Volume 7, Issue 1, January 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

K-D Indexing in Printed Trilingual Documen

[1]Mahesha D M, [2]Gopalan N P
[1] Dept. of Computer Science, Centre for Research and Development, PRIST University, Thanjavur, Tamilnadu, India
[2] Department of Computer Applications, NIT, Trichirappalli, Tamilnadu, India

DOI: 10.23956/ijarcsse/V6I12/0222

Abstract— In this work, we proposed indexing model for script identification. Rectangular White Space analysis algorithm is used to analyze and identify heterogeneous layouts of document images. To speed up the script identification, we focus on designing an indexing mechanism for tri-lingual scripts for optimizing the subsequent robust identification system. For representation, we extract features from Gabor responses and also using scale invariant feature transform. We considered a set of global features and index by Kd-tree. For experimentation, we have used our own database. Experimental results reveal that indexing prior to identification is faster than conventional identification method in terms of time for scripts.

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

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.

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. As a result, the width of the image has been divided into 'n' equal sections. 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 have been recorded as a result of this procedure. 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.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.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).

Once the horizontal and vertical edges are cropped, the areas uncovered by the white spaces could be easily extracted through rectangular formation procedure.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.

III. FEATURES EXTRACTION

A. Gabor Filter

In our proposed model, we have used Gabor features for representation of script recognition. Gabor features have been used for capturing local information in both spatial and frequency domain of the image, as opposed to other global techniques such as Fourier and Wavelet transform. One of the challenges, however, of such an approach is dealing with the tradeoff between the joint uncertainty in the space and frequency domains. Meaningful frequency based analysis cannot be localized without bound. An attractive mathematical property of Gabor functions is that they minimize the joint uncertainty in space and frequency. They achieve the optimal tradeoff between localizing the analysis in the spatial and frequency domains. The Gabor filters are orientation specific and this property allows us to analyze stroke directions in different scale and orientations. Also, the filtering output is robust to various noises since Gabor filters use information from all pixels in the kernel (Chen et al., [33]). Gabor filters are band pass filters which have both orientation selective and frequency selective properties and have optimal joint resolution in both spatial and frequency domains (Chen et al., [34]). By applying properly tuned Gabor filters to a signature image, local information about the signature can be extracted in detail. The signature analysis is accomplished by applying a bank of scale and orientation selective Gabor filters to an image (Newsam and Kamath, 2004).

B. Scale Invariant Feature Transform (SIFT) Descriptors

SIFT is one of the most widely used local approaches. It finds local structures that are present in different views of an image. It also provides a description of these structures reasonably invariant to image variations such as translation, rotation, scale, illumination and affine transformations. Moreover, several studies have shown that the SIFT descriptor performs better than others.

The first stage of the SIFT algorithm finds the coordinates of key points in a certain scale and assign an orientation to each one of them. The results of this guarantee invariance to image location, scale and rotation. Later, a descriptor is computed for each key point. This descriptor must be highly distinctive and partially robust to other variations such as illumination and 3D viewpoint. To create a descriptor, Lowe [30] proposed an array of 4×4 histograms of 8 bins. These histograms are calculated from the values of orientation and magnitude of the gradient in a region of 16×16 pixels around the point so that each histogram is formed from a sub-region of 4×4. The descriptor vector is a result of the concatenation of these histograms. Since there are $4 \times 4 = 16$ histograms of 8 bins each, the resulting vector is of size 128. This vector is normalized in order to achieve invariance to illumination changes. The distinctiveness of these descriptors allows us to use a simple algorithm to compare the collected set of feature vectors from one script to another in order to find correspondences between feature points in each flower.

Once the feature set is obtained, storing it in the database in an efficient manner so that only the potential candidates are selected for matching purpose is essential. This necessitates the use of backend tool called indexing mechanism for storing the data in some predefined manner such that only a few potential candidates are selected for matching purpose. Hence, multi-dimensional feature vectors obtained from scripts are indexed using Kd-tree in this proposed work. During signature identification, when a query feature vector of a signature is given, search is invoked using Kd-tree and top matches that lie within a predefined distance from the query are retrieved. These top matches are subsequently used for script identification.

C. Indexing

In the proposed indexing model, the obtained multi-dimensional feature vectors from tri-lingual scripts indexed separately using Kd-tree. The Kd-tree is one of the most prominent multidimensional space partitioning data structures for organizing points in a k-dimensional space and it is a useful data structure for searching a multidimensional key. The construction algorithm of Kd-tree is very similar to the planar case. At the root, we split the set of points into two subsets of roughly the same size by a hyperplane perpendicular to the x1-axis. In other words, at the root the point set is partitioned based on the first coordinate of the points. At the children of the root the partition is based on second coordinate and so on, until depth of d – 1 at which partition occurs based on last coordinate where d is the dimension of the feature space. After depth d, again, partitioning is based on first coordinate. The recursion stops only when one point is left, which is then stored at the leaf. Because a d-dimensional Kd-tree for a set of n points is a binary tree with n leaves, it uses O(n) storage with construction time being O(n logn). In addition to this, in Kd- tree there is no overlapping between nodes (Samet, [31]). Kd-tree is an appropriate data structure for biometric identification system particularly in the analysis of execution of range search algorithm. The proposed system decreases the search time as the Kd-tree is supporting the range search with a good pruning. When a query feature vector of dimension d is given, search is invoked using Kd-tree and top matches that lie within a specified distance from the query are retrieved. These top matches are subsequently used for script identification.

IV. EXPERIMENTATION

In this section, we present the results of the experiments conducted to demonstrate the effectiveness of the proposed indexing model. The proposed indexing approach is tested on our own tri-lingual database. For experimental purpose, we consider 400 tri-lingual scripts.

Table 1 shows the result of the classification when the SFS is 10. Here we have taken the three training percentage 30%, 50% and 70% and top 5 to top 30 samples for the experimentation and the corresponding results for logo retrieval has tabulated in the table. When the training percentage is 30%, 50% and 70%, we have got the best result in retrieving the top 25 samples and the result is marked using red, blue and green respectively rectangular box in the above table i.e. 65.2%, 76.2% and 78.3% respectively. Table 2 shows the result of the classification when the SFS is 20. When the training percentage is 30%, we have got the best result in retrieving the top 25 samples and the result is marked using red rectangular box in the above table i.e. 68.5%. In 50% we have got the best result in retrieving top 20 samples, and marked in blue rectangular box i.e. 75.5% and when the training percentage is 70% the best result in retrieving the top 30 samples i.e. 90.5%. After top 30 samples which the result is either reducing or same hence we stop the experimentation. Table 3 shows the result of the classification when the SFS is 30. When the training percentage is 30%, we have got the best result in retrieving the top 25 samples and the result is marked using red rectangular box in the above table i.e. 74.2%. In 50% we have got the best result in retrieving top 20 samples, and marked in blue rectangular box i.e. 80.3% and when the training percentage is 70% the best result in retrieving the top 25 samples i.e. 91%. After top 30 samples which the result is either reducing or same hence we stop the experimentation. Table 4 shows the result of the classification when the SFS is 40. When the training percentage is 30%, we have got the best result in retrieving the top 20 samples and the result is marked using red rectangular box in the above table i.e. 73.5%. In 50% we have got the best result in retrieving top 15 samples, and marked in blue rectangular box i.e. 80% and when the training percentage is 70% the best result in retrieving the top 25 samples i.e. 91%. After top 30 samples which the result is either reducing or same hence we stop the experimentation.

V. CONCLUSION

In this work, we proposed a Kd-tree based indexing mechanism for script identification. Gabor responses and SIFT based global features are used for representing tri-lingual scripts. It is observed that the indexing mechanism achieves relatively good classification accuracy when compared to any other available classifier. We have created our own database of documents. We conducted experimentation under varying database size and we studied its effect on identification accuracy.

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.

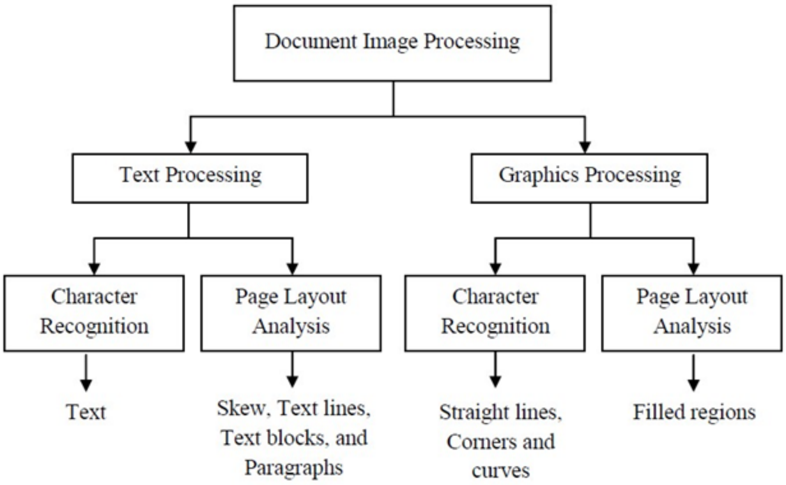

Figure 1. Hierarchy of document image processing with subcategories

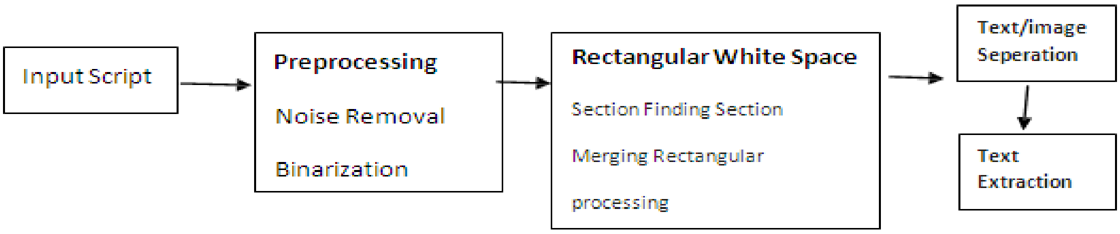

Figure 3. Steps involved in text extraction

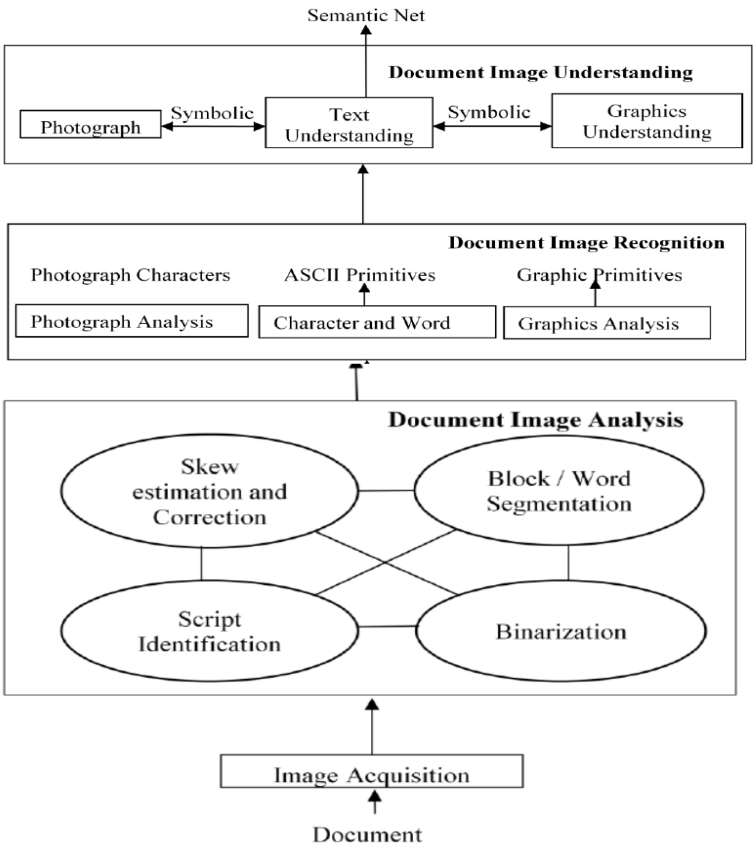

Figure 2. Steps involved in document image

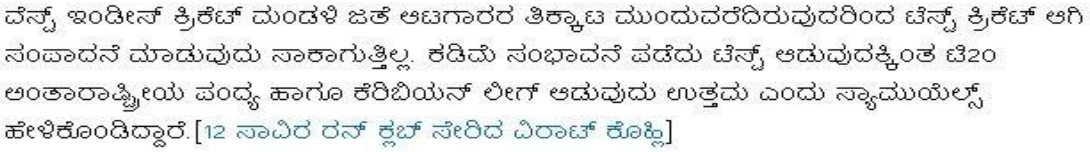
ವೆಸ್ಟ್ ಇಂಡೀಸ್ ಕ್ರಿಕೆಟ್ ಮಂಡಳಿ ಜತೆ ಆಟಗಾರರ ತಿಕ್ಕಾಟ ಮುಂದುವರೆದಿರುವುದರಿಂದ ಟೆಸ್ಟ್ ಕ್ರಿಕೆಟ್ ಆಗಿ
ಸಂಪಾದನೆ ಮಾಡುವುದು ಸಾಕಾಗುತ್ತಿಲ್ಲ. ಕಡಿಮೆ ಸಂಭಾವನೆ ಪಡೆದು ಟೆಸ್ಟ್ ಆಡುವುದಕ್ಕಿಂತ ಟಿ20
ಅಂತಾರಾಷ್ಟ್ರೀಯ ಪಂದ್ಯ ಹಾಗೂ ಕೆರಿಬಿಯನ್ ಲೀಗ್ ಆಡುವುದು ಉತ್ತಮ ಎಂದು ಸ್ಯಾಮುಯೆಲ್ಸ್
ಹೇಳಿಕೊಂಡಿದ್ದಾರೆ.[12 ಸಾವಿರ ರನ್ ಕ್ಲಬ್ ಸೇರಿದ ವಿರಾಟ್ ಕೊಹ್ಲಿ]

Figure 3. Input Image with text

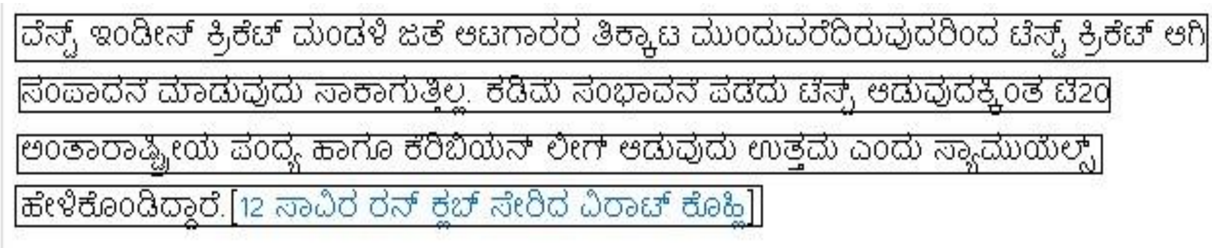
ವೆಸ್ಟ್ ಇಂಡೀಸ್ ಕ್ರಿಕೆಟ್ ಮಂಡಳಿ ಜತೆ ಆಟಗಾರರ ತಿಕ್ಕಾಟ ಮುಂದುವರೆದಿರುವುದರಿಂದ ಟೆಸ್ಟ್ ಕ್ರಿಕೆಟ್ ಆಗಿ
ಸಂಪಾದನೆ ಮಾಡುವುದು ಸಾಕಾಗುತ್ತಿಲ್ಲ. ಕಡಿಮೆ ಸಂಭಾವನೆ ಪಡೆದು ಟೆಸ್ಟ್ ಆಡುವುದಕ್ಕಿಂತ ಟಿ20
ಅಂತಾರಾಷ್ಟ್ರೀಯ ಪಂದ್ಯ ಹಾಗೂ ಕೆರಿಬಿಯನ್ ಲೀಗ್ ಆಡುವುದು ಉತ್ತಮ ಎಂದು ಸ್ಯಾಮುಯೆಲ್ಸ್
ಹೇಳಿಕೊಂಡಿದ್ದಾರೆ.[12 ಸಾವಿರ ರನ್ ಕ್ಲಬ್ ಸೇರಿದ ವಿರಾಟ್ ಕೊಹ್ಲಿ]

Figure 4. Input Image with elimination text ng

Table 1: shows the experimental result when SFS is 10.

Top samples	Training Percentage		
	30 (%)	50 (%)	70 (%)
5	60	70	75
10	62	71	76
15	64	73	78
20	64.7	76	78.1
25	65.2	76.2	78.3
30	65.1	76.2	78

Table 2: shows the experimental result when SFS is 20.

Top samples	Training Percentage		
	30 (%)	50 (%)	70 (%)
5	65	71	85
10	66	73	87
15	68	75	88
20	68.3	75.5	90
25	68.5	75.2	90.2
30	68	74	90.5

Table 3: shows the experimental result when SFS is 30.

Top samples	**Training Percentage**		
	30 (%)	**50 (%)**	**70 (%)**
5	70	78	90
10	71	79	90.2
15	73	80	90.5
20	74	80.3	90.9
25	74.2	80.2	91
30	74	80	91

Table 4: shows the experimental result when SFS is 40.

Top samples	**Training Percentage**		
	30 (%)	**50 (%)**	**70 (%)**
5	70.2	78.4	90.3
10	71.5	79.1	90.5
15	73.3	80.3	90.6
20	73.5	80.2	91
25	73.2	80	91
30	73.1	80.1	90

