Journal of Theoretical and Applied Information Technology
20th December 2015. Vol.82. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# ANALYSIS OF CANONICAL CHARACTER SEGMENTATION TECHNIQUE FOR ANCIENT TELUGU TEXT DOCUMENTS

**[1]N. VENKATA RAO, [2]Dr. A.S.C.S.SASTRY, [3]A.S.N.CHAKRAVARTHY, [4]A.V.SRINIVASA RAO**

[1]Research Scholar, Department of ECE , K.L.University, Vaddeswaram, Andhra Pradesh, India
[2]Professor, Department of ECE, K.L.University, Vaddeswaram, Andhra Pradesh, India
[3] Assoc. Professor, Department of CSE , University college of Engineering, Kakinada, Andhra Pradesh, India
[4]Assistant Professor, Department of ECE , Sasi Institute of Technology and Engg, Tadepalligudem, Andhra Pradesh, India

E-mail: [1]venkatnekkanti@rediffmail.com, [2] ascssatry @kluniversity.in,
[3]chakravarthy.cse@jntukucev.ac.in , [4]ad_bala@rediffmail.com

**ABSTRACT**

Character Recognition in ancient document images remains a challenging task. Initial scanning process deforms the document image, while aging process of document render it ancient which turns it to posses unwanted background noise. Segmentation includes an essential process in OCR. Complex scripts like derivatives of Brahmi, encounter various problems in the segmentation process. A hybrid model that entails segmentation in noisy images followed by binarization is proposed. In the first phase, segmentation technique for the ancient Telugu document image into meaningful units is proposed. Horizontal profile pattern is convolved with Gaussian kernel. The statistical properties of meaningful units are explored through an extensive analysis of the geometrical patterns of meaningful units. In the second phase, noisy documents are cleaned with the help of Modified IGT algorithm and then segmented by using conventional profile mechanism. The performance of the present hybrid technique is proved by the results of higher efficiencies for the cleaned documents. The efficiency analysis of segmentation carried out for the present hybrid technique reveals a threshold number of Vowels (V), Consonants(C), CV core characters to exhibit higher efficiencies. It also reflects upon the non-canonical features of any other marks of the Telugu document.

**Keywords:** *Segmentation, Profile, Gaussian derivative kernel, Modified IGT, Error Rate*

## 1. INTRODUCTION

India is a multilingual Asiatic country which possesses a rich collection of written ancient scripts. Optical Character Recognition (OCR) development is yet to take a commercial shape for many of these scripts. Ancient character segmentation is the first step of hybrid model that decomposes a document image into a sequence of sub images of individual symbols. Prior to that, binarization [11] is of greater significance. It is due to it being a first step of mechanization of OCR considered as an important preprocessing, step whose outcome directly affects further stages of OCR system.

In an automated historical document processing system, line segmentation engine is frequently used before character segmentation. The performance of a line segmentation engine is attached with significant influence on the accuracy of character segmentation and recognition. Different methods are proposed [12] in the literature. The horizontal profile technique creates [2,4,8] a histogram crossing on entire text block along a predetermined direction of the text line. The peak information between the lines is sufficient to separate lines. Hough transform based methods [3] are found to be identical to the projection profile method. However, Hough transform is noticed to be best suited for locating skewed text lines from the text sample. It is applied at a set of specified selected angles. Along each angle, straight lines are drawn with a metric for the fit. The best fit for the lines gives the skew angle and the location of the line. The other methods [5] in vogue explore the nearest neighbor clustering of connected components. Some of the existing methods are applicable for hand written line segmentation as well as machine printed text. In the method of Senior and Robinson, the process of locating text lines is based [6] on the gap between the lines,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

which is supposed to catch enough information to separate the lines. Zhixin Shi. et al, proposed [9] an Adaptive Local Connectivity Map(ALCM), in which the value of a pixel is the sum of the all pixels within a specified horizontal distance of that pixel. The ALCM method involves thresholding process which more or less resembles [1] Otsu's method, it connects the components to represent the probable regions for complete, or partial line of text. Manmatha, et al, used [10] a scale space technique for word separation that produces negotiable result on a large collection of George Washington's manuscripts, etc. Segmentation of text lines is performed using smoothed projection profiles, which is sufficient for the documents used in the tests. Nikos Nikolaou, et al, proposed [16] a novel Adaptive Run Length Smoothing Algorithm (ARLSA) in order to manipulate the problem of complex and dense document layout, detection of noisy areas and punctuation marks that are usual in ancient (aged) machine-printed documents. The detection of possible obstacles formed from background areas now to be identified in order to separate neighboring text columns or text lines, and to use the skeleton segmentation path for a possible isolation of connected characters. Vijaya Kumar Koppula et all., proposed [13]a method for text line extraction for Telugu text sample by clustering the connected components of a line using vertical spatial and nearest neighbor information, word extraction by the computation of space between two adjacent characters are clustered into word space. This method is found to be better suitable for segmentation of noise free Telugu text sample into text lines, words and characters under noise free environment. Negi et all., proposed[7] a novel approach for the location and extraction for Telugu script by using Hough transform, while involves the estimation of Sobel gradient magnitude in associated with the Recursive XY cuts to identify the paragraphs, lines and words. They adopted zoning and structural feature vectors(cavities) for the recognition of the isolated Telugu text patterns. But, this method is also found well worked on Telugu text in noise free environment. It seems that there is a greater necessity to evolve an OCR technique that takes care of the possible and inevitable noise that has been accumulated over times in case of ancient scripts.

In the wake of the work reported in the field of OCR of Devanagari scripts and abundance of invaluable knowledge base regarding the ancient scripts of Telugu lipi a humble attempt is made to evolve a model basing on segmentation of text into isolated patterns under the noisy environment.

### 1.1 Over view of Telugu Script

Out of 22 officially recognized languages in India, 9 languages have separate scripts(viz., Indic scripts) and the other languages are written [14] either in Perso-Arabic script or Devanagari script. Telugu is the official language of the state of Andhra Pradesh situated in southeastern India where it is spoken by close to 120 million people. Telugu is a highly developed language and happen to be the biggest linguistic unit in India. The Telugu script consists of vowels, consonants, consonant-vowel core formation and a large number of conjunct formations. For all these formations, there exist nasal sounds represented with the help of 'anuswara' sign' an addition. Vowels are constituted by 16 independent letters represented with individual glyph. Consonants are constituted by 36 individual letters with distinct glyph set. The vowel signs called as 'matras' play an important role in the formation of the glyph. Thus the character glyph formations for these combinations are logically arranged to 455. The shape of consonant-vowel formations and conjuncts dependent on the context and is affected by the order of consonants and vowels. But, Indic scripts provide different types of glyph orders for different languages, though the canonical structure is common.

A general technique for segmenting noisy Telugu text document image into lines by using horizontal profile which is convolved with Gaussian derivative kernel (of first order for identifying the zero crossing peaks) are used for line segmentation. The character segmentation is carried with the help of a vertical profile by selecting proper threshold. Cleaning of Noisy documents with the help of Modified Iterative Global Threshold algorithm. Based on the profile information cleaned document is segmented into lines and characters.

The paper is organized as 4 sections. In the first section, briefly discussed about the introduction, literature survey and problem definition. In the second section discussed the algorithm for line and character segmentation of noisy documents. In the third section, briefly discussed the experimental results and segmentation efficiency. Section four describes the conclusions and future scope of work.

Journal of Theoretical and Applied Information Technology
20th December 2015. Vol.82. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

## 2. METHODOLOGY

A novel technique (flow chart of as Figure 1) for the segmentation of noisy ancient Telugu documents with the details of their being cleaned and the corresponding stages of phased way of processing is described. Nevertheless a benchmark parameter is designed to estimate and discuss the efficiency of the present model is also illustrated in Figure 1.

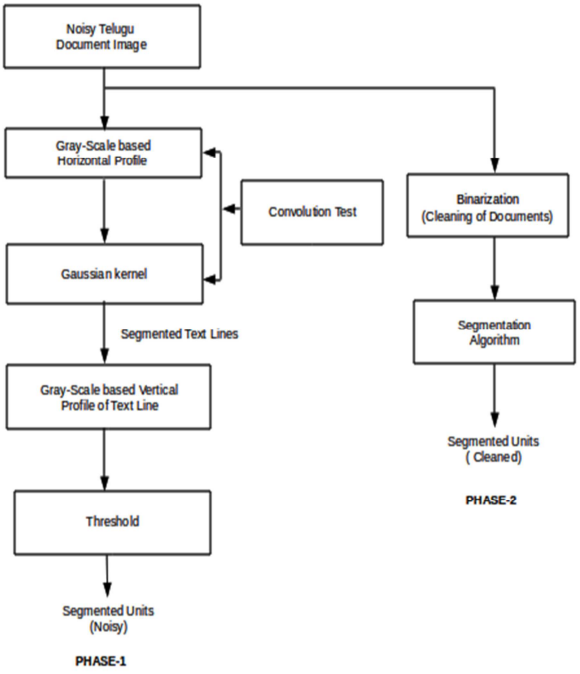



*Figure1: Flow Chart of the Proposed Model*

In the first phase of methodology, Segmentation is performed on Noisy Telugu text documents. Segmentation plays a major role in document image analysis. Segmentation of Telugu script into meaningful units is somewhat difficult because of cursive nature of the script. In this connection, segmentation of noisy document into syllables, still a challenging job till today.

In the first phase of methodology, Segmentation is performed on Noisy Telugu text documents. Segmentation plays a major role in document image analysis. Segmentation of Telugu script into meaningful units is some what difficult because of cursive nature of the script. In this connection, segmentation of noisy document into syllables, still a challenging job till today.

In this context, a pure gray scale image (which is represented by black and white intensities), provides information of foreground and background of the image. Black is designated as '0' and white is designated as '255'. During the process of segmentation, the segmentation of noisy document into meaningful units exclusively depends on the characteristics of the noise. Generally there is a slight change between the background pixel intensity and foreground pixel intensity of a noisy document. That difference is treated as noise. It, nevertheless contains gaps (between lines, characters etc. in the image) along with the foreground information. The foreground information itself changes its pixel intensity value due to the presence of noise. If the noise is non uniformly distributed in the image in (for whatever may be the reason), the segmentation of text document into segmentable units is difficult because noise dominates text information in some areas of the document.

An analysis of gray-scale based profile information of image would be useful for effective segmentation of text document into lines and characters. Width of the peak of intensity will provide basic information for separating lines and characters. It is noticed that the width of the peak gradually decreases between lines and characters. In order to perform line segmentation, the horizontal profile information is convolved with the Gaussian kernel of order-1 and sigma-3. The horizontal profile of the text sample is illustrated in Figure 2(a).

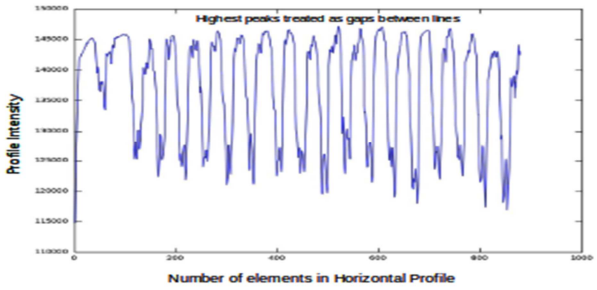



*Figure 2(a): Horizontal profile of the text sample*

Character segmentation is carried out based on a threshold, which is defined from the intensities of vertical profile in a text line. There would be significant change in the characteristics of vertical profile when compared with the horizontal profile information. The peaks of finite width are clearly identified in the horizontal profile, whereas in a vertical profile, the peaks are not clearly

Journal of Theoretical and Applied Information Technology
20th December 2015. Vol.82. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

identified. Further, they appeared at nonuniform intensities. The vertical profile of the text line is illustrated in Figure 2(b).

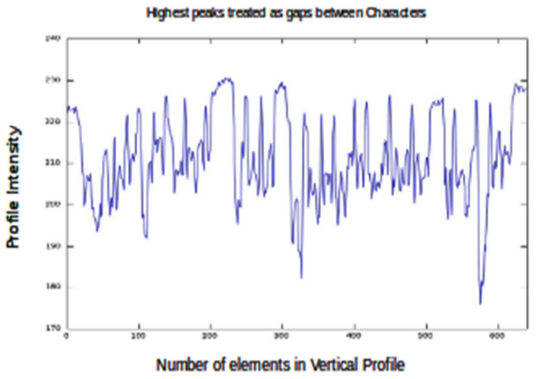



*Figure 2(b): Vertical profile of the text line*

In the second phase of methodology, segmentation is performed on cleaned Telugu text documents. In this connection, a Modified Iterative Global Threshold technique is used during cleaning of documents and segmenting the document into lines and words by using horizontal profile. The character segmentation is carried out by using the vertical profile. In case of any touching characters that exist, those are identified during character segmentation by using drop fall algorithm [15] for improving the segmentation accuracy.

### 2.1 Segmentation Algorithm for Noisy Documents

The series of sequential steps necessary for the segmentation algorithm suitable for noisy documents of a degraded text document (into lines and meaningful units), are viz., Extraction of degraded (noisy) document; Identification of the Horizontal Profile; Performing the convolution (between horizontal Profile and Gaussian kernel); Identification of the peaks (for line segmentation); Identification of the Vertical Profile (of the line); Defining a threshold intensity; Identifying the peaks for character segmentation.

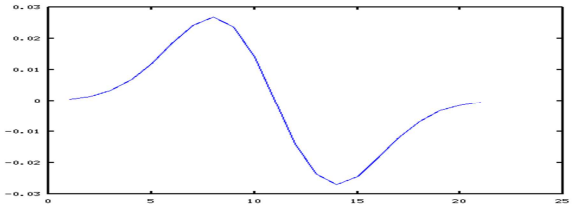
A noisy document is generally represented by I(n,m), where 'n' is the number of lines and 'm' is the number of columns. The horizontal profile I(n,m) is identified by considering the sum of all pixel intensities perpendicular to the Y-axis, and is represented by 'HP' of a specific size 'n' i.e.,

$$HP[i] = \sum_{j=1}^{m} I(i,j) \qquad (1)$$

In case of a direct horizontal profile for line separation, the identification of peaks and valleys are difficult, because the value of each pixel is large so that efining a threshold is a big task. Now the profile is convolved by Gaussian derivative kernel of order-1. Gaussian is self similar function. It is expressed by the Equation-1.

First order differentiation of Gaussian kernel is preferred for effective line segmentation. If we go for higher order, the Gaussian kernel information is deciphered would be one and same. However the higher order kernel additionally entails complexity of computations. Hence a first order Gaussian kernel is preferred for line segmentation. The width of Gaussian kernel for sigma-3 is found to be suitable for line segmentation illustrated in Fig-3(a) Segmentation efficiency during line segmentation process is found to possess varying values of order and sigma.

*Figure 3(a): Gaussian derivative kernel of order-1 & sigma-3*



Convolution with Gaussian kernel is a linear operation. Convolution is used to find the common area between the profile and the Gaussian kernel. The degree of shift in the Gaussian kernel during the convolution process linearly varies with horizontal profile information. So, this can be used to represent randomness in the profile and provides a zero crossing smooth curve, when it is convolved with the profile, represented by figure 3(b).

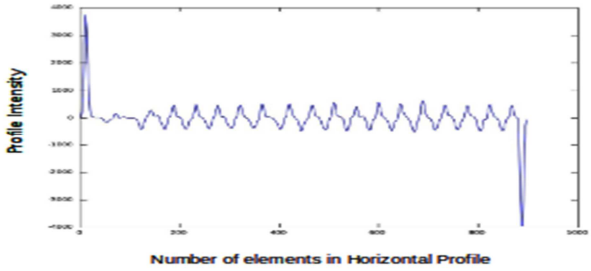



*Figure 3(b): Gaussian kernel convolved with horizontal profile*

The peaks which are above zero are treated as the gaps between the line. Based on this information, the line segmentation is performed.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The Gaussian kernel and the corresponding convolved profile is presented in Figure 3(a) and (b)

$$G = \left[\frac{1}{2\pi\sigma^2}\right] e^{\left[\frac{-x^2}{2\sigma^2}\right]} \quad (2)$$

Where σdetermines the width of the Gaussian kernel. As we have considered the Gaussian probability density function it may also be called as standard deviation σ The Square of σ represents its variance. The resultant equation after convolving the profile with Gaussian kernel is represented by the Equation 3.

$$C = \int G * HP * dt \quad (3)$$

From Figure 3(b) the Gaussian kernel provides smooth profile with a harmonic space between successive peaks. It contains both positive and negative peaks which represents the gaps between the lines and the foreground information. Due to this reason, the extraction of the lines from the text document becomes an easier process.

Character segmentation is carried out on the above segmented lines from their vertical profiles. Vertical profile of a line is generated by computing the sum of pixel intensities perpendicular to the X-axis, which is represented by 'VP' of line width 'd' and is defined by Equation 4

$$VP[j] = \sum_{i=1}^{n} I(i,j) \quad (4)$$

If we use the same Gaussian kernel for character segmentation results in a smooth profile with nonuniform peaks. Identification of nonuniform peaks with a specific intensity value involves difficulty during character segmentation. Hence a threshold for character segmentation is defined.

In the process of character segmentation (performed on the text line) finding a suitable threshold (from the vertical profile) happens to be a prerequisite condition. Threshold is calculated by means of maximum and minimum values of the vertical profile which is expressed by

$$I_{Th} = (VP)_{max} - \left[\frac{(VP)_{max} - (VP)_{min}}{3.6}\right] \quad (5)$$

### 2.2 Cleaning of Noisy Documents

Binarization is one of the several steps used in most document image analysis systems. It consists of labeling each pixel in an image as foreground and background. It provides an proper distinction between background and foreground. In this paper we used Modified Iterative Global Threshold algorithm for cleaning the documents which is proposed [19] by us.

### 2.3 Performance Evaluation

The segmentation efficiency of the noisy Telugu documents is estimated with the help of Gaussian kernel convolved with horizontal profile for line segmentation and I[th] from the vertical profile is for character segmentation. After cleaning the documents with the help of Modified Iterative Global Threshold algorithm, the segmentation efficiency is estimated based on the inter space between lines and characters from the profile information. The Segmentation efficiency of Lines and Characters under noisy and noise free environment is estimated by using the Equations 6 & 7. The error rate of Characters under noisy environment is measured by using the Equation8.

$$Percentage\ Efficiency\ for\ Character = \frac{NCSC}{TNC} * 100 \quad (6)$$

$$Percentage\ Efficiency\ for\ Lines = \frac{NCSL}{TNL} * 100 \quad (7)$$

$$Percentage\ Error\ Rate\ for\ Characters = \frac{TNC - NCSC}{TNC} * 100 \quad (8)$$

Where

NCSC – Number of Correctly Segmented Characters
NCSL – Number of Correctly Segmented Lines
TNC – Total Number of Characters
TNL – Total Number of Lines

## 3. RESULTS AND DISCUSSIONS

Segmentation algorithm is applied on a three sets of 94 noisy document images. They are collected from the Net (Telugu old book named "Thiagarajaswami Krithis" is published in 1933, at Kesari Printing Press, Chennapuri)[18], the scanned copies of story books(Telugu old book named "vydyula kathalu" is published in 1952 at Madras printing press) and first "Chandamama" story book[17] is published in 1947 which are of 60 to 70 years old. A typical noisy document is presented in Figure 4(a). After applying the defined algorithm for line segmentation, the resultant image is present

Journal of Theoretical and Applied Information Technology
20th December 2015. Vol.82. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

in Figure 4(b). In this resultant image, a dark horizontal line is marked for the identification of line separation. However the character segmentation is performed based a threshold value $I_{th}$ defined by the Equation 5. The lines which are extracted from the sample is presented in Figure 4(b). It is noticed that dark vertical lines are also marked for defining the boundaries of characters as illustrated in Figure 4(c). Segmentation accuracy (of characters) is generally found to depend on the threshold value. After a large experimentation on the vertical profile of a text line, the peaks which are obtained are found to be in the vicinity of a maximum value of the profile. This observation suggests for the definition of a threshold value. In Equation 5 the threshold value is noticed to be sensitive to the denominator. Hence, from the Equation 5 the denominator in the second part of equation is selected based on the trial and error method. Presently the denominator value taken to vary between 3.6 to 4 is found to give a better result. By defining the Gaussian kernel with different values of order and sigma for character segmentation (instead of the threshold value defined by Equation 5) the segmentation accuracy is found to change. The behavior of Gaussian kernel of order-1 and sigma-3 along with its effectiveness of segmentation is illustrated in Figure 4(d). The segmentation rate is found to be low, while it can be compared to the segmentation rate, estimated from threshold value. The specific case of Gaussian kernel of order-1 and sigma-1.8 along with its effectiveness of segmentation is illustrated in Figure 4(e). It is noticed that the segmentation rate gets drastically reduced, in comparison between the two cases.

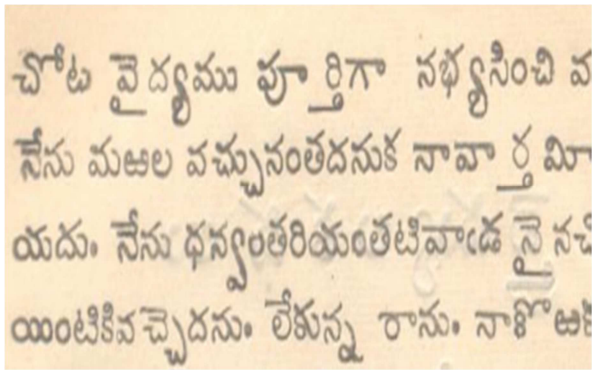


*Figure 4(a): Original Image*

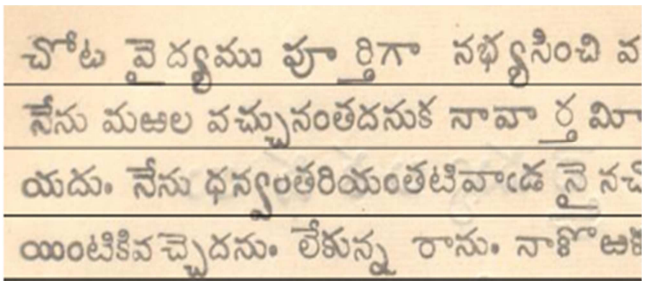


*Figure 4(b): Line segmented Image*

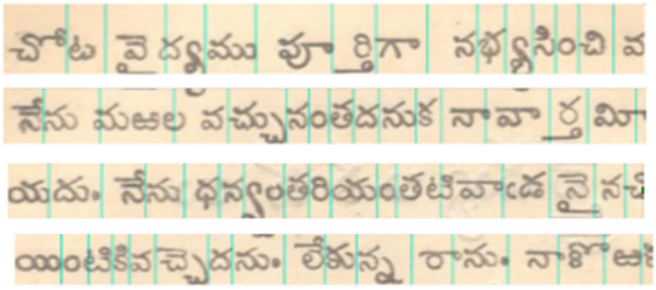


*Figure 4(c): Character segmented image using threshold*

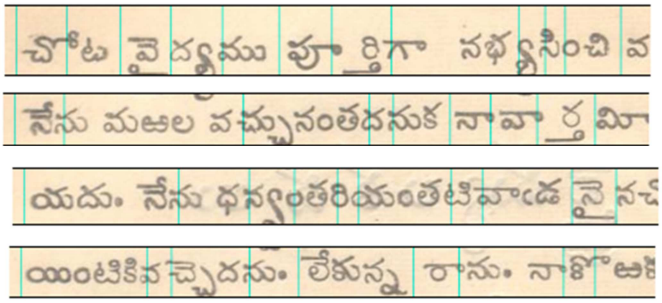


*Figure 4(d): character segmentation using Gaussian kernel with(sigma=3,order=1)*

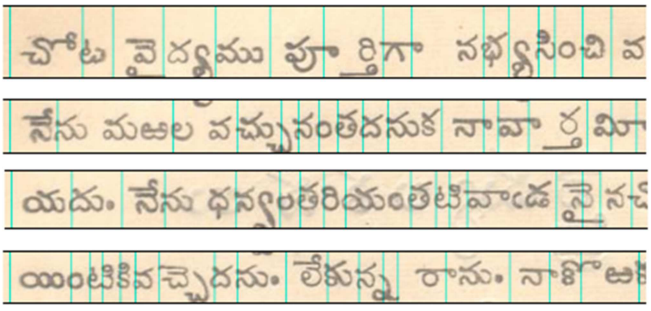


*Figure 4(e): character segmentation using Gaussian kernel with(sigma=1.8,order=1)*

Journal of Theoretical and Applied Information Technology
20th December 2015. Vol.82. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In the second phase of evaluation, the cited 94 documents are cleaned by using Modified Iterative Global Threshold algorithm and then segmented these text samples into lines and characters by using regular methods (profile). A typical noisy document is presented in Figure 4(a). The cleaned document by Modified IGT is presented in Figure 5(a). Segmented text document into lines, words and characters based on the inter space between lines, words and characters is illustrated in Figure 5(b).

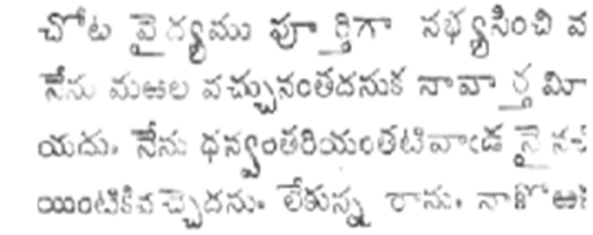


*Figure 5(a): Cleaned image of Figure 4(a)*

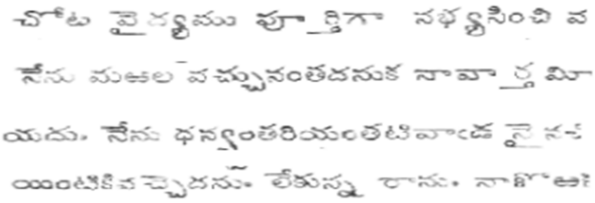


*Figure 5(b): Line segmented image of (a)*

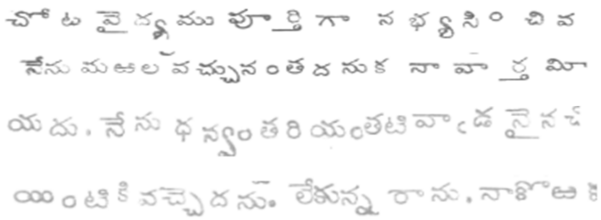


*Figure 5(c): Character segmented image of (b)*

### 3.1 Performance Evaluation

The efficiency and Error Rate estimated following procedure as detailed in section 2.3. It is noticed that the efficiency tabulated, corresponding to the segmentation process under noisy and non noisy conditions. The performance of the algorithm is tested over the 3 categories of text samples. The present case of a maximum of 100% efficiency corresponds to the line segmentation process exposed to 2nd and 3rd category of sample documents, while it is found to be 98.2% for 1st category.

*Table 1: Comparison of segmentation efficiency*

| S. No | Samples | No of Lines | Total no of Syllables | Correctly Segmented Syllables | Efficiency of noisy documents (%) | Error Rate (%) | Efficiency of cleaned documents (%) |
|---|---|---|---|---|---|---|---|
| **Samples from Thiagarajaswami Krithis book** | | | | | | | |
| 1 | 23 | 168 | 6121 | 2884 | 47.11 | 52.88 | 56.3 |
| **Samples from first Chandamama book** | | | | | | | |
| 2 | 45 | 680 | 20558 | 14194 | 69.04 | 30.95 | 71.87 |
| **Samples from story books** | | | | | | | |
| 3 | 26 | 382 | 8338 | 5887 | 70.6 | 29.39 | 95.59 |

The 382 lines from 26 text samples of category-1 are tested over the modified algorithm on this direction of lines, which are having 8338 characters. It is noticed that only 5887 are correctly segmented. Hence, an efficiency of 70.06 is found to be achieved during character segmentation with an error rate of 29.39. The 168 lines from 23 text samples of category-2 are tested, which are having 6121 characters. It is noticed that only 2884 are correctly segmented. Hence an efficiency of 47.11 is found to be achieved during character segmentation with an error rate of 52.88. In this category the error rate is very high due to the presence of noise dominates the object information. The 680 lines from 45 text samples of category-3 are tested, which are having 20558 characters. It is noticed that only 14194 are correctly segmented. An efficiency of 69.04 is found to be achieved during character segmentation with an error rate of 30.95. The case of wrongly segmented characters arises due to the, involving characters of which are touching with neighboring characters in any zone. Owing to this fact, the consideration of wrongly segmented characters is found that it never leads to achieve 100% efficiency for character segmentation. However, utility of this threshold value is originated during the process of character segmentation. Thus the Gaussian kernel used for character segmentation is found to lead the increasing efficiency, as it depends on the values of order and sigma. In case of cleaned documents the character segmentation efficiency is improved to 95.59 for category-1 samples, 56.30 for category-2 samples and 71.87 for category-3 samples. Further there is an additional 2% efficiency achieved by using drop fall algorithm for splitting the touching characters.

Journal of Theoretical and Applied Information Technology
20th December 2015. Vol.82. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The efficiency analysis of the present technique is carried out basing on the premise that Telugu document is inherent characteristics of its canonical structure. As such segmentation efficiency of Telugu text documents whether (ancient or cleaned) primarily depends on the canonical structure of the script. Ancient documents get noisy and the background is treated as noise environment. Hence efficiency of present technique is analyzed as a function of differently natured canonical features in any Telugu documents. The various groups of canonical structured features of the cited samples are presented in Table-2.

*Table 2: Various groups of Canonical Structure of the Samples*

| S. No | Samples Name | Vowels | Consonants | CV Group | others | Efficiency | |
|---|---|---|---|---|---|---|---|
| | | | | | | Noisy Documents | Cleaned Documents |
| 1 | **Thiagarajaswami Krithis book (23)** | 21 | 1618 | 2119 | 2363 | 47.11 | 56.3 |
| 2 | **First Chandamama book (45)** | 1133 | 3516 | 11733 | 4176 | 69.04 | 71.87 |
| 3 | **Story books (26)** | 224 | 2016 | 3913 | 2185 | 70.6 | 95.59 |

The efficiency is independently studied as a function of Vowels(Fig-6a), Consonants(Fig-6b), CV group(Fig-6c) and Others (grammar, punctuations etc.,)(Fig-6d) for both the noisy and cleaned documents. In all the cases the efficiency of cleaned document is observed to be higher, which reflects upon the additional specific advantages of the present technique over the existing techniques. The specific features of the efficiency analysis for the present model are:

1. The efficiencies estimated by Eqs-6,-7,-8 is found(Fig-6a,6b,6c) to be higher for cleaned documents, low for noisy documents to reflect the suitability of present technique for ancient Telugu documents.
2. Since efficiency of any (noisy or cleaned) document is found to attain maximum value only, when the document contains more than a threshold value for the number of vowels(Fig-6a), consonants(Fig-6b), CV compound(Fig-6c). This directs forward the optimization of the present technique contain a minimum specific number of canonical features.
3. The estimated efficiency of the present OCR technique for the consideration of other features (grammatical or punctuation marks etc.,) is found(Fig-6d) to respond with an opposite trend, which castes suspicious whether there other features truly represents the canonical features of the Telugu script.
4. The above features of efficiencies can also be visualized (Fig-7a,7b) through the 3D surfaces generated by plotting efficiencies as a function of either vowels/consonants for a common CV prevalence.

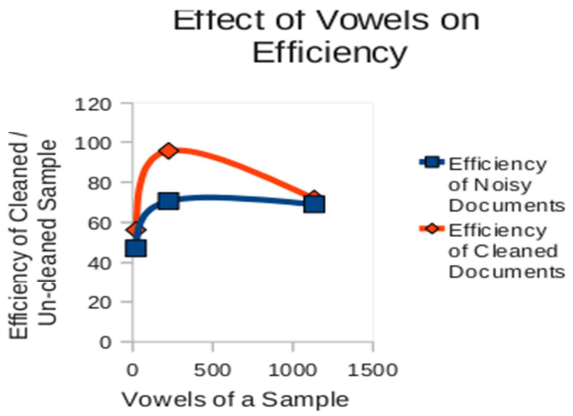



*Figure 6(a): Effect of owels on Efficiency*

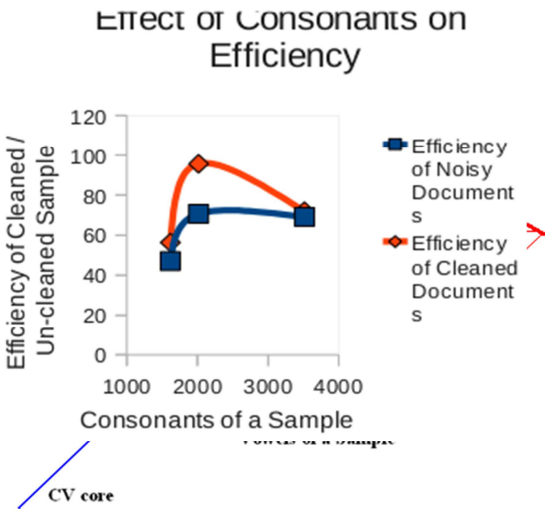



*Figure 6(b): Effect of Consonants on Efficiency*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

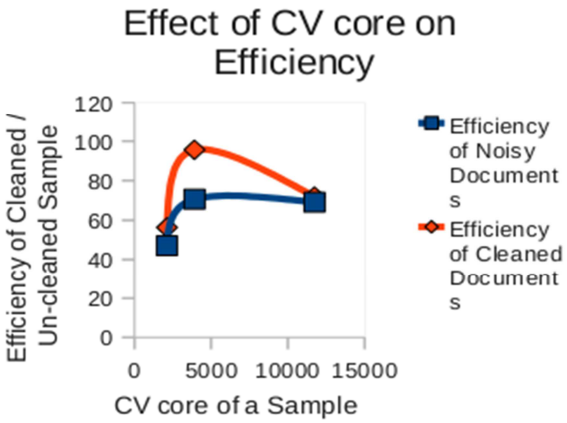



*Figure 6(c): Effect of CV core on Efficiency*

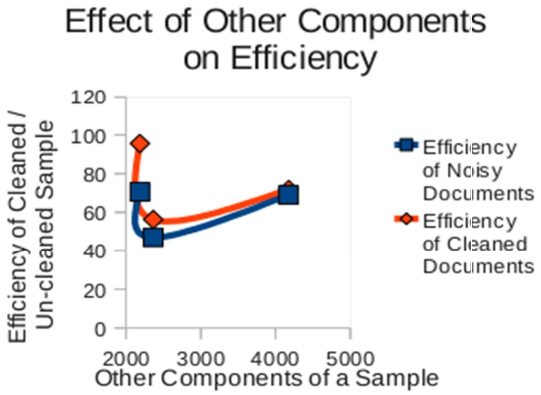



*Figure 6(d): Effect of other components on Efficiency*

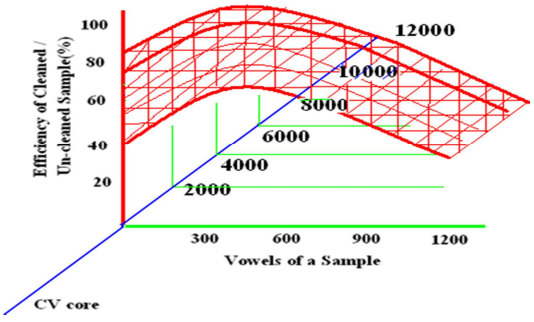



*Figure 7(a): Effect of Vowels & CV core on Efficiency*

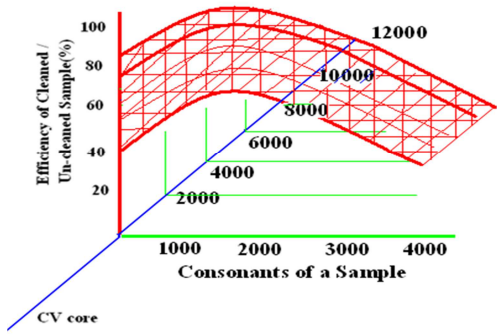



*Figure 7(b): Effect of Consonants & CV core on Efficiency*

## 4 CONCLUSIONS

- OCR of Ancient Telugu script can be implemented through the character segmentation of document images involving binarization.
- Noisy document image can be perceived by containing the background and foreground information.
- Cleaning of the noise bearing ancient documents can be carried out by implementing a modified IGT algorithm.
- Analysis of the efficiency suggests for the optimization of the present model in the domain of Vowels, Consonants, CV sounds and other (grammar and punctuation marks etc.,) considering them inclusively and/or exclusively.
- Line segmentation carried out by convolving horizontal profile with the help of Gaussian kernel (i.e., of higher orders) yield enhanced efficiency in the vicinity of 100%.

## REFERENCES

[1] N.Otsu, "A threshold selection method from a gray level histograms", IEEE Trans. Systems, Man, Cybernet., 9(1),1979, pp. 62- 66

[2] S.S.G.Nagy and S.Stoddard, "Document analysis with expert system," Procedings of Pattern Recognition in Practise II, June 1985.

[3] S.Srihari and V.Govindaraju, "Analysis of textual images using the hough transform," Machine Vision and Applications, vol.2, no.3,. Springer(1989 ), pp. 141-153.

[4] G.Ciardiello, G.Scanfuro, M.Degrandi, M.Spada, and M.P.Roccotelli, "An experimental system for office document handling and text recognition," patent no:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

US185813A in feb, 09, 1993

[5] L.O'Gorman, "The document spectrum for page layout analysis," IEEE Trans. Pattern Anal.Mach.Intell., vol. 15, no. 11, pp. 1162-1173, 1993.

[6] A.W.Senior and A.J.Robinson, "An off-line cursive hand-writing recognition system," IEEE Trans. Pattern Anal.Mach.Intell., 20(3): 309-321, March 1998.

[7] Atul Negi, K Nikhil Shanker, and Chandra Kanth Chereddi,"Localization, Extraction and recognition of Text in Telugu document Images," ICDAR 2003

[8] E.Kavallieratou, N.Dromazou, N.Fakotakis, and G.Kokkinakis,"An integrade system for hand written document image processing," International Journal of Pattern Recognition and Artificial Intellegence, 17(40), pp. 617-636,2003

[9] Z.Shi,S.Setlur and V.Govindaraju, "Text extraction from gray scale historical document images using adaptive local connectivity map. In 8th International Conference on Document Analysis and Recognition, ICDAR, volume 2, pp. 794-798, Seoul, Korea,August 2005.

[10] R.Manmatha, J.L. Rothfeder, "A Scale Space approach for automatically segmenting words from historical handwritten docuemnts," IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), pp. 1212-1225, August 2005.

[11] B.Gatos, I.Pratikakis, S.J.Perantonis, "Adaptive degraded document image binarization," Pattern Recognition vol 39, 2006 pp. 317-327

[12] L.L.Sulem, Abderrazak Zahour, Bruno Taconet, "Text Line Segmentation of Historical Documents: a survey " International journal on Document Analysis and Recognition, vol 9, pp.123-138, springer, 2007

[13] Vijaya Kumar Koppula, Negi Atul, Utpal Garain " Robust Text Line, Word and Character Extraction from Telugu Document Image," Proceeding ICETET 09' Proceedings of the 2009 Second International Conference on Emergign Trends in Engineering and Technology.

[14] D.Gosh, T.dube, A.P.Shivaprasad, "Script Recognition-A Review," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009

[15] AV Srinivasa Rao, M Subbarao, ASCS Sastry, N. Venkat Rao, L. Pratap Reddy "Segmentation of Touching Hand written Numerals and Alphabets" – Proceedings of IEEE International Conference ICCEE2009, Dubai, 28-30 December, 2009.

[16] Nikos Nikolaou a.b, Michael Makridis a, Basilis Gatos b, Nikolaos Stamatopoulos b,Nikos Papamarkos a "Segmentation of historical machine-printed documents using Adaptive Run Length Smoothing and skeleton segmentation paths," Image and Vision Computing 28 (2010) pp. 590–604

[17] http://www.4shared.com/get/AOeEmhST/First-Telugu-Chandamama-Book-J.html

[18] http://www.musicresearch.in/categorydetails.php? imgid=177

[19] N Venkata Rao, A V Srinivasa Rao, S Balaji, L Pratap Reddy,"Cleaning of Ancient Document Images using Modified Iterate Global Threshold, " IJCSI, Vol. 8, Issue 6, November 2011