International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 12 Issue 2 –JANUARY 2015.

SEVERANCE OF THE HANDWRITTEN TO MACHINE TYPED TEXT FROM DOCUMENT IMAGE

S.Dhanalakshmi[#1] and P.Sankar[*2]

[#]*M.E., CSE, Renganayagi Varatharaj College of Engineering, Salvarpatti, India*

[*] *B.E., CSE, Renganayagi Varatharaj College of Engineering, Salvarpatti, India*

***Abstract*— Handwritten Character recognition technique plays an important role in image processing and Pattern Recognizing. This paper provides the discrimination of handwritten text from the machine typed text for the Document images. Now a days it is most useful in many emerging application field, which requires further advanced methodologies. Extraction of handwritten text from machine typed text has four steps such as Pre-processing, segmentation, future extraction and classification. In preprocessing stage, apply Fuzzy Filters for removing noise from the Document Image then apply skew angle Correction. Then segment the noiseless document image into lines, words and characters by using Connected Component. Then extract the features such as density, vertical transition rate, horizontal transition rate and Major vertical edge. Then the handwritten and machine printed text identified by using SVM classifiers. This method provides best performance based on detection rate and recognition accuracy**

***Index Terms*— Fuzzy Filters, Segmentation, Skew angle Detection and correction, SVM Classifiers.**

I. INTRODUCTION

A document documents something. It is a representation serving as evidence for some purpose. The most ancient and well known document type is of course the written document [1]. Researchers introduce the technique of converting the physical documents into images for further processing in future. That document images may contain the mixture of machine typed and handwritten, such as admission form, bank & postal services and etc. Hence it is necessary to separate these two texts before feeding them in to respective OCR (Optical Character Recognition) system.OCR is very popular research field since 1950`s. It is a electronic conversion of scanned images of handwritten, typed or printed text into machine encoded text.

II. ASSESSMENT OF HANDWRITTEN AND MACHINE PRINTED CHARACTERS

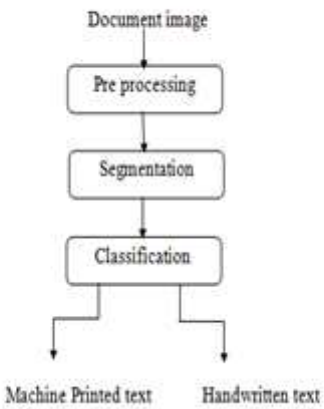
Some features of machine printed and handwritten characters which can be used for separation are given as follows:

1. Machine printed characters are written in proper alignment they have larger regularities in projection profile but the handwritten characters do not have regularities in projection profile because of varying style of different authors and the environments [8].
2. Machine printed characters are written straight whereas handwritten characters may or may not written straight. A large proportion of printed text be linear and aligned properly either horizontally or vertically while edges in the handwritten characters may not be linear [1].
3. Machine printed characters have proper spacing and are less likely to overlap, whereas handwritten characters may have overlapping and touching character which results in major challenge to preprocessing and segmentation step.
4. A machine printed text line has relatively stable height compared to handwritten text line, and the mean and variance of width of each character is consistent.
5. Horizontal run and gradients are uniform in machine printed text. If the text is repeated then it have stroke in same direction in all occurrences [10].

III. THE PROPOSED APPROACH

Separation of machine printed and handwritten is challenging task and also it become tougher in the documents in which handwritten texts overlapping with the printed text and another major problem is the presence of noise and skewing in the document images.

Figure 1.1 flow Diagram



The Separation of Machine and Typed text have following stages. The Fig. 1.1 Shows the Flow diagram.

A. Pre processing
B. Segmentation
C. Classification

A. *Pre processing*

Preprocessing is improving the quality of the image and smoothening the input image. So for improving quality we have to remove noises from an input image. Several filters are available for removing noise from an image[3]. Such as mean filter, median filter, 3*3 median filters etc, In proposed method we use fuzzy filter for removing noise [2]. Because fuzzy filter performs both preservative and smoothing. After removing noise, apply skew angle detection and correction[4] [5] . Fig. 1.2 a and b Shows input image and skew corrected and noise removed images.

Figure 1.2.a Input Image b. Skew corrected and noise removed image

Skew angle correction is performed using centroid of image. These have 2 steps such as base line identification and skew angle correction. Base line is the line that is base for the calculation of center of gravity. It is used to detect the skew in the document. The angle is measured which gives the angle by which the word or document is rotated. The algorithm is as follows:

Algorithm:

Step 1: Determine the farthest point in all the direction.

Step 2: Find the centroid using these four points, so the previous 4 points represents the polygon corners and polygon centers (COG) can be calculated by using the equations,

$$c_x = 1/6A \sum (x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$$

$$c_y = 1/6A \sum (y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$$

where $A = 1/2 \sum_{i=1}^{n-1} x_i y_{i+1} - y_{i+1} y_i$

Step 3: To get the baseline, join centroid to the origin.

Step 4: Find the skew angle $\vartheta = \arctan\left(\frac{c_x}{c_y}\right)$

Step 5: Rotate the document in the reverse direction (clockwise direction) by the skew angle.

After skew angel correction we need to convert the Document image into binary image. Binarization have two types such as local and global. Global binarization calculates one threshold value for whole image while local binarization calculates threshold value for each region. Select the threshold (T) value for each region. Let I(x,y) is the original image, the binary image B(x,y) is calculated as

$$B(x,y) = \begin{cases} 1, if\ I(x,y) \leq T \\ 0, if\ I(x,y) > T \end{cases}$$

Threshold value selection is important feature for binarization. Here the local threshold technique is used for binarization because it selects the threshold value for each region of the image.

B. *Segmentation*

Segment the binarised image based on the fast connected component (CC) labeling. Segmentation subdivides the document in constituent objects or regions [6] . It is used to identify the basic objects in the documents. The result of segmentation is splitting up of the image from the connected areas. There are 3 types of segmentation in document image. Such as line segmentation, word segmentation and character segmentation.

Line segmentation is done at the preprocessor stage by calculation of base line. Segment the word and character by using connected component labeling [7]. First scans the image from the left corner. When it encounters the first pixel it identifies the complete character through connected component. Each connected component is enclosed in a box. The midpoint of the character is computed. Similarly the second character is identified and midpoint value is computed.

The Euclidean distance between the midpoints is computed to know whether the character belongs to the same line or next line. This is determined based on the threshold which is based on the assumption that the space between the text lines is greater than that between the characters[11] . So from this each character is segmented.

C. *Classification*

Classifications are performed based on the density, Vertical Projection variance, Major horizontal projection difference, Major vertical Edges of each connected component [8]. The following formulas used for the calculation of the above mentioned:

Density D (CC) = $\frac{No.of\ foreground\ pixels\ Fn(CC)}{Total\ no.of\ pixels\ in\ the\ Bounding\ Box\ (H(cc)W(CC))}$

CCs are considered as noisy elements and are eliminated if $H(CC) < 2$ or $W(CC) < 2$ or $D(CC) < 0.05$ or $D(CC) > 0.9$ or $E(CC) < 0.08$ [11].

Vertical Projection Variance (VPV) : The vertical projection of black pixels inside the CC is evaluated . Then, the variance of the vertical coordinates of the profile of this vertical projection is calculated as a measure of homogeneity of the projection profile.

Major Horizontal Projection Difference (MHPD) : For computing this feature, the horizontal projection of the pixels inside the CC is executed. Then the major absolute difference of the abscissas of adjacent pixels of this profile is computed.

Major Vertical Edge(MVE) : The vertical edges inside the CC obtained of last feature are used here. However, only the bigger vertical edge (that is, the one with greater number of pixel) is considered to compute this feature. Thus

the relation between the number of pixels of the major vertical edge inside the CC and its height is a new feature, which we call ”Major Vertical Edge”.

If {(Major Vertical Edge <= 0.422) and (Vertical Projection Variance <=99)} then the result is handwritten. If{(Major Vertical Edge > 0.422) and (Major Horizontal Projection Difference > 27)and (Pixels Distribution <= 594)} then the result is Printed If (Major Horizontal Projection Difference <=20) then the result is handwritten[11]. The Fig. 1.3 shows this process.

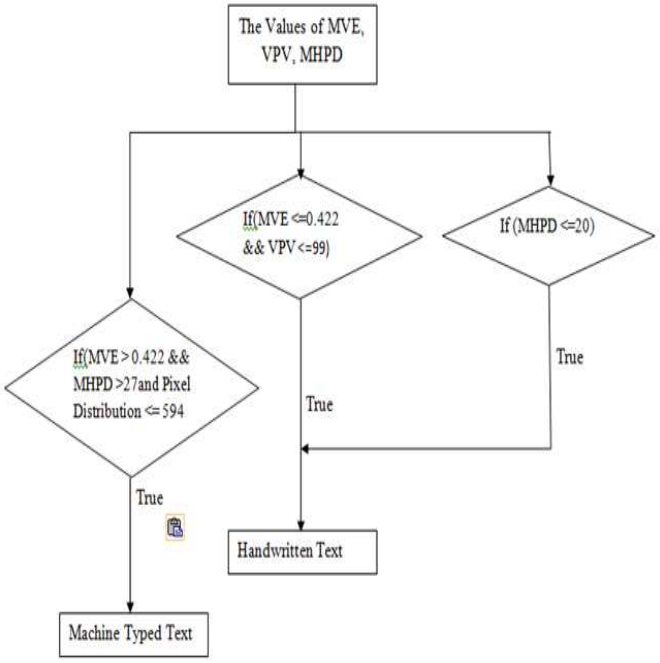


Figure 1.3 Separation of Handwritten and Machine printed text

SVM is one of the best known methods in pattern classification and image classification [9]. For the separation of handwritten from machine text 2 SVM classifiers are used [9]. The first (SVM1) deals with the handwritten text problem against all the other and the second (SVM2) deals with the machine printed text problem against all the other. There are four outcomes.

1. If the SVM1 output is *TRUE* and SVM2 is *FALSE* then the block contains handwritten text.
2. If the SVM1 output is *FALSE* and SVM2 is *TRUE* then the block contains machine printed text.
3. If the SVM1 and SVM2 output is *FALSE* then the block contains noise.
4. If the SVM1 and SVM2 output is *TRUE* then the distance of the block descriptor with the closest Support Vector for each SVMi is calculated.

Finally, among those two distances the SVMi that is related to the maximum distance defines the class of the block. Another advantage of the proposed approach is the training of only two SVMs instead of three SVMs. This reduces the computational cost and considerably increases the speed of the process.

IV. CONCLUSION

This paper produces best performance based on detection rate and recognition accuracy. The detection rate of this technique is 97.5%.

In the future it is planned to implement other classification techniques (such as minimum distance, neural networks, and those based in on fuzzy logic) for comparisons with the proposed classification techniques. More over. Other future improvement could be hybrid text segmentation for segmenting. Finally, it is to be considering improving this system for performing the documents with tables, figures, graphs and other elements.

REFERENCES

[1] David doermann, Karl tombre editors, “Handbook of Document Image processing and recognization”, *Springer Reference*, 2013.

[2] C.Mythili and Dr V. Kavitha. “ Efficient Technique for Color Image Noise Reduction” ,*The Research Bullet in of Jordan ACM*,Vol II(III) 2011.

[3] 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, March 13 - 15, 2013, Hong Kong

[4] Naazia Makkar and Sukhjit Singh, “ A Brief tour to various Skew Detection and Correction Techniques” *International Journal for Science and Emerging Technologies with Latest Trends* Vol 4(1): 54-58 -2012.

[5] R.J.Ramteke, Imran Khan Pathan, S.C.Mehrotra, “ Skew Angle Estimation of Urdu Document Images: A Moments Based Approach” *International Journal of Machine Learning and Computing*, Vol.1, No. 1, April 2011

[6] Vassilis Papavassiliou, Themos Stafylakis, Vassilis Katsouros, George Carayannis, “Handwritten document image segmentation into text lines and words” *Journal of Elsevier in Pattern Recognition* 43 (2010) 369 – 377 2010.

[7] Sunanda dixit, Dr.N.H.Suresh, “South Indian tamil language handwritten document text line segmentation technique with aid of sliding window and skewing operations” *Journal of Theoretical and Applied Information Technology* Vol. 58 No.2 2013.

[8] D. Abdel Bela, K.C Santosh, Vincent Poulain d'Andecy , “Handwritten and Printed Text Separation in Real Document” *Cornall university Library in the subject of Computer Vision and Pattern Recognition cite as* arXiv:1303.4614 [cs.CV],2013.

[9] Konstantinos Zagoris, Ioannis Pratikakis, Apostolos Antonacopoulos, Basilis Gatos, Nikos Papamarkos, “Handwritten and Machine Printed Text Separation in Document Images using the Bag of Visual Words Paradigm” *International Conference on Frontiers in Handwriting Recognition* 2012.

[10] Richa Goswami, O.P.Sharma, " A Review on Character Recognition Techniques" *International Journal of Computer Applications* (0975 – 8887) Volume 83 – No 7 2013

[11] Lincoln Faria da Silva and angel sanchez," Automatic Discrimination between printed and handwritten text in documents"

First Author- S.Dhanalakshmi received M.E. degree in Computer science and engineering from dhanalakshmi Srinivasan Engineering College in 2009. Currently she is working the assistant Professor in the department of Computer science and engineering at Ranganayagi Varatharaj College of Engineering, Sivakasi. Her research interests involve Document Image Processing and Image segmentation.

Second Author- P.Sankar doing his B.E. in the department of Computer Science and Engineering from renganayagi varatharaj college of engineering.