Text Extraction from Document Images Using Edge Information and Translation Using EBMT

Arun Kumar. N
Department of computer Science and IT
Amrita School of arts and Sciences
Kochi, India

Abstract - **In this paper, I propose a simple edge based feature to perform text detection and extraction. The texts got sharp edges as compared to images. We identify these edges and separating it from images. Empirical rules are used to eliminate false positive based on geometrical properties. The extracted text information is transformed into a translation engine, which uses Example Based Machine Translation (EBMT) Technique. In this system, the extracted English sentence is matched with the contents in the database. Then use an automatically generated rule base together with a Morphological Analyzer for translating it into corresponding Malayalam sentence. The system learns new words from sentences and creates a word corpus, and thereby increasing the probability of translation. This method is applied to a large group of document images and obtains promising results.**

Keywords - **Extraction, MGD, MT, Translation.**

I. Introduction

In this method pixels are segmented into text or non text on the basis of text information computed in the neighborhood of each pixel[1][2][3].The gradient magnitude have high values in edges of the characters , even when the text is embedded in images. In this the algorithm traces the feature points in different entities and groups those edge points of textual regions. It uses line approximation and layout categorization. Using this technique it can successively retrieve directional based text blocks. Finally, feature based connected component merging was introduced to gather homogeneous textual regions together within the scope of its bounding rectangles. Text has the following characteristics.

1. Text posses certain frequency and orientation information.
2. Text shows spatial cohesion-characters of the same text strings are of similar height, orientation, and spacing.

Therefore the most intuitive characteristics of text are its regularity. Printed text consists of characters with approximately the same size and line thickness, are located at a regular distance from each other[4][5]. Such regularities can also be observed from edges being detected on textual boundaries. we may easily find the elongation tendency when take run-length smearing operation on textual areas. The regular alignment of characters into lines and columns let it be classified from other functional regions distinctly. As a result, our approach tried to find critical characteristics of text and then make use of its edge features to segment the image into text or non-text regions as best as possible, accordingly higher accuracy can be achieved when OCR system do recognition focusing on identified regions. In short, our method includes several major steps such as Preprocessing for noise reduction; Edge detection to form directional edge plane; Edge merging and bounding line approximation; Region classification by spatial grouping.

Documents in which text is embedded in complex colored backgrounds are increasingly common today, for example, in magazines, advertisements and web pages. Robust detection of text from these documents is a challenging problem [6][7]. Text extraction has a vast number of applications:

- Text searches in Images - Currently, Image searches deliver inaccurate results as they do not search the image content. Text extraction would enable better searching by extracting the content of an image.• Content based Indexing - For the purpose of archiving and indexing documents, the content of the document is required in the digital format. Knowledge about the text content of documents can help in the building of an intelligent system which archives and indexes the printed documents.
- Reading foreign language text - One of the common problems faced by a person in foreign land is that of communication, understanding road signs, signboards etc. The proposed method, aims to alleviate such problems by reading the text information from the image scenes which are captured by a camera.[8]
- Archiving documents - Archives of paper documents in offices or other printed material like magazines and newspapers can be electronically converted for more efficient storage and instant delivery to home or office computers. In this paper, we will demonstrate that simple texture measures based on edge information provide very useful information for text detection from complex document images[9][10]. Detected texts are given for Malayalam translation. Machine Translation often referred to as MT, translates sentences from one language to another [11]. The Example Based Machine Translation (EBMT) [12][13][14] is characterized by the use of bilingual corpus . The basic units of EBMT are sequences of words and the basic techniques are the matching of strings against strings in the Database. There are several EBMT techniques available and we

chose Morphological Analysis [15] for implementation purpose.

Morphological Analysis was developed by Fritz Zwicky [16]. It is a method for identifying patterns of word formation within and across the language.

As we compare the context of Simple English and Malayalam sentences, English Sentences (ES) consists of a Noun Phrase (NP) and a Verb Phrase (VP) [17].One way of representing it is as follows,

ES->NP+VP (1)

Each of these constituents can then in turn be described according to its own internal constituent's structure as follows:

NP->(DET)+N (2)

NP-> (DET)+ADJ+N (3)

VP->Vt+NP (4)

Where DET(Determiner)(eg: the, my etc) is an optional constituent in the structure of the Noun Phrase, which may be a noun by itself or a Determiner and a Noun.

Malayalam Language has got very vast and extensive set of Grammar Rules. It consists of 53 letters including 20 vowels (long and short) and the rest are consonants. Malayalam belongs to the family of Dravidian Languages. The Malayalam language and its Grammar System are closely related to the Dravidian Languages Sanskrit and Tamil.

II. Proposed Methodology

This text detection method has three steps: detection and localization, boundary refinement, false positive elimination.

A. Text detection and localization

Text regions typically have a large number of discontinuities like transition between text and background. Although, text regions show high contrast values. It is because they produce high peaks in horizontal projection. The input image is converted to grey scale and filtered by a 3x3 laplacian mask to detect the discountinuity in four directions: horizontal,vertical,Up-left and Up-right. Since we are interested in edges, it is natural to detect them in gray-scale images. By forming a weighted sum of R, G, B components help to convert a color image into gray-scale. The laplacian filtered image consists of both positive and negative values. The transition between these values corresponds to the transition between text and background. Inorder to capture the relatonship between positive and negative values maximum gradient difference(MGD) is used. if f is the laplacian filtered image, then the MGD value at pixel(i,j) is computed as follows,

MGD (i, j) = max(f(i,j-t))-min(f(i, j-t))

Where, t = – N/2[-1, (N-1)/2]

The MGD map is obtained by moving the window over the image. Text regions typically have larger MGD values than non-text regions. At the end each connected component in the text clusture is a candidate text region.

B. Boundary Refinement

The binary sobel edge map SM of the input image is computed. The horizontal and vertical projection profile is defined as follows,

$$HP(i) = \sum_{j} SM(i, j)$$

$$VP(j) = \sum_{I=i1}^{i2} SM(i, j)$$

If HP(i) is greater than a certain threshold, row i is part of a text line and if VP(j) is greater than a certain threshold, column j is a part of a text line; otherwise it is part of the gap between different words. Finally, different words on the same text line are merged if they are close to each other.

C. False positive elimination

We eliminate false positives based on geometrical properties. Let W-width,H-Height,AR-Aspect Ratio,A-Area, and EA-Edge Area of the text block(B).

$$AR = W / H$$

$$A = W \times H$$

$$EA = \sum SM(i, j)$$

$$(i, j) eB$$

If AR< T1 or EA/A<T2, the candidate text block is considered as false positive; otherwise it is accepted as a text block. The first rule checks whether the aspect ratio is below a certain threshold. The second rule assumes that a text block has a high edge density due to the transition between text and background. The extracted texts are given for Malayalam translation. In this Translation System, first a database with two fields is created: one field for the English sentences and another for its corresponding Malayalam sentences. Then, design an interface for inputting entries into the database. As we supply sample sentences in to the database, the database automatically expand by comparing similar sentences [18]. An Example of this is given below,

This is a tree. (Ithu oru maram aakunnu) and That is a book. (Athu oru pusthakam aakunnu) Four alternative translations are formed from the above two simple sentences. This is a tree ,That is a book, That is a tree(Athu oru maram aakunnu) and This is a book(Ithu oru pusthakam aakunnu). When new sentences are added, the system searches the database for similar sentence, learns automatically and expands the database.

If another sentence is added like That is a tree(Athu oru maram aakunnu), then from these three sentences, This is a tree . (Ithu oru maram aakunnu) That is a book.(Athu oru pusthakam aakunnu) and That is a tree(Athu oru maram aakunnu) the System learns that the meaning of "That" as "Athu" , "is a tree" as "oru maram aakunnu", "tree" as "maram", "book" as "pusthakam" and "is a book" as "oru pusthakam aakunnu". i.e., the machine learns these samples and insert into a new corpus, and eliminating redundancy.

Table 1: English to Malayalam Syntactic Transfer Rule

English Sentence(ES)	Malyalam Sentence(MS)
ES->NP+PRIN+NP+PP	MS->NP+PP+NP+PRIN
ES->NP+AP+NP+PP	MS->NP+PP+NP+AP
ES->NP+PRIN+OBJ1+OBJ2+PP	MS->NP+OBJ1+PP+OBJ2+PRIN
ES->NP+AP+OBJ1+OBJ2+PP	MS->NP+OBJ1+PP+OBJ2+AP

The system uses the syntactic rule shown in Table 1 for translating English sentences into the corresponding Malayalam sentences. When we get an extracted English Sentence to translate, it searches the database for a direct match. If there is a direct match, the corresponding Malayalam Sentence is retrieved from the database. If there is no match, then English parse tree [19] is generated corresponding to the input sentence. After that the corresponding Malayalam Parse Tree is generated and arranged based on predefined syntactic rules.

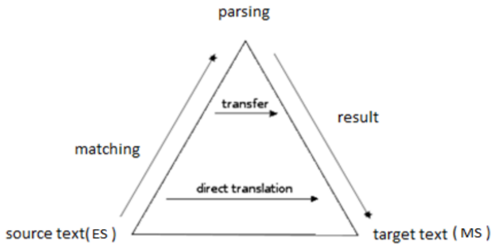

Fig.1. Representation of the Translation System as a Pyramid

The comparative depth of translation is shown in Fig. 1. In this Figure, the bottom level searches the sentence for an exact match (direct matching). Otherwise, Parsing of input sentences is performed and corresponding Malayalam parse tree is generated in the next level. Then Morphological Analysis is applied to obtain the corresponding Malayalam sentence.

D. Description of algorithm steps

One most important step of the algorithm is to find approximate locations of text lines in a gray-scale image. The main idea of this algorithm can be explained by considering a printed page with Manhattan layout. If we compute the spatial variance along each horizontal line over the whole image, we see that regions with high variance correspond to text lines, and regions with low variance correspond to background or other textures. Moreover, through run-length smearing operation we can find that edges from textual regions may elongate to form the approximate encircling rectangles. Text lines can then be found by extracting the rows between two parallel edges of the spatial variance – in specific distance of average character height. This heuristic can be applied to a more complex image, assuming that the spatial variance in the background is lower than in the text. Since we need to locate both the row coordinates of a text line and the column coordinates of its beginning and end, the spatial variance must be computed for each pixel over a local neighborhood in the horizontal direction. This results in an image of horizontal spatial variances with the same size as the input image. From this image we need to find significant horizontal edges and then pair the edges with opposite directions into lower and upper boundaries of a text line.

III. Experimental Results

We evaluate a huge set of test pictures, scanned from newspapers and magazines. There were pictures taken to test for usual text, that is white paper and black text. Then there were pictures taken to test with different colored background as well as vertical text and obtain promising results.

IV. Conclusion

The texts got sharp edges as compared to images. Identify these edges and separating it from images. Empirical rules are used to eliminate false positive based on geometrical properties. The major disadvantage of using this method is when the gradient of intensities of text and image are quite similar. Finding a generalized value which can work on every kind of image also needs some working. The extracted text information is transformed into a translation engine, which uses Example Based Machine Translation (EBMT) Technique. This translation system translates extracted English sentences into corresponding Malayalam sentences. The major factor affecting the accuracy was the samples produced. As we supply more samples, we got more accuracy. The system learns new words from sentences and creates a dictionary and thereby increasing the probability of translation. This method is applied to a large group of document images and obtains promising results.

References

[1] L. O. Gorman and R. Kasturi, *Document Image Analysis*. Los Alamitos, California, USA: IEEE Computer Society Press, 1995.

[2] D. Wang and S. N. Srihari, “Classification of newspaper image blocks using texture analysis,” *Computer Vision Graphics, and Image Processing*, vol. 47, pp. 327–352, 1989.

[3] T. Pavlidis and J. Zhou, “Page segmentation and classification,” *Computer Vision Graphics, and Image Processing*, vol. 56, no. 6, pp. 484– 496, 1992.

[4] [4] Q. Yuan and C. L. Tan, “Text extraction from gray scale document images using edge information.”

[5] A. K. Jain and S. Bhattacharjee, “Text segmentation using gabor filters for automatic document processing.”

[6] [6] D. F. Dunn and N. E. Mathew, “Extracting colour halftones from printed documents using texture analysis,” *Pattern Recognition*, vol. 33, no. 3, pp. 445–463, 2000.

[7] M. I. C. Murguiu, “Document segmentation using texture variance and low resolution images,” in *Proceedings of IEEE Southwest Synipoium on Image Analysis and Interpretation*, Tucson, Arizona, USA, 1998, pp. 164–167.

[8] L. Clique, L. Lombardi, and G. Mazini, “A multirestoration approach for page segmentation,” *Pattern Recognition Letters*, vol. 19, no. 2, pp. 217–225, 1998.

[9] K. Etemad, D. S. Doermann, and R. Chellappa, "Multiscale segmentation of unstructured document pages using soft decision integration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 1, pp. 92–96, 1998.

[10] A. K. Jain and Y. Zhong, "Page segmentation using texture analysis," *Pattern Recognition*, vol. 23, no. 2, pp. 743–770, 1996.

[11] John Hutchins . W and Harold L. Somers :" An Introduction of Machine Translation", Academic Press, London, UK,1992, ISBN 0-12-362830-X, 362 pp.

[12] Turcato D, Popowich F(2003)."What is Example-Based Machine Translation?". In Carl and Way(2003), pp 59-81

[13] Sumita E(2003) "An example-based machine translation system using DP matching between word sequences". In:Carl and Way(2003) pp 189-209

[14] Brown.Ralf.D.(1996)." Example-Based Machine Translation in the Pangloss System". In proceedings of the 16thInternational Conference on Computational Linguistic (pp. 169-174)

[15] Aaron .B.Philips,Violetta Cavalli-Sforza,Ralf D.Brown," Improving Example Based Machine Translation Through Morphological Generalization and Adaptation". Machine Translation Summit XI, Copenhagen, Denmark, September 2007.

[16] Zwicky, F. (1969). "Discovery, Invention, Research - Through the Morphological Approach". Toronto: The Macmillian Company.

[17] Shah Asaduzzaman, "A comprehensive study on MT towards development of Bangla-EnglishTranslation system". Thesis at CSE, Bangladesh University of Engineering and Technology, September 1999.way of machine translation from English to Bengali" .1999

[18] M.Nagao "A framework of a mechanical translation between Japanese and English by analogy principle". In proceedings of the international NATO symposium on Artificial and human Intelligence,pages173-180,1984.

[19] Antony P.J, Santhanu P. Mohan, Soman K.P., "SVM Based Part of Speech Tagger for Malayalam," itc, pp.339-341, 2010 International Conference on Recent Trends in Information, Telecommunication and Computing, 2010.