ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 4, April 2016

# *Detection of Word by Inter-Intra Gap Technique for Handwritten Documentes.*

**Komal Kadam , Dipali Phadatare , Aarti Mali ,Pallavi Nimbalkar ,Prof.Priti Gode**

*Abstract* — **Segmentation of physically composed record photographs into substance material follows and words is an essential undertaking for optical character affirmation. In any case, for the reason that elements of manually written record are not proper and different depending at the character, it's far viewed as a testing issue . A decent approach to manage the bother, we define the word division issues as a parallel quadratic undertaking issue that considers pair-wise relationships among the holes and the probabilities of individual distinction. Regardless of the way that disjoin a parameters are secured in this itemizing, we gage all parameters considering the based Bayesian classifier so that the proposed strategy highlights splendidly paying little appreciate to creating styles and made lingos without benefactor described parameters. In this paper we are utilizing inter_gap,and intra hole for finding the crevice between two words. Test consequences for ICDAR 2009/2013 handwriting split databases display that proposed technique performs the high caliber in wonderfulness execution on Latin-fundamentally based and Indian languages.**

***Keywords***- *Handwritten documents, Bayesian classifier, word segmentation.*

## I. INTRODUCTION

Segmentation of record photographs into substance takes after and words is an imperative step for the record information. In any case, not like system revealed reports, the division of deciphered chronicles is still considered a troublesome issue because of (i) bizarre scattering among expressions and (ii) assortments of making styles depending at the man or woman. With respect to ICDAR 2009 and 2013 handwriting division challenge affects, the substance line division counts had been produced to a degree, in any case, there is in light of current circumstances an extraordinary arrangement space for changes inside the case of word segmentation systems. For the word division, report pictures are at first divided into substance lines. By then, the word segmentation computation (for a single substance line) is associated with individual substance lines. Given a singular substance line, the customary word segmentation computations involve two stages: the underlying step is to focus contender for between word openings (word-separator) and the accompanying step is to mastermind the hopefuls into intra/between word gaps. For the candidate period, a given substance line is addressed with a game plan of super-pixels (where a super-pixel as a rule analyzes to a letter or a social occasion of letters) and their openings are viewed as contender to be gathered. This is a parallel gathering issue that doles out a name, where 0 infers that the hole is an intra-word opening and 1 exhibits it is a between word gap [1].

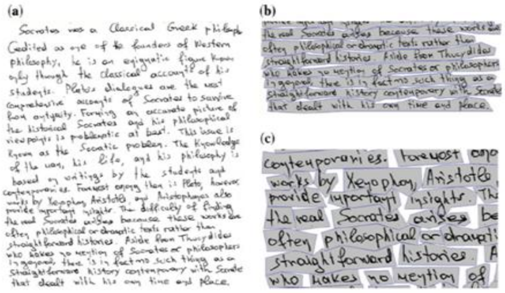


Figure 1 : (a) Handwritten document image (b)text line (c) Word segmentation

In the Fig.1, A handwritten document image has been taken as a sample for text line and word segmentation where writer has performed text and word segmentation manually. From the outcome appeared in Fig. 1, we can extremely well reason that there were no rules given to the scholars, the yield pictures contain every difficult issue for written by hand record division, e.g. contrast in the skew edge between content lines or along the same content line, presence of neighboring content lines or words touching, presence of characters with various sizes and variable intra-word gaps.

Disregarding the way that the qualities of between expression openings are changing over all through (or even in) documents, there are solid associations (e.g., scale) between them in an artistic substance line. Regardless, it has been hard to abuse these connections in the standard systems, where the class is made openly in light of the homes of every hole. That allows you to help these issues. We broaden a novel structure that considers these associations despite neighborhood discernments (i.e., the spots of every opening). To be particular, we characterize the word division as a change issue that expands the likeness among between expression openings and the difference in the middle of word and intra-word cleft, despite the probabilities.

Like other standard systems, we consider the word division issue as a naming issue that assigns an engraving (intra-word/between word gap) to every hole in a given substance line. Along these lines, we first present systematized super-pixel representation structures that think a blueprint of hopeful openings in every substance line. By then, we portray the task issue as a parallel quadratic issue, which permits to consider pair-wise relations besides neighborhood properties.

## II. LITERATURE REVIEW

Page layout analysis is a traditional document processing technique proposed by L. O'Gorman to determine the format of a page. The document spectrum is a method for structural page layout analysis based on bottom-up, nearest neighbor clustering of page components. The method yields an accurate measure of skew, within line, and between-line spacing and locates text lines and text blocks [2]. In practical scenario it is very difficult to build such type of architecture coz it is expensive and lengthy process. Cursive script word recognition is the problem of transforming a word from the iconic form of cursive writing to its symbolic form. Several component processes of a recognition system for isolated offline cursive script words are described. A word image is transformed through a hierarchy of representation levels: points, contours, features, letters, and words. A unique feature representation is generated bottom-up from the image using statistical dependences between letters and features. Ratings for partially formed words are computed using a stack algorithm, several novel techniques for low- and intermediate-level processing for cursive script are described, including heuristics for reference line finding, letter segmentation based on detecting local minima along the lower contour and areas with low vertical profiles, simultaneous encoding of contours and their topological relationships, extracting features, and finding shape-oriented events [5].

Word extraction from handwritten text lines usually involves the calculation of a line specific threshold which separates the gaps between words from the gaps inside the words in that line. T. Varga and H. Bunke show that traditional approach can be improved if the decision about a gap is not only made in terms of a threshold, but also depends on the context of that gap, i.e. if the relative sizes of the surrounding gaps are taken into consideration[10]. For this purpose, they developed a model to build a structure tree of the text line, whose nodes represent possible word candidates. Such a tree is traversed in a top-down manner to find the nodes that correspond to words of the text line. The problem of separating words in a handwritten line is made difficult by the presence of non uniform spacing between words and between characters within a word. A central sub-problem in word separation is the estimation of gaps between adjacent components in a line. U. Mahadevan and R. Nagabushnam has introduced a technique to estimate inter-component distances that is based on the gap between their convex hulls [6].

The projection profile of a text line is a one-dimensional array that demonstrates the quantity of pixels for every even position. Subsequently, the zero-run (the length of back to back zeros) of projection profile has been misused for the word division of machine-printed archives. In any case, in manually written archives, zero-run highlights turn out to be less notable since letters in various words are liable to touch one another and the skew (or bend) of a content line might degenerate the zero-keep running in the projection profile [12]. ICDAR 2009 and ICDAR 2013Handwriting Segmentation Contest was organized in the context to record recent advances in off-line handwriting segmentation. The contest includes handwritten document images produced by many writers in several languages (English, French, German and Greek). These images are manually annotated in order to produce the ground truth which corresponds to the correct text line and word segmentation result. Two benchmarking datasets, one for text line and one for word segmentation, were created in order to test and compare all submitted algorithms [3][4].

## III. MATHEMATICAL MODEL

Let S is the Whole System Consist of

S= {I, P, O}

I = Input.

I = {U, IMG, D}

U = User

U = {u1,u2….un}

IMG = Image

IMG = {img1, img2…imgn}

D = Dataset.

P = Process:

P = {TS, P, HM, PP, WS, DC, GC, SR, Op}

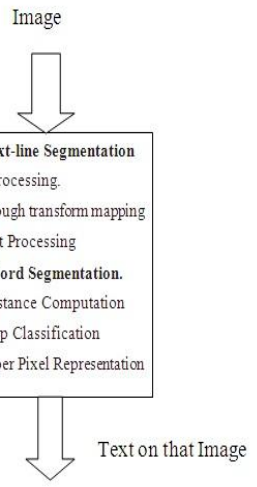


Fig 3.Mathematical Flow of System

TS= Text-line Segmentation
P=Pre Processing.
HM=Hough transform mapping
PP= Post Processing

WS= Word Segmentation.

DC= Distance Computation
GC= Gap Classification
SR= Super Pixel Representation
Op= Text on that image.

### 1.1. Algorithm For Proposed System

**Cutting Plane Algorithm:**

1. First solve the LP-relaxation.
2. Optimize using primal simplex method.
3. The optimal solution is fractional.
4. Generating an objective row cut.
5. A new slack variable is added.
6. The new cut is added to the dictionary.
7. Re-optimize using dual simplex method.
8. A new fractional solution has been found.
9. Generating an objective row cut.
10. The second cutting plane.
11. Add a new slack variable.
12. The new cut is inserted into the optimum dictionary.
13. The new optimum solution is integral.

### 1.2.Naïve Bayes

Step 1: Each gap from given text line will be considered to classify as being InterWordGap, IntraWordGap

Step 2: we will determine priori probability for each class: P(InterWordGap) = Number of InterWordGap/ Number of gap P(IntraWordGap) = Number of IntraWordGap/ Number of gap

Step 3:We will determine the probability for each gapto be in class InterWordGap,IntraWordGap

Step 4:The posteriori probability that the gap to be in class InterWordGap is: P(G|InterWordGap) = Number of InterWordGap in the vicinity of G / Number of InterWordGap

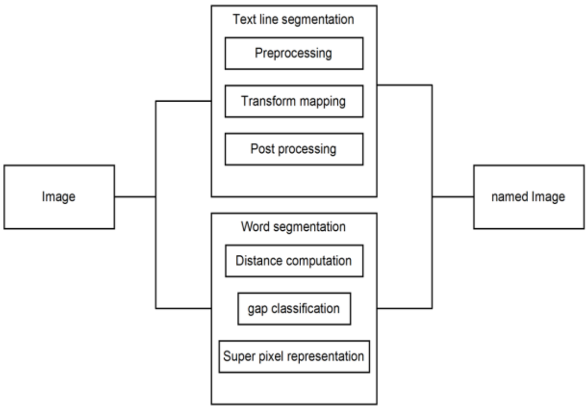
Step 5:The posteriori probability that the gapto be in class IntraWordGap is: P(G|IntraWordGap) = Number of IntraWordGap in the vicinity of G / Number of IntraWordGap

Step 6:Finally we will determine the probability P for each gapand assignto the class of InterWordGap orIntraWordGapthat has the greatest probability.

P(InterWordGap|G)=P(G|InterWordGap)*P(InterWor Gap)

P(IntraWordGap|G)=P(G|IntraWordGap)*P(IntraWord Gap)

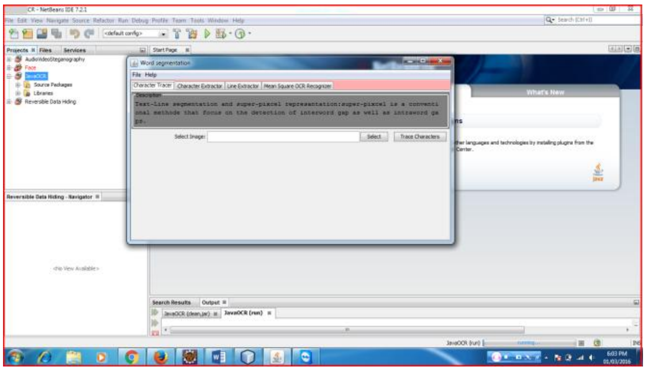
## IV. SYSTEM ARCHITECTURE



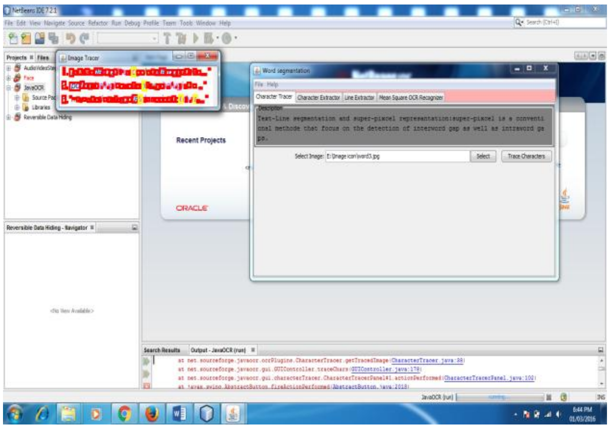
## V. CONCLUSION AND FUTURE WORK

In this paper, we create is a word segmentation strategy for manually written by method for hand file photographs. We demonstrate the division trouble as a twofold quadratic programming and assessed the parameters with the arranged picking up learning of technique. In view of the proposed definition, we should remember the pair savvy similarities between word-separators and furthermore unary properties inside the word segment. Moreover, because of the Bayesian classifier all parameters are assessed in a principled way and its miles related on that our gadget can be easily extended to dataset.

## VI RESULT ANALYSIS



Text-Line segmentation and super-pixcel representation:super-pixcel is a conventional methods that focus on the detection of interword gap as well as intraword gaps.

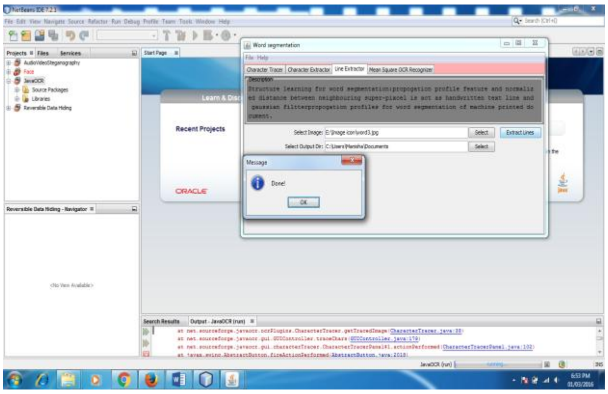


Word segmentation

Description
Text-Line segmentation and super-pixel representation:super-pixel is a conventional methode that focus on the detection of interword gap as well as intraword gap.

Select Image: E:\Image icon\word3.jpg

```
at net.sourceforge.javaocr.ocrPlugins.CharacterTracer.getTracedImage(CharacterTracer.java:88)
at net.sourceforge.javaocr.gui.GUIController.traceChars(GUIController.java:178)
at net.sourceforge.javaocr.gui.characterTracer.CharacterTracerPanel$1.actionPerformed(CharacterTracerPanel.java:102)
at javax.swing.AbstractButton.fireActionPerformed(AbstractButton.java:2018)
```



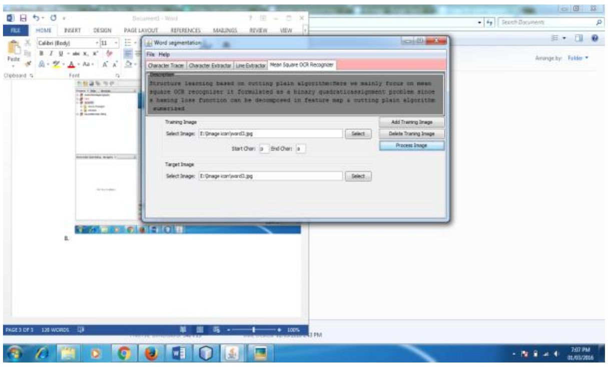
Word segmentation

Description
Structure learning for word segmentation:propogation profile feature and normalized distance between neighbouring super-pixcel is act as handwritten text line and gaussian filtterpropogation profiles for word segmentation of machine printed document.

Select Image: E:\Image icon\word3.jpg
Select Output Dir: C:\Users\Manisha\Documents

Message
Done!

```
at net.sourceforge.javaocr.ocrPlugins.CharacterTracer.getTracedImage(CharacterTracer.java:88)
at net.sourceforge.javaocr.gui.GUIController.traceChars(GUIController.java:178)
at net.sourceforge.javaocr.gui.characterTracer.CharacterTracerPanel$1.actionPerformed(CharacterTracerPanel.java:102)
at javax.swing.AbstractButton.fireActionPerformed(AbstractButton.java:2018)
```



Word segmentation

Description
Structure learning based on cutting plain algorithm:Here we mainly focus on mean square OCR recognizer it formulated as a binary quadraticassignment problem since a haming loss function can be decomposed in feature map a cutting plain algorithm sumarized

Training Image
Select Image: E:\Image icon\word3.jpg
Start Char: a End Char: a

Target Image
Select Image: E:\Image icon\word3.jpg

## REFERENCES

[1] Jewoong Ryu, Hyung Il Koo, NamIk Cho "Word Segmentation Method for Handwritten Documents based on Structured Learning" *IEEE Signal processing Letters*,vol.22,no. 8, Aug 2015

[2] L. O'Gorman, "The document spectrum for page layout analysis,"*IEEE Trans. Patt.Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1162–1173,Nov. 1993.

[3] B. Gatos, N. Stamatopoulos, and G. Louloudis, "ICDAR 2009 handwritingsegmentation contest," in *Proc. Int. Conf. Document Analysisand Recognition (ICDAR)*, 2009, pp. 1393–1397.

[4] N. Stamatopoulos, B. Gatos, G. Louloudis, U. Pal, and A. Alaei,"ICDAR 2013 handwriting segmentation contest," in *proc. Int. Conf.Document Analysis and Recognition (ICDAR)*, 2013, pp. 1402–1406.

[5] R. Bozinovic and S. Srihari, "Off-line cursive script word recognition,"*IEEE Trans. Patt.Anal. Mach. Intell.*, vol. 11, no. 1, pp. 68–83, Jan.1989.

[6] U. Mahadevan and R. Nagabushnam, "Gap metrics for word separationin handwritten lines," in *proc. Int. Conf. Document Analysis andRecognition (ICDAR)*, 1995, pp. 124–127.

[7] G. Seni and E. Cohen, "External word segmentation of off-line handwrittentext lines," *Patt.Recognit.*, vol. 27, no. 1, pp. 41–52, Jan. 1994.

[8] V. Papavassiliou, T. Stafylakis, V. Katsouros, and G. Carayannis,"Handwritten document image segmentation into text lines andwords," *Patt.Recognit.*, vol. 43, no. 1, pp. 369–377, Jan. 2010.

[9] T. Stafylakis, V. Papavassiliou, V. Katsouros, and G. Carayannis,"Robust text-line and word segmentation for handwritten documentsimages," in *proc. IEEE Int. Conf. Acoustics, Speech and SignalProcessing (ICASSP)*, 2008, pp. 3393–3396.

[10] T. Varga and H. Bunke, "Tree structure for word extraction from handwritten text lines," in *proc. Int. Conf. Document Analysis and Recognition(ICDAR)*, 2005, pp. 352–356.

[11] S. H. Kim, S. Jeong, G. S. Lee, and C. Y. Suen, "Word segmentation inhandwritten Korean text lines based on gap clustering techniques," in*Proc. Int. Conf. Document Analysis and Recognition (ICDAR)*, 2001,pp. 189–193.

[12] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, "Text line and word segmentation of handwritten documents," Patt.Recognit., vol. 42, no. 12, pp. 3169–3183, Dec. 2009.

[13] R. Manmatha and J. L. Rothfeder, "A scale space approach for automatically segmenting words from historical handwritten documents," IEEE Trans. Patt.Anal. Mach. Intell., vol. 27, no. 8, pp. 1212–1225, 2005.

Miss. Komal Chandrakant Kadam received Bachelor of Computer Engineering from ALARD COLLEGE OF ENGINEERING & MANAGEMENT MARUNJE, SavitribaiPhule Pune University

.

Miss. Dipali Dadaso Phadatare received Bachelor of Computer Engineering from ALARD COLLEGE OF ENGINEERING & MANAGEMENT MARUNJE, SavitribaiPhule Pune University

.

Miss. Aarti Mali received Bachelor of Computer Engineering from ALARD COLLEGE OF ENGINEERING & MANAGEMENT MARUNJE, SavitribaiPhule Pune University

.

Miss. Pallavi Nimbalkar received Bachelor of Computer Engineering from ALARD COLLEGE OF ENGINEERING & MANAGEMENT MARUNJE, SavitribaiPhule Pune University

.

Prof. Priti Gode Received master of Computer Engineering, SavitribaiPhule, Pune University, ALARD COLLEGE OF ENGINEERING & MANAGEMENT MARUNJE, SavitribaiPhule sPune University

.