I.J. Image, Graphics and Signal Processing, **2012, 6, 8-14**
Published Online July 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2012.06.02

Segmentation of Ancient Telugu Text Documents

Srinivasa Rao A.V
ECE Department, AKRGCET, Nallajerla, Andhra Pradesh, India
E-mail: ada_bala@rediffmail.com

Abstract— OCR of ancient document images remains a challenging task till date. Scanning process itself introduces deformation of document images. Cleaning process of these document images will result in information loss. Segmentation contributes an invariance process in OCR. Complex scripts, like derivatives of Brahmi, encounter many problems in the segmentation process. Segmentation of meaningful units, (instead of isolated patterns), revealed interesting trends. A segmentation technique for the ancient Telugu document image into meaningful units is proposed. The topological features of the meaningful units within the script line are adopted as a basis, while segmenting the text line. Horizontal profile pattern is convolved with Gaussian kernel. The statistical properties of meaningful units are explored by extensively analyzing the geometrical patterns of the meaningful unit. The efficiency of the proposed algorithm involving segmentation process is found to be 73.5% for the case of uncleaned document images.

Index Terms— Segmentation, Profile, Line, Syllable, Gaussian derivative kernel

I. INTRODUCTION

India is a multilingual Asiatic country which possesses a rich collection of written ancient scripts. OCR development is yet to take a commercial shape for many of these scripts. Ancient character segmentation is the first step of OCR system that decomposes a document image into a sequence of sub images of individual symbols. Prior to that, binarization [12, 18] 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 Optical Character Recognition (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In an automated historical document processing system, line segmentation engine is frequently used before character segmentation.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 [13] in the literature. The horizontal profile technique creates [2, 4, 7] 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 method [7] in vogue explores 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, which is supposed to catch enough information to separate the lines. Zhixin Shi. et al. proposed [10] 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 [11] 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 [17] 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. Negi et al. proposed[8] 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. Lakshmi et al. [9] proposed the concept of basic symbols in the script. Simple moment features are used for around 386 basic symbols in the training process. Vijaya Kumar Koppula et al. proposed [16]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.

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.

II. 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 [15] 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. Vowels are constituted by 16 independent letters represented with individual glyph. Consonants are constituted by 35 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.

III. FEATURES OF TELUGU SCRIPT

Telugu script is an abugida from the Brahmic family of scripts. The writing systems that employ Devanagari and other Indic scripts constitute a cross between Syllabic writing systems and Phonetic writing systems. The effective unit of these scripts is the Orthographic Syllable constituting of a consonant and vowel (CV) Core and optionally, one are more preceding consonants with a canonical structure of ((C)C)CV. The orthographic syllable need not correspond exactly with a phonological syllable especially when a consonant cluster is involved. But writing systems is built on phonological principles and tends to correspond quite closely to pronunciations. In Brahmi based scripts, characters can combine or change shape depending on their context. A character's appearance is affected by its ordering with respect to other characters. Though vowels and consonants are defined separately, each consonant is defined with an inherent vowel /A/. When the same consonant combines with other vowels (CV Core), the shape of the character is dependent on vowel modifiers. Apart from CV core, a large number of conjunct formations are found in all Indic Scripts. The conjunct formations are nothing but a combination of one or two consonants preceding the CV core, which provide more effective association with phonetic syllables. A detailed analysis on Canonical structure of characters in Telugu Script is discussed [14] by Srinivas et al. As per the analysis, the basic structure is decomposed into vowels, consonants, CV core Conjunct formations and dead consonants. For all these formations there exists a nasal sound represented with the help of ‘anuswara’ sign as an addition. Few character combinations are found with a special symbol ‘Visarga’, which is a rare occurrence. In the real usage of script, the above character combinations are found with certain percentages of occurrence.

A general technique for segmenting 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.

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.

Presently we described a novel technique figure 1 for segmentation of ancient (i.e., noisy) Telugu documents. The flow chart for the proposed model evolved currently with details of phased processes of the script and the benchmark for its efficiency are illustrated in the figure 1.

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.

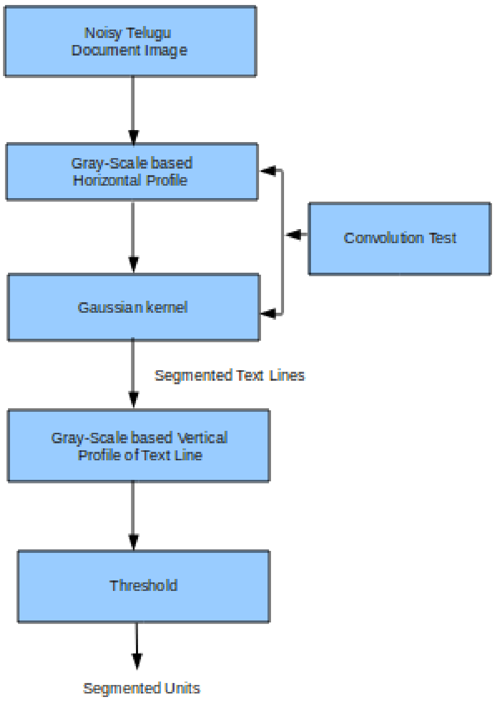
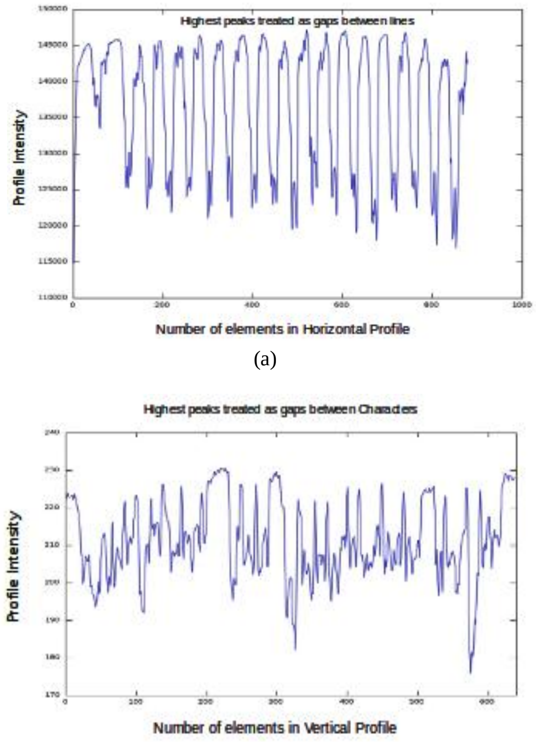

Noisy Telugu Document Image
Gray-Scale based Horizontal Profile
Convolution Test
Gaussian kernel
Segmented Text Lines
Gray-Scale based Vertical Profile of Text Line
Threshold
Segmented Units

Figure 1. Flow Chart of the Proposed Model

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).

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



(a)

(b)

Figure 2. (a) Horizontal profile of the text sample (b) Vertical profile of the text line

IV. 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.,

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

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 defining 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 "(2)".

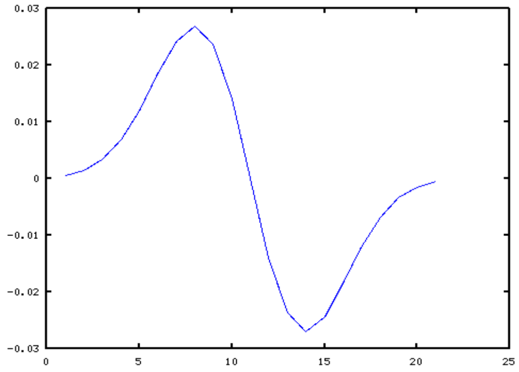
First order differentiation of Gaussian kernel is preferred for effective line segmentation. If we go for higher order, the Gaussian kernel information is deciphered would be one and same. However the higher order kernel additionally entails complexity of computations. Hence a first order Gaussian kernel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

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 'C'. The peaks which are above zero are treated as the gaps between the lines. Based on this information, the line segmentation is performed. The Gaussian kernel and the corresponding convolved profile is presented in figure 3(a) & (b)

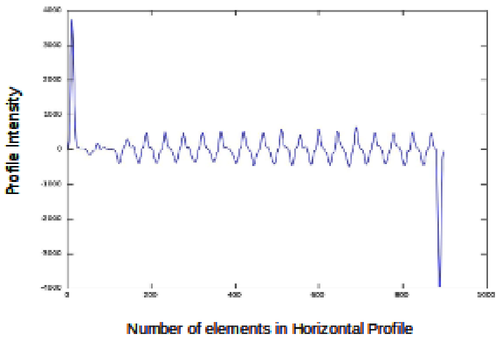
$$G = \left[\frac{1}{2\pi\sigma^2}\right] e^{\left[\frac{-x^2}{2\sigma^2}\right]} \qquad (2)$$

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

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



(a)



(b)

Figure 3. (a) Gaussian derivative kernel of order-1 & sigma-3
(b) Gaussian kernel convolved with horizontal profile

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 Y-axis, which is represented by 'VP' of size 'm' and is defined by "(4)".

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

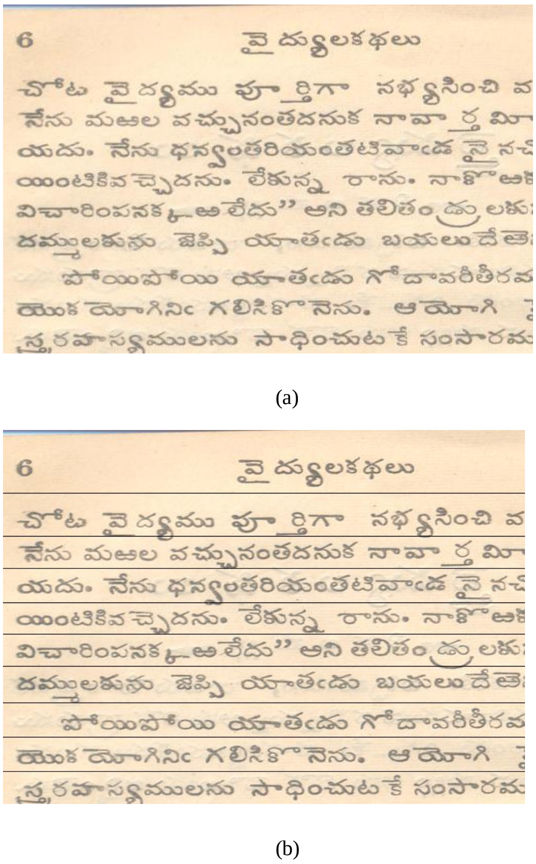
If we use the same Gaussian kernel for character segmentation results in a smooth profile with non-uniform peaks. Identification of non uniform peaks with a specific intensity value involves difficulty during character segmentation. Hence a threshold for character segmentation is defined.

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

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

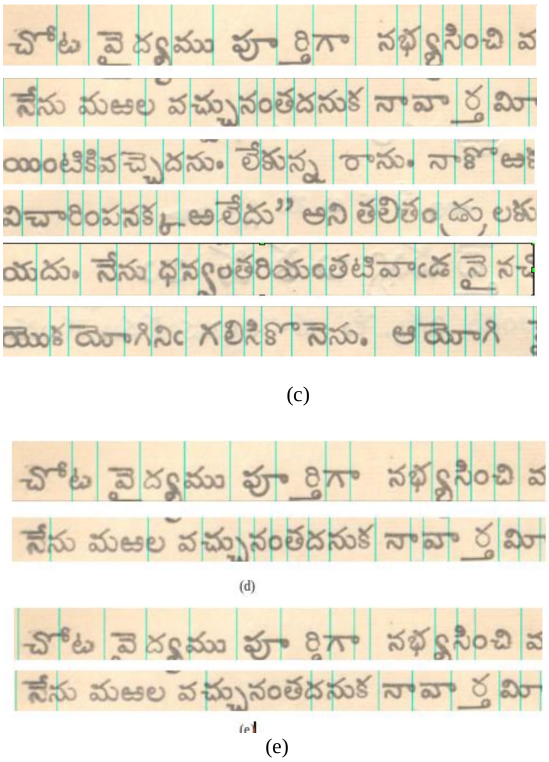
V. RESULTS AND DISCUSSIONS

Segmentation algorithm is applied on a set of 20 noisy document images. They are collected from the Net (Telugu old book named "Thiagarajaswami Krithis" is published in 1933, at Kesari Printing Press, Chennapuri) and the scanned copies of old story books(Telugu old book named "vydyula kathalu" is published in 1942 at Madras printing press) which are of 50 to 60 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 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 “(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 “(5)” the threshold value is noticed to be sensitive to the denominator. Hence, from the “(5)” the denominator in the second part of equation is selected based on the trail 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 “(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.



(a)

(b)



(c)

(d)

(e)

Figure 4. (a) Original Image (b) Line segmented Image © Character segmented image using threshold
(d) character segmentation using Gaussian kernel with(sigma=3,order=1)
(e) character segmentation using Gaussian kernel with(sigma=1.8,order=1)

VI. PERFORMANCE EVALUATION

The data of efficiency (i.e., estimated by involving the line segmentation and character segmentation over a degraded image) along with that of the allied parameters estimated by using the Gaussian kernel and I_{Th} on the set of considered images are tabulated in the table 1. The efficiency tabulated, corresponding to the segmentation process under noisy conditions. Performance of the algorithm is tested over the 20 text samples result for 100% line segmentation efficiency. The 296 lines from 20 text samples are tested over the defined algorithm on this direction of lines, which are having 5974 characters. It may be noticed that only 4396 are correctly segmented. 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, never leads to achieve 100% efficiency for character segmentation. Hence an efficiency of 73.58 is found to be achieved during character segmentation. However, utility of this threshold value is originated during the process of character segmentation. If, thus Gaussian kernel used for character segmentation is found to lead the increasing efficiency, as it depends on the values of order and sigma also.

TABLE I. Comparison of segmentation efficiency

S. No	Samples	No of Lines	Total no of Characters	Correctly Segmented Characters	Efficiency (%)
Segmentation Rate of Noisy Documents					
1	20	296	5974	4396	73.58

VII. CONCLUSIONS

Character segmentation in Ancient Telugu text document images is proposed in the present work without involving binarization phenomena. Document image is perceived as background and foreground information containers. Background information reflects the characteristics of noise where as the foreground information is embedded with text information with lower intensity values tending towards black pixels. The characteristic of noise is analyzed in the horizontal and vertical profiles. Line segmentation is carried out by convolving horizontal profile with Gaussian kernel of order-1 and sigma-3. The maximum efficiency of 100% is observed even under deformations. Text information within the script line is identified between two successive peaks of the resultant profile. However individual character segmentation from the script line is found to be ineffective with this approach due to high degree of non uniform intensities. An extensive analysis is carried out to identify the relation between maxima and minima of the vertical profile. Thresholding approach is adopted while segmenting the individual characters in the script line. An efficiency of 73.58% is observed with this approach without losing any information in the ancient document image. A hybrid approach of combining the statistical behavior of foreground text information with background noise characteristics is in progress.

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: 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] 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

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

[9] C V Lakshmi & C Patvardhan, '*Optical Character Recognition of Basic Symbols in Printed Telugu Text'*, IE (I) Journal-CP, Vol 84, November 2003 pp. 66-71.

[10] 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.

[11] 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.

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

[13] 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.

[14] A.V.S.Rao, N.V.Rao, A.S.C.S.Sastry, L.P.Reddy," Canonical Syllable Segmentation of Telugu document images," Procedings of International Conference TENCON 2008, Hyderabad, 18-21 November, 2008.

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

[16] 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

[17] 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

[18] A.V.S.Rao,G.Sunil,N.V.Rao,T.S.K.Prabhu, A.S.C.S. Sastry, L.P.Reddy,"Adaptive Binarization of Ancient Documents," Procedings of International Conference on Machine Vision, 978-0-7695- 3944-7/10, 2010.

Adabala Venkata Srinivasa Rao obtained his B.Tech degree in Electronics and Communication Engineering from JNT University, Kakinada, India, AMIE Electrical from Institute of Engineers (India), Kolkotta,India. and M.Tech in Instrumentation and Control Systems from JNT University, Kakinada, India. He was worked in various engineering colleges at different positions. He is currently working as an Associate Professor in AKRG college of engineering and technology, Nallagerla, WG Dist, Andhra pradesh, India. He has 9 years of teaching and 4 years of industrial experience. He has 13 publications in various International Conferences and journals. His area of interests include Pattern Recognition , Image Processing and VLSI. He is an active member in professional bodies like AMIE, IACSIT