The Seybold Report

ISSN 1533-9211

DOI 10.5281/zenodo.6891896

HYBRID APPROACH FOR FRAGMENTING DEVNAGARI DOCUMENT IMAGES TO CHARACTER

Dr. SARIKA T DEOKATE

Computer Engineering, Dr. D. Y. Patil Institute of Technology Pimpri Pune, India.

Abstract:

In any OCR, segmenting the manuscript needs special attention as there are numerous issues associated in the Devnagari script. Printed or handwritten manuscripts have diverse concerns, which need to be studied deeply. To fragment the manuscript which is handwritten needs a lot of pre-processing tasks. As this type of manuscript contains diverse strokes, ink variations, extra drawings on the manuscript, slant in the writing of lines, words and many more issues. Printed manuscript processing is also facing many issues, as it contains many font types, font sizes, degraded manuscripts, available datasets etc. In this work, we tried the fragmentation method which generally works for both printed and handwritten documents. By using our system lines and words are fragmented to a superior extent, in character fragmentation approximately 85% of characters are fragmented correctly. Some of the characters may not get fragmented correctly and may remain partially together due to some noise, overlapping characters etc. In future, we will work on these issues.

Keywords: Classification,Dilation, Erosion, NLP, OCR, Segmentation, Vertical Projection

1. Introduction

The field of Natural Language Processing (NLP) is considered as the field which is providing the communiqué amid natural languages of Human and machine. To perform this, NLP takes the advantage of AI artificial intelligence. Day by day the research work is extending towards further excellence for a variety of applications in real life. In various diverse commercial applications, it is not practical and promising approach to perform the operations like compression or recognition on an intact manuscript image unswervingly. That's why diverse algorithms are designed and executed for the fragmentation of the intact images. Fragmented image is utilized to categorize, making the cluster or group of the image parts called as segments, in accordant to the different available features of that image[1], [2]

The accomplishment of any OCR is dependent on the accurateness of fragmentation of words and then characters. Many researchers worked on the applications like postal manuscripts, books, cheques, forms at diverse regions for their provincial languages[3]–[5].

Image fragmentation is utilized to fragment the intact image into the more useful and consequential segments with identical properties and features. This is done to obtain the best outcome at the time of categorization for the analysis and identification. Till date, lots of fragmentation approaches are utilized in the applications for the processing of various types of images, speeches, audio and video. Based on these methodologies, these algorithms are grouped as edge based fragmentation, histogram based, region based fragmentation and region based fragmentation[6].

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6891896

The significant fields where fragmentation of image is utilized prominently are: CBIR i.e. content based image retrieval, medical field, satellite, traffic manage system, sentiment analysis, object discovery in various fields, speech recognition and many more. After segmentation, diverse classification techniques has been studied and applied to recognize and classify these characters[7]–[9]

In this system we learnt about the different algorithm utilized by the researchers for the English, Chinese, Devnagari script e.g. Hindi, Bangla and Marathi etc. The different scripts have different features, characteristics and depending on these types the researchers utilized the segmentation algorithm[4], [10], [11]. It is very necessary to choose the right algorithm with right factors to get the good outcome for the proposed system.

2. Proposed System

The extraneous spurious entities are removed in pre-processing to avoid falsification categorization[12]. To perform the fragmentation of lines, characters and words, the existing histogram approach with morphological operations has been utilized, but with some enhanced approaches. In this fragmentation, instead of fragmenting the characters, upper modifiers and lower modifiers separately, the characters are separated with the modifiers. To check the perfect characters even Shirorekha i.e. header line is eradicated and bounding region is estimated. The whole process is explained in thefollowing section.

Morphological operations are applied in computer vision for the several applications like smoothing, thinning, shape identification, noise eradication, contour and edge discovery, object discovery and analysis. These operations are utilized for detecting the objects, lines, column discovery[13], [14]. In this system, we utilized the erosion with dilation operation and horizontal profile for detecting the manuscript lines.

2.1 Line Extraction

The source image is taken as the input for the extraction of the lines. Smoothing of image is done using the morphological dilation then erosion approach which is also termed as opening technique. Horizontal morphology has been applied on the image to reduce the noise. Horizontal dilation has been performed for discovering the lines. In dilation it inserts the pixels at the borders of the entity within an image.Dilation is performed on the image S utilizing the structuring element T which constructs the new bi-level image$\mathbf{f} = \mathbf{S} \oplus \mathbf{T}$.

Dilation can be termed as:

$$\text{Dilation (S, T)} = S \oplus T = \cup_{\beta \epsilon T}(S + \beta) \text{Where} - T = \{-\beta \,|\beta \epsilon T\} \qquad (1)$$

The pixel value of the output is decided utilizing the largest values from all the pixels that exist in the structuring element. To perform this, the horizontal kernel has been utilized. Horizontal projection has been performed and then noise is removed using the opening operation. In this system the horizontal projection is estimated for the image. Using the horizontal histogram, the upper rising and lower falling (min) peak points are estimated.These entities can be taken out from the background by using some threshold

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6891896

range. Means if any pixel point is above the specified threshold then it is considered as entity else considered as background.

Projection profile works mainly with the quantity of the entity pixels. These entity pixels are represented on the X and Y axis. Each point on this axis represents the quantity of the pixels which are above the specified threshold value[3]. The vertical and horizontal type projection profile is depicted as below:

$$\text{VerticalPP(Y)} = \sum_{1\le x\le i} g(x,y) \qquad (2)$$

and

$$\text{HorizontalPP(X)} = \sum_{1\le y\le j} g(x,y)` \qquad (3)$$

Where (i) is the number of rows and (j) is the number of columns.

Bi-level image or Greyscale image can be depicted by utilizing a f(x, y) function i.e. the pixel terms can be indicated in a vector plane on X and Y axis. Histogram is utilized as one of the scheme in the structural analysis. Histogram method utilized in both manuscript line and manuscript word fragmentation.

2.2 Word extraction

The extracted lines are fetched one by one for the word segmentation.If there are any breaks in the lines and words or matras, these are joined by utilizing the morphological structuring element. Rectangular element of different dimension is tried. Vertical dilation is utilized to connect the broken parts of the words.Then tiny white noise is eradicated by utilizing the erosion operation by using 5 by 5 kernels.

Erosion for image S and structural element T can express as following:

$$\text{Erosion(S, T)} = \mathbf{S} \ominus (-\mathbf{T}) = \cap_{\boldsymbol{\beta}\epsilon\mathbf{T}}(\mathbf{S} - \boldsymbol{\beta}) \text{Where, } -B = \{-\beta \mid \beta \epsilon B\}. \qquad (4)$$

Here both the closing and opening operation is utilized. Closing performs the closing of the small broken holes and opening performs the eradication of the noise. For both these operation, rectangular structuring element is utilized. Closing is nothing but erosion succeeded by dilation. It is mainly utilized to fill the tiny dark spots over the forefront entities or tiny holes within the entities.

Opening on image S and structural element T can be expressed as following:

$$\text{opening(S, T)} = \text{S} \circ \text{T} = \text{Dilation(Erosion(S, T), T)} \qquad (5)$$

Closing on image S and structural element T can be expressed as following:

$$\text{closing(S, T)} = \text{S}\bullet\text{T} = \text{Erosion (Dilation(S} - \text{T), } -\text{T)} \qquad (6)$$

The contours of the extracted lines have been estimated. These contour points are utilized to find the external contour point of the words. Using these points the perfect height and width of the bounding box is decided. Also the leftmost y and x pixel point values are preserved. Then these heights are sorted and median of these height points is estimated. Here median of

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6891896

the bounding box is estimated and ROI i.e. region of interest is calculated and then words are extracted from the original line images. Due to this estimation, as shown in Fig. 4 if some of the lines which are fragmented together also works well to segment the words. This results with good outcome for line and word segmentation also.

2.3 Character Fragmentation

Many researchers utilized the vertical profile projection approach for the fragmentation of character. In this system, we utilized the header line extraction and nearing region estimation approach. We also filled the gap amid the letters/characters and its modifiers. The extracted word list is accepted and then header line is detected for these words. The height and width of these words are checked. If the width of the word is less than 32 then these entities will be small unwanted things and will be discarded as it may not have header line. It will not be considered for the fragmentation.

Vertical projection profile has been performed to identify the lines on the words.Maximum value of the horizontal line is estimated and then it is considered as the “Shirorekha” i.e. header line and then masked with zeros using line function. The upper modifier and letters are separated. Bresenham Line is utilized to mask the header line with dark pixels.This method is applicable to the images which are having minimum or negligible skew. So if there is high skew, it is very necessary to reduce the skew at word level also. Again the unwanted entities are detected from this image and removed it.

Two images of words are preserved here, one with header line and other without header line and applied for the character fragmentation. Next vertical projection profile has been estimated for the non-header word image. The left and right nearing neighbours are checked within the tolerance region. The right and left nearing region is checked using the estimated vertical histogram. If the tolerance factor is matching, then new width and height is estimated. And characters with modifiers are extracted. If some of the characters are not segmented, corrected its width is checked above specific range and again it will be sent for the fragmentation. Thus we are achieving the segmentation at good extent. Thus we targeted the fragmentation of the lines, words and characters successfully.

3. Result and Discussions

Morphological dilation executed on the source image for taking out the every line. Applied the horizontal form of projection to estimate and execute it lucratively. The unrequited entities are eradicated by checking the threshold. The zone of the image lines are taken out as detached lines.

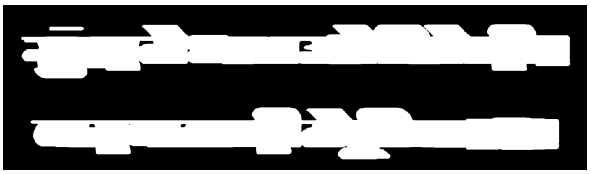

Figure 1: Dilated Outcome

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6891896

The taken out lines are showed in the following diagram for the considered source image.

Figure2: Extracted Line 1 from source image

Figure 3; Extracted Line 2 from source image

We done the analysis on many images of newspaper, acquired graphitize images from the internet, poem images, book pages etc. Some newspaper outlets are scanned and preserved. Any image format can be utilized in this work. This also works well for the English manuscripts,some handwritten manuscripts which are written on proper Shirorekha and two lines are not mixed with each other. For such type though line fragmentation not works well, word fragmentation and then character fragmentation tries to separate out the line, words and then characters correspondingly. We trailed and tallied our system on many images with diverse patterns. In few cases, it segments the two lines together. E.g. the two lines are detached together for one of the image as shown beneath

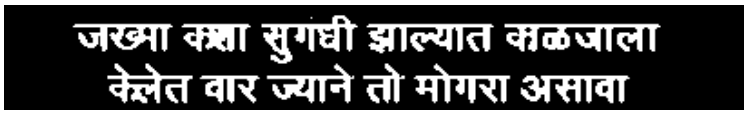

Figure 4: Fused Outcome of the Line Fragmentation

But still our word fragmentation algorithm works in superior manner and extracts these lines into words. This is depicted in the beneath outcome.

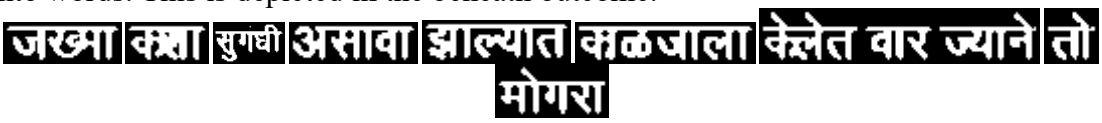

Figure 5: Word Fragmentation of fig. 4

Rectangular nature bounding box has been utilized to take out the word images. Here the height of every contour is checked before performing the fragmentation. Utilizing these heights, the average median height is estimated to take out the words perfectly with its upper modifiers i.e. matras, Velanties and lower modifier i.e. Ukar and other form of modifiers. Then the boundary of every word is decided and accordingly fragmented. That is why though some of the lines are not properly fragmented, this approach does the word fragmentation effectively. The outcome of the word fragmentation for our source image is showed in the beneath diagram.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6891896

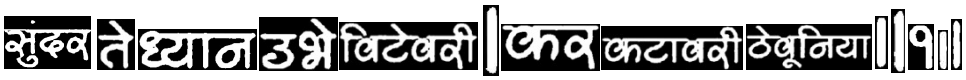

Figure 6: Source Image Fragmented Words

As we already illustrated in above cases, word fragmentation is done flawlessly for the extracted lines of the source image. But issue of these connected words too resolved at some extent at the time of character fragmentation as shown in fig. 7 (b).

(a) 

(b) 

Figure 7: Outcome of Collectively fragmented words

In our system we eradicated the tiny entities having the size less than 32 pixels by testing its height and width. To perform the correct fragmentation here the mixture of vertical projection and Shirorekha removal approach has been utilized. This is proved to be better.Bresenham line discovery algorithm utilized to discover the flat line Shirorekha and then mapped to dark pixel to remove it. The boundary of the characters is discovered utilizing the non-shirorekha image with Shirorekha image. It is proved to be better as compared to purely vertical projection approach. The concluding outcome of the character fragmentation for the source image is depicted in the following diagram.

As we can see, one image is fragmented with two letters. To re-fragment these types of images, particular image is thinned and then again character fragmentation is performed. In character fragmentation we got at the par result as compared to the previous approaches.

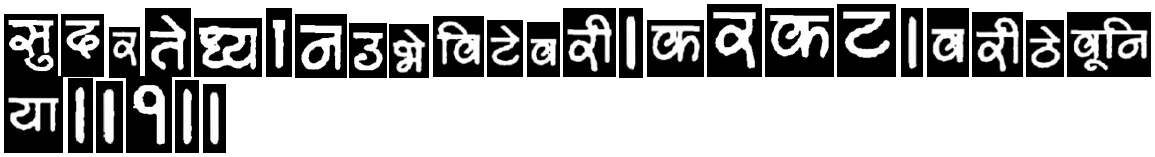

Figure 8: Fragmented symbols

We tried the distinct images with diverse form to check the precision of line fragmentation, word and letters fragmentation. Rigorous evaluation has been done and depicted that our system is performing well for the images which are having some spurious entities, objects, tiny skew in it. We also examined our system for the newspaper articles which are having two column forms.Still the fragmentation is done with superior precision. Some of the news contains the photos or some object which are also taken out successfully and eradicated at the time of performance of the character segmentation.

4. Conclusion

In this system, we learnt diverse fragmentation approaches utilized in the literature. We illustrated our designed algorithm for the fragmentation of the lines, words and then characters.This system works well for the images which are having less skew, tiny entities or

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6891896

some graphical images. In future concentration will be on the overlapping character fragmentation and try to implement a generic system for printed and handwritten documents.

References

1. V. Bansal and S. R. M. K., "Segmentation of touching characters in Devanagari," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 1655, pp. 151–156, 1999, doi: 10.1007/3-540-48172-9_13.
2. U. Pal and B. B. Chaudhuri, "Machine-printed and hand-written text lines identification," Pattern Recognit. Lett., vol. 22, no. 3–4, pp. 431–441, 2001, doi: 10.1016/S0167-8655(00)00126-4.
3. R. Jayadevan, S. R. Kolhe, P. M. Patil, and U. Pal, "Database development and recognition of handwritten Devanagari legal amount words," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, pp. 304–308, 2011, doi: 10.1109/ICDAR.2011.69.
4. U. Pal, T. Wakabayashi, N. Sharma, and F. Kimura, "Handwritten numeral recognition of six popular Indian scripts," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2, pp. 749–753, 2007, doi: 10.1109/ICDAR.2007.4377015.
5. A. K. Bhunia, P. P. Mukherjeea, Subham Sain, Aneeshan Bhattacharyya, Avirup Bhunia, Ankan Kumar Roy, and U. Pal, "Indic Handwritten Script Identification using Offline-Online Multimodal Deep Network," 2018.
6. S. T. Deokate and N. Uke, "Various Traditional and Nature Inspired Approaches Used in Image Preprocessing Sarika," in Techno-Societal 2016, 2018, pp. 345–352, doi: 10.1007/978-3-319-53556-2.
7. S. T. Deokate and N. J. Uke, "CNN Classification Approach For Analysis And Recognition Of Marathi Manuscript," Int. J. Sci. Technol. Res., vol. 9, no. 2, p. 2, 2020, [Online]. Available: www.ijstr.org.
8. S. T. Deokate and N. J. Uke, "Review on Deep Learnable Approach for Categorization," JASC J. Appl. Sci. Comput., vol. 6, no. 2, pp. 2530–2533, 2019.
9. S. S. Shelke and S. S. Apte, "A Multistage Handwritten Marathi Compound Character Recognition Scheme using Neural Networks and Wavelet Features," … Pattern Recognit., vol. 4, no. 1, pp. 81–94, 2011, [Online]. Available: http://www.earticle.net/Article.aspx?sn=148423.
10. M. Mathew, A. K. Singh, and C. V. Jawahar, "Multilingual OCR for Indic Scripts," Proc. - 12th IAPR Int. Work. Doc. Anal. Syst. DAS 2016, pp. 186–191, 2016, doi: 10.1109/DAS.2016.68.
11. S. T. Deokate and N. J. Uke, "Hybrid methods for Segmenting and Identifying the Marathi Text," 2019 IEEE 5th Int. Conf. Converg. Technol. I2CT 2019, pp. 1–5, 2019, doi: 10.1109/I2CT45611.2019.9033923.
12. B. Gatos, I. Pratikakis, and S. J. Perantonis, "Adaptive degraded document image binarization," Pattern Recognit., vol. 39, no. 3, pp. 317–327, 2006, doi: 10.1016/j.patcog.2005.09.010.
13. X. Ye, M. Cheriet, and C. Y. Suen, "A generic method of cleaning and enhancing handwritten data from business forms," Int. J. Doc. Anal. Recognit., vol. 4, no. 2, pp. 84–96, 2001, doi: 10.1007/s100320100056.
14. Y. Hsiao, C. Chuang, J. Jiang, and C. Chien, "A Contour based Image Segmentation Algorithm using Morphological Edge Detection," 2005 IEEE Int. Conf. Syst. Man Cybern., vol. 3, pp. 2962–2967, 2005, doi: 10.1109/ICSMC.2005.1571600.