DOCUMENT CLASSIFICATION USING HYBRID EXTREME LEARNING MACHINE

[1]M. Dhivya M.Sc., M.Phil and [2]V. Bakyalakshmi M.Sc, M.Phil

Research Scholar, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore, India

Assistance Professor, Research Scholar, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts And Science, Coimbatore, India

[1]mdhivya.mphill@gmail.com, [2]vbakyalakshmi@yahoo.co.in

Abstract: Document image segmentation is one of the critical phases in image processing application of character recognition. Accuracy is determined by the correct segmentation of document image segmentation. Correct segmentation of individual characters decides the accuracy of the recognition system. It is used to decompose the sequence of characters into individual characters to segmenting text lines and then words. Ancient Tamil scripts documents consist of vowels, consonants and various modifiers. Hence proper segmentation algorithm is required. In existing methods, segmentation of overlapping lines and characters are difficult. In order to overcome this problem, two methods are proposed one for line segmentation and another for character segmentation, first method uses projection profile and PSO for segmentation. Then the classification of document segmentation is obtained by using HYBRID ELM.

Keywords: *Document Segmentation, Hybrid Extreme Learning Machine, Median Filter, Particle Swarm Optimization Algorithm and Gray Level Co-Occurrence Matrix.*

1. INTRODUCTION

Document clustering is one of the basic operations and it is used in unsupervised document organization, automatic topic extraction and information retrieval. An image is divided into regions or objects by using segmentation process. Usually, document segmentation seek to extract the basic part of the script these are certainly characters. This type of characters is needed because it is given as a input to the document classification analysis [1]. Segmentation phase is also crucial in contributing to this error due to touching characters, which the classifier cannot properly tackle. Even in good quality documents, some adjacent characters touch each other due to inappropriate scanning resolution [2].

Number of segmentation process are handled for perfect text extraction, one of the application for text extraction is OCR. Many techniques are used to segment the text, one of the most popular method is X-Y cut algorithm for top-down approaches for document segmentation [3] principal of this segmentation is to detect the white space using horizontal and vertical projections. Upon horizontal and vertical, bottom –up technique also used to segment the document, this is done by using Run Length Smearing Algorithm (RLSA), in this algorithm usually use the region growing of characters to detect the text regions [4, 5]. k-nearest neighbor clustering is used in Docstrum algorithm [6], it is also one of the another approach to detect the text regions in document segmentation. Next one method to detect the text region is implemented by using thresholding method [7].

One of the most traditional methods for text segmentation is Otsu's method [8] it is used to split the document histogram into object and background with the help of threshold pixels. Several methods are implemented from the modified Otsu's method [7, 9]. In [10, 11] use a local thresholding approach, obtained entropy information from the global thresholding [12] moment preserving approach [13]. Some of the comparison are handled in [14] for text segmentation.

The organization of this paper is follows. In section II discusses about some related work for document segmentation and Section III describes about POS segmentation and IV tells about HYBRID ELM classification. Section V gives the experimental result for the proposed approach.

2. RELATED WORK

Text line extraction is generally seen as a preprocessing step for tasks such as document structure extraction, printed character or handwriting recognition. Many techniques have been developed for page segmentation of printed documents (newspapers, scientific journals, magazines, business letters) produced with modern editing tools [15]. The segmentation of handwritten documents has also been addressed with the segmentation of address blocks on envelopes and mail pieces [16], and for authentication or recognition purposes [53] [17].

Document Classification allows the automatic distribution or archiving of documents. For example, after classification of business letters according to sender and message type (such as order, offer or inquiry), the letters are sent to the appropriate departments for processing [18].

Non-textual elements around the text such as book bindings, book sides, parts of fingers (thumb marks from someone holding the book open f.i.) should be removed upon criteria such as position and intensity level. On the document itself, holes, stains, may be removed by high pass filtering [19]. Other non-textual elements (stamps, seals) but also ornamentation decorated initials, can be removed using knowledge about the shape, the color or the position of these elements [20]. Extracting text from figures (text segmentation) can also be performed on texture grounds [21][22] or by morphological filters [24][23]. Linear graphical elements such as big crosses (called "St Andre's crosses") appear in some of Flaubert's manuscripts. Removing these elements is performed through GUI by Kalman filtering in [25].

A robust approach to segment text from color images was put forth in [26]. The proposed algorithm uses the multiscale wavelet features and the structural information to locate candidate text lines. Then a SVM classifier was used to identify true text from the candidate text lines. This approach mainly included four stages. In preprocessing step text blocks were enhanced by using cubic interpolation to rescale the input text blocks and a Gaussian filter to smooth the text blocks and remove noises. These image blocks were split into connected components and non-text connected components were eliminated by a component filtering procedure. The left connected components were merged using K-means clustering algorithm into several text layers, and a set of appropriate constraints were applied to find the real text layer. Finally, the text layer was refined through a post-processing step

3. METHODOLOGY

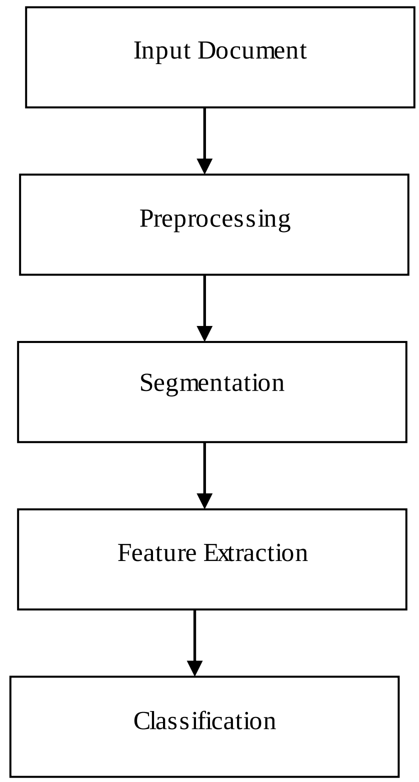

Figure 1: Block Diagram for Document Segmentation

3.1 Preprocessing

In the input image some noises are present. To obtained the document with fulfilled by using filters. Normally filters are used to remove the noise in the image or document. Filters are two types they are linear filter and nonlinear filter. Linear filters provides blurred in the images or signal, even though it preserve images, but it blurred the details of the pixels and sharp edges. So its preservation of the images is poor. Next nonlinear filters it removes the noise without remove the sharp edges and fine details of the pixels. From the conclusion nonlinear preserve the image is better than the linear filter.

Based on this in this paper removal of noise in the input document is handled by nonlinear filter that is median filter.

Median filtering could be a widespread technique of the image improvement for removing noise without effectively reducing the image sharpness [27].

Median filter is kind of common as a result of it provides excellent noise-reduction talents, with primarily less blurring than similar size linear smoothing filters. Here, the median method was performed by simply a 3×3 windowing operator over the image. It considers each pixel and its neighbors in pictures to search out whether or not it's an illustration of the environment. It replaces the value of component with the median of the neighboring pixel components. Tend to calculate the median by sorting the whole component values from the neighborhood into numeral sort then replaced the component being studied with the middle component worth. If the neighborhood below condition constitutes a good pixel worth, the common of the 2 middle component values is that the median.

3.2 Document Segmentation by using PSO algorithm

The space between the lines is used to separate the lines. Normally the distances between two lines are larger than the distances between words, thus lines can be segmented by comparing this distance against a suitable threshold. To determine an optimal threshold, Particle Swarm Optimization technique is used. It is known from literature, Particle Swarm Optimization (PSO) algorithm is used to solve many of difficult problems in the field of pattern recognition [28]. Hence, PSO is used to compute an optimal value.

Let X and V denote the particle's position and its corresponding velocity in search space respectively. At iteration K, each particle i has its position defined by $X_j^k = (x_{j1}, x_{j2}, \dots, x_n)$ and a velocity is defined by $V_j^k = (v_{j1}, v_{j2}, \dots, v_{jn})$ in search space n. Velocity and position of each particle in next iterations can be calculated using following equation (1) and (2)

$$V_{ij}^{k+1} = wv_{ij}^k + C_1 r_1 (pbest_{ij}^k - x_{ij}^k) + C_2 r_2 (gbest_{ij}^k - x_{jji}^k) \quad (1)$$

$$x_{ij}^k = x_g^k + v_g^k \quad (2)$$

Where k is the current iteration number, w is inertia weight, v_{ij} is then updated velocity on the i^{th} dimension of the j^{th} particle, C_1 and C_2 are acceleration constants, C_1 and C_2 are positive constant parameters, usually $C_1 = C_2 = 2$. r_1 and r_2 are the real numbers drawn from two uniform random sequences of U (0, 1).

The algorithm starts by generating randomly initial population of the PSO. In PSO, every particle is initialized with locations and velocities using the equations (1) and (2).These locations consist of the initial solutions for the optimal threshold.

The procedure of the proposed PSO algorithm is described as follows:

Step 1: Initialize N particles with random positions $x_1, x_2, \dots, x_n$ according to Eq. (1) and velocities V_i where i = 1, 2...N.

Step 2: Evaluate each particle according to equation 3

$$f(t) = w_0(t) \times w_1(t) \times (\mu_0(t) - \mu_1(t))^2 \quad (3)$$

Where, t is a gray level between 0 and 255 which can be obtained through the particle's position

Step 3: Update individual and global best positions. If $f(pbest_i) < f(x_i)$, then $pbest_i = x_i$ search for the maximum value f_{max} among $f(pbest_i)$, if $\max f(gbest) < f_{max}$ then $gbest = x_{max}$, x_{max} is the particle associated with f_{max}.

Step 4: Update velocity: update the i^{th} particle velocity using the Eq. (2) restricted by maximum and minimum threshold v_{max} and v_{min}.

Step 5: Update Position: update the ith particle position using Eq. (1) and (2).

Step 6: Repeat step 2 to 5 until a given maximum number of iterations is achieved or the optimal solution so far has not been improved for a given number of iteration.

3.3 Feature Extraction using Gray Level Co-occurrence Matrix (GLCM)

Co-occurrence features [17, 18, 19] are a popular and effective texture descriptor using statistical approach. Given an image of n gray levels, s, characteristics of images are estimated from the second-order statistical features by considering the spatial relationship of pixels in the image. A GLCM element $P_{\theta}, d\ i, j$ is the joint probability of the gray level pairs I and j in a given direction θ seperated by distance of d units. For each region of interest (ROI) in this work, five features are determined for texture discrimination: Energy (ENR), Entropy (ENT), Sum Entropy (SEN), Difference Entropy (DEN) and Standard Deviation (STD). Each subdivided block is an independent ROI. Multi-distance and multi-direction can be used to extract a large number of features. Extract the GLCM features using one distance $d = 1$ and four direction $\{\theta = \theta^0, 90^0, 180^0, 270^0\}$, which result in 20 that is $1 \times 4 \times 5$) features extracted for each block.

$$\mathrm{ENR} = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} P_d^2\ i,j \quad (4)$$

$$\mathrm{ENT} = - \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} P_d\ i,j\ \log_2 P_d\ i,j \quad (5)$$

$$\mathrm{SEN} = - \sum_{k=0}^{2n-2} P_{x+y}\ k\ \log_2 P_{x+y}(k) \quad (6)$$

$$\mathrm{DEN} = - \sum_{k=0}^{n-1} P_{x-y}\ k\ \log_2 P_{x-y}(k) \quad (7)$$

$$\mathrm{STD} = \frac{\sqrt{\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} P_d\ i,j - \mu^2}}{n \times n} \quad (8)$$

Where

$$\mu = \frac{\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} P_d(i,j)}{n \times n} \quad (9)$$

$$P_{x+y}\ k = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} P_d\ i,j \quad (10)$$

For $i + j = k$, here k=0,1,…,2n-2.

$$P_{x-y}\ k = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} P_d\ i,j \quad (11)$$

for $i - j = k$, k=0,1,…,n-1

GLCM features extraction (GFE) can be expressed as

$$\mathrm{GFE} = \mathrm{Img} \rightarrow \mathrm{Img} \times R^p \quad (12)$$

Where

$$\mathrm{GFE}\ b_{\Pi} = F\ \mathrm{GLCM}\ b_{\Pi} = b_{\Pi}, f_p \quad (13)$$

Where f_p is the feature vector and p is the number components (features extracted from b_{Π})

The GLCM function is introduced in (13) can be defined as

$$\mathrm{GLCM}: \mathrm{Img} \rightarrow M_{n \times n}\ N \quad (14)$$

Where $M_{n \times n}\ N$ is the set of square matrices with the dimension n × n. The function in (14) takes a block b_{I_i} of the image Img and returns its GLCM, given a direction and a distance.

The F function introduced in equation (13) can be defined a

$$F: M_{n \times n}\ R \rightarrow Img \times R^p \quad (15)$$

where the function of equation (15) takes a GLCM of a block b_{I_i} and returns b_{I_i}, f_p .

3.4 Document classification using Extreme Learning Machine

Let the given training set $N = xi, ti$, where the training sample $xi = [xi1, xi2, \ldots, xin]^T \in R^n$ and the corresponding target value $ti = [ti1, ti2, \ldots, tim]^T \in R^m$. SLFNs with N hidden neurons and activation function f(x) are mathematically modelled as

$$\sum_{i=1}^{N} \beta i.f\ wi.xj + bi = oj, j = 1, \ldots, N \quad (16)$$

Here $wi = wi1, wi2, \ldots, win^T$ is the weight vector connecting the ith hidden neuron and the input neurons and bi denotes the bias of ith hidden neuron.

Then $\beta i = \beta i1, \beta i2, \ldots, \beta in^T$ is the weight vector connecting the ith hidden neuron and the output neurons.

$wi.x_j$ defines the inner product of $wi\ and\ x_j$.

The fact that standard SLFNs with N hidden neurons each with activation function f(x) can approximate these N samples with zero error, means

$$\sum_{i=}^{N} o_j - t_j = 0 \quad (17)$$

That there exist $\beta i, wi\ and\ bi$ such that

$$\sum_{i=1}^{N} \beta i.f\ wi.xj + bi = t_j, j = 1, \ldots, N \quad (18)$$

Here the N equation can be rewritten as

$$H\beta = T \tag{19}$$

In this equation H is denoted as

$$(w_1, \ldots, w_N, b_1, \ldots, b_N, x_1, \ldots, x_N) = \begin{bmatrix} f(w_1.x_1 + b_1) & f(w_N.x_1 + b_N) \\ f(w_1.x_N + b_1) & f(w_N.x_N + b_N) \end{bmatrix}_{N \times N} \tag{20}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \ldots \\ \beta_N^T \end{bmatrix}_{N \times m} \tag{21}$$

$$T = \begin{bmatrix} t_1^T \\ \ldots \\ t_N^T \end{bmatrix}_{N \times m} \tag{22}$$

The SFLN can be solved by using a gradient based solution by finding suitable values of $w', b'\ and\ \beta'$ satisfying the model as:

$$||H(w_1', \ldots, w_H', b_1', \ldots, b_H')\beta' - T|| = \min_{w_i, b_i, \beta} ||H(w_1, \ldots, w_H, b_1, \ldots, b_H)\beta - T||$$
(23)

A gradient based learning algorithm can be used to minimize the $H\beta = T$ by adjusting the parameters $w_i, b_i and\ \beta_i$, when the H hidden layer matrix is unknown iteratively. In [34] proved that single layer feed forward neural network with randomly assigned input weights and hidden layer biases and with almost any nonzero activation function can universally approximate any continuous functions on any input data sets. In [35] suggested an alternate way to train a SHLFN by finding a least square solution β' of the linear system represented in equation 28. The unique minimum norm least square (LS) solution is modelled as

$$\beta = H^+T$$
(24)

In this equation H^+ is the MP generalized inverse of matrix H. this type of inverse is used to attain the good generalization performance with dramatically increased learning speed. HYBRID ELM algorithm is given in below:

Given a training set $N = \{(xi, ti) | xi \in R^n, ti \in R^m, i = 1, \ldots, N\}$, kernel function f(x), and hidden neuron N

Step 1: Select suitable activation function and number of hidden neurons N for the given problem.

Step 2: Assign arbitrary input weight w_i and bias $b_i, i = 1, \ldots, H$

Step 3: Calculate the output matrix H at the hidden layer

$$H = f.(w.+x + b) \tag{25}$$

Step 4: Calculate the output weight β

$$\beta = H^+T \tag{26}$$

4. EXPERIMENTAL RESULT

To evaluate the proposed work which is implemented here is analyzed and results produced from this algorithm are shown here. To prove its efficiency here it is compared with existing classification SVM technique. The document images are collected from freely available database from websites and this technique is implemented in MATLAB software of R2012a version environment.
The performance of this work is measured using accuracy of correctly classified and execution time taken for this work to complete the process.

4.1 PERFORMANCE COMPARISON:

The performance of this proposed technique is given to SVM and HYBRID ELM classifier. The obtained results are shown here,

4.2 SVM Classification Results

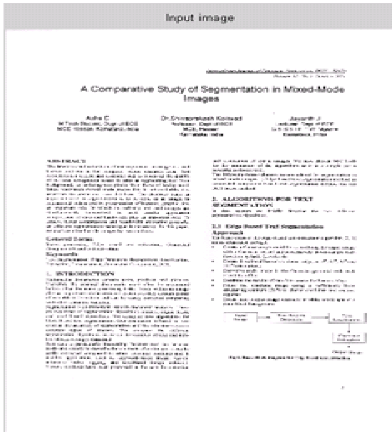

Figure 1: Input Image

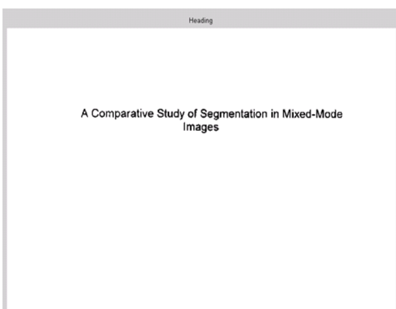

Figure 2: Detected Header from the Image

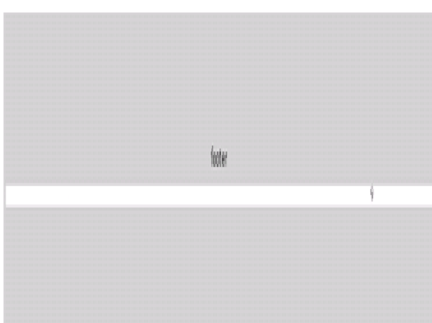

Figure 3: Detected Footer

4.3 HYBRID ELM classification results

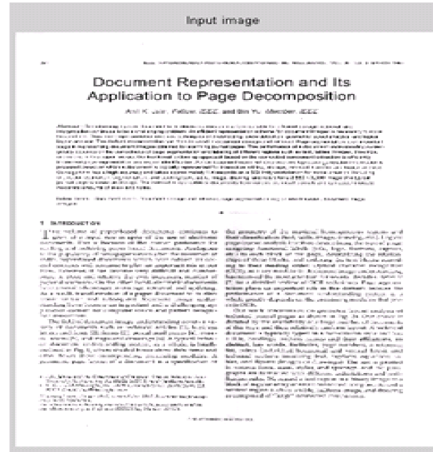

Figure 4: Input Image

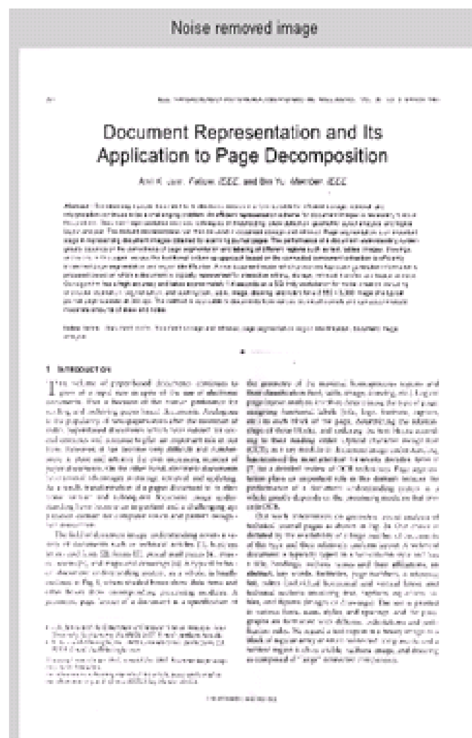

Figure 5: Noise Removed Image

Figure 6: PSO Segmented Output

Figure 7: Result of Connected Components

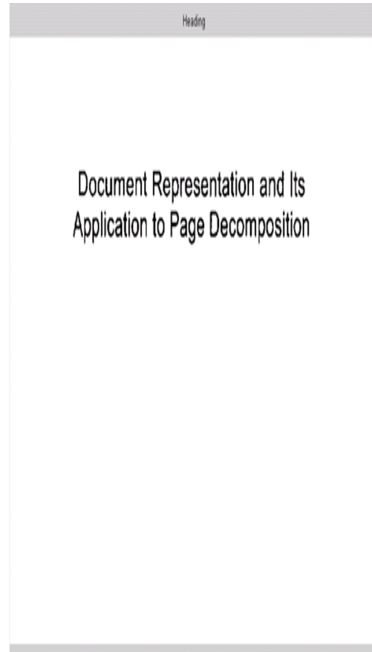

Figure 8: Detected Heading from the Image

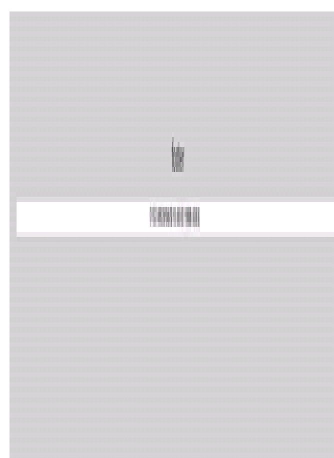

Figure 9: Detected Footer from the Image

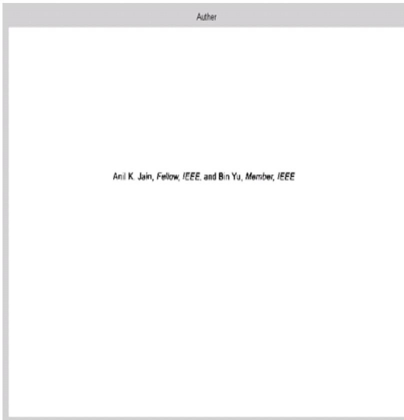

Figure 4.10: Detected Authors Name

4.4 Accuracy and Execution Time Comparison:

The accuracy of correctly detected text from the document image is shown in graph for the SVM and HYBRID ELM classifier.

Table 1: comparison report for accuracy with existing algorithm

Methodology	Accuracy for classification %
SVM	62.5
HELM	87.5

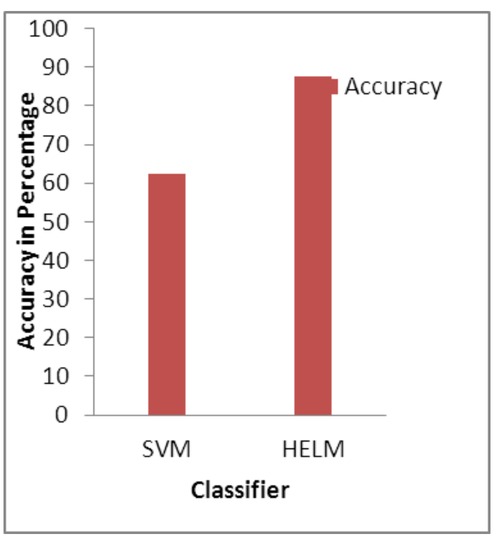

Figure 11: Comparison Accuracy of SVM and HYBRID ELM

The execution time taken by the classification technique of SVM and HYBRID ELM is shown below.

Table 2: comparison report for Classification time with existing algorithm

Methodology	Time taken for classification(sec)
SVM	15.58
HELM	10.55

Time taken for the execution of this document classification is compare to svm the proposed hybrid extreme leaning machine efficient form the existing system the proposed

The simulation result shows that the designed hybrid extreme leaning machine for the original classification over's a good compromise between the simplicity of the classification and good accuracy performance compared to the Support vector machine (SVM)

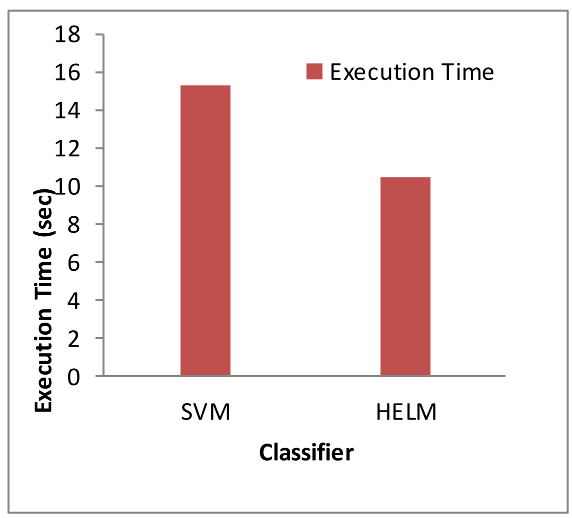

Figure 12: Comparison of Execution Time Taken by SVM and HYBRID ELM

From the figure 4.16 and 4.17 , we can conclude the HYBRID ELM produces better accuracy than SVM in reduced time and proved as best classifier for document Image Classification.

5. CONCLUSION

Document classification is an active research area at present which helps to extract needed information easily and to analyze the document images. Many researchers contribute their work towards this classification and produced the result but some of them fail in extracting them in noisy environment. Here this work extracts the document image from the noisy environment and segment the text using PCO algorithm and finally classification is done using HYBRID ELM classifier. The performance of this approach is compared with SVM classification and experimental results proved that HYBRID ELM technique suits for document image classification.

e future work can be extended by detecting images, postal codes, handwritten and printed documents by adding some more features and improving the classification algorithm with other algorithms.

REFERENCES

[1] Raghuraj Singh. S. Yadav and Prabhat Verma" Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network" International Journal of Computer Science & Communication, Vol. 1, No. 1, January-June 2010, pp. 91-95.

[2] Vijay kumar and Pankaj K. Sengar, "Segmentation of Printed Text in Devanagari Script and Gurmukhi Script", International Journal of Computer Applications,Vol 3, No.8, June 2010, pp. 24-29.

[3] G.Nagy, S. Seth and M. Viswanathan, "A prototype document image analysis system for technical journals", Computer, vol.25, no.7, pp.10-22, 1992.

[4] K.Y. Wong and F.M. Wahi, "Document analysis system", IBM J. Res, Develop., Vol. 26, pp.647-656, 1982.

[5] J.Fisher, " A rule-based system for document image segmentation", in Proc. 10th Int. Conf. Pattern Recognition, 1990. Pp.567-572.

[6] L.O,. Gorman, "The document spectrum for page layout analysis", IEEE Trans. Pattern Analysis. Mach. Intell., vol.15, n0.11, pp. 1162-1173, Nov.1993.

[7] Y.Chen and B.Wu, "A multi-plane approach for text segmentation of complex document images", Pattern nRecognition., vol.42, no.7, pp.1419-1444, 2009.

[8] N. Otsu, "A threshold selection method from gray-level histograms",
IEEE. Trans. Syst., Man Cybern., vol.9, no.1, pp.62-66. Jan.1979.

[9] H. Oh, K. Lim and S. Chien, " An important binarization algorithm based on a water flow model for document image with inhomogenous backgrounds", Pattern Recognition., vol. 38, no.38, pp.26112-2625, 2005.

[10] W. Nibalck, "An important to digital image processing", Upper saddle River, NJ"; Prentice-Hall. 1996.

[11] J. sanvola and M. Pietaksinen, "Adaptive document image binarization", Pattern recognition., vol.33, no.2, pp.236-335, 2000.

[12] J. Kanpur, P.Sahoo and A. Wong, " A new method for gray-level picture thresholding using the entropy of the histogram", Graph Models Image Process., vol. 29, no.3, pp.273-285, 1985.

[13] W. Tasi, " Momenet-preserving thresholding: A new approach", Graph Graph Models Image Process., vol.29, no.3, pp.377-393, 1985.

[14] P.Sathis, E.Kavallerieratou and N. Papamarkos, " An evaluation survey of binarization algorithm on historical documents", In Proc. 19th Conf. Pattern Recognition., 2008. Pp.1-4.

[15] Antonacopoulos A. (1994), Flexible Page Segmentation Using the Background, Proc. 12th Int. Conf. on Pattern Recognition (12th ICPR), Jerusalem, Israel, October 9-12, vol. 2, pp. 339-344.

[16] Downton A., Leedham C. G. (1990), Preprocessing and presorting of envelope images for automatic sorting using OCR, Pattern Recognition, 23(3-4):347-362.

[17] Zhang B., Srihari S.N., Huang C. (2004), Word image retrieval using binary features, SPIE Conference on Document Recognition and retrieval XI, San Jose, California, USA, Jan 18-22.

[18] Brukner, T., Suda, P., Block, H and Maderlechner, G, " In-house mail distribution by automatic address and content interpretation", In: Proceedings of the 5th Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, USA, April 1996, pp.67-75, 1996.

[19] Feldbach M. (2000), Generierung einer semantischen reprasentation aus abbildungen handschritlicher kirchenbuchaufzeichnungen, Diplomarbeit, Otto von Guericke Universitat Magdeburg.

[20] Gusnard de Ventadert, J. André, H. Richy, L. Likforman-Sulem, E. Desjardin, (1999), Les documents anciens, Document numérique, Hermès, Vol. 3, no 1-2, june, pp. 57-73

[21] Kim I.-K, Jung D.-W., Park R.-H. (2002), Document image binarization based on topographic analysis using a water flow model, Pattern Recognition 35:265-277.

[22]Mello C. A. B., Cavalcanti C. S.V.C., C. Carvalho (2004), Colorizing paper texture of green-scale image of historical documents, Proceedings of the 4th IASTED Conference on Visualization, Imaging and Image Processing, VIIP, Marbella, Spain.

[23] Mengucci M., Granado I. (2000), Morphological Segmentation of text and figures in Renaissance books

(XVI Century), in Mathematical Morphology and its applications to image processing, J. Goutsias, L. Vincent, D. Bloomberg(eds), Kluwer, pp. 397-404.

[24] Granado I., Mengucci M., Muge F., (2000) Extraction de textes et de figures dans les livres anciens à l'aide de la morphologie mathématique, Actes de CIFED'2000, Colloque International Francophone sur l'Ecrit et le Document, Lyon , pp. 81-90.

[25] L. Likforman-Sulem (1998), Extraction d'éléments graphiques dans les images de manuscrits, Colloque International Francophone sur l'Ecrit et le Document (CIFED'98), Québec, pp. 223-232.

[26] Y. Zhan, W. Wang, W. Gao (2006), "A Robust Split-And-Merge Text Segmentation Approach For Images", International Conference On Pattern Recognition,06(2):pp 1002-1005.

[27] Chan, R.H., C.W. Ho and M. Nikolova, Salt-and pepper noise removal by median-type noise detectors and detail-preserving regularization. IEEETrans. Image Proc., 14: 1479-1485. 2005.

[28] Oliveira .S.L., S. A. Britto, and R. Sabourin, " Optimizing Class-Related Thresholds with Particle Swarm Optimization", Proceeding of International Joint Conference on Neural Networks, IEEE, Montreal, Canada, July 31 – August 4, 2005,pp. 1511 – 1516.