



AN UNSUPERVISED CLASSIFICATION TECHNIQUE FOR RECOGNITION OF SCRATCHED AND NON-SCRATCHED WORDS IN PRE-PRINTED DOCUMENTS

¹N. SHOBHA RANI, ¹VASUDEV T, ²VINEETH .P, ²DEEPTHA AJITH

¹Maharaja Research Foundation, University of Mysore, Mysore, Karnataka

²Department of Computer Science, Amrita Vishwa Vidyapeetham, Amrita University, Mysore, India

E-mail: n.shoba1985@gmail.com, vasu@mitmysore.com, vineeth.p.nair777@gmail.com, deepthaajith@gmail.com

ABSTRACT

Pre-processing of document images is the most variant factor from one type of document image to another. In general, especially document images require more intensive pre-processing procedures than other type of images; one of such categories is pre-printed form images. Pre-processing of such documents is different from other type of images containing simple text and free from graphical components. This paper proposes a generic pre-processing algorithm adaptable for pre-printed application form images. The work supports specifically on problem of detection and removal of scratched words inherent in the text, since these elements are interpreted neither by humans nor by machines. The algorithm exploits the features like Euler's number, number of connected components and area covered by holes with in a text block for detection of scratched out text blocks. The algorithm has yielded reasonably good results with an overall efficacy of around 96.5%.

Keywords: *Irrelevant Information, scratched words, non-scratched words, Morphological Operations, Pre-printed forms, unsupervised learning.*

1. INTRODUCTION

Pre-processing of document images is an essential pre-requisite stage for the transformation of the raw input image into a more suitable form for subsequent processing stages. Moreover it is the preliminary stage of Optical Character Reader (OCR) [1], the outcome of this stage may improve the scope of attaining high accuracies during recognition of characters. Pre-processing is concerned with removal of irrelevant portions in the document which is not necessary for subsequent operations. The pre-processing procedure generally varies from one type of images to another. The variety of document images exists in practice includes bank cheques, postal documents, text books, handwritten scripts, machine printed, type-written documents, certificates and application forms etc. The category of the documents that possess both printed as well as handwritten text is pre-printed forms. There are wide variety of documents that falls under this category, includes forms in Govt./Private offices, admission forms in schools/colleges, cheques or other miscellaneous

forms employed in banks, insurance agencies and other commercial organizations.

In comparison with variety of document images that are processed by OCR, the *application form* images require more specific pre-processing to considerable extent. Especially the *application form documents* that are adapted for serving needs of various authorization in municipal offices, educational institutions and other corporation sections requires an extensive pre-processing. These pre-printed *application form documents* possess different types of graphical elements like scratched words, horizontal and vertical grid like structures, symbols, logos and emblems, etc [2]. These graphical elements are the barriers for error free recognition by OCR. In plain text documents these kinds of graphical elements are not found and eliminates the need of performing the most intensive and specific pre-processing. The pre-printed *application form* images are of different kinds and layout varies from one type of business system to the other. Depending on the type of document layout the pre-processing procedures are to be chosen to eliminate the irrelevant objects

which are not preceded for subsequent processing by OCR.

In general pre-processing step for textual images involve operations like thresholding [3], binarization [4], skew detection and correction [5], noise removal [6], layout elimination [7] etc. Attributing to pre-printed *application form* images, we have both machine printed text blocks as well as the handwritten text blocks. The handwritten text blocks are filled by users in the space provided with respect to the various fields of data entry. If user wishes to strike out any text which is incorrectly filled by him, then user scratches the block of text and generates unwanted marks resulting into a scratched word. Scratched words are generally the irrelevant, unwanted information for the user as well as the machine which are to be ignored. The presence of scratched words may lead to misrecognition, when the scratch is not clear and legible. On the other side, it introduces ambiguity during the process of classification and recognition. In addition more the scratched words in the documents invariably increases the ambiguity during classification and recognition stages. The demand for identification and removal of irrelevant information has motivated us to detect scratched text blocks in documents. Figure 1 depicts a typical pre-printed *application form document* containing few scratched words.

The generation of scratch words in the document highly depends upon the writing style as well as tendency of user to make mistakes while filling the details. Since each user has a unique writing style and qualities a variety of scratch marks are generated in a system. Scratch marks possess two characteristics, type and length respectively. The type represents its appearance, as a series of slant lines, strike-throughs, encircled text blocks, blobs and scribbles etc. The length of scratch mark is proportional to its width in terms of number of columns or the number of characters that persists in a scratched text block. In some cases the entire word is scratched and in other cases only a part of it is scratched. Therefore the scratch can exist either at word level or at character level blocks. The detection of scratch marks has to be performed both at word level and character level. The figure 2 represents few samples of scratch marks at word level and character level.



Figure 2: Types Of Scratched Words

In the proposed work, our interest is to focus on the identification of scratch marks at word level and support the elimination of irrelevant information in the pre-printed *application form documents*. We concentrate on detection of scratched words both handwritten and printed text regions in the document. The escalations of scratched words in the document images aggravate the complexity during segmentation, classification and recognition steps of OCR. It is very obligatory to identify the disproportion between the scratched and non-scratched words during the pre-processing procedure. Therefore the proposed work restricts in devising an algorithm for the identification of scratched and non-scratched words in pre-printed *application form documents*.

2. LITERATURE SURVEY

Numerous attempts are reported in the literature on pre-processing procedures for document images. Some of the approaches relevant to the proposed work are discussed in brief.

Shobha et. al [8], proposed a generic line elimination methodology for removal of horizontal

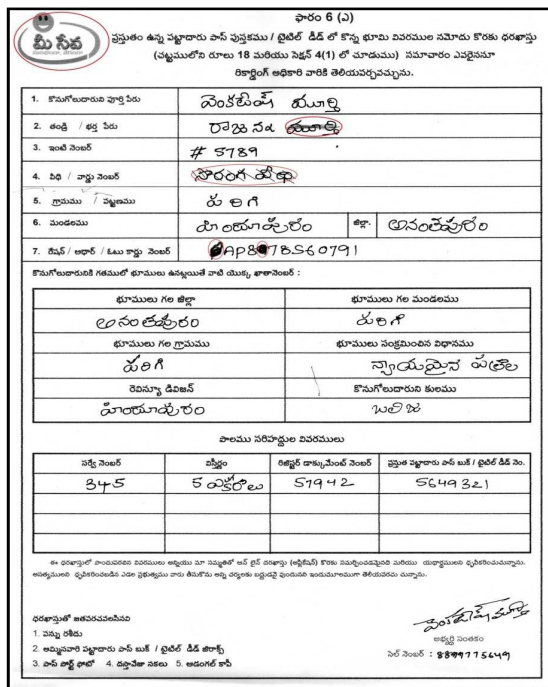


Figure 1: A Sample Of Pre-Printed Application Form

grid like structures using circular structuring element for pre-printed *application form* images and had achieved an accuracy more than 90%. Shazia et.al [9], provided a detailed overview of various document pre-processing methodologies like thresholding, binarization, skew detection and correction, normalization and also mentioned about the structural layout of documents in research articles. Dipti et.al [10], proposed some pre-processing approaches like image registration, image masking and other image enhancement techniques that makes the handwritten characters in pre-printed *application form documents* non-sensitive for recognition by neural network classifier. The form templates considered for experimentation are specialized forms used for OCR input i.e., each character is separately bound with in a rectangle. Maya et. al. [11] had evaluated the performance of various approaches for binarization based on the different thresholding algorithms. Deivalakshmi et.al [12] had proposed an algorithm for line removal from documents using shift, difference and coordinate logic operations. The algorithm had shown good results, however the document type considered is simple documents consisting of lines rather than application form documents composed of complex structures. Gatos et. al. [13] had proposed an algorithm for automatic table detection in documents using line length and line width estimation by using edge detection operators. Yefeng et. al [14] had contributed an algorithm to detect the severely broken parallel lines in handwritten document images based on directional single connected chain method using three parameters called skew angle, vertical line gap and vertical translation. The experimentation had produced results of around 94% for Arabic documents. Christian wolf [15] had contributed an algorithm for the removal of ink bleeds in handwritten text using hidden markov random fields and Bayesian MAP estimation for the document datasets of 18th century. Sandhya N et. al [16] had outlined the various types of image noise reduction approaches with respect to different types of noises that are present in document images. Reza et. al [17] had performed the classification and evaluation of document image based on matching algorithms and also evaluated the performance of various algorithms using various functional metrics. The classification is done based on various graphical elements in the images. Mohammed et. al. [18] had proposed various approaches for document image retrieval based on the detection of logos, graphical icons, signatures and layouts etc.

Lawrence et. al [19] had provided a detailed insight into various approaches available for the pre-processing of document images.

Most of the reported works concentrate on removal of noise in the document images which can improve the quality of text and makes it legible for further processing. On the other side, there are algorithms focusing on classification of graphical components like logos symbols, ink bleeds etc using morphological tools and various other feature extraction and classification algorithms. To the best of our knowledge, we have reviewed the various pre-processing tasks on document images in literature and we could not find specific procedures specially developed to detect scratched words filled in pre-printed forms designed for Govt./Private organizations. This initiated us to attempt to handle the variety of scratch words found in pre-printed *application form documents* which bottlenecks the process of character recognition.

3. PROPOSED METHODOLOGY

Detection of scratched words in the pre-printed *application form documents* is a significant pre-processing task. The classification of scratched and non-scratched words will reduce the massive computations during segmentation and classification stages of OCR. The proposed methodology performs the detection of scratched words and non-scratched words in pre-printed *application form documents*.

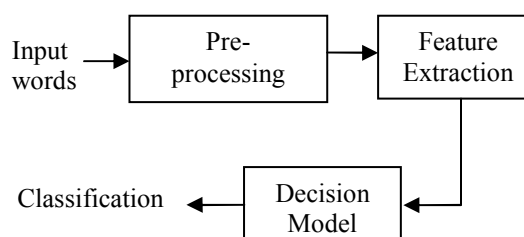


Figure 3: Architecture Of Proposed System

The algorithm assumes the words for classification are segmented from document as input. Initially the features are extracted from input word and finally the decision model accomplishes the classification. Figure 3 depicts the block diagram of proposed work.

3.1 Pre-Processing of Handwritten Word Blocks

The algorithm triggers by acquisition of segmented binarized handwritten word blocks from

the pre-printed *application form documents*. The handwritten word blocks are subjected to morphological operation ‘bridge’ [20]. The bridge operation sets a pixel to ‘1’ if it has two non-zero neighbors which are not connected, so it bridges the gaps between broken characters. Especially, for the low resolution images it provides an enhancement in gradient information of the characters formed. To achieve the connectivity between the edge pixels contributing in the formation of a character or scratches which are very thin, the morphological bridging operation is employed in the proposed methodology. The results of morphological bridge operation on a handwritten word block are as detailed in figure 4.

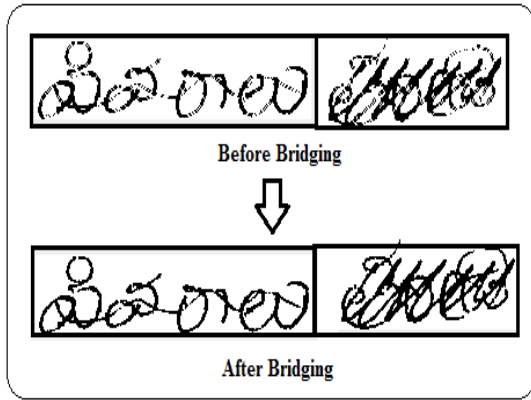


Figure 4: Result Of Bridge Operation On Scratched Handwritten Words

3.2 Computation of Features

The bridged handwritten words are further subjected to feature extraction. In the proposed algorithm, three different features are computed from the input word. The Euler’s number, number of connected components and area covered by holes in each word block are computed. The subsections 3.2.1 through 3.2.3 explore more about the features computed.

3.2.1 Euler’s number

Euler’s number is employed as one of the feature for scratched word detection in the proposed work. The South Indian scripts like Telugu and Kannada are better structural in representation and consist of number of objects within a single character. The structural diacritics of the characters may result in formation of holes and many isolated objects in a word. In regard to this, a scratched word possesses less number of objects and holes than a non-scratched word. Thus Euler’s number is employed for the detection of scratched word

block. Particularly when the scratches are very thin, these features are providing much differentiation for the classification while the other features like connected components and area non-covered by holes are not much reliable.

The Euler’s number of a binary image is the difference of number of objects and number of holes in those objects of the image [21]. The Euler’s number for a binary image is defined by equation (1),

$$E = N - H \tag{1}$$

where E , N and H are Euler’s number, number of objects and number of holes respectively.

Figure 5 represents the Euler’s features computed for 121 handwritten words including 32 Non-scratched and remaining as scratched words.

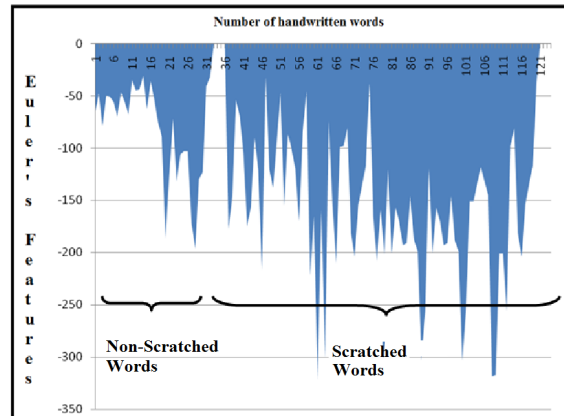


Figure 5: Euler’s Features – Scratched Vs. Non-Scratched Words

It is noticed that Euler’s features may not be efficient enough in discrimination of scratched to non-scratched words except for few cases of thin scratched words when the other features fail. Thus the other features are combined to add more reliability to the system.

3.2.2 Number of connected components

The number of connected components [22] in a handwritten word is determined for the scratched word detection. Connected components represent a sequence of adjacent pixels which are generally belonging to the single object. The connected components of a text block for a non-scratched will be obviously less in number compared to a scratched text block. Thus the number of connected components seems to be a

useful feature in identification of scratched words from non-scratched words. The number of connected components features gradually increases as the amount the scratch marks increases in the word. Thus the connected component features are employed in the detection of scratched words. The experimental analysis of scratched and non-scratched words representing number of connected components is presented in figure 6.

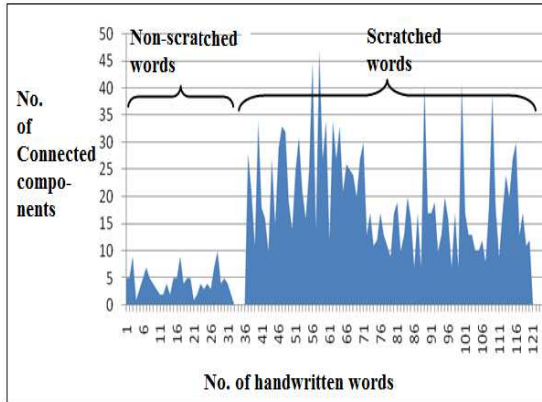


Figure 6: Number Of Connected Components For Scratched And Non-Scratched Words

The number of connected components is more robust features in comparison with the Euler’s features. The number of connected components in scratched words are very much high especially in densely scratched words. However the cases of thin scratched words are exception with respect to connected component features and can be justified with the Euler’s features.

3.2.3 Area non-covered by holes

The filling of closed regions [23] in binary images perform flood fill operations on the background pixels of the images. Further the area covered by the filled regions is estimated. The area of the filled regions in binary image corresponds roughly to the total number of pixels in the image, which is a scalar value. The area features are represented by black pixels in the each word is the non-covered regions by holes in the text block. Figure 7 shows the area features of a scratched and non-scratched word.

The area non-covered by holes with in the non-scratched word is comparatively more than the scratched words. Figure 7 clearly depicts that area covered by filled regions gradually decreases as the amount of scratch marks increases in the image.

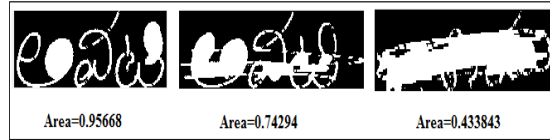


Figure 7: Area Features For Scratched And Non-Scratched Words

Thus the area of covered holes is a desirable feature for classification of scratched words and non-scratched words. Figure 8 details the various features with respect to scratched and non-scratched words.

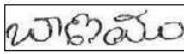
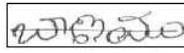

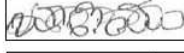

	Euler's Number	Number of connected components	Area covered by holes
	-89	5	1.71134
	-85	14	1.28815
	-46	47	0.433843
	-188	17	0.91325
	-199	7	1.23328

Figure 8: Various Features For Different Types Of Scratches

The various features with respect to different types of scratches obviously explicate the differentiation from scratched word to a non-scratched word. As mentioned in above sections, the number of connected components and area covered by filled regions are more robust compared to Euler’s number. In some cases where the scratches are very thin, the Euler’s number features are more robust compared to other features. Thus the features computed are reliable enough in classification of scratched words and non-scratched words.

3.3 Classification of Handwritten Words

The features computed from the handwritten words are analyzed to derive the threshold for classification of scratched and non-scratched words. Let E , N_c and A represents the features Euler’s number, number of connected components and area non-covered by holes respectively. The number of connected components and area non-covered by holes are used to generate a combined scratch factor R . The scratch factor R is defined as the area non-covered by holes A to

the number of connected components N_C in a handwritten word image H_T and is given by equation (2).

$$R = \frac{A}{N_C} \quad (2)$$

3.3.1 Determination of threshold for classification

The determination of user-defined threshold with respect to features E, N_c, A and R is identified by considering a handwritten word with very thin scratches. In the proposed methodology the threshold for classification of scratched and non-scratched word is estimated by computing the average of the features of around 50 handwritten words where scratches are very thin and with different types of scratches. The figure 9 demonstrates some of the thin scratched words used to deduce the average of threshold with respect to every feature.

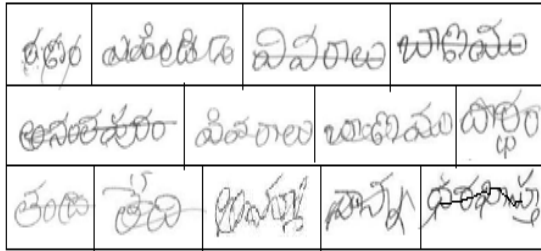


Figure 9: Instances Of Few Scratched Words Used For Threshold Computation

Let δ_e, δ_{cc} and δ_R represents the thresholds of features E, N_c and R respectively. Let $E_1, E_2, E_3 \dots E_n$ indicates the Euler number features of handwritten words with thin scratches, and then the threshold δ_e is given by equation (3).

$$\delta_e = \frac{1}{n} \sum_{i=1}^n E_i \quad (3)$$

If $N_{c1}, N_{c2}, N_{c3} \dots N_{cn}$ represents the number of connected components features of handwritten words with thin scratches, and then the threshold δ_{cc} is given by equation (4).

$$\delta_{cc} = \frac{1}{n} \sum_{i=1}^n N_{ci} \quad (4)$$

If $R_1, R_2, R_3 \dots R_n$ represents the features of ratio of area covered by filled regions to number of connected components of handwritten words with thinner scratches, then the threshold δ_R is given by equation (5).

$$\delta_R = \frac{1}{n} \sum_{i=1}^n R_i \quad (5)$$

Thus the average of non-densely scratched words is considered for determination of threshold and for the classification of scratched words to non-scratched words.

The algorithm Scratch_Detect depicts the working of proposed methodology.

Algorithm Scratch_Detect

1. Read the word image H_T .
2. Compute the Euler's number E of H_T .
3. Compute the number of connected components N_C of H_T .
3. Flood fill the background pixels of H_T to obtain F_{H_T} .
5. Estimate the area A covered by F_{H_T} .
6. Let $\delta_e, \delta_{cc}, \delta_a$ and δ_R be thresholds w.r.t E, N_c and A .
7. Compute the scratch factor $R = A/N_c$
8. if $E \geq \delta_e$
 - if $N_c \geq \delta_{cc}$
 - if $R \geq \delta_R$
 - Output " Non-Scratched word"
 - else
 - Output " Scratched word"
 - end
 - else
 - Output " Non-Scratched word"
 - end
 - else
 - Output "Scratched word"
 - end
9. Stop

4. EXPERIMENTAL ANALYSIS

The classification of scratched words and non-scratched words is performed on around 200 non-scratched words and 800 scratched words. The

datasets employed for the investigation is extracted from the Telugu pre-printed *application form documents* belonging to the regions of Anantapur district, Andhra Pradesh. Some of the scratched words are synthetically generated with the help of around 50 different users. The experimental accuracy in the proposed system is defined in two ways. One is the number of scratched words and non-scratched words recognized correctly to the total number of handwritten words extracted.

Let N_s and N_{ns} represents the number of scratched words and non-scratched words in the document. If N indicates the total number of handwritten words employed for analysis, which is given by sum of scratched and non-scratched words as mentioned in equation (6).

$$N = N_s + N_{ns} \quad (6)$$

Let the number of scratched words recognized correctly is given by N_{sc} and number of non-scratched words recognized correctly is given by N_{nsc} . Then the accuracy of the proposed system is given by the empirical relation (7).

$$Accuracy = \frac{N_{sc} + N_{nsc}}{N} \quad (7)$$

The other way in which the efficiency of proposed system is computed is as follows. Let FS^+ represents the false positives with respect to scratched words. In the proposed system, the false positives FS^+ are the number of non-scratched words that are recognized as scratched words. The Fns^+ denotes the false positives with respect to non-scratched words. The Fns^+ are the number of scratched words that are recognized as non-scratched words. The error rate of scratched word classification in the proposed system is the total number of false positives with respect to scratched words to the total number of true positives TS^+ . The true positives TS^+ represents the number of incorrectly recognized scratched words. The error rate with respect to scratched words is given by equation (8).

$$Error\ rate\ E_s = \frac{FS^+}{TS^+} \quad (8)$$

Similarly the error rate with respect to non-scratched words is given by the equation (9)

$$Error\ rate\ (E_{ns}) = \frac{Fns^+}{Tns^+} \quad (9)$$

where Fns^+ is the number of false positives with respect to the non-scratched words. Tns^+ is the number of incorrectly recognized non-scratched words.

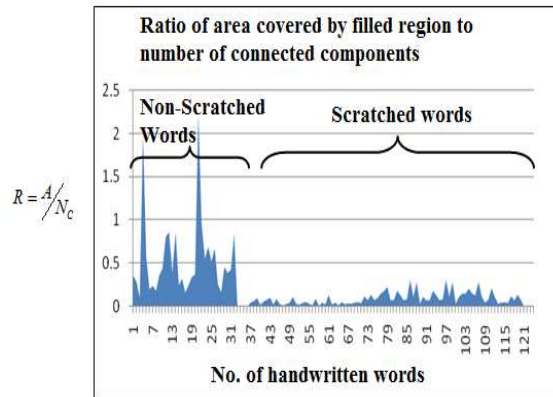


Figure 10: The Ratio Features Vs. Number Of Handwritten Words

The figure 10 provides an overview of ratio of combined features generated from area and number of connected components features to the number of handwritten word images.

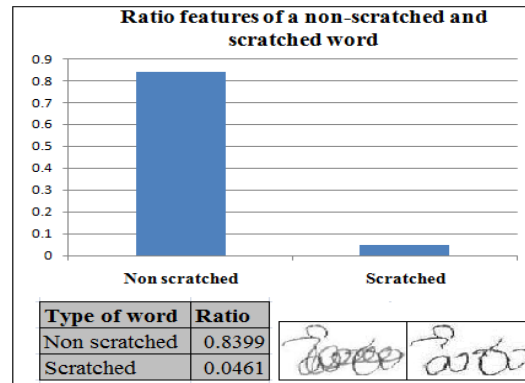


Figure 11: Ratio Features For Non-Scratched Word And Scratched Word

The ratio of area covered by filled regions to the number of connected components in a handwritten word proves to be really efficient enough in discrimination of scratched and non-scratched words. Figure 11 shows the details of ratio features with respect to a single scratched word to a non-scratched word.

Thus the ratio features from the figure 11 are very obvious in differentiating a scratched word from a non-scratched for an averagely scratched word. This implies that the ratio features includes adequate efficiency along with other features in the proposed system. The table 1 depicts the accuracies obtained for the proposed methodology.

Table 1: Experimental statistics of scratch detect algorithm

Words Type	True positives (TP)	False positives (FP)	Total	Accuracy
Non-Scratched	194	6	200	97.00%
Scratched	771	29	800	96.37%
Total	965	35	1000	96.50%

The experimentation was also extended for the few of the instances of printed words in both English and Telugu script. The outcome of experimentation are as depicted in figure 12.

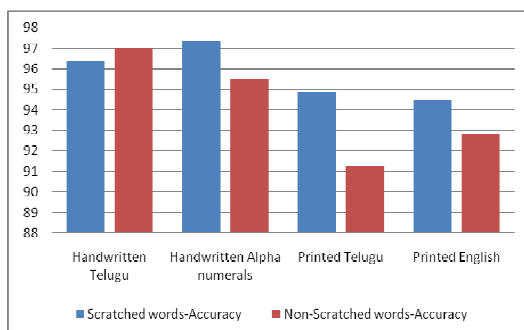


Figure 12. Efficiency Of Scratch Detect Algorithm Vs. Types Of Scripts

It is evident from the experimental statistics that the accuracies obtained are consistent in the proposed system for classification of non-scratched words and scratched words.

5. CONCLUSIONS

The proposed algorithm for automatic detection and classification of scratched and non-scratched words employs the Euler's features, number of connected components and area non-covered by holes with in the handwritten word. The hybrid features are the ratio of the area non-covered by holes to the number of connected components in the handwritten word. The Euler's features are efficient in the cases of thin scratched words and the number of connected components and ratio

features are more robust for average and densely scratched handwritten words. Further the algorithm can be extended for detection of scratch marks present at character level and also can be made portable for other language scripts which are under investigation. The proposed work can be a supporting stage for detection and removal of scratched irrelevant information and will improve the overall performance of the OCR due to the reduced computational risks in the stages of segmentation and recognition.

REFERENCES:

- [1] Kasturi, R., O'gorman, L., &Govindaraju, "Document image analysis: A primer", *Sadhana*, Vol. 27(1), 2002, pp.3-22.
- [2] Raval Ajay A, Jalwani Ayoob A, Karathiya Manoj B, "Document Image Analysis", *International journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2(5), 2012, pp. 346-349.
- [3] Ray Smith, Chris Newton, Phil Cheatle, "Adaptive Thresholding for OCR: A Significant Test", Personal Systems Laboratory, HP Laboratories Bristol, HPL-93-22, March, 1993.
- [4] Sergey Milyaev, Olga Barinova, Tatiana Novikova, Pushmeet Kohli, Victor Lempitsky, "Image binarization for end-to-end text understanding in natural images", Lomonosov Moscow State University, Moscow, Russia, Microsoft Research, Cambridge, UK.
- [5] Atallah Mahmoud Al-Shatnawi and Khairuddin Omar, "Skew Detection and Correction Technique for Arabic Document Images Based on Centre of Gravity", *Journal of Computer Science 5* (5): 363-368, 2009.
- [6] Atena Farahmand, Abdolhossein Sarrafzadeh, and Jamshid Shanbehzadeh, "Document Image Noises and Removal Methods", *Proceedings of the International Multi Conference of Engineers and Computer Scientists 2013* Vol I, IMECS 2013.
- [7] Ray Smith, "Hybrid Page Layout Analysis via Tab-Stop Detection", *10th International Conference on Document Analysis and Recognition*, 2009.
- [8] Shobha Rani N, Vasudev T, "A Generic Line Elimination Methodology using Circular Masks for Printed and



- Handwritten Document Images”, *Proceedings of second international conference on emerging research in computing, information, communication and applications, Elsevier science and technology*, 2014, ISBN: 9789351072638.
- [9] Akram, S., Dar, M. U. D., & Quayoum, A. “Document Image Processing-A Review”. *International Journal of Computer Applications*, 10(5), 35-40.
- [10] Deodhare, D., Suri, N. R., & Amit, R. “Preprocessing and Image Enhancement Algorithms for a Form-based Intelligent Character Recognition System”. *IJCSA*, 2(2), 2010, pp. 131-144.
- [11] Gupta, M. R., Jacobson, N. P., & Garcia, E. K. (2007). OCR binarization and image pre-processing for searching historical documents. *Pattern Recognition*, 40(2), 389-397.
- [12] Deiva lakshmi, S., Harinivash, B., & Palanisamy, P. “Line removal technique for document and non document images”. *11th International Conference on Hybrid Intelligent Systems (HIS), IEEE*, 2011, pp. 534-539.
- [13] Gatos, B., Danatsas, D., Pratikakis, I., Perantonis, S. J. (2005). Automatic table detection in document images. In *Pattern Recognition and Data Mining* (pp. 609-618). Springer Berlin Heidelberg.
- [14] Zheng, Y., Li, H., & Doermann, D. “A model-based line detection algorithm in documents”. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference, IEEE*. August 2003, pp. 44-48.
- [15] Wolf, C. “Document ink bleed-through removal with two hidden markov random fields and a single observation field. *Pattern Analysis and Machine Intelligence*”, *IEEE Transactions on*, 2003, 32(3), 431-447.
- [16] Sandhya, N., Krishnan, R., Babu, D. R. “A language independent Characterization of Document Image Noise in Historical Scripts”. *International Journal of Computer Applications*, 2012, 50(9).
- [17] Tavoli, R., “Classification and Evaluation of Document Image Retrieval System”, *WSEAS Transactions on Computers*, 11(10), 329-338, 2012.
- [18] Keyvanpour, M., Tavoli, R., “Document image retrieval: Algorithms, analysis and promising directions”. *International Journal of Software Engineering and Its Applications*, 7(1), 93-106, 2013.
- [19] O’Gorman, L., & Kasturi, R. “Document image analysis”. Los Alamitos, CA: IEEE Computer Society Press. VOL 39, 1995
- [20] Edward R Dougherty, Roberto A Lotufo. “Hands on Morphological Image Processing”, 2003, eISBN: 9780819478665.
- [21] C R Dyer (1980), “Computing the Euler number of an image from its quad tree”, *Computer Graphics Image Processing*, Vol. 13(3), pp 270-276.
- [22] C. A. Bouman (2015), “Connected component analysis-Digital Image Processing”.
- [23] Gray S.B, “Local Properties of Binary Images in Two Dimensions”, *IEEE Transactions on Computers*, Vol. C-20(5), 1971, pp. 551-561.