Multidisciplinary Review Journal

Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 2, 2020

Review Article

CLASSIFICATION OF PRINTED TEXT AND HANDWRITTEN CHARACTERS WITH NEURAL NETWORKS

K. Bramara Neelima[1], Dr.S. Arulselvi[2]

[1]Research Scholar, Bharath Institute of Higher Education and Research.
[2]Research Supervisor, Bharath Institute of Higher Education and Research.

Received: 19.11.2019 ***Revised: 20.12.2019*** ***Accepted: 21.01.2020***

ABSTRACT
Various scanned documents consist of machine printed text as well as handwritten alphabet. The classification of handwritten alphabet and machine printed alphabet in document images is the required process afore character recognition. We can distinguish these two types of alphabet images by their shape structural, statistical and visual difference features. This work proposes the artificial neural network based classification technique for machine printed and handwritten text classification at character level using a set of new features which are combination of statistical, shape structural and visual impression features. The projected technique attained remarkable classification efficiency on two databases; IAM dataset and prepared dataset.

Keywords: Document Image, Handwritten Text, Classification, Neural Networks.

DOI: http://dx.doi.org/10.31838/jcr.07.02.25

INTRODUCTION

Printed alphabet and handwritten alphabet are the two distinguish characters appeared in diverse documents such as financial documents, bank cheques, scanned documents, graphical texts, and other types. The classification of printed and handwritten alphabet areas is the advanced research process in the document image understanding. Characters on the document image can be further divided into text characters and numerical characters. Text characters are formed with the set of alphabets of numerous languages. Numerical characters are formed with the number set of zero to nine, total ten numerals. On the processing of the document image these characters give the impression and well thought-out as optical characters. Processing of the optical characters, with the advent use of digital image processing algorithms, is known as optical character recognition.

Numerous algorithms and software are generated to achieve optical character recognition with greater percentage of extraction performance. From an image, machine printed characters can be easily recognized and extracted with various uncomplicated optical character recognition processing methods. Recognition and extraction of handwritten characters from images is challenging process because of the variations in handwriting styles, characters set used, method of character connectivity and several handwritten features which are different from person to person. The feature set of printed characters and handwritten characters are dissimilar categorically.

Theoptical character recognition methods are generally processed in four stages namely, preprocessing, segmentation, feature extraction and, classification. The block diagram of general purpose optical character recognition process is shown in Figure1. Preprocessing stage comprises the quite a lot of image processing techniques such as skew and orientation correction, noise elimination, background subtraction, base line removal, and a few morphological operations. The preprocessed images are segmented using line segmentation, word segmentation, character segmentation, and paragraph segmentation methods, into a set of individual character images. The character image obtained as the image segmentation output may be a printed text character or a handwritten text character. The characteristics of each character are obtained by feature extraction stage.

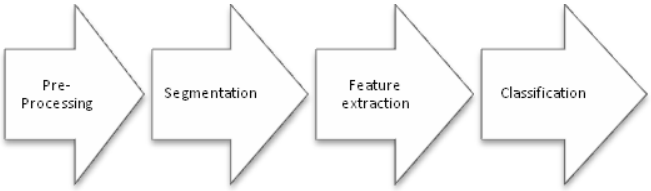

Figure 1: Generalized OCR Model

Feature extraction is the fundamental process in character recognition. Corresponding feature vector is attained by applying appropriate feature extraction technique on segmented character image. These feature vectors are made use of in the classification stage for classifying the segmented character image into the printed text character or the handwritten text character. The dissimilarities in the characteristics of printed and the handwritten alphabet are the basic essentialities in the classification stage.

In this article an artificial neural network (ANN) is developed and conquered for the process of printed and handwritten alphabet classification. The neural network is trained with handwritten digits and printed digits based on supervised learning. Further test document images are classified conferring to the handwritten or printed character.

RELATED WORK

The existence of machine printed and handwritten alphabet in the similar text image is an significant obstacle to the automation of the optical character recognition (OCR) procedure [1]. In literature the works presented maybe classified into four different levels of text separation paragraph, text line, word and character level. Also handwritten recognition, character recognition and text localization are included as part of the study in many works [2-4]. Zheng et al. [5] projected a method of text identification in noisy document images with relative outcomes at all levels. Imade et al. [6] performed a classification using neural networks on mixed document image and classified handwritten character, printed character, graphic image region, and picture. Fan et al. [7] extracted word block using X-Y cut algorithm from a document image and then printed and handwritten words are classified using spatial feature and block variance.

Shape orientation is one of the feature used to separate machine printed and handwritten text. Shape structural properties obtained by Radon transform are considered as a set of features used in the classification of printed and handwritten text in [8].A decision tree classifier has been considered based on boundaries of six element gray level feature set values for classification of handwritten and printed text in [9]. In [10] shape structural properties of given text are derived using chain code method used to classify the printed text which contains straight strokes and the handwritten text which has inclined strokes in various directions. In [11], a new set of features are projected and data mining techniques are used to classify handwritten and printed text.

In [12], G-means based classification and Markov Random Field based classification are used for patch level separation and pixel level separation of identified three categories of classes which are machine printed text, handwritten text and overlapped text. Tangila Saba [13] presented a technique based on structural and statistical features of text lines and proposed a set of classification rules to classify multilingual text lines into handwritten and printed text. In [14], a novel integrated classification technique is proposed to distinguish between handwritten text and machine printed text. The classification is taken place inside the intelligent character recognition cell integrated with several texture features of text lines.

In [15], a combination of binary SVM classifiers are used for characterization of blocks of interest in the document image as handwritten or machine printed text. In this paper the feature descriptor is obtained from bag of visual words of each block. Noise is also attained as one of the classification class. In [16], K-Nearest Neighbor (KNN) classifier is used to discriminate between handwritten and machine printed text at word level. Statistical texture features of word, namely, smoothness, mean, standard deviation, moment, entropy, uniformity, are included in feature vector for classification. In [17], shape features such as area, perimeter, form factor, roundness, compactness, minor and major axes are consumed to differentiate between printed text and handwritten text. Bala Mallikarjunarao, et.al, [18] proposes a machine print and handwritten text classification approach at word level using several intensity and shape structural features and their combinations on IAM dataset. Samir Malakara, et.al,[19] proposes a decision tree based handwritten and printed word classification by prioritizing the features extracted from gray scale images.

PROPOSED APPROACH

The current work centered on the classification of machine printed and handwritten text in document images. The process is as follows in figure 2. The proposed method describes the type of the document selected, its digitization method means acquiring as image; applied various image preprocessing techniques on acquired document image to make it prepared for further processing; segmenting involves localization of text regions and further segmenting into lines, words, and individual characters; the process of extraction of features; the classification process performed by the system based on the features defined; finally, post-processing using suitable recognition techniques for character recognition. The overall proposed system block diagram is shown in Figure2.

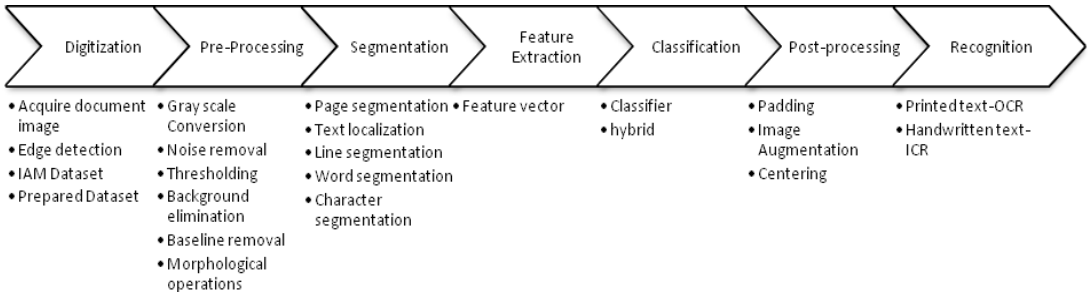

Figure 2: Proposed Classification Methodology

Document Digitization and Data Collection

Some of the documents consist of several elements such as the blanks, lines, logos, figures, tables, graphs, printed text, type written text, hand written text and other elements. These documents can be classified into the text region and the other regions, by classifying the document image. The proposed system consists of the documents with text regions. These types of document images are available in the IAM dataset. Hence, here considering the two different datasets, one is IAM dataset and other is our own dataset. The prepared dataset is formed from documents such as bank documents like overdrafts and cheques, logistic questionnaires, application forms, and other similar forms, which contains text regions. These documents are scanned using flat-bedded scanner with 300dpi resolution.

Pre-Processing

The five image pre-processing techniques are accomplished on the acquired document image to prepare it for segmentation. Firstly, all the acquired images are converted into gray scale images to reduce the memory usage. Secondly, during the scanning or digitization, there appears the noise; which will be eliminated by 5x5 median filters. Then, binarize the document using the Otsu's threshold method, which automatically separates the text from background. Next, by applying simple horizontal line extraction technique, baselines are removed. Finally, morphological operations are applied, further remove the unwanted components such as dots, colons, question marks, reminiscent noises and pliable the vertical contours of text.

Segmentation

The segmentation initially embroils text localization method on binary input image, which obtained as the outcome of image processing step. Text localization is achieved with the process given in [20]; vertical text zones are identified by analyzing vertical rule lines as well as vertical white runs of document image, later horizontally restrict the zones based on connected component analysis. The text line segmentation is accomplished by the combination of connecting component analysis and Hough transform method. Ensuing, in order to localize a word from text line and a character from word, the proposed system achieves the abstraction of connected components and which pass through the bounding box technique. Individual characters are identified by bounding boxes and words are formed by uniting the overlapping boxes or boxes in same text line with space less than the half of the average size of boxes. The average size of the boxes is calculated as the ratio of sum of sizes of all boxes in image with the number of boxes. Mathematically expresses as $Average\ size = \sum_i Bi/n$, where n is the total number of boxes and Bi is the size of each box.

Also, we performed manual segmentation on the prepared dataset at word level and character level by cropping the document images. 5000 words and 200 characters are cropped manually and labeled them as printed or handwritten. Out of 5000 words, 2500 words are printed/typewritten words and 2500 words are handwritten words. Out of 200 cropped characters, 100 characters are handwritten characters and remaining are printed characters.

Feature Extraction

The properties of the handwritten text and machine printed text characters are utilized for feature extraction. The shape structural features and statistical features are discerned in the handwritten texts and machine printed. Shape structural features can be signified by the aspect ratio, shape energy, intrinsic features, stroke width, density and variance of intensity values, pixel distribution, pixel intensity uniformity, physical size, etc. Statistical features can be represented by mean, standard deviation, entropy, otsu's threshold, intensity value distribution, local maxima, etc. The visual difference among printed and handwritten text is more obvious in terms of texture(feel and appearance), pixel intensity, more straightness, font type unique character(text shape), size(height, width), number of black pixels per character, line straightness, variance in intensity value, stroke width, density of pixels, smoother (at curved shapes), variations(at circle, eclipse),etc. The combination of significant shape structure, statistical and visual difference features are considered and the feature vector is extracted with ten features. These features are given as input to the classifier to perform classification concerning handwritten text and machine printed texts.

Classifier:

In this work, artificial neural network is used as the classification method. The neural network is trained with the collected feature vectors and manual segmented labels. During training, the gradient set as e-10, the target vector is normalized in the range of [-1 to 1] and number of iterations 1000 for efficient training. The trained neural network classifier is used to categorize the test the feature vectors and their corresponding labels as machine printed text character [label= -1] and handwritten text character [label = 1]. The validation of the proposed model is substantiated using the performance accuracy.

Post-processing and Recognition:

In order to make the individual characters and single words, which are the outcome of classification step, more compatible with the recognition models, some post-processing and augmentation techniques are applied on them. Since the words are of different heights and widths, the respective images are also of different sizes. To make all the images of the same size, white spaces are added on both sides of width and height evenly; this process is called as zero padding. Image augmentation is performed on the padding images with a small angle tilt to right side, and image centering is performed on each image by subtracting the mean pixel value from the pixel values of image. This modified image dataset is applied to the suitable recognition model for recognizing the characters and words. Various technologies are invented to extract information from images and convert into machine encoded format. Optical Character Recognition (OCR) is functioned for recognition of machine printed or type written printed text and Intelligent Character Recognition (ICR) is utilized for recognition of hand printed or hand written text.

EXPERIMENTAL RESULTS

Experimental results of the proposed method are discussed here. Tests have been implemented by using two databases; they are IAM database and prepared database. These databases contain the documents consisting of both printed and handwritten text. These text documents are segmented into word and character images. These segmented images are used in our experiment. Feature vector is extracted from these images and features are used to analyse the differences among handwritten and machine print texts. To differentiate the handwritten text from machine printed text, the following features are proposed.

Aspect Ratio: Size represents the height and width of the image, whose distributions are stable in machine printed text and very diverse in the handwritten text. The aspect ratio is defined as the ratio of width to height of the image. Let consider w is the width of the given character image and h is the height of the character image, the aspect ratio is defined as $a = w/h$. The aspect ratio variations in handwritten and printed texts are showed in figure 3.

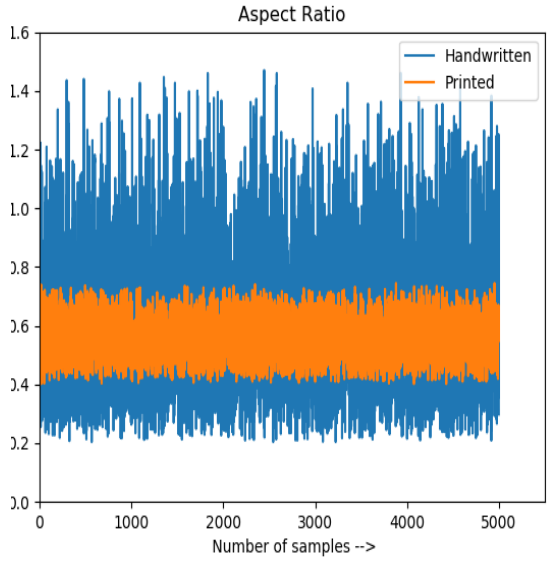

Figure 3: Aspect Ratio Distribution of Handwritten and Printed Characters

Pixel Density: Pixel density represents the area occupied by the black binary pixels in the total image area. This feature is important for classification efficiency.

$$pixel\ density = \frac{total\ number\ of\ black\ pixels}{total\ pixel\ area\ of\ image}$$

The pixel density variations in handwritten and printed texts showed in figure4.

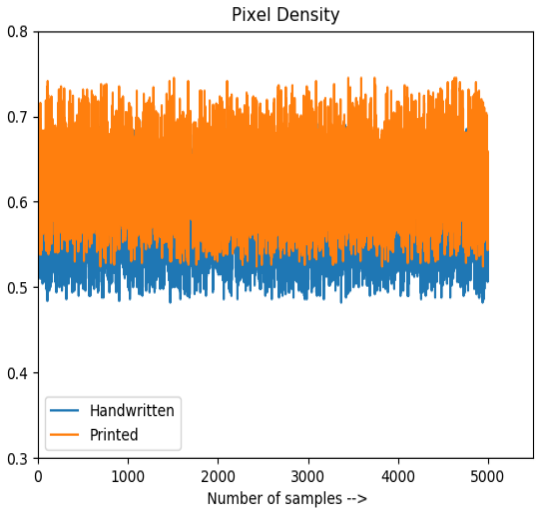

Figure 4: Pixel Density Variations of Handwritten and Printed Characters

Pixel Distribution: Pixel distribution is defined as the modulus difference between density of upper part and density of lower part. The distribution of the upper part and lower part of the handwritten text is quite diverse, whereas for the machine printed text the distribution is uniform and stable.

$$pixel\ distribution = |\ upper\ part\ density - lower\ part\ density$$

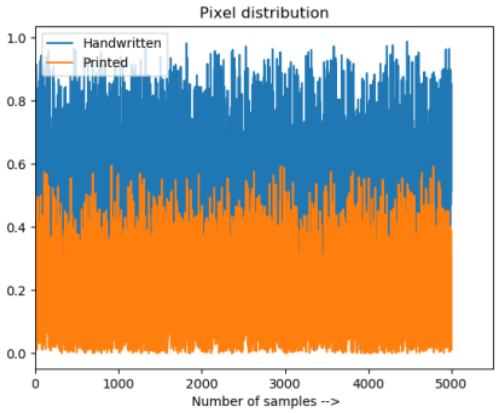

Figure 5: Pixel distribution variations of Handwritten and Printed Characters

Horizontal projection profile: The horizontal projection of black pixels in the image is executed and the difference of the adjacent pixels horizontal coordinates of the projection profile is calculated as a feature. The horizontal projection profile is much more obvious in handwritten text than the printed text.

Vertical projection profile: The vertical projection of black pixels in the image is calculated and the variance of the vertical coordinates of the projection profile is calculated as a feature. The vertical projection profile is most homogenous in machine printed text than the handwritten text.

Elasticity: It is a shape structural feature and its value is acquired as the first order derivative of each pixel in the image. Mathematically represents as in the equation.

$$elasticity = \sum_{i=0}^{n-1} [(x_{i+1} - x_i)^2 - (y_{i+1} - y_i)^2]$$

Where x and y are the coordinates of the pixel and i is the order of pixel.

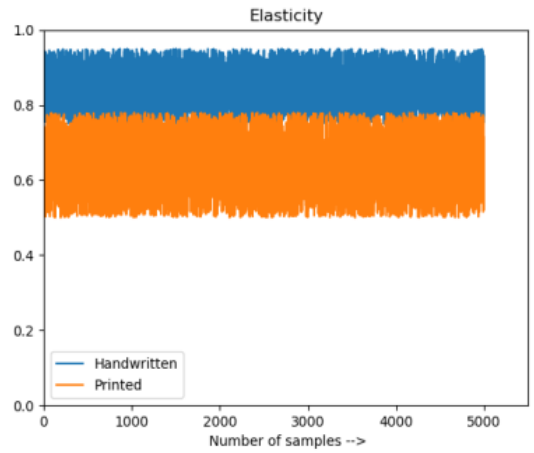

Figure 6: Elasticity Variations of Handwritten and Printed Words

Curvature: It is a shape energy structural feature and is obtained as the second order derivative of each pixel in the image. Mathematically, it can represents as in the below equation.

$$curvature = \sum_{i=2}^{n-2} [(x_{i+1} - 2x_i + x_{i-1})^2 - (y_{i+1} - 2y_i + y_{i-1})^2]$$

Where x and y are the coordinates of the pixel and i is the order of pixel.

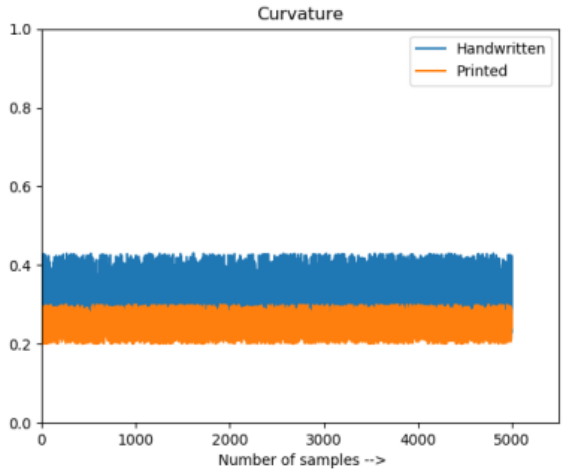

Figure 7: Curvature Variations of Handwritten and Printed Characters

Mean: The mean intensity values of handwritten images are significantly higher than the printed images. Let $I(x,y) \in [0,1]$ where $0 \le x < w$, $0 \le y < h$ is a given image with w and h are width and height respectively. The mean pixel intensity value is calculated as,

$$mean\ (\mu) = \frac{\sum_{x=0}^{w-1} \sum_{y=0}^{h-1} I(x,y)}{w \times h}$$

The graphical representation of the mean pixel intensity values of handwritten and printed characters are shown in figure8. It is quite observed from the figure8, that the mean helps to classify the machine printed text and the handwritten text.

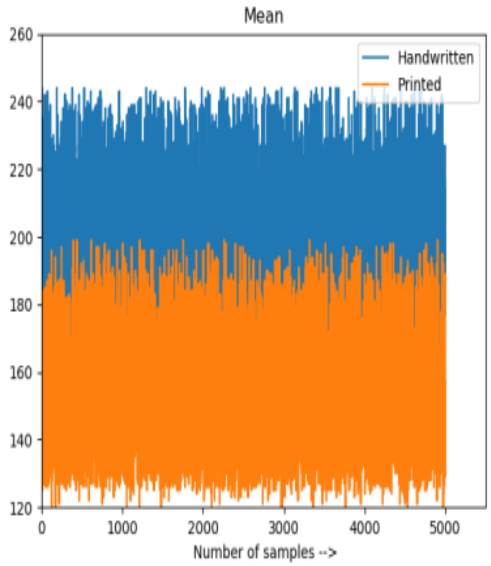

Figure 8: Mean Distribution of Handwritten and Printed Characters

Standard Deviation: It shows how much dispersion exists from mean pixel intensity value. The standard deviation is defined as below.

$$Standard\ deviation\ (\sigma) = \sqrt{\frac{\sum_{x=0}^{w-1}\sum_{y=0}^{h-1}(I(x,y)-\mu)^2}{w \times h}}$$

The standard deviation of the handwritten and printed text is shown in figure9. There is a clear separation between handwritten and machine printed text standard deviation values.

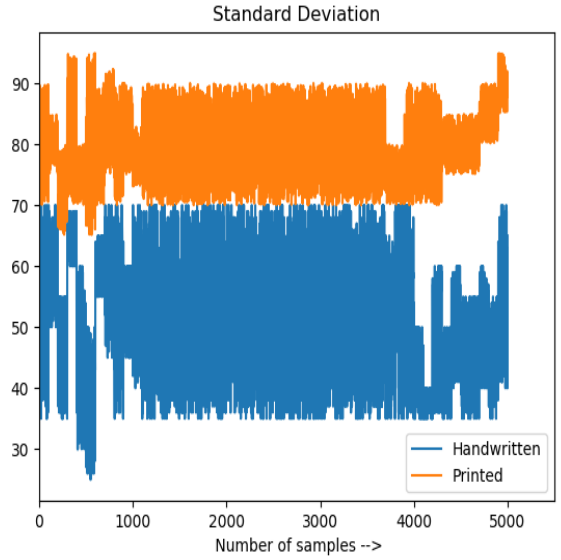

Figure 9: Standard Deviations of Handwritten and Printed Words

Black pixels of character: The number of black pixels per character in boundary box is calculated and divided by its height. The sum of results is used as a feature in feature vector.

Local maxima: The number of local maxima represents the straightness variations in a given image. The local maxima of handwritten and machine printed text are showed in figure10. It's the most useful feature for the classification of print and handwritten texts.

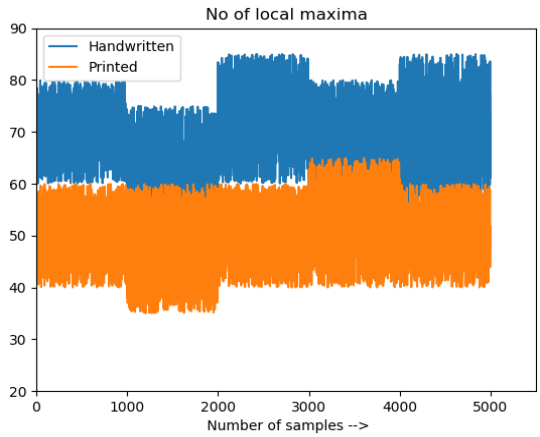

Figure 10: No. of Local Maxima of Handwritten and Printed Words

All these features are concatenated to represent each character image in form of feature vector. The feature vectors of each image of two databases are calculated and these feature vectors are given as the input to the classifier. The classification accuracy achieved by the proposed model is 97.6% and recognition accuracy of 99.2%.

CONCLUSION

A text classification model was proposed, which is able to discriminate between machine printed and handwritten text characters. We used two databases, one is IAM database and the other is our own prepared database. In this work, we proposed a set of new features to be extracted, which are a combination of shape structural, statistical and visual difference features derived from character image. An artificial neural network is constructed for classification process with the help of feature values. According to the results, the proposed features provide the significant differences between handwritten and machine printed texts. The overall system classification accuracy achieved is 97.6%.and of recognition is 99.2%.

In future work, we plan to implement other classification methods (SVM, fuzzy logic, kNN) for comparisons with the proposed classification technique and also based on different text areas(word, sentence). To implement techniques that would handle the document images with graphics and variable fonts.

REFERENCES

1. E. Kavallieratou, S. Stamatatos, H. Antonopoulou, "Machine-Printed from Handwritten Text Discrimination", *IWFHR-9 2004, 9th Intern. Workshop on Frontiers in Handwriting Recognition*, pp. 312-316, 26–29 Oct., 2004.
2. X. Peng, S. Setlur, V. Govindaraju, R. Sitaram, Handwritten text separation from annotated machine printed documents using Markov Random Fields, *in International Journal on Document Analysis and Recognition* pp. 1-16 (2013)
3. A. Belaïd, K.C. Santosh, V.P. D'Andecy. Handwritten and Printed Text Separation in Real Document, in Machine Vision Applications, version 2 (2013).
4. Roy, K., S. Kundu Das, and SkMd Obaidullah. "Script identification from handwritten document." Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2011 *Third National Conference on. IEEE,* 2011.
5. Y. Zheng, H. Li, D. Doermann, "Text identification in Noisy Document Images Using Markov Random Field", *Proceedings of 7th ICDAR,* 2003, pp.599-603.
6. S. Imade, S. Tatsuta and T. Wada, Segmentation and Classification for Mixed Text/Image Document Using Neural

Network, in *Proceedings of 2nd International Conference Document Analysis and Recognition,* pp. 930-934 (1993).
7. K.C. Fan, L.S. Wang and Y.T. Tu, Classification of Machine-Printed and Handwritten Texts Using Character Block Layout Variance, in *Pattern Recognition,* vol. 31, No. 9, pp. 1275-1284(1998).
8. Zemouri, ET-Tahir, and Youcef Chibani. "Machine printed handwritten text discrimination using radon transform and svm classifier." Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on. IEEE, 2011.
9. Samir Malakara, Rahul Kumar Das, Handwritten and Printed Word Identification Using Gray-scale Feature Vector and Decision Tree Classifier, *International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA)* 2013
10. Chanda, Sukalpa, Katrin Franke, and Umapada Pal. "Structural handwritten and machine print classification for sparse content and arbitrary oriented document fragments." *Proceedings of the 2010 ACM Symposium on Applied Computing. ACM,* 2010.
11. Lincoln Faria da Silva, Aura Conci, Angel Sanchez, Automatic Discrimination between Printed and Handwritten Text in Documents, 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing.
12. Peng, Xujun, et al. "Handwritten text separation from annotated machine printed documents using Markov Random Fields." *International Journal on Document Analysis and Recognition (IJDAR)* 16.1 (2013): 1-16
13. Tanzila Saba, and A Nikolaidis. "Language Independent Rule Based Classification of Printed & Handwritten Text". *Proceedings of the IEEE 2015 tenth workshop on Multimedia Signal Processing,* pp.393-398, 2015.
14. Abhishek Jindal, and Mohd Amir. "Language Independent Rule Based Classification of Printed & Handwritten Text". *Proceedings of the IEEE* 2015 *tenth workshop on Multimedia Signal Processing,* pp. 393-398, 2015.
15. Konstantinos Zagoris, and Ioannis Pratikakis. "Automatic Classification of Handwritten and Printed Text in ICR Boxes". *Proceedings of the IEEE* 2014.
16. M. Hangarge, K.C. Santosh, S. Doddamani, R. Pardeshi, Statistical Texture Features based Handwritten and Printed Text Classification in South Indian Documents, in *Proceedings of International Conference on Emerging Trends in Electrical, Communications and Information Technologies, Elsevier,* pp. 215-221 (2012)
17. P.L, Upasana, and M. Begum, Word Level Handwritten and Printed Text Separation Based on Shape Features, in *International Journal of Emerging Technology and Advanced Engineering (IJETAE),* vol. 2, Issue 4 (2012)
18. Bala Mallikarjunarao Garlapati, Srinivasa Rao Chalamala, "A System for Handwritten and Printed Text Classification", *UKSim-AMSS 19th International Conference on Modelling & Simulation, IEEE,* 2017, pp. 50-54.
19. Samir Malakara, Rahul Kumar Das, Handwritten and Printed Word Identification Using Gray-scale Feature Vector and Decision Tree Classifier, *International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA)* 2013
20. Basilis Gatos, Georgios Louloudis and Nikolaos Stamatopoulos, "Segmentation of Historical Handwritten Documents into Text Zones and Text Lines", *14th International Conference on Frontiers in Handwriting Recognition, IEEE* 2014, pp.no 464-469.
21. S. Swati, K. Sowjanya, R. Lakuma, S. A. Sunaina, G. Srividya, V. Rohitha. "Epidermodysplasia Verruciformis-A Genetic Disorder." *Systematic Reviews in Pharmacy* 8.1 (2017), 71-75. Print. doi:10.5530/srp.2017.1.12