International Journal of Research in Engineering, Science and Management
Volume-2, Issue-6, June-2019
www.ijresm.com | ISSN (Online): 2581-5792

Text Line Segmentation of Handwritten Documents

Abhishek Kotrayya Hiremath[1], Rashmi Athanikar[2]
[1]*Student, Department of Computer Science and Engineering, SDM College of Engineering, Dharwad, India*
[2]*Assistant Professor, Dept. of Computer Science and Engineering, SDM College of Engineering, Dharwad, India*

Abstract*: Segmentation is an essential work of any Visual Text identification system. It is the picture of the words document within lines, texts and characters. The veracity of Optical Character Recognition scheme principally depends on the distribution theorem being provided. Distribution of typewritten word of some more Indian languages like Kannada, Telugu, Assamese is harder when related within classic languages located being of its fundamental intricacy and more character phase. It consists of vowels, consonants and compound aspects. Few of the text line may overlay composed. Despite assorted favorable tasks in OCR all accomplished the world, accumulation of OCR tools in Indian languages is still being as growing process. Text line segmentation plays a significant role amidst incorrectly distributed text lines are not likely to be understood exactly. In this paper, a distribution method for segmenting typewritten Kannada script within lines, words and characters using external happenings and projection profiles is considered. The technique was done on totally abandoned Kannada literature, which pays much demand and hardness is due to the complex problems involved in the literature. Making use of the morphology made separating text lines efficiently by an average separation rate of 94.5%. Because of the differing inter and intra text spaces an average distribution rate of 82.35% and 73.08% for words and characters respectively is achieved.

***Keywords*: Optical Character Recognition, Segmentation, Text Line Segmentation, Visual Identification System.**

1. Introduction

Image Processing: Picture conversion is a method to convert an image into digital form and perform some operations on it, in order to get an upgrade picture or to select some effective data from it. It is a type of signal disbursement in which information is image, like video frame or snapshot and turnout may be image or attributes related with that picture. Usually picture conversion system includes treating pictures as two dimensional signals while although previously stated signal processing techniques to them. It is among fast spreading automation today, with its applications in various purposes of a business. Picture conversion forms core research area within engineering and computer science methods too.

Picture conversion follows three different steps

- Moving an Image from the visual scanner or by digital design.
- Studying and shaping the image which includes data Reduction, image enrichment and stippling figures which are not visible to human eyes.
- Output is the last stage in which result can be revised picture or description that is based on image study.

A. Purposes of image processing:

It is classified into 5 types

- *Visualization* - Monitor the objects that aren't visible to your eyes.
- *Image sharpening and recovery* - To create a better image
- *Image renewal* - Seek for the image of importance.
- *Evaluation of pattern* – Measures various objects in an image.
- *Image acceptance* – Separate the objects in an image.

B. Types of image processing:

The objective of document image study is to notice the text and visual factors in images of documents, and to extract the expected knowledge as a human would. Two categories of document image study can be defined (see Figure1.1). Wording processing dealt with the word parts of a document picture. Some tasks here are: Conclusive the skew (any tilt at which the document may have been scanned into the computer, finding columns, paragraphs, text lines, and words, and finally rectifying the text (and possibly its characteristics such as size) by visual character identification. Visual processing deals with the non-textual line and symbol components that makeup line diagrams, planning straight lines between text sections, company logos etc. Pictures are a third major component of documents, but except for identifying their location on a page, further study of these is usually the task of other image processing and machine vision techniques. After uses of these text and graphics study methods, the several megabytes of beginning data are extracted to return a much more brief correct definition of the document.

Consider three precise instances of the requirement for document study explained here.

1. Typical documents in today's office are computer-generated, but even so, necessarily by different computers and software such that even their electronic layout are incompatible. Some include both planned

text and tables as well as handwritten entries. There are various volume, from a business card to a greater engineering art. Document study systems notice types of documents, uphold the evocation of their functional parts, and translate from one computer made layout to another.

2. Computerized mail-sorting machines to perform sorting and address perception have been used for several decades, but there is the need to process more mail, more quickly and-more correctly.
3. In an usual library, loss of material, disorganize, limited numbers of each copy and even degeneration of materials are common problems, and may be improved by document study. All these instances serve as purposes to keep it ready for the hidden explications of document image study. Document study systems will become increasingly more evident in the form of everyday document systems. For instance, OCR systems will be more widely used to store, search, and except from paper-based documents. Page-layout study methods will recognize a particular form, or page format and allow its replication.

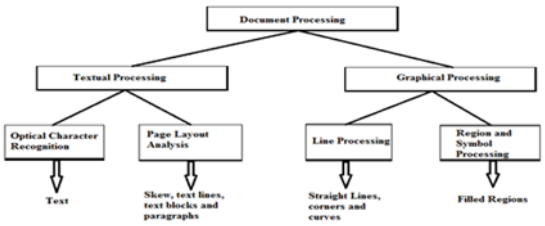

Fig. 1. Hierarchy of document processing

Fig. 2. Sequence of steps for document analysis

Figure 2, instances a common array of steps in document image study. After data, taking the image endures pixel-level processing and feature analysis and then text and visuals are treated separately for the perceptions of each. I portray these tracks briefly in the following sections; the reader is referred to the book, Document image study, for details (O'Gorman & Kasturi 1997). I conclude this paper by considering the demands in studying multilingual documents which is especially necessary in the substance of Indian language document study.

C. Optical character recognition

The computerization techniques involves the document image analysis(DIA) that interests with the automatic apprehension of document into text, graphics, drawings etc OCR refers to a process of achieving a character information by visual means, like scanning, for identifying in following phases by which a imprinted or handwritten text can be indoctrinated to a form which a computer can discover and operate. Optical Character Recognition, usually abstracted to OCR, is the mechanical or electronic paraphrasing of images of handwritten, copied or engraved text into machine- text. The pictures are usually picked up by a scanner. However, over the text, i would be implying to imprinted text by OCR. Data Entry through OCR has rapid speed, more definitiveness, and mostly more quantified than keystroke data entry. The relevance of the intricacy rises as i move beyond text line segmentation process from imprinted text images nearing a handwritten text images.

D. OCR methods

The OCR techniques consists of a number of phases listed below.

- Binarization
- Skew discovery & correction

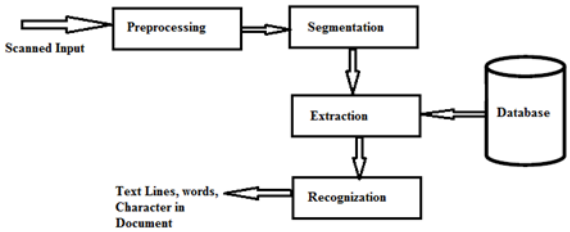

Fig. 3. Process of OCR

System first scan the image of Kannada script, using scanner then to pre-processing step, binarization, is the technique of changing a gray scale image (0.255 picture element charges), into binary (0 & 1 element charges) by beginning. Binary document needs less space to store, this method removes the majority of the noise. Noise introduced during scanning or due to poor quality of a page may also contain blur image has to be cleared before further processing. Thinning minimizes the size of the character by using designing the image by different technique. Skew detection and correction, aggregate scalar products of windows of text blocks with the Gabor permeates at other positions are calculated. Maximum additive scalar products give the skew angle.

2. Proposed methods

1) Pre-processing & data augmentation

Before instructing our models with the dataset, we have tested various pre-processing and data enhancement methods on dataset in order to make our data more adaptable with the models and to make the dataset more sturdy to real life situations.

2) Padding images

As suggested above, the dataset contains the images of individual words only. Moreover, the images are of various sizes because various words are of various lengths and heights. To evaluate the image of the word 'error' has a lesser width than the image of the word 'congratulations' because of the length of the words. Similarly, the image heights varied between images due to the heights of their characters. For instance, the image of the word car has a lesser height than the image of the buy since the characters of the word 'buy' increase above and below with 'b' and 'y'. Our architectures, however, affected the input images to be of the same size just like any other Convolution neural network architecture. This is necessary as the weights of the layers are managed according to the first input image, and the model would not process as well if weights were not steady, or replaced structures, for various inputs. Thus, we determined to make all the images of the same structure.

3) Rotating images

Even though my dataset contains images of each word separately, some words between these images were lightly sloped. This was because the colleagues of the dataset were requested to write on barren paper with no lines and some of the words were written in a more sloped pattern. This fortuity happens very rarely in real life whether or not the page has lines, thus i determined to make my data more vigorous to this argument by rotating an image almost the right by a very small angle with irregular plausibility and adding that image to my intent. This data enhancement method helped me make my model more sturdy to some minor yet so common details that might come up in our test set.

3. Experimental results and comparative study

In this section, I have represented the experimental results of both the proposed methods and they are compared with existing methods.

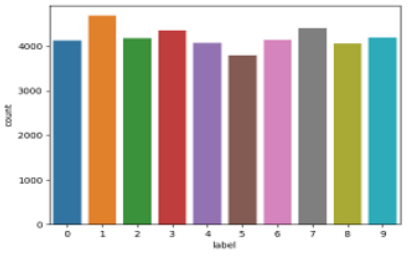

Fig. 4. Graph representing datasets for 0 to 9 digits with minimum count 5000.

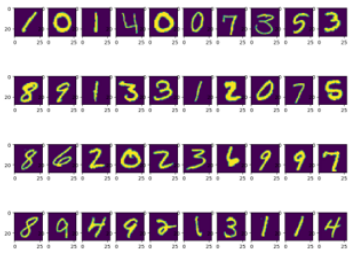

Fig. 5. Represents the character segmentation of digits

The following diagrams indicate the Confusion matrix and the graph showing the accuracy using Deep Learning and Convolution Neural Networks.

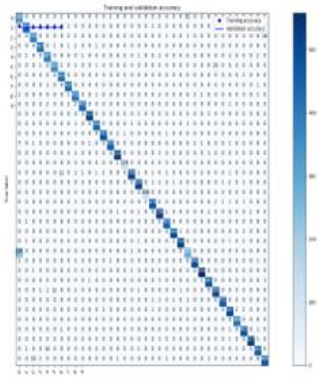

Fig. 6. Confusion matrix for the digits 0-9 and the Upper case Alphabets A-Z

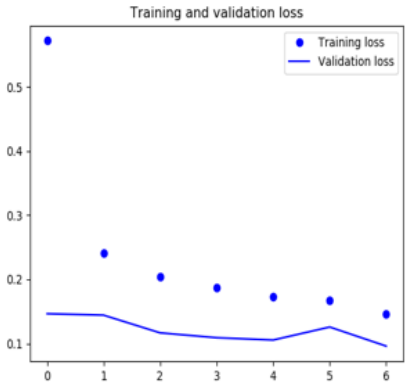

Fig. 7. Graph representing the Accuracy for the Text line segmentation of Handwritten Documents

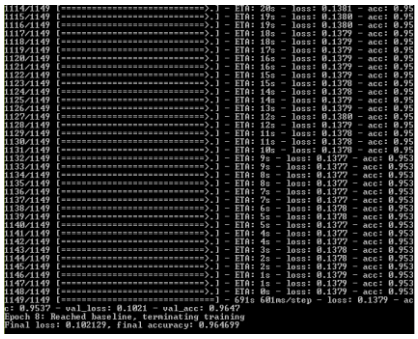

Fig. 8. Image showing the Validation loss, Validation accuracy, Final loss and Final accuracy

Fig. 9. Predicting the output image to an input image using the Text Line Segmentation of Handwritten Documents

The word "CAT" is an input image and using the Convolution Neural Network from Deep Learning the output is being predicted using the Text Line Segmentation of Handwritten Documents.

4. Conclusion and future work

Section-1 signifies the concise opening about OCR and its technique, nature of an image and document image converting, selection and types of separations. Section-2 entitles literature survey that has been fashioned during this training. In section-3, I discussed the demands in the segmentation of the text-lines and inspiration that immediate us in suggest distribution techniques. In section-4, I briefly described the propose methods of segmentation of handwritten document into lines. Firstly, to have more vigorous and hefty training, I could handle additional pre-processing methods such as delaying. I could also partition every pixel by its equivalent standard deviation to arrange the data. Then, provided time and budget vitalities, I was restricted to 20 training instances for each provided word in order to handily compute and edit my design. Another technique of mending my character distribution model would be to move further a greedy search for the most bent solution. I would access this by seeing a more comprehensive but still regular decoding algorithm such as beam search. I can provide a character/word-based language-based technique to count a penalty/benefit score to each of the available to the final beam search candidate paths, along with their tangled distinctive softmax characteristics, showing the characteristics of the series of characters/words. If the language model reveals perchance the most likely candidate word confer to the softmax coating and beam search is very unlikely given the framework so as far as to some other likely candidate words, then my design can correct itself properly.

References

[1] Antonacopoulos, A., Karatzas, D.: Document image analysis for World War II. personal records. In: Workshop on Document Image Analysis for Libraries. pp.336–341. IEEE (2004).
[2] Avidan, S., Shamir, A.: Seam carving for content-aware image resizing.In: ACM SIGGRAPH 2007 Papers. SIGGRAPH ’07, ACM (2007).
[3] Bluche, T.: Joint line segmentation and transcription for end-to-end hand written paragraph recognition. In: Advances in Neural Information Processing Systems(NIPS). pp. 838–846 (2016).
[4] Bluche, T., Louradour, J., Messina, R.: Scan, attend and read: End-to-end hand written paragraph recognition with MDLSTM attention (04 2016).
[5] Boiangiu, C.A., Tanase, M., Ioanitescu, R.: Handwritten documents text line segmentation based on information energy. International Journal of Computers, Communications and Control (IJCCC) 9, 8–15 (12 2014).
[6] Bunke, H., Bengio, S., Vinciarelli, A.: Offline recognition of unconstrained handwritten texts using HMMs and statistical language models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 26(6), 709–720 (2004).
[7] Diem, M., Kleber, F., Fiel, S., Gr¨uning, T., Gatos, B.: cBAD: ICDAR2017 competition on baseline detection. In: 14th International Conference on Document Analysis and Recognition (ICDAR). pp. 1355–1360. IEEE (2017).
[8] Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. CoRR abs/1312.2249 (2013).
[9] Frinken, V., Fischer, A., Mart´ınez-Hinarejos, C.D.: Handwriting recognition in historical documents using very large vocabularies. In: 2nd International Workshop on Historical Document Imaging and Processing (HIP). pp. 67–72. ACM (2013).
[10] Graves, A., Fern´andez, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling un segmented sequence data with recurrent neural networks. in: 23rd International Conference on Machine Learning. pp. 369–376. ACM.