Implementation of Multilevel Segmentation using Cognitive Approach - A Case study on Devanagari Script

Sirisha Badhika*

* Research Scholar, Shri Jagdish Prasad Jhabarmal Tibrewala University, India, sirisha.badhika@gmail.com

Abstract- Optical character recognition strategies are concentrated towards improving recognition efficiencies by adapting post processing techniques. OCR errors of complex scripts are due to improper segmentation as well as inappropriate rendering. Topological features of a script as a global knowledge and geometrical features of isolated patterns as local knowledge are combined together while segmenting the basic unit, syllable. A multilevel segmentation was implemented to adopt the idea of human visual perception capability into the proposed model to use the distinct features exhibited by Devanagari script. The proposed method is based on implementing cognitive approach in segmentation phase by dealing with the syllable as a meaningful unit of information right from segmentation phase itself instead of isolated pattern on document images of Devanagari script.

Index Terms- OCR, segmentation, syllable, Cognitive approach , Topology, Geometry.

I. INTRODUCTION

Optical Character Recognition (OCR) is a convenient and efficient tool for office Automation and information retrieval, and is becoming more and more important in today's office and library environment. Current OCR research and development is mostly centered around isolated patterns of pixels . In a general OCR model, the key step is to segment the scanned image. Segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as superpixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. As every individual word is traditionally built by juxtaposition of letters which are recognized using a combination of strokes and curves. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. Each of the pixels in a region is similar with respect to some characteristic or computed property, such as color, intensity, or texture. Adjacent regions are significantly different with respect to the same characteristics. Several general-purpose algorithms and techniques have been developed for image segmentation. Since there is no general solution to the image segmentation problem, these techniques often have to be combined with domain knowledge in order to effectively solve an image segmentation problem for a problem domain. When the script is associated with a simple syntactic rule of defining isolated pattern as equivalent to a character, then the job of the recognizer becomes easy[1]. Under the influence of syntactic rules with shape variations and contextual formations, the job of the recognizer becomes highly complex. The cumulative errors of all the above stages are addressed in the post processing stages with the help of knowledge sources like dictionary, phrase database etc. Error detection itself is a complex phenomenon in this context. Adaptability of syntactic rules at this stage is yet to progress. In certain scripts like Telugu, Devanagari, etc., (Bhrami derived) development of knowledge sources is still at a primitive stage. Shape variations within the structure is well defined for all the scripts which is an important knowledge source from the angle of syntactic rules[2]. Even though these rules are bounded by the limited approach at sub-word level pattern, it is possible to combine grammar rules in this unit with structural pattern recognition. The current research was motivated to adopt the idea of human visual perception capability into the proposed model to use the distinct features exhibited by each script. So, the target of this paper is to identify the script type of the texts without reading the contents of the document. The present research work concentrates on implementing cognitive approach in segmentation phase by dealing with the syllable as a meaningful unit of information right from segmentation phase itself instead of isolated patterns.

II. LITERATURE REVIEW

A. Related Work:

The development of OCR has incremental improvements in individual phases. The segmentation phase broadly follows three approaches: dissection method, recognition based segmentation or hybrid approach. Few of the Dissection methods used are discussed here. Initial systems for segmenting machine-printed characters are based on two simple features: white space and pitch [4]. Classifying the isolated patterns into identifiable glyphs is the major approach attempted till recent period. Dissecting the word into possible isolated patterns in Indic scripts is a continuation of the effort made in [7]. Even though isolated patterns are considered to be primitives of feature extraction stage, large numbers of errors are reported in the segmentation process. Sometimes the errors are due to in appropriate formulation of association rules among recognizable unit. Interestingly all these errors are attempted at the post processing stage [8]. The topological and geometrical features of a syllable are explored extensively in [9]. The concept of zone is introduced while defining syllable patterns in a script line. Three zones are Top, Middle and Bottom zones and are identified within the script line using difference profile algorithm on Horizontal Projection Profile. Pal and Chaudhuri [6] explored a feature

based algorithm for automatic recognition of Devanagari script. A structured binary-tree classifier was used by implementing global features for upper tiers of the tree, based on condensed run length approach, for classification and the lower tiers of the tree uses local features, estimated as moments, for classification. Pratap Reddy et. al. [11] proposed a classical approach in the segmentation of Canonical Syllables of Telugu document images. This model is based on the canonical definition of the Telugu script. The relation between zones and components is established in the segmentation process of canonical syllable. The components in the script are classified into six different classes. An individual component associated with either top zone or middle zone or bottom zone is classified as Top Zone (TZ class), Middle Zone (MZ class) and Bottom Zone (BZ class) respectively. A.S.C.S Sastry et al. [12] proposed
a zone extraction method for Telugu text which uses 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.

B. Segmentation Challenges:

With isolated scripts, characters are written to be separable (although they may touch due to degradation) while connected scripts cannot be easily segmented. Scripts can also be broadly classified based on word composition into syllabic and non-syllabic. In non-syllabic scripts, characters are horizontally (or vertically) separable glyphs whereas in syllabic scripts, glyphs appear as syllables, which are in turn a complex combination of one or more characters. Sometimes, characters fuse together to form new shapes,The presence of language-specific constructs, in the domain of non-Latin scripts, such as shirorekha (Devanagari), modifiers (south-east Asian scripts), writing order, or irregular word spacing (Arabic and Chinese) requires different approaches to segmentation. While conventional vertical or horizontal profiling methods fail to segment characters directly from words, character segmentation from syllables using only connected component analysis itself is a complex task which is highly correlated with the script characteristics [3].
Casey and Lecolinet define four methods of character segmentation [4] as follows:

- The first is dissection based where the image is decomposed into classifiable units before feature extraction and classification. Due to its dis-connectivity from the later modules, the feedback mechanism is expensive.
- The second method tries to classify subsets of spatial features collected from a word image as a whole. Segmentation hypotheses are generated and choosing best hypothesis along the word gives best recognition result. The challenge of this approach is to come up with minimal number of possibly correct hypotheses.
- The third involves over-segmenting of the word image i.e., one single character could be segmented into two or more parts. The over-segmentation in the vertical direction only happens to long characters such as ए and ह [3].
- The fourth method recognizes an entire word as a unit and is a holistic strategy. A major drawback of this class of methods is that their use is usually restricted to a predefined lexicon.

III. INDIAN SCRIPT FEATURES

A. Devanagari Script characteristics:

India is a multilingual country. A significantly large number of scripts are used to represent these languages. Indian languages are broadly divided into two categories . The Northern Languages are stroke based and southern are cursive scripts. Most of the Northern languages are derived from Devanagari script. As any other script Devanagari script has its own specified composition rules for combining vowels, consonants, and modifiers. Devanagari script has 13 vowels [see Fig. 1(a)] and 34 consonants [see Figure 1(b)] along with 14 modifiers of vowels and of "rakar", as shown in Figure 1(a)] symbols Some of them can be combined with their type[5]. Devanagari script consonants are written in a linear left-to-right order, vowel signs are positioned non-linearly above, below or either side of consonants. A modifier can be attached to a vowel or to a consonant. Consonants may have a half form when they are combined with other consonants Except for some characters, the half forms of consonants are the left part of original consonants with the right part removed.

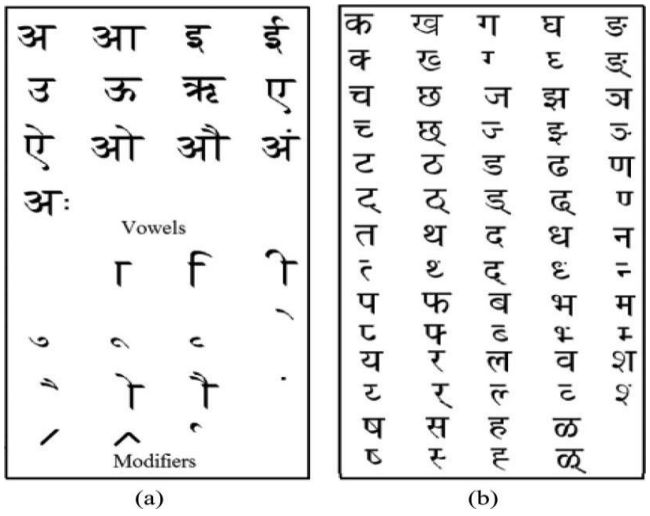


Figure 1 :(a) Vowels and modifiers of Devanagari script. (b) Consonants and their corresponding half forms (shown below the consonants) in Devanagari script.

Another distinctive feature of Devanagari script is the presence of a horizontal line on the top of all characters. This line is known as header line or "shirorekha" (see Figure 2). The words can typically be divided into three strips: top, core, and bottom, as shown in Fig. 3. The header line separates the top and core strips and a virtual base line separates the core and lower strips. The top strip generally contains the top modifiers, and bottom strip contains lower modifiers. When two or more characters

appear side by side to form a word in Devanagari, the header lines touch and generate a bigger header line[5].

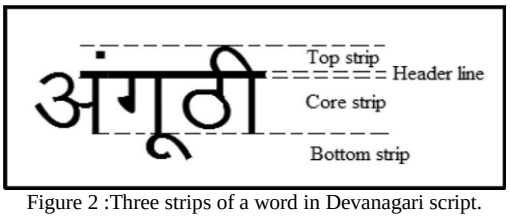


Figure 2 :Three strips of a word in Devanagari script.

The distinct visual appearance of every script is due to the presence of the segments like – horizontal lines, vertical lines, upward curves, downward curves, descendants and so on. The presence of such segments in a particular script is used as visual clues for a human to identify the type of even the unfamiliar script.

B. Language Dependent and independent features:

All languages have basic common features of representation. Identification of the characters in Latin and Han based scripts and easy because each character is independent by itself, However dealing with complex scripts(Bhrami based) it is very difficult as it is a combination of independent and dependent consonants. For example Devanagari characters hang from a horizontal line(called the head stroke) written at the top of the character, unlike English letters which are written up from the line below them. It is very important to learn the correct stroke order for Devanagari characters. In English Devanagari is often called a syllabary, rather than an alphabet, because each Devanagari character normally represents a consonant and a vowel combination or a vowel on its own. Each Devanagari character normally represents a complete syllable and a syllable can be combination of full form or half form. In this paper an exploration of cognitive science is to be proposed to identify the language dependent features.

IV. PROPOSED METHOD – *SYLLABLE SEGMENTATION*

The current research was motivated to adopt the idea of human visual perception capability into the proposed model to use the distinct features exhibited by each script. Syllable is a fundamental unit of information in any script based language, derived from the means of phoneme. The structural property of a syllable is dependent on the rules defined by the script. One or more isolated character is associated with a syllable in English. The patterns of individual vowels and consonants are fixed in nature. However the consonant-vowel cluster formation follows a predefined set of rules in combination with patterns of vowel modifier and consonant modifier [2].However it is difficult in complex scripts like Bhrami and Devanagari scripts. This can be explained clearly with two examples one in English(Latin derived) and another in Hindi(Devanagari derived). The word 'vijay' consists of 5 isolated patterns 'v', 'i', 'j' , 'a' and 'y'. Each one of them map on a unique code which generate a glyph to render them correctly. Whereas the word विध्या consists of 4 meaning full units(2 syllables). The first syllable alphabet वि is complex in nature with a combination ध्या consists of 4 meaning full units(2 syllables). The first syllable alphabet वि is complex in nature with a combination of a consonant(/YA/) with a dependent consonant(/DYA/). Adding to this complexity the order in which they are rendered is also important.. A comparison between the Isolated Pattern approach and Syllable approach is presented in the below Figure 3(a) and (b).

Word	IP1	IP2	IP3	IP4	IP5	IP6
मुस्कान	म	ु	स्	क	न	
भाषा	भ	ा	ष	ा		
विधाता	ि	वे	ध	ा	त	ा
वास्तव	व	ा	स्	त	व	

Figure 3 : (a) Isolated Pattern Approach

Word	IP1	IP2	IP3	IP4
मुस्कान	मु	स्का	न	
भाषा	भा	षा		
विधाता	वि	धा	ता	
वास्तव	वा	स्त	व	

Figure 3 : (b) Syllable Approach

V. SEGMENTATION PHASES:

After scanning the document, the document image is subjected to pre-processing for back ground noise elimination, skew correction and binarization to generate the bit map image of the text. The pre-processed image then go through number of segmentation phases .

A. Line Segmentation:

In the first phase the preprocessed image is segmented into lines by using horizontal projection profile. The horizontal projection profile is the histogram of the number of ON pixels along every row of the image. White space between text lines is used to segment the text lines. The projection profile will have valleys of zero height between the text lines. Line segmentation is done at these points.

B. Word segmentation:

The spacing between the words is used for word segmentation. For Devanagari script, spacing between the words is greater than the spacing between characters in a word. The spacing between the words is found by taking the vertical projection profile of an

input text line. Vertical projection profile is the sum of ON pixels along every column of the image.

C. Syllable segmentation:

For Devanagari script, character segmentation involves the removal of sirorekha/Header line. In this third phase topological information of the script line is used to divide the syllable into three zone's, the top zone, bottom zone and middle zone, horizontally[9]. The middle zone, through its horizontal difference projection profile, is then used to determine the candidate boundaries of the syllable. Syllable is treated as a basic unit of information as equivalent to one meaningful unit as per the cognitive approach. Syllable is composed of one or more components[2]. These candidate boundaries are confirmed or modified based on the application of the knowledge source about its geometry(inter syllable distance,aspect ratio, length, etc.).

Algorithm used for syllable segmentation:

1. The horizontal projection profile of the meaningful unit is first extracted for every syllable.
2. The difference profile of the same is then computed from the above.
3. To determine the left boundary traverse the difference profile starting from the middle towards the beginning and determine at what point the maximum peak occurs. This point represents the left boundary of the core component.
4. Similarly the right boundary is located by starting from the middle of the difference profile and moving towards the end. The minima in this traversal indicates the right boundary of the core component.
5. In both traversals exempt the extreme points as they represent the start and end of the meaningful unit or syllable.
6. Once the boundaries of the core component are established this divides the syllable into three regions horizontally namely left region, center region and right region.
7. Now combine these 3 regions with the 3 vertical zones(top, middle and bottom) to segment the syllable into nine new regions namely Top-left(TL), Top-Center(TC), Top-Right(TR), Middle-Left(ML), Middle-Center(MC), Middle-Right(MR), Bottom-Left(BL), Bottom-Center(BC) and Bottom-Left(BL) regions.

Script analysis is then used as a knowledge source to indicate which of these regions are essential i.e., contain significant information for that script. Other regions may contain information but that would be redundant information. This information also forms part of the knowledge source. The vowel modifiers and the consonant modifiers are treated as non-core components. The geometric representation of core components will always be within the boundaries of middle zone. The geometrical features of non-core components might spread over all regions with a condition that their placement is around the core component only. The presence of another core component will be treated as a new syllable. An example is showed in Figure 4.

Figure 4 : Devanagari Script after syllable segmentation

Similar exercise may done on other scripts and identified essential regions can be evaluated. An attempt is made to customize the application depending on the script.

VI. RESULTS OF SYLLABLE SEGMENTATION

The segmentation process starts from dissection of lines. A test document image is presented in figure 5.document image is presented in Figure 5.

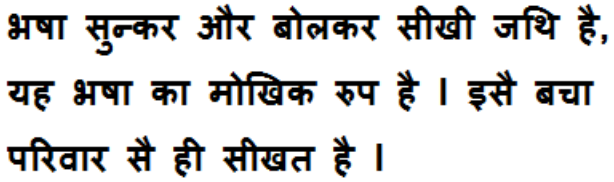
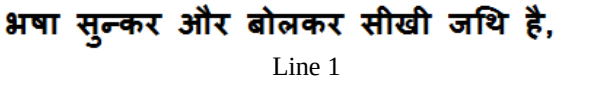
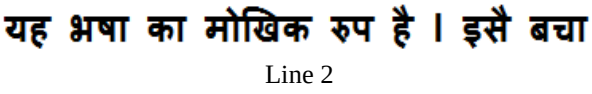


Figure 5 : Sample Devanagari Script

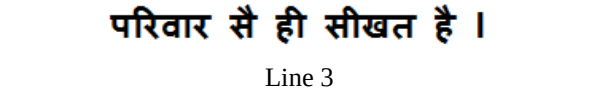
Line Segmentation: Script line segmentation is carried out with the help of profile plot of horizontal plane. The horizontal projection profile is the histogram of the number of ON pixels along every row of the image. White space between text lines is used to segment the text lines. Figure(6) shows the segmented lines .



Line 1



Line 2



Line 3

Figure 6 : Devanagari i Script Segmented into lines

Simple profile information is found to be insufficient for further processing. segmentation of syllables is more convenient if the middle zone is made the basis for syllable segmentation as shown below in Figure(7). Starting of a line is identified with the help of 5 non-zero values of difference profile proceeded by 2 zeros. The starting point(S) is identified as the first non-zero quantity. Ending(E) of the line is identified with the reverse phenomena in which 2 non-zero elements are to be followed by 5 zeros. The

peak value in the positive plane & valley in the negative plane will provide us the labelling of Top line(T) & Bottom line(B) for each identified script lines.

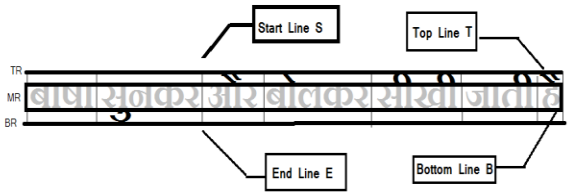


Figure 7 : Script line segmentation profiles with zones

The middle zone information of each script line is highlighted for the purpose of readability. The labeled information corresponding to zones is explored further in the second phase of segmentation model.

Word Segmentation: Word Level segmentation within the script line is the next step carried out and the results are presented in the Figure(8)

Figure 8 : Line1 Word Segmentation

The profile information of the vertical plane of the respective script line is explored during word segmentation. The same phenomena of identifying non zero quantities of difference profile yielded efficient results. However the spacing between words is much less when compared to line spacing we adopted a lower threshold of 2. Fixing of threshold is dependent on font size. For Devanagari script, character segmentation involves the removal of sirorekha /Header line .This can be achieved by computing the horizontal projection of the word image box. The row containing maximum number of black pixels is considered to be the header line. Figure 9 shows the constituent components getting separated after the removal of the sirorekha (the horizontal bar) from a word.

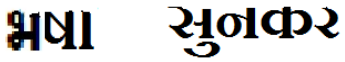


Figure 9 : Removal of the sirorekha from words

Sub-word level segmentation is carried out in the further step which is a more complex task. We first segment candidate patterns of syllable from the difference profile of a word as shown in figure(10). We can observe the peaks and valleys will appear successively in the difference profile of a word. We can observe the peaks and valleys will appear successively in the difference profile.

Figure 10 : Syllable Segmentation for Line1

In the present work, we optimized the number of regions of a syllable to four instead of nine for Devanagari script. The region MC is essential. Among other regions, consonant modifiers are recognized with MR and BC regions. The vowel modifier is associated with TC region in majority of the syllables. Few exception cases are handled separately amounting to less than .05 percent of the total syllables. Script analysis is then used as a knowledge source to indicate which of these regions are essential i.e., contain significant information for that script. Other regions may contain information but that would be redundant information. This information also forms part of the knowledge source.

Four regions are found to be significant and other regions possess minute information or redundant information. For Devanagari script only four regions namely, middle-center(MC), top-center(TC), bottom-center(BC) and middle-right(MR) are sufficient and complete labeling will be made easy.

These rules are adopted in syllable segmentation process as presented in Figure 11.

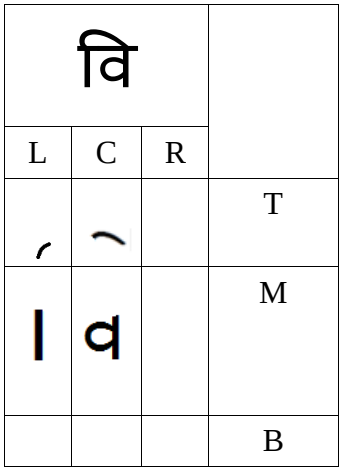


Figure 11 : Syllable segmented into four Regions

VI. CONCLUSION

In this paper, an improved character identification is proposed. Especially in complex scripts where syllable representation is associated with a cluster of code points, the proposed framework will help in adapting script related locale features. Syllable segmentation approach is adopted in the present work where the wholistic unit, Syllable, is segmented using topological features of the script and the syllable is further segmented into 9 regions while exploring the geometrical features. The proposed model considers script dependent rendering(syntactic) rules. It is possible to extend this approach to any script in the world – to be carried out as a future task. The proposed method is designed in such a way that, appropriate knowledge source can transform it into the respective Devanagari derived script OCR system. Extension of the present model towards Bilingual and Multilingual OCR is the major future scope.

ACKNOWLEDGMENT

I would like to record my sincere thanks to my research supervisor **Dr. L. PRATAP REDDY,** Director R&D Cell, Prof. in ECE Department, JNTUH College of Engineering . His vision, breath of knowledge, perseverance and patience have been the motivating factors behind this work .

REFERENCES

[1] Pritee gupta, Vandana Malik, Mallika Gandhi,"Implementation of Multilevel Threshold Method for Digital Images Used In Medical Image Processing", 2012,volume 2,Issue 2 .

[2] Satya Prasad Lanka, Akbar Hussain Darga, Kishore TVS , Pratap Reddy L "A Framework for Improving Recognition Efficiency of Characters Based on Grapheme Characterisation" , 2011 Int. J. Of Computer and Electronics Engineering , ISSN No. 0975-4202 Vol. 3, No. 2, July –December.

[3] Mudit Agarwal, Hvanfeng Ma and Davin Doreman , "Generalization of Hindi OCR Using Adaptive segmentation and font files ", 2009 Advances in Pattern Recognition ,springer – Verlog London limited.

[4] Richard G. Casey; Eric Lecolinet; "A survey of methods and strategies in character segmentation", Pattern Analysis and Machine Intelligence, IEEE Transactions, July 1996, volume 18, issue 7, pp 690-706.

[5] R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and Umapada Pal,"Offline Recognition of Devanagari Script: A Survey",2010 IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C:APPLICATIONS AND REVIEWS.

[6] U. Pal and B. B. Chaudhuri(1997), "Printed Devanagari script OCR system", Vivek, Vol. 10(1), pp. 12-24.

[7] U.Pal and B. U. Chaudhur -"A complete printed Bangla OCR system "- Pattern Recognition, 1998. Vol. 31 ,issue 5, pp-531-549.

[8] Azam Beg, F.Ahmed, Piers Campbell," Hybrid OCR Techniques for Cursive Script Languages - A Review and Applications", 2010 , Proc.Computational Intelligence, Communication Systems and Networks (CICSyN), Pg101 – 105

[9] L. Pratap Reddy, T. Ranga Babu, N. Venkata Rao, B. Raveendra Babu, " Touching Syllable Segmentation using Split Profile Algorithm", IJCSI, International Journal of Computer Science Issues, Vol. 7, Issue 3, No 9, May 2010 17 – 26.

[10] U. Pal and B. B. Chaudhur, "Printed Devanagari script OCR system", 1997 Vivek, Vol. 10(1), pp. 12-24.

[11] L.Pratap Reddy, Chandrasekhara Sastry.A.S, A.V.Srinivasa Rao, N.Venkata Rao,"Canonical Syllable Segmentation of Telugu Document Images", TENCON 2008,2008 IEEE Region 10 Conference Publication, Non-2008, pp 1-5.

[12] Chandrasekhara Sastry.A.S, Satya Prasad Lanka, L.Pratap Reddy, "Middle Zone Component Extraction and Recognition of Telugu Document image", Ninth International Conference on Document Analysis and Recognition, ICDAR 2007,Volume 2, Issue, 23-26 Sept. 2007 pp. 584–588.

[13] C. Neerugatti Vishwanath, Anil Kumar Gogi, P Prem Kishan, SK. Khamuruddeen, S.V.Devika " Classification Of Scripts Using Vertical Stroke Feature", S.V.DEVIKA/International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 2,Mar-Apr 2012, pp.1168-1175

AUTHORS

First Author – Sirisha Badhika, Research Scholar, Shri Jagdish Prasad Jhabarmal Tibrewala University, India, sirisha.badhika@gmail.com.
Lecturer, University of Technology, Kingston, Jamaica.

Correspondence Author – Sirisha Badhika, c_sirisha2k@yahoo.com, +91 9441053929.