle θs and to rotate the image by θs in the opposite direction.

2.3 Segmentation

Next step for OCR is the Segmentation of the image. In This paper we propose a segmentation algorithm, in which text is easily segmented into Lines and Words using the traditional vertical and horizontal projection [6].

2.3.1 Line Segmentation

Once the image of the text cleaned, the text is segmented into lines. This is used to divide text of document into individual lines for further preprocessing. For this, we used analysis techniques of horizontal projection histogram of the pixels in order to distinguish areas of high density (lines) of low-density areas (the spaces between the lines) (see Fig.2). These techniques were often used to extract lines in printed texts [1].

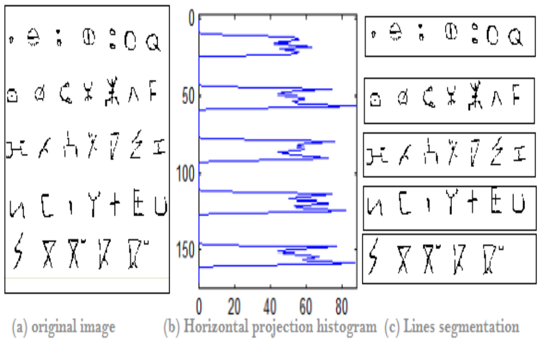


Fig.2 Lines segmentation

2.3.2 Characters Segmentation

We used in this part the vertical projection histogram to segment each text line of characters. Fig.3 shows a text line, the vertical histogram and the result of segmentation into characters [2].

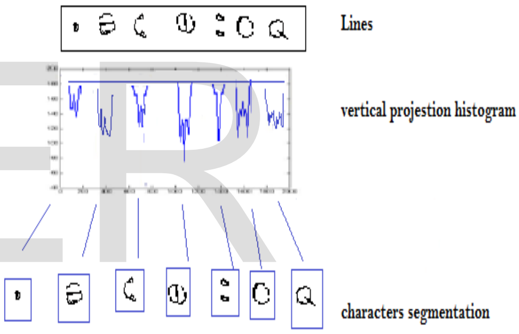


Fig.3 Characters segmentation

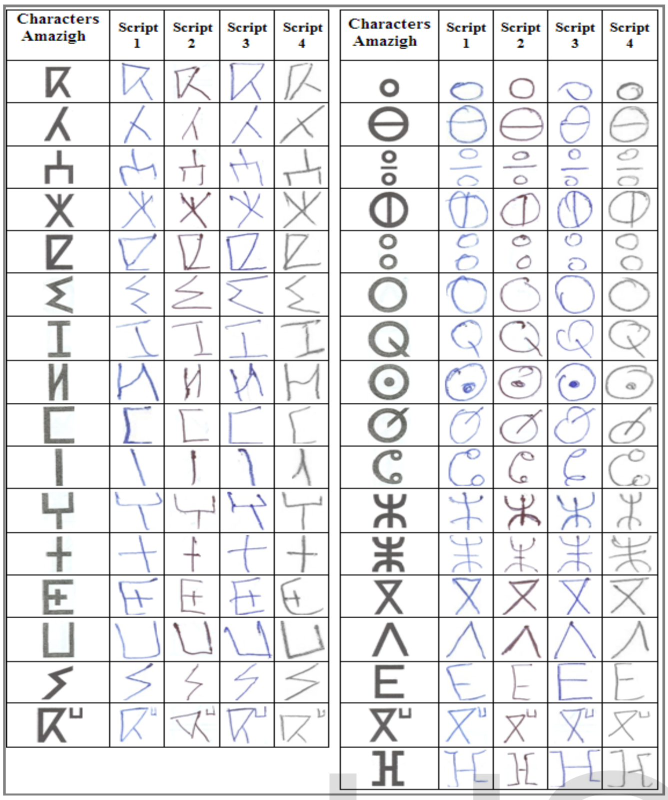
3 BASE TIFINAGH (CHARACTERS AMAZIGH)

To evaluate the performance of the proposed method, the experiments were performed on both bases of the Amazigh patterns: a printed database of Amazigh characters and another one for handwritten characters created locally. The first is a database of printed Amazigh patterns of different fonts and sizes. It contains a total of 12 fonts and sizes of 10 points to 28 points for each model.

Our database of isolated Amazigh handwritten characters was collected from 25 people. The samples were collected by asking the participants to write on a form 4 examples for each Amazigh character, in total, we obtained 100 variations for each character led to an overall of 3300 isolated Amazigh handwritten characters (100×33). Some Amazigh handwriting samples are given in Table 1.

Table1. Some Amazigh Handwriting Samples

ISSN 2229-5518



4 FEATURES EXTRACTIONS

In This part we present some feature extraction methods for recognition of segmented (isolated) characters [15]. Selection of a feature extraction method is probably the single most important factor in achieving high recognition performance in character recognition systems. Different feature extraction methods are designed for different representations of the characters, such as solid binary characters, character contours, skeletons (thinned characters) or gray-level sub-images of each individual character[15], In this paper, we have tested four methods: Zoning Method, Diagonal based feature extraction, Horizontal method and Zernike Moments.

4.1 Zoning

We have proposed a statistical feature extraction method named zoning [3]. In this proposed method, resized individual image of size 50*60 pixels is divided into 30 equal zones or blocks each of size 10*10 pixels. The features are extracted by counting the number of black pixels in each zone.

4.2 Diagonal Based Feature Extraction

These features are extracted from the pixels of each zone by moving along its diagonals. Following algorithm describes the computation of Diagonal Features for each character image of size 50*60 pixels having 5*5 zones and thus each zone having 10*10 pixel sizes [4]. Each of these zones is having 9 diagonals. The number of foreground pixels along each diagonal are summed up to get 9 features from each zone, then these features for each zone are averaged to extract a single feature from each zone [4].

4.3 Horizontal Based Feature Extraction

These features are extracted from the pixels of each zone by moving along its horizontal. Following algorithm describes the Computation of Horizontal Features for each character image of size 50*60 pixels having 5*5 zones and thus each zone having 10*10 pixel sizes.

4.4 Zernike Moments

Moment descriptors have been studied for image recognition and computer vision since the 1960s [5]. Teague first introduced the use of Zernike moments to overcome the shortcomings of information redundancy present in the popular geometric moments [6, 7]. Zernike moments are a class of orthogonal moments and have been shown effective in terms of image representation.

The Zernike moments [8] of order n and repetition m are defined as follows of an image I (x, y):

$$\mathbf{Z_{mn}} = \frac{\mathbf{m+1}}{\boldsymbol{\pi}} \iint_{\mathbf{xy}} \mathbf{I(x,y)[V_{mn}(x,y)]dxdy} \qquad (1)$$

Where $V_{mn}(x,y)$ is represented in polar coordinates as follows:

$$\mathbf{V_{mn}(r,\theta) = R_{mn}(r)e^{-jn\theta}} \qquad (2)$$

Where $R(r)$ is the orthogonal radial polynomial given as:

$$\mathbf{R_{mn}(r)} = \sum_{\mathbf{s=0}}^{\frac{\mathbf{m-|n|}}{\mathbf{2}}} (\mathbf{-1})^{\mathbf{s}} \frac{\mathbf{(m-n)!}}{\mathbf{s!}\left(\frac{\mathbf{(m+|n|)}}{\mathbf{2}}-\mathbf{s}\right)!\left(\frac{\mathbf{(m-|n|)}}{\mathbf{2}}-\mathbf{s}\right)!}\mathbf{r^{m-2s}}$$

(3)

4.5 Gabor filters

As a powerful feature, the Gabor filters [21] have been successfully applied in numerous pattern recognitions including face recognition fingerprint recognition …, as well as optical characters recognition. The Gabor filters are defined by a complex sinusoidal modulated by a Gaussian envelope described as follows:

$$G(x,y,\theta,f)=e^{-\frac{1}{2}\left(\frac{R_1^2}{\sigma_x^2}+\frac{R_2^2}{\sigma_y^2}\right)}\cos(2\pi f x_\theta) \quad (4)$$

Where:

$$R_1 = x\cos(\theta)+y\sin(\theta)$$
$$R_2 = y\cos(\theta)-x\sin(\theta)$$

f represents the frequency of the Sinusoidal plane wave along the direction θ, and (σx,σy) explain the standard deviations of the Gaussian envelope along x and y directions [21].

Table 2. Combination of the different feature vectors

Feature Method	Contained Feature	Size
FM1	Zoning	30
FM2	P.Diagonal	30
FM3	P.Horizental	30
FM4	Zernike Moments	32
FM5	Gabor Filters	32

5 CLASSIFICATIONS

In the complete process of system recognition of forms, the classification plays an important role by pronouncing on the membership of a shape in a class. The main idea of the classification is to attribute an example (A form) not known about one Class predefined from the description in parameters of the form. Several surrounding areas of classification are used in the field of recognition of forms which are more or less good adapted to the recognition of the writing.
In litterateur, there are many types of classifiers that have been implemented in handwritten optical character Amazigh recognition problems. Among them, in this paper we have used two classifiers: the Support Vector Machine (SVM) and Random Forest.

5.1 Support vector machines (SVM)

The Support vector machines (SVM) are a category that belongs to the supervised machine learning models that can be used to classifications or regressions problems, they proposed by Vapnik [22]. SVM modeling was originally used to optimize the linear hyper plane which separate two classes, That is to say, the empty region around the decision boundary determined by the distance to the nearest training pattern [22].

We consider the problem of classification the group of training data (xi,yi) i=1,…,l into two classes, where y i ∈{-1,+1} is the class (label) and xi is a feature vector. In the classification with SVM model a label yi will be assigned to a feature vector xi by evaluating:

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i K(x_i,x)+b\right) \quad (5)$$

Where αi corresponding to the weighs and b was the bias, these two variables are called SVM parameters and adopted into training by maximizing:

$$L_D=\sum_i x_i-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j S_i S_j K(P_i,P_j) \quad (6)$$

Within the constraints:

$$\sum_i x_i S_i = 0 \quad \text{And} \quad 0\le\alpha_i\le c \quad (7)$$

Where C is a positive constant, and K (Pi,Pj) is named the Kernel function of the SVM model. In this paper, we use the polynomial Kernel; it's given by the following form:

$$K(P_i,P_j)=(P_i * P_j+1)^d \quad (8)$$

In the case where the classes are not linearly Separable, it necessary to combine several SVMs models to solve a multi-class classification problem. The strategy is the "one against all" which involves building a SVM per class, so each classifier trained to distinguish the data of his class from those of all other classes. Another classic strategy is to use "one against one" which building a SVM classifier for each pair of classes. In this work, we have selected "one against one" strategy for multi-class classification.

5.2 Random Forest

Random forest is an ensemble training algorithm that constructs multiple decision trees. It suppresses over- fitting to the training samples by random selection of training samples for tree construction in the same way as is done in bagging(Breiman, 1996)[18],(Breiman, 1999)[19], resulting in construction of a classifier that is robust against noise. Also, random selection of features to be used at splitting nodes enables fast training, even if the dimensionality of the feature vector is large [1].

- **Algorithm**

$z=\{(x_1,y_1),\dots(x_n,y_n)\}$ learning sample, x_i describes nominal variables p explanatory [20]:

1. for b=1 to B (B number of trees)

(a) Draw a boostrap sample z_b of size N from the training data

(b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

i. Select m variables at random from the p variable

ii. Pick the variable/split-point among the m

iii. Split the node into two daughter nodes

2. Output the ensemble of tree $\{T_b\}_1^B$

To make a prediction at a new point x:

Regression:

$$\boxed{\hat{f}^B_{\mathrm{rf}}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)} \quad (4)$$

Classification: let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then

$$\boxed{\hat{C}^B_{\mathrm{rf}}(x) = majority\ vote\ \{\hat{C}_b(x)\}_1^B} \quad (5)$$

Why Random Forest works [20]

Mean Squared Error = Variance $+\text{Bias}^2$

If trees are sufficiently deep, they have very small bias

How could we improve the variance?

$$\mathrm{var}\left(\frac{1}{B}\sum_{i=1}^{B} T_i(c)\right) = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B} Cov(T_i(x), T_j(x)) \quad (6)$$

6 EXPERIMENTAL RESULTS

The database contains 100 samples of 33 classes, collected from 5 different writers. As a result the database consists of 3300 samples. The samples are divided randomly into two set, one for training stage, we have used 85 % (2805 samples) and the other for testing stage, we have used 15 % (495 samples).

We have tested the proposed system on database of handwritten Amazigh characters acquired by camera phone the SAMSUNG Galaxy S4 of this characteristics; 13 Megapixels (4 128×3 096 px).

All the experiments are carried out in MATLAB 7.9 environment with using a PC with windows 7 as an operating system equipped with Intel (R) Core (TM) i7-3337U processor 1.80 GHz and 4 Go RAM. For classification stage we have used two classifiers: the Support Vector Machine (SVM) and the Random Forest and for each classifier we employed a set of different features extraction methods.

The Zoning feature extraction provides higher recognition and learning rate, with the achievement of a rate of 93.17 % and 97.75 % as recognition accuracy, respectively for SVM and Random Forest. But in the case of Zernike moment we have the recognition rate for SVM is 51.09 and 56.43% for Random Forest.

Table.3 Results (1) of different single feature vectors using Support Vector Machine and Random Forest classifiers

Classifier Feature Vector	Support Vector Machine (SVM)		Random Forest	
	Learning R.	Recognition R.	Learning R.	Recognition R.
FM1	93.21%	93.17%	97.98%	97.75%
FM2	96.59%	94.64%	97.74%	95.53%
FM3	96.42%	96.07%	97.76%	97.56%
FM4	57.61%	51.09%	71.20%	56.43%
FM5	75.36%	73.43%	94.82%	89.19%

We made the choice of the platform Weka (Witten and Frank, 2005) to realize training. They have work by SVM for compare the results of random forest .

All the results obtained using the two classifiers are compared in Table 3. In fig.4 and fig.5 we present the learning and recognition rate by Random Forest and SVM

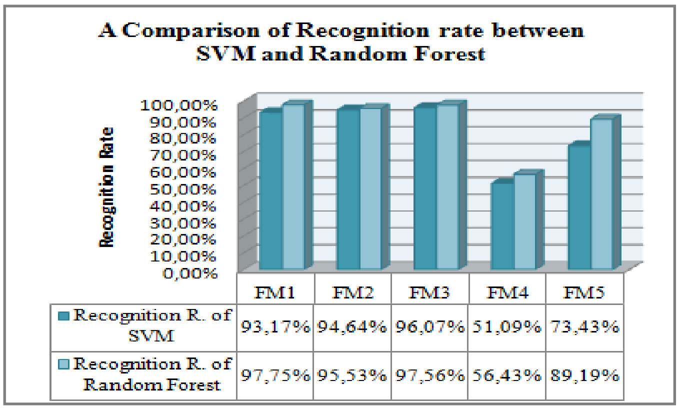


Fig.4.A Comparison of Recognition rate between SVM and Random Forest

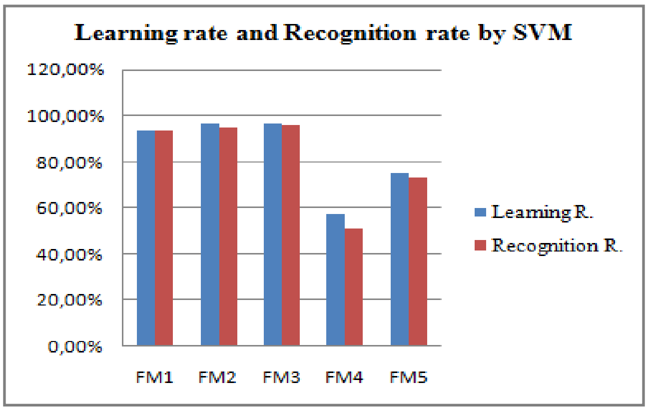


Fig.5Learning rate and Recognition rate by SVM

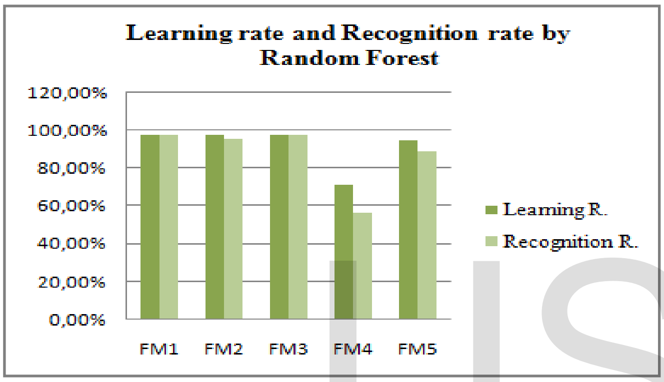


Fig.6 learning rate and Recognition rate by Random Forest

According to these results, it can be noticed that the best results always are obtained with Random Forest (number of trees is 500).

In SVM classifier, we obtained the highest result with the use of polynomial Kernel and the value 1 for cost parameter C. the test shown that its performance is mediocre at out recognition problem.In term of performance and efficiency, we can conclude that Random Forest classifier with Zoning feature method and P.Horizental feature are powerful tools for solving the problem of handwriting optical characters recognition

7 Conclusion

In this paper, we have presented a system of handwriting Amazigh character recognition based on the method Random Forest. Several features have been studied and compared; as a result we've chosen Otsu method duo its ability to remove the noise.

The experiments carried out in database were performed on a database obtained by camera phone with applying different classifiers and for each classifier we have tested a set of single feature methods. The results obtained in this paper that has Been compared and analyzed have shown that Random Forest with Zoning feature and P.Horizental are the best in terms of recognition accuracy rate.

In future work, we will add other features methods that improve the results for some.

.

REFERENCES

[1] Svetnik, A. Liaw, C. Tong, J. Culberson, R. Sheridan, and B. Feuston. Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43:1947–1958, 2003.

[2] Hassan El Bahi, Zouhir Mahani and Abdelkarim Zatni "A robust system for printed and handwritten character recognition of images obtained by camera phone" .*Published in WSEAS Transactions on Signal Processing, Volume 11*, 2015, pp. 9-22

[3] Elima Hussain, Abdul Hannan, Kishore Kashyap, "A Zoning based Feature Extraction method for Recognition of Handwritten Assamese Characters" *IJCST Vol. 6, Issue 2*, April - June 2015

[4] Anita Jindal, RenuDhir and Rajneesh Rani, " Diagonal Features and SVM Classifier for Handwritten Gurumukhi Character Recognition ", *International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 5*, May 2012 ISSN: 2277 128X .

[5] Teh, C. and Chin, R.T. On Image Analysis by the Methods of Moments.*IEEE Trans*. on PAMI, *10 (4).* 496-513

[6] Lipscomb, J.S. A Trainable Gesture Recognizer. Pattern Recognition, 24 (9).895-907.

[7] Teague, M.R. Image Analysis via the General Theory of Moments. *Journal of the Optical Society of America, 70 (8). 920-930*

[8] S. K. Hwang, W. Y .Kim, "A novel approach to the fast computation of Zernike moments", Pattern Recognition (36), pp. 2065– 2076, 2006

[9] Chesner D'esir, Simon Bernard, Caroline Petitjean, Laurent Heutte "One Class Random Forests" Pattern Recognition, *Elsevier, 2013, 46*, pp.3490-3506

[10] N.Otsu "A threshold selection method from gray-level histograms", IEEE Trans. Sys,Man., Cyber, vol. 9, 1979, pp. 62–66

[11] H.Yan, "Skew correction of document images using interline cross-correlation", *CVGIP: Graphical Models Image Process 55*, 1993, 538-543.

[12] T. Pavlidis and J. Zhou, "Page segmentation and classification", *Comput. Vision Graphics Image Process. 54*, 1992, 484-496.

[13] D. S. Le, G. R.Thoma and H. Wechsler, "Automatic page orientation and skew angle detection for binary document images", *Pattern Recognition 27*, 1994, 1325-1344.

[14] Archana A. Shinde, D.G.Chougule, "Text Pre-processing and Text Segmentation for OCR" *IJCSET |January 2012| Vol 2, Issue 1,810-812*

[15] Ivind Due Trier , Anil K. Jain, TorfinnTaxt, "Feature extraction methods for character recognition a survey" Revised July 19, 1995

[16] K.W. Wong, C.S. Leung & S.J. Chang, "Handwritten digit recognition using multilayer feedforward neural networks with periodic and monotonic activation functions", *ICPR, vol.3,*2002, pp. 106–109.

[17] S. Singh, and A. Amin, "Neural Network Recognition of Hand Printed Characters", *Neural Computing and Applications*, *vol. 8, no.1*, 1999, pp. 67-76.

[18] Breiman, L. (1996). Bagging predictors. In Machine Learning, Springer.

[19] Breiman, L. (1999). Using adaptive bagging to debias regressions. In Technical Report. Statistics Dept. UCB.

[20] G.Louppe, P.Geurts "Understanding Random Forests" July 2014, University of Liège.

[21] J.G. Daugman, " Two-dimensional spectral analysis of cortical receptive field profile, "Vision Research, 20, pp. 847-856, 1980.

[22] Vapnik. V. 1995. The Nature of Statistical Learning Theory. Springer, N.Y. ISBN 0-387-94559-8.