Volume 7, Issue 2, February 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Segmentation of Tri-Lingual Documents

[1]Mahesha D M, [2]Bhavya D N, [3]Nandini H M
[1]Department of Studies in Computer Science, Karnataka State Open University, Mysore, India
[2]Department of Studies in Computer Science, Karnataka State Open University, Mysore, India
[3]Department of Studies in information technology, Karnataka State Open University, Mysore, India
DOI: 10.23956/ijarcsse/V6I12/0221

Abstract*— Physical layout analysis intends to study the arrangement of layouts or locations of the regions present in a document image before understanding it. Before extracting the text or information from a document image, page segmentation (layout analysis) techniques need to be applied to identify the exact layout (area) where the text or image resides. In Page Segmentation, Top-down methods are simple and efficient but fail in non Manhattan layouts. In contrast, Bottom-up approaches adapt non Manhattan layouts easily than the top down approaches, but heavily depend on the threshold, parameters and extensive computations for layout identification. On the other hand, Hybrid methods (Bruel [31], Bruel [32]) suits well for layout identification by eliminating the dependency on threshold and parameters. But this analyzes the white background of the image with small white rectangles and merges them to locate the content blocks. Merging of small white rectangles makes the identification process tedious since large number of small white rectangles gets involved in the image. In addition, this approach heavily relies on heuristics for merging operations, which affects the segmentation rate considerably. In all the above reported methods (Bottom up and Hybrid approaches), connected component analysis (requires more number of pixel visits) is required to identify black and white components from the image. Therefore, the above shortcomings motivated this research towards designing a White Space analysis technique which eliminates the usage of the connected component analysis (to identify white spaces), heuristics, threshold and prior knowledge. As a result, in this thesis, Rectangular White Space Analysis (RWSA) technique has been proposed to grab all the white spaces over the image in a single scan over the image with minimum pixel visits, and the white spaces are merged together without the assumptions of heuristics and threshold to segment the layouts. Moreover, two statistical properties have also been proposed in this thesis, to separate the text blocks and images from the identified layouts and this hybrid approach has been explained in the subsequent section.

Keywords*— Segmentation; Section finding; Section Merge; Feature Extraction; Indexing

I. INTRODUCTION

Anything which conveys information is known as a document. Generally, a document is a knowledge container. Most of the times we acquire knowledge from documents such as Newspapers, Textbooks, Scientific journals, Magazines, Technical reports, Office files, Postal letters, Bank cheques, Application forms etc. (Tang et al., [1]). To understand the huge information, an extensive amount of manual processing is required and such a manual processing is very much time consuming. To overcome this difficulty, it is essential to automate the manual process which needs efficient algorithms. This automation process is considered as document image processing (DIP). In general, the document image processing is divided into text processing and graphics processing. Text processing is further divided into character recognition and page layout analysis. Graphics processing is further divided into line processing and region processing as shown in Figure 1.

A. Stages in Document Image Processing

The document image processing involves three basic steps at conceptual levels, which are document image analysis, document image recognition and document image understanding. Within these three levels, there are several other interacting modules such as image acquisition, binarization, block segmentation, block classification, logical block grouping, character and word recognition, picture processing and analysis, graphic analysis, picture understanding, text understanding and graphics understanding. The interactions between these processes and data flow between levels are shown in Figure 2.

1) Document Image Analysis

Document image analysis is a process of recovering syntactic and semantic information from images of documents, prominently from scanned versions of paper documents. There are two distinct tasks in document image analysis. The first has a syntactical goal consisting of the identification of basic components of the document, the so-called document objects. The second has a semantic goal consisting of the identification of the role and meaning of the document objects in order to have an interpretation of the whole original document. The structural analysis, on the other hand involves usage of layout clues to identify headlines, locate different lines, etc. In general, image analysis involves

the extraction and use of attributes and structure relationships in the document in order to label its components within contextual rules dictated by the document class. Analysis of printed documents obviously involves skew angle estimation and correction which is a very challenging task.

2) Document Image Recognition

Document Image Recognition (DIR), a very useful technique in office automation and digital library applications, is to find the most similar template for any input document image in a prestored template document image. Nowadays a large amount of existing paper documents are transformed to digital document images through scanners and cameras. However, the next step is to analyze a document and segregate text blocks, graphic block, picture block, etc, so as to facilitate labeling of the blocks. This process of labeling the blocks is said to be document image recognition or identification.

3) Document Image Understanding

Document image understanding is a component which extracts the logical relationships between the respective blocks of a document. Logical document structure is a hierarchical representation of semantics of the given document. The same logical document structure is formatted in varieties of physical layouts by changing the variables such as number of pages and font sizes, spacing between paragraphs and between sections, number of columns, etc. In all these layouts, the semantics of the document remains unaltered. Logical structure analysis determines the document's semantic structure and provides data appropriate for information retrieval.

B. Script Recognition

The OCR technology for Indian documents is in emerging stage and most of these Indian OCR systems can read the documents written in only a single script. As per the Indian constitution, every state Government has to produce an official document containing a national language (Hindi), official language (English) and state language (or regional language). According to the three-language policy adopted by most of the Indian states, the documents produced in an Indian state Karnataka, are composed of texts in the regional language- Kannada, national language Hindi and the world wide commonly used language-English. In addition, majority of the documents found in most of the private and Government sectors of Indian states, are Trilingual type (a document having text in three languages). So, there is a growing demand to automatically process these Trilingual documents in every state in India, including Karnataka

The monolingual OCR systems will not process such multi-script documents without human involvement for delineating different script zones of multi-lingual pages before activating the script specific OCR engine. The need for such manual involvement can result in greater expense and crucially delays the overall image to text conversion. Thus, an automatic forwarding is required for the incoming document images to handover this to the particular OCR engine depending on the knowledge of the intrinsic scripts. In view of this, identification of script and/ or language is one of the elementary tasks for multi-script document processing. A script recognizer, therefore, simplifies the task of OCR by enhancing the accuracy of recognition and reducing the computational complexity.

Script Recognition approaches can be broadly classified into two categories, namely, local and global approaches. The local approaches(Pal and Chaudhury [2], Pal et al [3]) analyze a list of connected components (Line, word, char) in the document images, to identify the script(or class of script). In contrast, global approaches (Joshi [4]) employ an analysis of regions (block of text) comprising atleast two lines (or words)without finer segmentation. In general, global approaches work well based on texture measurement, but this relies heavily on a uniform block of text (Buschet al [5]), and extensive preprocessing (to make the text block uniform) is required to measure the texture. Even though local approaches rely on the accuracy of character segmentation or connected component analysis, it could work well on the documents irrespective of their quality or uniformity in the block of text.

In the literature, many works have been reported for script recognition at the document, line and word levels, using local approaches. In this context, researchers have made a number of attempts to discriminate the Han and Latin script (Spitz [6], Lu and Tan [7]) at the document level and

exploited many Indian scripts at line level and word level (Pal and Chaudhury [8], Pal and Chaudhury [9], Padma and Nagabhushan [10], Dhandra et al [11]). However, all the techniques reported in the literature are script dependent. Since this research is intended to develop an classification system for kannada document images, Script Recognition, to discriminate the kannada from English scripts in bilingual document images is becoming important. In this connection, few local approaches are reported in the literature, such as spatial spread analysis (Dhanya et al [12]), Aspect Ratio (Tan et al [13]), Structural features (Pal and Chaudhury [14]), and Water Reservoirs (Pal et al [15]). However, all the above mentioned techniques produce a low discrimination rate due to its incapability in exploration of thescripts. Global approaches (Pati et al [17], Pati and Ramakrishnan [18]). S Chaudhury et al., [19] has proposed a method for identification of Indian languages by combining Gabor filter based technique and direction distance histogram classifier considering Hindi, English, Malayalam, Bengali, Telugu and Urdu.

G D Joshi et al., [20] have presented a script identification technique for 10 Indian scripts using a set of features extracted from logGabor filters. Dhanya et al., [21] have used Linear Support Vector Machine (LSVM), K-Nearest Neighbour (K-NN) and Neural Network (NN) classifiers on Gabor-based and zoning features to classify Tamil and English scripts. Hiremath [22] have proposed a novel approach for script identification of South Indian scripts using wavelet based co-occurrence histogram features. S R Kunte and S Samuel [23] have suggested a neural approach in on-line script recognition for Telugu language employing wavelet features. Peeta et al., [24] have presented a technique using Gabor filters for script identification of Indian bilingual documents.

II. SEGMENTATION

A. System Method

A hybrid layout analysis framework along with a text image separation technique has been proposed in this thesis to identify simple and complex images and to separate the text and images from the layouts. The impact of the hybrid layout analysis has been motivated in this thesis to design a Rectangular White Space Analysis technique, since white spaces are a generic layout delimiter. In general, publishers use white spaces to separate blocks of text since they are constrained by common printing technology. The RWSA technique is parameter free and threshold free and adapts to various heterogeneous structures through layout gaps rather than connected component analysis and heuristics. Also, two Statistical properties such as the Black run length, and Transition rate have been proposed in this thesis to isolate the textual areas from images in the segmented blocks.

The architecture of the Hybrid layout analysis using RWSA is depicted in Figure 3.1. This kind of a layout analysis is necessary before understanding the text in all applications. When document images are given as input to this system, they undergo Noise removal and Binarization in the preprocessing phase.

RWSA technique consists of White space Section Finding, Section Merging, Cropping Extraneous section and Rectangular Formation phases to identify layouts. The input image to the system could be a color or a gray scale image obtained by scanning the newspapers. Documents containing text, graphics, figures, maps and tables are taken as input. Scanned input images passes through the following phases to identify the layouts.

1) Pre-processing

The preprocessing of the document images includes the process of Noise removal and Binarization. Noise could occur in the document images due to many sources such as aging, photocopying etc. and the application of filters reduces noises in these images. Here noise has been suppressed in the document image by using a median filter, since median filters smear the character image strokes. For this median filter, a 3*3 mask has been chosen and it is applied over the image, which replaces nine pixels by the intensity of the center pixel over this mask. As a result of pulling the median filter output to the gray level of the center pixel, the shapes of the character strokes can be preserved. Binarization has been applied after noise removal. Binarization is a technique by which the color and gray scale images are converted into binary images. The most common method is to select a proper threshold for the image and convert all the intensity values above the threshold into an intensity value representing as ‘white’ and below the threshold as ‘black’ value. All intensity values below a threshold are converted to one intensity level and intensities higher than this threshold are converted to the other chosen intensity.

2) Section Finding

After noise removal, RWSA technique has been applied over the image which contains Section Finding as the initial process. Since all the books and magazines use white spaces as a separator within and between the texts, the observation of small white spaces becomes mandatory to identify the text area. Therefore, in Section Finding, white spaces are used as delimiters and observed for analysis. Variable length white spaces exist inside the text in both the directions, apart from the white spaces surrounding the textual zones. Due to the existence of non-uniform, small white gaps in the image apart from the column separators, a careful analysis is required to observe and record the white spaces (There may be a possibility of a small white space separator inside and across the paragraphs). As a result, in this thesis, the width of the image has been divided into ‘n’ equal sections (The Total number of sections is defined as the ratio of the Width of the image to the Section length; the Section length is experimentally fixed as 5 pixels and this length suits all kind of images). Since connected component analysis has been eliminated, a single horizontal scan has been performed over the image to grab the white spaces. After an entire horizontal scan of an image, all the sections which appear as white spaces are reported and their positions with the corresponding row number (section numbers along with their row numbers) have been recorded as a result of this procedure.

3) Section Merging

It is hard to process various white space section numbers to identify the layout gaps if the merging procedure has been avoided. Once all the white space section numbers based on their row number have been indicated, the merging of adjacent sections in both the directions is required to form horizontal and vertical white space rectangles which are done through the Section Merging phase.

The Section Merging phase consists of two processes: Horizontal Section Merging and Vertical Section Merging. Initially, horizontal section merging accepts all the white space section numbers with their corresponding row numbers as the input and produces a series of within-line or row-wise white space clusters as output(i.e.), subsequent white space sections in each row gets merged together to produce a series of row-wise white space sections. Since all the white spaces (section-wise) are identified and merged properly, the chance of getting under-segmentation has been completely eliminated.

B. Rectangular Analysis

The rectangular analysis phase consists of Cropping and the Rectangular formation process. After the identification of horizontal and vertical white space rectangles, finding the areas which are uncovered by the white space rectangles could yield the layout. Deviated edges or edges which do not have intersections over them, in horizontal and vertical white space rectangles must be trimmed or cropped properly to obtain the areas which are uncovered by the white spaces. In order to crop the non intersecting portion of the edges, a cropping procedure has been designed and applied over the horizontal and vertical sides of the rectangles.

1) Cropping Procedure

The Cropping procedure acts over the white space rectangles in both the directions by accepting the horizontal edges of each Horizontal White Space Rectangle (HWSR) and the vertical edge of each Vertical White Space Rectangle (VWSR). In a horizontal orientation, this procedure attempts to identify the two boundary vertical edges which pass through the edges of each HWSR and crops the extraneous portion of the horizontal edges of each HWSR, which appears apart from the intersecting boundary of the vertical edges. If no two boundary vertical edges pass through the horizontal edge of the HWSR, then the total horizontal edge would be removed for further processing.

2) Rectangular Formation

Once the horizontal and vertical edges are cropped, the areas uncovered by the white spaces could be easily extracted through rectangular formation procedure. Recursively, this procedure takes two cropped horizontal edges of each HWSR and checks with each pair of the vertically cropped edges for the formation of a rectangle. If the integration of these pairs of horizontal and vertical edges coincides,

C. Text/ Image Analyzer

Once the content blocks have been identified, the next step attempts to separate the textual blocks from the images and pictures, since textual blocks are required for further processing. Once the homogeneous regions are obtained, each region gets passed into the text image analyzer to identify the text component.

Two statistical properties called as Black Run Length (BRL) and White Black Transition Count (WBTC), which spans in the horizontal direction of the image have been used here to identify the textual blocks. Black run length corresponds to the ratio of the total number of black pixels in a row to the total transition (black-white disposition) count in that row. The White Black Transition count corresponds to the ratio of the total number of transitions in a row to the total number of pixels in that row. It is concluded that if the mean black run length appears to be more, and the Mean white black transition count of all the rows appears to be lesser than the threshold, it is concluded as image and not as a text.

The Black Run and Transition count of each row in the region is computed as stated in equations (3.1) and (3.2). The Black run BR (black density), of the i^{th} row of the region is represented in equation (3.1) and Transition count TC of the i^{th} row of the region has been computed as stated in equation (3.2)

$$BR(f(x,y)) = \begin{cases} br = br+1 & if\ f(x_i, y)=0 \\ br & otherwise \end{cases} \quad (3.1)$$

$$TC(f(x_i, y)) = \begin{cases} tc\ tc\ 1 & if\ (f(x_i, y)\ 1)\&((f(x_{i\ 1}, y))\ 1 \| (f(x_{i\ 1}), y)\ 1) \\ Tc & otherwise \end{cases}$$

where x_i represents i^{th} row and $(x1\ x_i\ x2)$ (3.2)

where f(x,y) is the 2-D array of the pixel region with the coordinates ranging from $x_1,y_1 \ldots x_2,y_2$ and BR represents black run , TC represents the transition count, 1 represents the presence of a white pixel and 0 represents the presence of a black pixel.

The Black Run Length (BRL) of the i^{th} row has been computed as the ratio of Black runs in a row to the total transition count in that row as stated in equation (3.3) and White to Black Transition Count (WBTC) of the i^{th} row corresponds to the total number of transitions in that row to the total number of pixels in that row.

$$BRL_i(f(x_i, y)) = \frac{BR(f(x_i, y))}{TC(f(x_i, y))} \quad (3.3)$$

With this, the Average Black Run Length (ABRL) and Average White Black Transition count (AWBTC) of the content blocks has been computed using the BRL and WBTC values of every row (Mean of all BRL and WBTC). It is observed that the black run length appears to be more and the transition count appears low for the image regions. The ABRL produces a greater ratio (experimentally threshold has been fixed as greater than 0.4) for the image regions rather than the text regions. In contrast, the ABRL of the text regions appears to be lower than that of the image regions since the black run length is low while the transition count appears to be more. Apart from this, the AWBTC appears to be more for the text regions than for the images due to the large number of transitions in the text regions.

As a result, it is concluded that if the average black run length appears to be less (lesser than threshold-experimentally fixed and results are evaluated), and the average white black transition count appears to be more (AWBTC corresponds to 15-40% of total area for the text whereas it lies within 5-15% of the total area for the images), it is a text and vice versa for the images.

The results of the Text/Image separation system has been depicted in Figure 3.2, which show the image region in a different color. Details of the experiments conducted, data collected and the experimental evaluation of the RWSA system along with the text/image separation scheme is discussed in the following subsections.

III. EVALUATION OF SEGMENTATION METHODS

The performance of the proposed segmentation model is evaluated by the use of the measures such as Probabilistic Rand Index (Pantofaru and Hebert [29]), Variation of Information (Rubner et al., [30]), Global Consistency Error (Rubner et al., 2000) and Boundary Displacement Error (Schmid [31]).

A. Evaluation of Segmentation Approaches

Evaluation results vary significantly between different evaluators, because each evaluator may have distinct standards for measuring the quality of the segmentation.

Rand Index

Consider two images, say ground truth and segmented respectively: S1 and S2 of N points $X = \{x_1, x_2, x_3, ..., x_N\}$; that assigned labels $\{l_i\}$ and $\{l_i^{'}\}$ respectively to point x_i. The Rand Index can be computed as the ratio of the number of pairs of vertices having the compatible label relationship in S1 and S2. It can be defined as:

$$R(S_1, S_2) = \frac{1}{2^N} \sum_{\substack{i,j \\ i \neq j}} \left[I\left(l_i = l_j \wedge l_i^{'} = l_j^{'}\right) + I\left(l_i \neq l_j \wedge l_i^{'} \neq l_j^{'}\right) \right] \quad (1.1)$$

Where, I is the identity function, and the denominator is the number of possible unique pairs among N data points. This gives a measure of similarity ranging from 0 to 1.

Variation of Information

It measures the sum of information loss and information gain between the two clustering, and thus it roughly measures the extent to which one clustering can explain the other. For segmentations, it can be interpreted as the average conditional entropy of one segmentation given the other.

$$VI\left(S_{test}, S_K\right) = H\left(S_{test} \mid S_K\right) + H\left(S_K \mid S_{test}\right) \quad (1.2)$$

The first term in the above equation measures the amount of information about S_{test} that we lose, while the second term measures the amount of information about S_K that we have to gain, when going from segmentation S_{test} to ground truth S_K. Where, H (.|.) is the conditional entropy.

Global Consistency Error

Measures the extent to which the regions in one segmentation are subsets of the regions in second segmentation (i.e. the refinement). Let $R(S, p_i)$ be the set of pixels in segmentation *S* that contains pixel p_i , then the local refinement error is defined as:

$$E\left(S_1, S_2, p_i\right) = \frac{\left|R(S_1, p_i) \setminus R(S_2, p_i)\right|}{R(S_1, p_i)} \quad (1.3)$$

This error is not symmetric $\left(i.e., E\left(S_1, S_2, p_i\right) \neq E\left(S_2, S_1, p_i\right)\right)$ w.r.t. the compared segmentations, and takes the value of zero when S_1 is a refinement of S_2 at pixel p_i. Global Consistency Error is then defined as:

$$GCE\left(S_1, S_2\right) = \frac{1}{n} \min\left\{ \sum_i E\left(S_1, S_2, p_i\right), \sum_i E\left(S_2, S_1, p_i\right) \right\} \quad (1.4)$$

where, n is the number of pixels.

Boundary Displacement Error (BDE)

The BDE is a boundary based metric to evaluate the segmentation quality. It defines the error of one boundary pixel as the distance between the pixel and its closest pixel in the other boundary image. Let *B*1, *B2* represent respectively the boundaries of segmentation and Ground truth. The BDE can be computed using the minimum absolute difference from arbitrary point *x* in *B*1 to all the boundary points in *B*2. A near-zero mean and small standard deviation of BDEs computed for all the points in *B*1 indicate the quality of the image segmentation.

i.e., $BDE = \min(|x - y_i|)$, where x ∈ B1 and $y_i \in B2, i = 1, 2, .., n$, n is the number of boundary points in B2.

IV. EXPERIMENTATION

To evaluate the segmentation results produced by different algorithms we have compiled a database, containing 300 corridor images along with ground truth segmentations. The corridor images was taken in twenty different buildings exhibiting a wide variety of different visual characteristic. Segmentation method is evaluated by assessing its consistency with the ground truth segmentation given by the human expert.

The segmentation is evaluated by assessing its consistency with the ground truth segmentation. Any evaluation metric desired should take into account the following effects: Over-segmentation where region of the reference is represented by two or more regions in the examined segmentation. Under segmentation were two or more regions of the reference are represented by a single region in the examined segmentation. In accurate boundary localization the ground truth is usually produced by humans that segment at different granularities. And finally in different number of segments one needs to compare two segmentations when they have different numbers of segments.

Table 1 shows the parameter values of different segmentation methods. The PRI value should be higher for an image and VOI, GCE. BDE values must be lower for an image. Each parameter is described by ground truth and

proposed method. Each row is represented by average of each class totally about 100 images Form the table 1 the proposed method achieves values of PRI 0.9725, VI 2.23, GCE 2.14 and BDE 1.24, We can understand that proposed method achieves good results. From this evaluation, it is found that Region merging segmentation is well suited for the corridor images.

Table 1. Shows the segmentation results

Images no.	PRI		VI		GCE		BDE	
	Ground Truth	Proposed Method	Ground Truth	Proposed Method	Ground Truth	Proposed Method	Ground Truth	Proposed Method
1	0.9844	**0.9725**	0.9199	1.0986	2.4053	3.3784	0.2163	0.3480
2	0.9814	0.9668	0.8788	1.4768	3.2799	3.3368	0.2886	0.3087
3	0.9763	0.9766	0.9459	1.5755	**2.2390**	3.1061	0.2333	0.4429
4	0.9780	0.9699	0.8770	1.6447	2.4230	2.9160	0.2678	0.4030
5	0.9862	0.9702	0.8671	1.8866	3.1828	**2.5884**	0.3283	0.3421
6	0.9815	0.9665	0.8748	1.4235	2.4569	3.4137	0.2568	0.3388
7	0.9802	0.9652	0.8229	1.6247	2.9266	3.3850	0.2914	0.4066
8	0.9809	0.9694	0.8348	1.5332	2.7568	3.2784	0.2774	0.3917
9	0.9896	0.9633	0.9257	1.1464	2.7651	3.2690	0.1104	**0.2144**
10	0.9858	0.9669	0.9590	1.5422	3.2584	3.3537	0.1555	0.3495
11	0.9819	0.9628	0.9359	1.2741	2.6944	3.5492	0.2544	0.2803
12	0.9852	0.9631	**0.9575**	1.1895	3.2977	3.3552	0.1940	0.3639
13	0.9811	0.9615	0.8671	1.5554	2.7562	3.5547	0.3065	0.4117
14	0.9793	0.9625	0.9174	1.6896	2.5537	3.4230	0.3046	0.3925
15	0.9767	0.9693	0.9281	1.7558	2.6725	3.3752	0.2913	0.3946
16	0.9758	0.9628	0.9016	1.6596	2.4516	3.6604	0.2473	0.2882
17	0.9747	0.9631	0.9060	1.8337	2.6995	3.4166	0.2438	0.3454
18	0.9840	0.9615	0.9274	1.1838	2.7629	3.4586	0.2207	0.3583
19	0.9845	0.9625	0.8426	1.2308	2.3854	3.3339	0.2221	0.2488
20	0.9819	0.9693	0.8310	1.4975	2.7491	3.2971	0.2898	0.2541

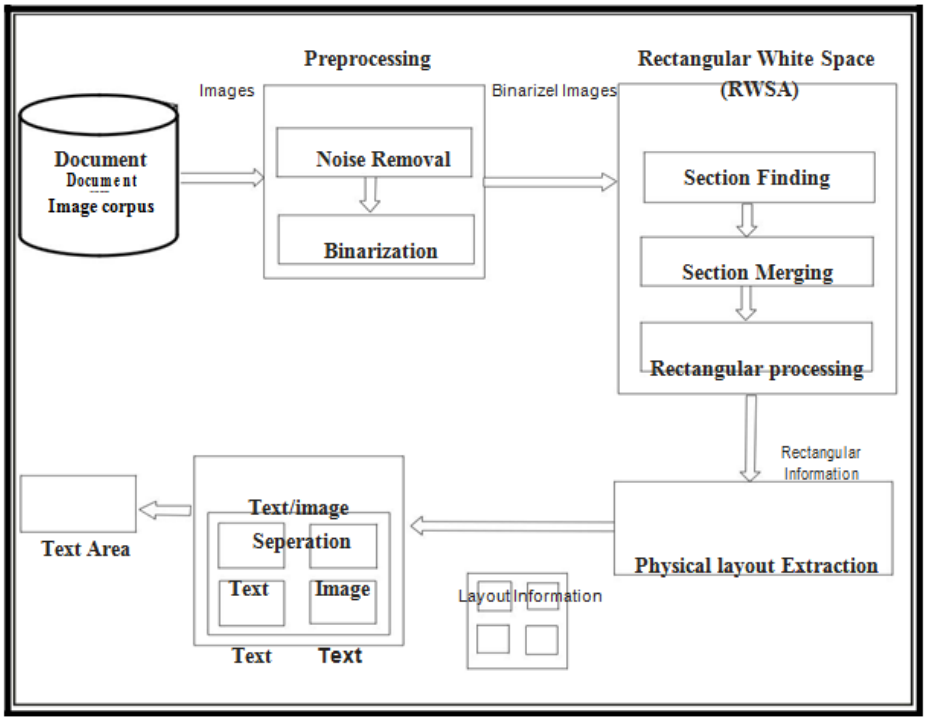

Figure 3 Hybrid Layout Analysis Architecture

REFERENCES

[1] Tang Y. Y., Lee S. H and Suen C. Y., 1996. Automatic document processing: a survey. Pattern recognition, Vol. 29, No.12, pp. 1931-1952.

[2] Pal U. and Chaudhuri B.B. (2001), 'Automatic identification of English, Chinese, Arabic, Devnagari and Bangla script line', Proceedings of the International Conference on Document Analysis and Recognition, pp. 790-794.

[3] Pal U., Sinha S. and Chaudhuri B.B. (2003), 'Word-Wise Script Identification From A Document Containing English,Devnagari And Telugu Text', Proceedings of the Document Analysis and Recognition, pp. 213-220.

[4] Joshi G., Saurabh G. and Jayanthi S. (2006), 'Script Identification from Indian Documents', Proceedings of the Seventh IAPR workshop on Document Analysis Systems, LNCS 3872, pp. 255-267.

[5] Busch A., Boles W.W. and Sridharan S. (2005), 'Texture for Script Identification', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.27, No.11, pp.1720-1732.

[6] Spitz A.L. (1997), 'Determination of script, language content of document images', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.19, No.3, pp. 235-245.

[7] Lu S. and Tan C.L. (2008), 'Script and Language Identification in Noisy and Degraded Document Images', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, No. 1, pp. 14-24.

[8] Pal U. and Chaudhuri B.B. (1997), 'Automatic separation of words in multi-lingual multiscript Indian documents', Proceedings of the International Conference on Document Analysis and Recognition, pp. 576-579.

[9] Pal U. and Chaudhuri B.B. (1999), 'Script Line Separation from Indian Multi-Script Documents,' Proceedings of the International Conference on Document Analysis and Recognition, pp.406- 409.

[10] Padma M.C. and Nagabhushan P. (2003), 'Identification and Separation of text words of Kannada, Hindi and English languages through discriminating features', Proceedings of the National Conference of Document Analysis and Recognition, pp. 252-260.

[11] Dhandra B.V., Mallikarjun H., Ravindra H. and Malemath.V.S. (2007), 'Word Level Script Identification in Bilingual Documents through Discriminating Features', Proceedings of International Conference on Signal processing, Communications and Networking, pp. 630-635.

[12] Dhanya D., Ramakrishnan A.G. and Peeta Basa P. (2002), 'Script Identification In Printed Bilingual Documents,' Sadhana, Vol. 27, Part-1, pp. 73-82.

[13] Tan C.L., Leong P.Y. and He S. (1999), 'Language Identification in Multilingual documents', Proceedings of the International Symposium on Intelligent Multimedia and Distance Education.

[14] Pal U. and Chaudhuri B.B. (1999), 'Script Line Separation from Indian Multi-Script Documents,' Proceedings of the International Conference on Document Analysis and Recognition, pp.406- 409.

[15] Pal U., Sinha S. and Chaudhuri B.B. (2003), 'Word-Wise Script Identification From A Document Containing English, Devnagari And Telugu Text', Proceedings of the Document Analysis and Recognition, pp. 213-220.

[16] Pati P.B., Sabari Raju S., Pati N. and Ramakrishnan A.G. (2004), 'Gabor filters for document analysis in Indian Bilingual Documents', Proceedings of the International Conference on Intelligent Sensing and Information Processing, pp.123-126

[17] Pati P.B. and Ramakrishnan A.G. (2006), 'HVS inspired system for script identification in Indian multi-script documents', Seventh IAPR Workshop on Document Analysis Systems, LNCS, Vol. 3872, pp. 380-389.

[18] Santanu Chaudhury, Gaurav Harit, Shekar Madnani, Shet R.B., (2000), Identification of scripts ofIndian languages by Combining trainable classifiers", Proc. of ICVGIP, India.

[19] Gopal Datt Joshi, Saurabh Garg, and Jayanthi Sivaswamy, (2006), Script Identification from Indian Documents, DAS 2006, LNCS 3872, 255–267.

[20] Dhanya D., Ramakrishnan A.G. and Pati P.B., (2002), Script identification in printed bilingualdocuments, Sadhana, vol. 27, 73-82.

[21] Hiremath P S and S Shivashankar, "Wavelet Based Co-occurrence Histogram Features for TextureClassification with an Application to Script Identification in a Document Image", PatternRecognition Letters 29, 2008, pp 1182-1189.

[22] Srinivas Rao Kunte R. and Sudhakar Samuel R.D., (2002), A Neural Approach in On-line ScriptRecognition for Telugu Language Employing Wavelet Features, National Workshop onComputer Vision, Graphics and Image Processing (WVGIP), 188-191.

[23] Peeta Basa Pati, S. Sabari Raju, Nishikanta Pati and A. G. Ramakrishnan, "Gabor filters forDocument analysis in Indian Bilingual Documents", 0-7803-8243-9/04/ IEEE, IClSlP, pp. 123-126, 2004.

[24] Newsam, S. D., and Kamath, C.: Retrieval using texture features in high resolution multi-spectral satellite imagery. In SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology VI(2004).

[25] Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons, International Journal of Computer Vision 43(1):29-44, (2001).

[26] Schmid, C.: Constructing models for content-based image retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pages 39-45(2001).

[27] Geusebroek, M., Smeulders, W. M., Weijer J.: Fast Anisotropic Gauss Filtering. IEEE Transactions on Image Processing, 12(8):938-943(2003).

[28] Varma, M., Zisserman A.: A statistical approach to texture classification from single images, International Journal of Computer Vision: Special Issue on Texture Analysis and Synthesis, 62(1--2):61—81(2005).

[29] Wasserman P D., Advanced Methods in Neural Computing, New York: Van Nostrand Reinhold (1993), pp. 155-61, and pp. 35-55(1993).

[30] Qasem, S. N., and Shamsuddin, S. M. :Generalization Improvement of Radial Basis Function Network Based on Multi- Objective Particle Swarm Optimization, Journal of Artificial Intelligence(2009).
[31] Lowe, David G. (1999). "Object recognition from local scale-invariant features". Proceedings of the International Conference on Computer Vision. pp. 1150–1157.
[32] Samet H., 1994. The Design and Analysis of Spatial Data Structures. Addison – Wesley.
[33] Chen J., H. Cao, R. Prasad, A. Bhardwaj and P. Natarajan, P., 2010. Gabor features for offline arabic handwriting recognition. Proceedings of the Ninth IAPR International Workshop on Document Analysis Systems, pp. 53 – 58.
[34] Chen X. and J. Zhang, 2012. Optimized discriminant locality preserving projection of gabor feature for biometric recognition. International Journal of Security and Its Applications, vol. 6, no. 2, pp. 321 – 328.

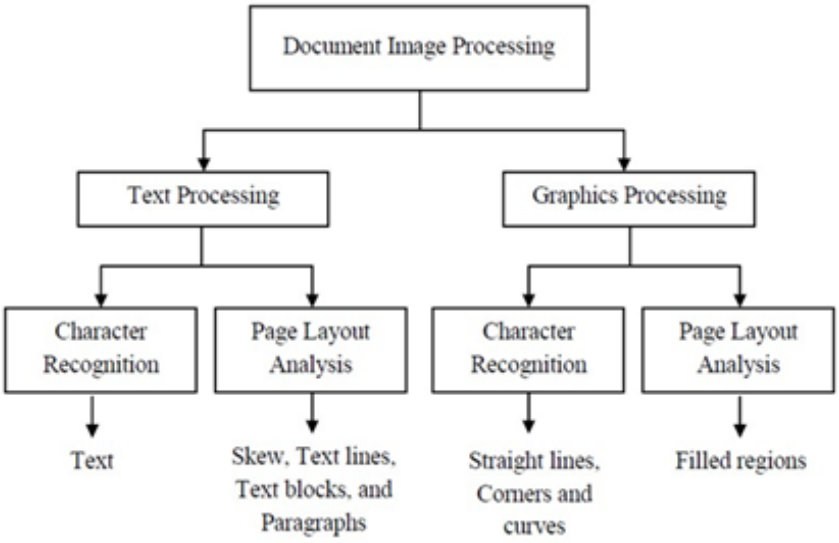

Figure 1. Hierarchy of document image processing with subcategories

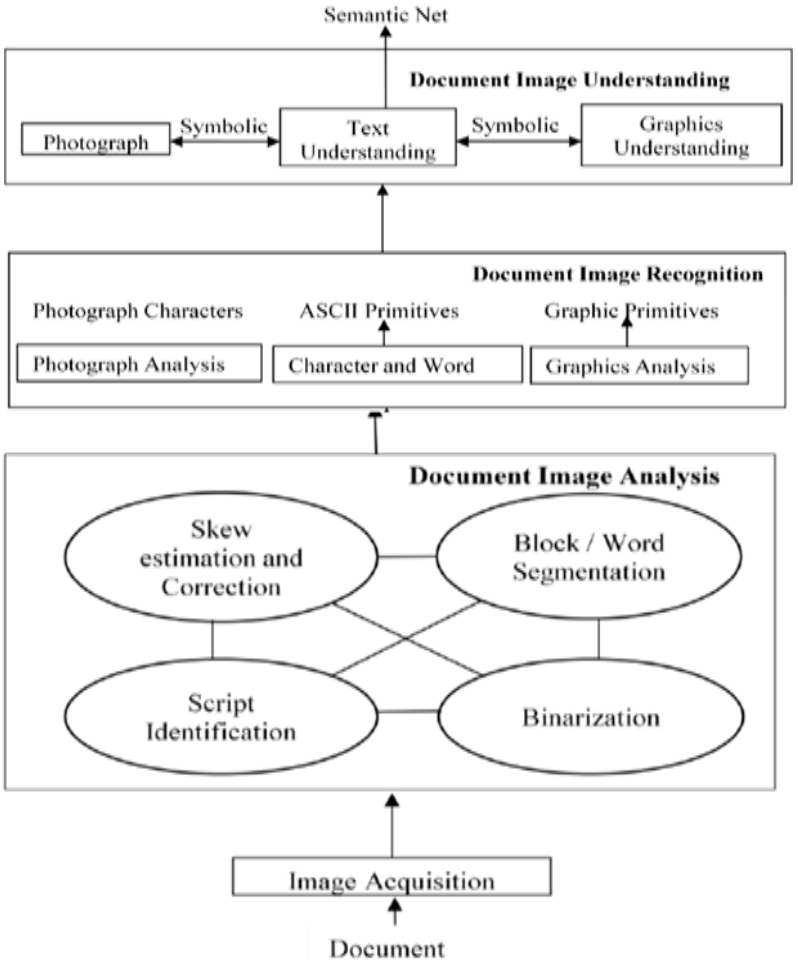

Figure 2. Steps involved in document image processing

ವೆಸ್ಟ್ ಇಂಡೀಸ್ ಕ್ರಿಕೆಟ್ ಮಂಡಳಿ ಜತೆ ಆಟಗಾರರ ತಿಕ್ಕಾಟ ಮುಂದುವರೆದಿರುವುದರಿಂದ ಟೆಸ್ಟ್ ಕ್ರಿಕೆಟ್ ಆಗಿ
ಸಂಪಾದನೆ ಮಾಡುವುದು ಸಾಕಾಗುತ್ತಿಲ್ಲ. ಕಡಿಮೆ ಸಂಭಾವನೆ ಪಡೆದು ಟೆಸ್ಟ್ ಆಡುವುದಕ್ಕಿಂತ ಟಿ20
ಅಂತಾರಾಷ್ಟ್ರೀಯ ಪಂದ್ಯ ಹಾಗೂ ಕರಿಬಿಯನ್ ಲೀಗ್ ಆಡುವುದು ಉತ್ತಮ ಎಂದು ಸ್ಯಾಮುಯೆಲ್ಸ್
ಹೇಳಿಕೊಂಡಿದ್ದಾರೆ.[12 ಸಾವಿರ ರನ್ ಕ್ಲಬ್ ಸೇರಿದ ವಿರಾಟ್ ಕೊಹ್ಲಿ]

Figure 3. Input Image with text

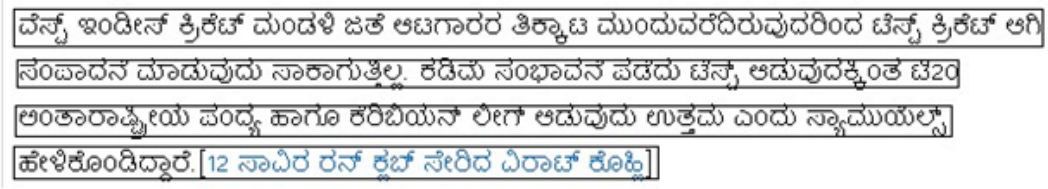

Figure 4. Input Image with elimination text ng

