RESEARCH ARTICLE OPEN ACCESS

Improvement of the Recognition Rate by Random Forest

Youssef Rachidi *, Zouhir Mahani **
** (Laboratoire Image et Reconnaissance de Formes – Systèmes Intelligents et Communicants (IRF – SIC), Université Ibn Zohr Agadir, Maroc*
*** (Laboratoire Des Sciences de l'Ingénieur et Management de l'Energie, Université Ibn Zohr Agadir, Maroc*

ABSTRACT
In this paper; we introduce a system of automatic recognition of 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 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 multi-layer perceptron (MLP) 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.
Keywords: Handwritten Character Recognition, Mobile phone, Random Forest, Zoning, Zernike Moments.

I. INTRODUCTION

The automatic recognition of handwritten or printed characters remains 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 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], Horizontal 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, 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]. To rememdy these limits, we suggest a new method based on the Random Forest which should render the increase of the rate of recognition possible.

II. 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]. 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 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 II, 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 feature extraction procedure adopted in the system is detailed in the section III. Section IV describes the classification and recognition using propagation neural network and Random Forest. Section V presents the experimental results and comparative analysis. Finally, the paper is concluded in section VI.

2.1 Binarization and Noise Remo- oval

We used the Sauvola method for binarization [10] this method of thresholding is performed as a preprocessing step to remove the background noise from the picture prior to extraction of characters and recognition of text. Fig.3 (a) shows a sample input handwritten character image and Fig.3(b) shows the binarized image after the thresholding step using Sauvola method.

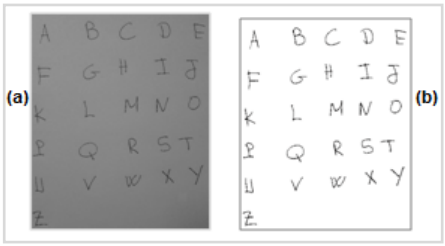


Fig3. (a) Example of an input image, (b) Thresholded image with Sauvola method.

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.4). These techniques were often used to extract lines in printed texts [1].

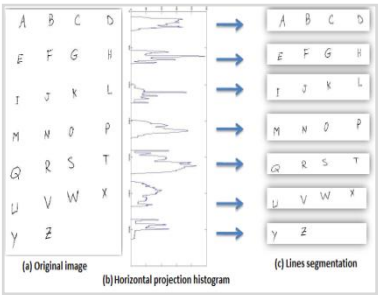


Fig.4 Lines segmentation

2.3.2 Characters Segmentation

We used in this part the vertical projection histogram to segment each text line of characters. Fig.5 shows a text line, the vertical histogram and the result of segmentation into characters [2].

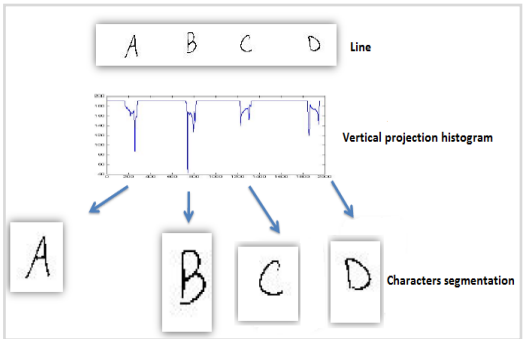


Fig.5 Characters segmentation

III. FEATURE EXTRACTION

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.

3.1 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.

3.2 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].

3.3 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.

3.4 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):

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

Where $Vmn\ (x,y)$ is represented in polar coordinates as follows:

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

Where $R_{mn}(r)$ is the orthogonal radial polynomial given as:

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

IV. CLASSIFICATION

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 recognition problems. Among them, in this paper we have used two classifiers: the multi-layer perceptron (MLP) artificial neural network and Random Forest.

4.1 Artificial Neural Network

Artificial Neural network (ANN) is Parallel distributed processes which allow the learning and the recognition. They knew a big success from the Nineties.

The main idea is that a formal neuron is capable of doing elementary calculations as separation vector in two classes, every class being determined by the weight of the neuron. The problem is then to choose Coefficients to be allocated to the weights to realize an optimal separation. There are many types of neural networks; in our work we have selected the Multilayer Perceptron (MLP) [16] [17].

Multi-layer perceptron architecture (MLP) using the back propagation with momentum learning scheme. The multilayer perceptron architecture have three layers of neurons, input layer of information processing nodes, the last is the output layer determined by the total number of classes' recognition, and at least a hidden layer with hidden nodes[1]. The Fig.9 presents the Artificial Neural Networks (ANN) Architecture [16] [17].

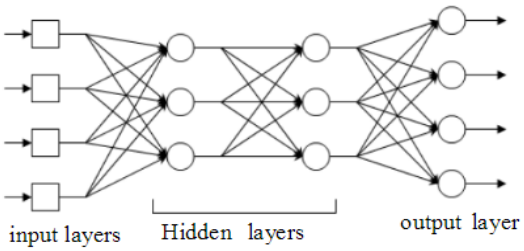


Fig.9 Presents the Artificial Neural Network Architecture

In this paper, the network learned on the entire training set using the back propagation method, and was then tested (with the validation set) for its performance with a limit of 700 epochs. The network is trained with 0.3 learning rate and 0.2 momentums constant.
We worked for activation function the sigmoid function defined by the following:

$$f(x) = \frac{1}{1+e^{-x}} \quad (3)$$

4.2 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**

$z=\{(x_1, y_1), \ldots (x_n, y_n)\}$ 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:

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

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

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

Why Random Forest works [20]
Mean Squared Error = Variance + Bias^2
If trees are sufficiently deep, they have very small bias
How could we improve the variance?

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

V. EXPERIMENTAL RESULTS

The database contains 100 samples of 26 classes, collected from 4 different writers. As a result the database consists of 2600 samples.

The samples are divided randomly into two set, one for training stage, we have used 85 % (2210 samples) and the other for testing stage, we have used 15 % (390 samples). We have tested the proposed system on database of handwritten characters acquired by camera phone the SAMSUNG Galaxy S4 of this characteristics; 13 Megapixels (4 128×3 096 px).

For classification stage we have used two classifiers: the Multilayer perceptron (MLP) 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 97.43 % and 100 % as recognition accuracy, respectively for MLP and Random Forest. But in the case of Zernike moment we have the recognition rate for MLP is 48.22 and 47.69% for Random Forest.

Table.1 Results of different single feature vectors using Multilayer Perceptron and Random Forest (NumTree = 10) classifiers

	MLP (Num.Epoch =500)		RandomForest (NumTree=10)	
	Learning Rate	Recognition Rate	Learning Rate	Recognition Rate
Zoning	97,72%	97,43%	99,76%	100%
P.Diagonal	97,59%	92,56%	99,69%	92,30%
P.Horizontal	97,22%	97,31%	99,58%	99,61%
M.Zernike	71,29%	48,20%	99,66%	47,69%

We made the choice of the platform Weka (Witten and Frank, 2005) to realize training and testing the method suggested.
For Recognition rate increases, we can go about reducing variance by increasing number of trees B. So we have (6):

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

$$= \frac{1}{B^2}\sum_{i=1}^{B}\left[\sum_{j\neq 1}^{B} Cov(T_i(x), T_j(x)) + \mathrm{var}(T_i(x))\right]$$

$$= \frac{1}{B^2}\sum_{i=1}^{B}((\mathrm{B}-1)\,\sigma^2.\rho + \sigma^2)$$

$$= \frac{\mathrm{B}(\mathrm{B}-1)\rho\sigma^2 + \mathrm{B}\sigma^2}{\mathrm{B}^2}$$

$$= \rho\sigma^2 + \sigma^2\frac{1-\rho}{\mathrm{B}}$$

Decreases, if ρ decreases, i.e., if m decreases

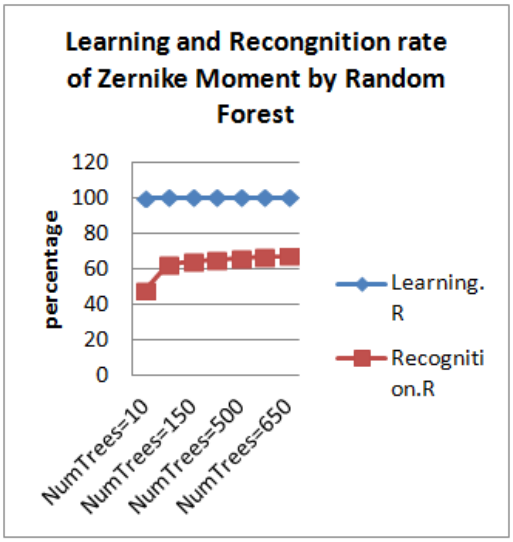
Decreases, if number of trees B increases (irrespective ofρ)

As the size of the ensemble gets arbitrarily large, i.e., as $\mathrm{B} \to \infty$, the variance of the ensemble reduces to $\boldsymbol{\rho\sigma^2}$.Under the as-sumption that randomization has

some effect on the predictions of randomized models, i.e., assuming **□(x) < 1**, the variance of an ensemble is therefore strictly smaller than the variance of an individual model. As a result, the expected generalization error of an ensemble is strictly smaller than the expected error of a randomized model. As such, improvements in predictions are solely the result of variance reduction, since both noise(x) and bias² (x) remain unchanged. Additionally, when random effects are strong, i.e., when **□(x) □ 0**, variance reduces to $\sigma^2 \frac{1}{B}$, which can further be driven to 0 by increasing the size of the ensemble. On the other hand, when random effects are weak, i.e., when **□(x) □ 1,** then variance reduces to σ^2, and building an ensemble brings no benefit. Put otherwise, the stronger the Random effects, the larger the reduction of variance due to ensembling, and vice-versa.
In Zernike Moment, If we change NumTree =10 to 110,150,300,500,600 and 650, will we get an increase in rate of recognition from 47.69% to 67.12%.

Table.2 Improvement of the Recognition Rate by Random Forest of Moment Zernike

		M. Zernike
Random Forest (NumTrees = 110)	Learning.R	**100 %**
	Recognition.R	**62.30 %**
RandomForest (NumTrees = 150)	Learning.R	**100 %**
	Recognition.R	64.02 **%**
RandomForest (NumTrees = 300)	Learning.R	**100 %**
	Recognition.R	64.87 **%**
RandomForest (NumTrees = 500)	Learning.R	**100 %**
	Recognition.R	65.64 **%**
RandomForest (NumTrees = 600)	Learning.R	**100 %**
	Recognition.R	66.92 **%**
RandomForest (NumTrees = 650)	Learning.R	100 **%**
	Recognition.R	67.12 **%**



There is no change in recognition rate of 700 trees (if NumTree > 650 Then Recognition rate = 67.17%), the rate remain stable When number of trees is more than 650. So we increased the recognition rate of Zernike Moment by Random Forest and the same others feature extraction

Table.3 Results of different single feature vectors using Multilayer Perceptron and Random Forest (NumTree = 650) classifiers

	MLP (Num.Epoch =500)		RandomForest (NumTree=650)	
	Learning Rate	Recognition Rate	Learning Rate	Recognition Rate
Zoning	97,72%	97,43%	99,98%	100%
P.Diagonal	97,59%	92,56%	99,92%	95,38%
P.Horizontal	97,22%	97,31%	100,00%	99,84%
M.Zernike	71,29%	48,20%	100,00%	67,12%

So, they have improve the recognition rate of Zernike moment by random forest from 47,69 to 67,17 (they have improve the rate by 50 %) , also there is a small change concerning other methods when Numtree= 10 by Numtree =650 was changed
They have work by MLP for compare the results of random forest by that of MLP.
All the results obtained using the two classifiers are compared in figure 10. In fig.11and fig.12 we present the learning and recognition rate by Random Forest and MLP

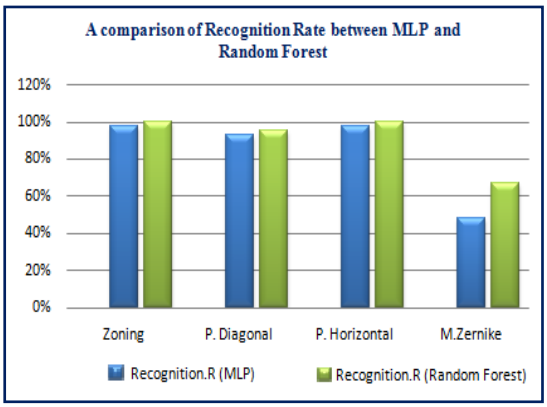


Fig.12 A Comparison of Recognition rate between MLP and Random Forest

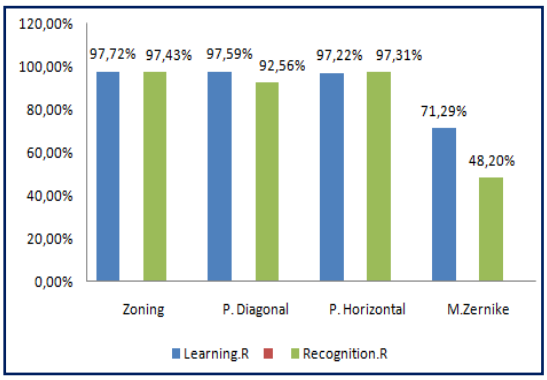


Fig.11 Learning rate and Recognition rate by MLP

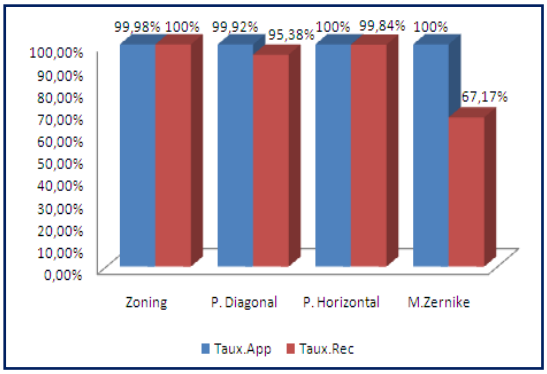


Fig.12 Learning rate and Recognition rate by Random Forest

According to these results, it can be noticed that the best results always are obtained with Random Forest.

VI. CONCLUSION

In this paper, we have presented a system of handwriting character recognition based on the method Random Forest. Several features have been studied and compared; as a result we've chosen Sauvola [10] 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 and other results of Zernike Moment show a significant improvement in recognition rate when using random forest in 650 of NumTree.

In future work, we will add other features methods that improve the results for some characters for example, minimize the length of execution of program which to calculate the recognition rate.

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] J. Sauvola and M. Pietikainen, "Adaptive Document Image Binarization," Pattern Recognition 33(2), pp. 225–236, 2000.
[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.