International Journal of Engineering Research in Computer Science and Engineering (IJERECE) Vol 3, Issue 3, March 2016

Statistical Analysis of Difference Profile of Word Patterns

[1] Bhaskara Rao Perli, [2] D.Vara Prasad [3] R.K.Singh
[1][2][3][4] Electronic & Communication Engineering, St.Peter's Engineering College, Hyderabad, A.P, India
[1] mail2bhaskarp@yahoo.com[2] vara222@gmail.com, [3] ad.rk@gniindia.org

Abstract: **Optical character recognition with segmentation of isolated pattern and recognition of each pattern, at the same time the relation between patterns are very important while translating document image into equivalent editable form. It is quite interesting that the human learning process starts with word level recognition at the first instance and later improved up to the level of isolated pattern and there relations in between Indic scripts are more specific and also unique due to a large number of convict formations. The inter relations among various syllable may vary from script to script. Another important feature in this regard is identified as zone information model which is adopted in a large number of attempts made by various researches. The present work is aimed at exploring zone information with regard to text line and at the same time the statistical variation among word pattern. The concept of differential profile which is used in the earlier literature will be explored for the analysis of word pattern in southern scripts.**

***Keywords*— Pattern; scripts; syllable; segmentation**

I. INTRODUCTION

Most of the information that is available in the world is in printed medium. This has hindered storing, exchanging and processing of the information electronically. Converting them into electronic medium is time consuming, laborious and expensive as well. These factors have motivated people to develop automated systems to perform this task. Optical Character Recognition is the process of converting scanned images of machine printed or hand-written text into a computer processable format. Though a great amount of work has been carried out on middle zone extraction of Indian scripts, very few works are reported in literature at script independent level. Also, the great demand for automatic processing of multi-lingual documents show that much more work needs to be carried out on script independent level. So, this thesis focuses on middle zone extraction of Telugu, Tamil, Kannada, Malayalam, Bengali, Hindi and English scripts.

Research in Optical Character Recognition (OCR) [1,2,3,4,5,6,7] is popular for its application potential in banks, post-offices and defense organizations. Other applications involve reading aid for the blind, library automation, language processing and multi- media systems design. They are proposed for Bangla script at text line level. Bangla characters contain a horizontal line at the top of the middle zone. This line is called matra. A.G.Ramakrishnan and Kaushik Mahata [8] are proposed for Tamil text at line level. They reported that the text lines of any Tamil text will have three zones as upper, middle and lower zones.

A.S.Chandrasekhara Sastry, Satyaprasad Lanka, P.Paul Clee and L.Pratap Reddy [9] are proposed for Telugu text using statistical properties of peaks and valleys of the profile vector. In this the horizontal profile of a text line was first obtained and row with peak in the first half is upper bound of middle zone. Find the slope of the valley from the row with peak in lower half of profile with respect to the successive row. The row with the maximum slope of the valley is identified as the lower bound of middle zone. M.C. Padma and P.A Vijaya [10] are proposed for Kannada, Hindi and English languages. They projected by partitioning the text line using the four lines that are obtained from the top-profile and the bottom-profile of each text line.

The present work focus on Matched word in a line which is independent of the language in a document image. Many languages that are involved from the Roman culture and also indict valley civilization concentrated on linear structures of various stroke based as well as curved based components of an image. This is due to fact that writing styles are found on stone bricks and palm leaves. A common observation made these scripts during the learning style of script which is defined as zones. Keeping in a view of above perspective a generic method is attempted in the present work while identifying zone segmentation induces. Many algorithms adopted various zone

segmentation algorithms where as presented aim at unify many of these approaches. Two models are proposed while addressing this problem after analyzing the statistical properties of the scripts of a whole line. Horizontal profile (which reflects the above property) of a text line image is considered as basis in the proposed segmentation approach. Features like, peaks and valleys are adopted in the zone identification process.

II. LINE PROFILE VS WORD PROFILE

A. Zone Extraction Model

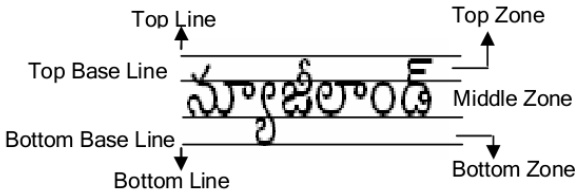


Fig.1. Structure of a text line

B. Differential Profile Based Matched Word Identification Algorithm

Step1: Convert the RGB image into Gray scale image.
Step2: Extract horizontal profile vector for the text line image.
Step3: Find the row with maximum value of black pixel density and it is the line for dividing the profile in to two halves along the length.
Step4: Find the slope with reference to the row in step3 to rows of the first half of the profile.
Step5: Find the row with minimum value in step4 is marked as Headline.
Step6: Find the slope with reference to the row in step3 to rows of the second half of the profile.
Step7: Find the row with maximum value in step5 is marked as Baseline.
Step8: Extract the Middle zone of the text line image.
Step9: Divide the line into words by using vertical profile
Step10: apply the steps 3-7 to extract middle zone of each word
Step11: find the match words in the line

C. Difference Profile Based Matched Word Identification Algorithm

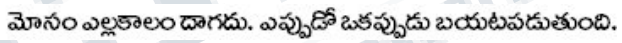
Step1: Convert the RGB image into Gray scale image.
Step2: Extract horizontal profile vector for the text line image.
Step3: Compute the linear difference of a profile.
Step4: Divide the linear difference of a profile vector two halves along the length.
Step5: Find the row with peak in the first half of difference profile vector is marked as Headline.
Step6: Find the row with valley in the second half of difference profile vector is marked as Baseline.
Step7: Extract Middle zone of text line image.
Step8: Divide the line into words by using vertical profile
Step9: apply the steps 3-7 to extract middle zone of each word
Step10: find the match words in the line.

III. RESULTS

A. Evaluation Of Differential Profile

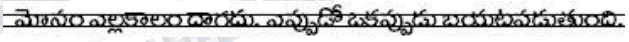
In this consider the Differential profile model in which take valley in first part and peak in second part as Headline and Baseline respectively is tested on various font types with font size 12, 14, 16 and 19 of Telugu, Tamil, Kannada, Malayalam, Bengali, Hindi and English languages. Sample output images of above mentioned languages text lines are shown in Fig:2-5 and details of results obtained are tabulated in Table 1-4 respectively.

Input image

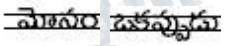


Telugu text line

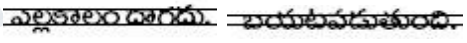
Output:



Middle zone of Telugu text line



Middle Zone Matched Word Of Telugu Text Line



Middle Zone Unmatched Word Of Telugu Text Line

Table .1: Middle Zone Extraction of Telugu Language using Differential Profile

Font size	matched	Base only matched	Head only matched	unmatched	Total words
16	42	30	16	12	100
18	30	21	33	16	100

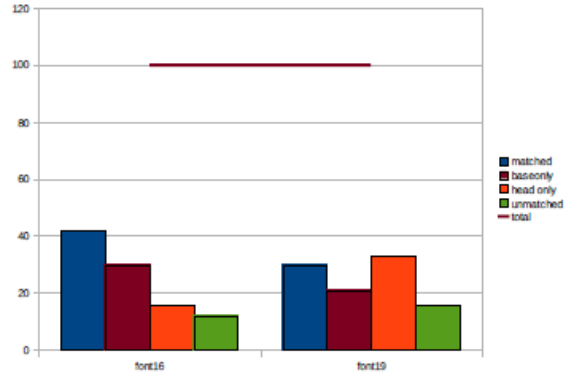
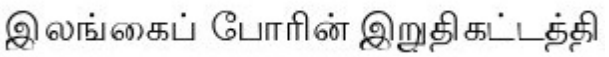


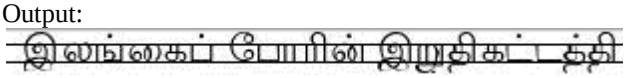
Fig. 2. Sample output images of Telugu text line.

Input Image

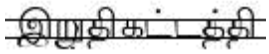


Tamil text line

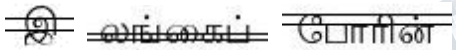
Output:



Middle zone Of Tamil Text Line



Middle zone Matched Word Of Tamil Text Line



Middle zone Unmatched Word Of Tamil Text Line

Table.2. Middle Zone Extraction of Tamil Language using Differential Profile

Font size	matched	Base only matched	Head only matched	unmatched	Total words
16	39	19	10	32	100
18	30	14	36	20	100

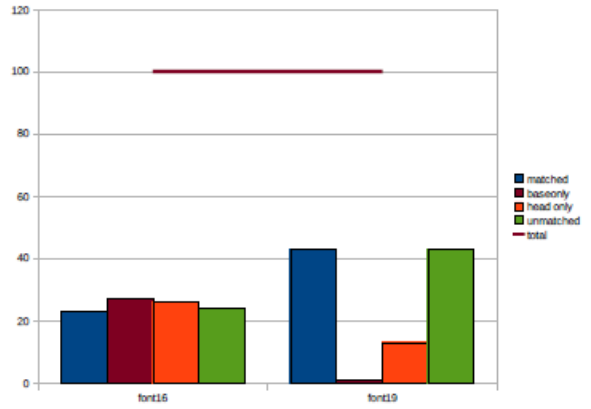
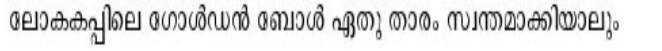
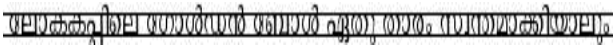


Fig. 3. Sample output images of Tamil text line.

Input Image

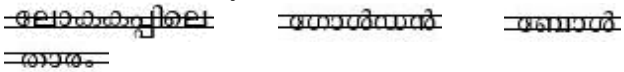


Malayalam text line



Output:

Middle zone Of Malayalam Text Line



Middle zone matched Word Of Malayalam Text Line

Middle zone Unmatched Word Of Malayalam Text Line

Table. 3. Middle Zone Extraction of Malayalam Language using Differential Profile

Font size	matched	Base only matched	Head only matched	unmatched	Total words
16	23	27	26	24	100
18	43	1	13	43	100

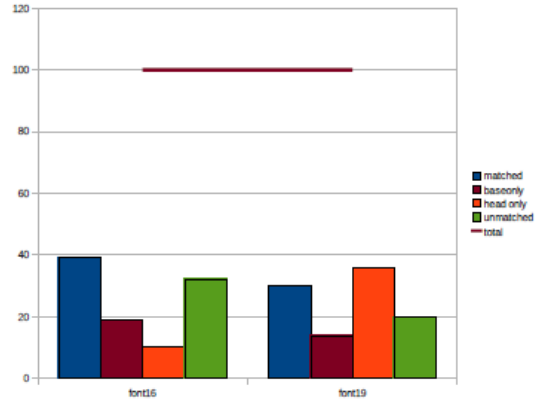
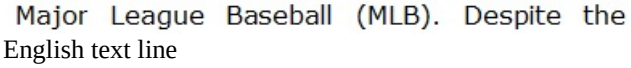


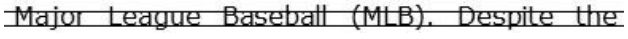
Fig:4. Sample output images of Malayalam text line.

Input Image

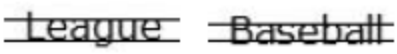


English text line

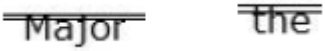
Output:



Middle zone Of English Text Line



Middle zone Matched Word Of English Text Line



Middle Zone Unmatched Word Of English Text Line

Table. 4. Middle Zone Extraction of English Language using Differential Profile

Font size	matched	Base only matched	Head only matched	unmatched	Total words
16	46	5	26	23	100
18	40	12	21	27	100

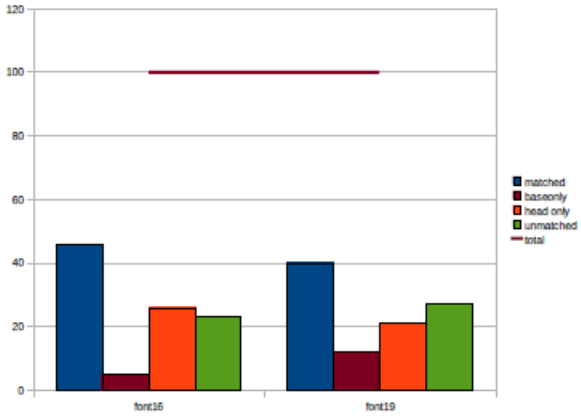
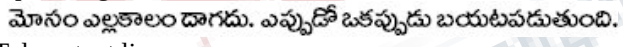


Fig. 5. Sample output images of English text line

B. Evaluation Of Difference Profile

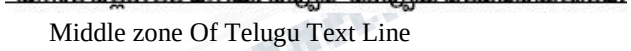
In this consider the first order linear difference profile model in which take peak in first half and valley in second half as Headline and Baseline respectively is tested on various font types with font size 12, 14, 16 and 19 of Telugu, Tamil, Kannada, Malayalam, Bengali, Hindi and English languages. Sample output images of above mentioned languages text lines are shown in Fig:6-9 and details of results obtained are tabulated in Table 5-8 respectively.

Input Image

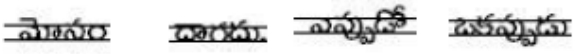


Telugu text line

Output:



Middle zone Of Telugu Text Line



Middle Zone Matched Word Of Telugu Text Line

Middle Zone Unmatched Word Of Telugu Text Line

Table. 5. Middle Zone Extraction of Telugu Language using Difference Profile

Font size	Matched words	Base only matched	unmatched	Total words
16	57	16	27	100
18	57	14	29	100

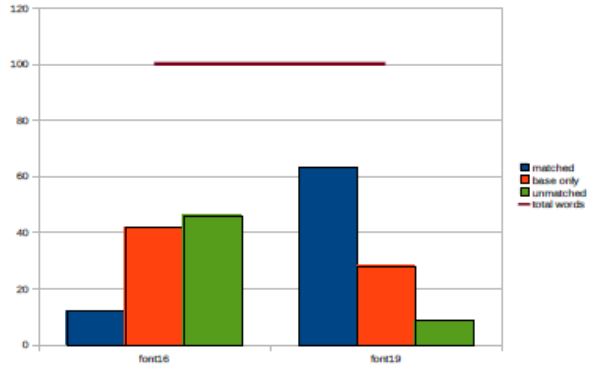
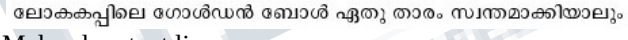


Fig. 6. Sample output images of Telugu text line.

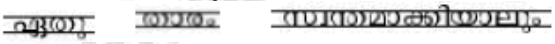
Input Image



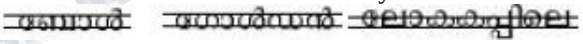
Malayalam text line

Output:

Middle Zone Of Malayalam Text Line



Middle zone Matched Word Of Malayalam Text Line



Middle Zone Unmatched Word Of Malayalam Text Line

Table.6: Middle Zone Extraction of Malayalam Language using Difference Profile

Font size	Matched words	Base only matched	unmatched	Total words
16	12	42	46	100
18	63	28	9	100

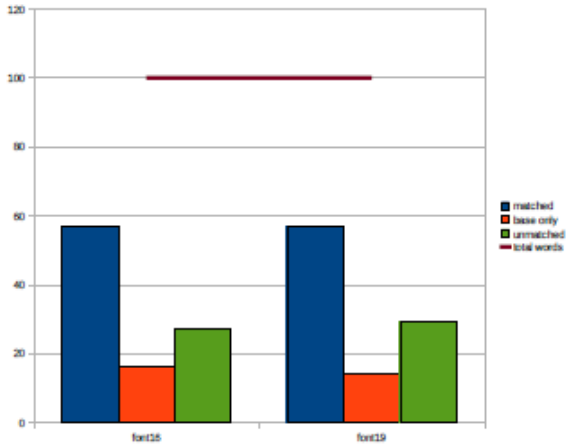


Fig. 7. Sample output images of Malayalam text line.

Input Image

नेहरूगांधी फैमिली ने सीधे कुछ कहने से प

Output:

नेहरूगांधी फैमिली ने सीधे कुछ कहने से प

Table.7: Middle Zone Extraction of Hindi Language using Difference Profile

Font size	Matched words	Base only matched	unmatched	Total words
16	63	24	13	100
18	34	51	15	100

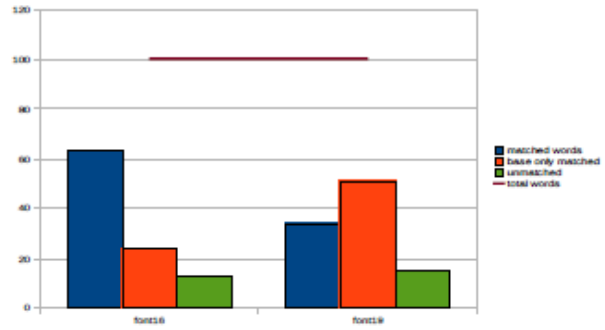


Fig. 8. Sample output images of Hindi text line.

Input Image

Major League Baseball (MLB). Despite the

English text line

Output:

Major League Baseball (MLB). Despite the

Middle zone Of English Text Line

Middle zone Matched Word Of English Text Line

Middle Zone Unmatched Word Of English Text Line

Table.8. Middle Zone Extraction of English Language using Difference Profile

Font size	Matched words	Base only matched	unmatched	Total words
16	70	12	18	100
18	63	17	20	100

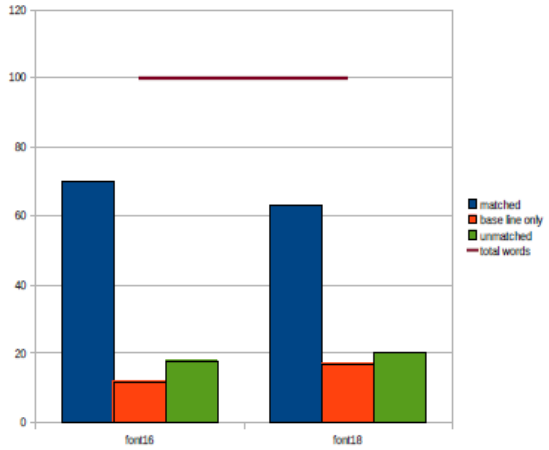


Fig. 9. Sample output images of English text line.

Table.9.Percentage of Matched Words for Languages in Differential profile

Languages	Percentage of matched words
TAMIL	33
ENGLISH	43
MALAYALAM	34.5
TELUGU	36

Table.10: Percentage of Matched Words for Languages in Difference Profile

Languages	Percentage of matched words
HINDI	48.5
ENGLISH	66.5
MALAYALAM	57
TELUGU	37.5

V. CONCLUSIONS

The present work focus on Matched word in a line which is independent of the language in a document image. Differential profile is the first model which analyzes the pixel distribution with reference to the maxima. The differential valley of the first half of the profile, differential peak of the second half of the profile reflect the zone separation in majority of the scripts. However this property inconsistent with regard to font styles. The evaluation is carried on 50 samples with an observed average matching word efficiency of 43% english, 36% telugu, 33% tamil, 34.5% malayalam and respectively. The horizontal profile is analyzed with an assumption that the profile reflects variations of pixel distribution among successive rows of the text line. Using this logic Difference profile algorithm is proposed. The relative variations of pixel distribution among the successive rows entire data is analyzed for each half of the horizontal profile. Peak of the first half and valley of the second half resulted in zone separation lines of top and bottom respectively. Evaluation of this algorithm on the same target samples is carried out with observed average matching word efficiency of 66.5% english, 37.5% telugu, 48.5% hindi, 57% malayalam and respectively. Extension of the proposed algorithm on the other languages with an improvement up to 100% can be taken in future scope. Statistical behavior of horizontal and vertical profiles can be extended in the segmentation process of meaningful units of the respective language is another task for future.

REFERENCES

1) U.Pal and B.B.Chaudhuri, OCR in Bangla: “an Indo-Bangladeshi Language”, pp.269-273, 1994.

2) U Garain and B.B.Chaudhuri, “Segmentation of Touching Characters in Printed Devanagari and Bangla Scripts using Fuzzy Multifactorial Analysis”, IEEE Transactions on Systems, Manand Cybernetics, Part C, Vol. 32, No. 4, pp. 449–459, 2002.

3) R. Casey and G. Nagy, ‘Recognition of printed characters’, IEEE Trans Electron, Comput., vo1.15, pp. 91 - 101, 1991.

4) R. Chandraeekaran, M. Chandrasekaran and G. Siromony, ‘Computer recognition of Tamil, Malayalam and Devnagari characters’, J. Inst. Eletron. Telecom. Engg.(India), 30, pp. 150 - 154,1984.

5) B. Chaudhuri and U. Pal, ‘Recognition of Bangla Printed Script’, Proc. Int Conf. on Application of Information Technology in South Asian Language, New Delhi Feb 25-26,1994.

6) V. K. Govindan and A. P. Shivaprasad, ‘Character Recognition - a Survey’ Pattern recognition, vol-23, no-7, pp. 671 - 683, 1990.

7) R. M. K. Sinha, ‘ Rule based contextual post processing for Devnagari text recognition’, Pattern Recognition , vol.-20, no-5, pp. 475 - 485, 1985.

8) K. Y. Wang, R. G. Casey and F. M. Wahl , ‘Document Analysis System’, IBM J. Res. Development , vol.-26, no-6, pp. 647 - 656,1982.

9) A.G. Ramakrishnan and Kaushik Mahata, “A complete OCR for printed Tamil text”, pp.151-156, 2000.

10) A.S.Chandrasekhara Sastry, Satyaprasad Lanka, P.Paul Clee and L.Pratap Reddy, “Combining Spatial and Transform Features for the Recognition of Middle Zone Components of Telugu”, pp.1-5, 2008.

11) M.C. Padma and P.A Vijaya, “Script Identification of Text Words from a Tri Lingual Document Using Voting Technique”, IJIP, Vol.4, pp. 35-52, 2010.