# EXTRACT THE PUNJABI WORD WITH EDGE DETECTOR FROM MACHINE PRINTED DOCUMENT IMAGES

GAURAV SINGLA
M.Phil in Computer Application (Research Scholar), Department of Computer Application
Guru Kashi University
Talwandi Sabo, Punjab, India

Dr. PARMOD KUMAR
Assistant Professor, Department of Computer Application
Guru Kashi University
Talwandi Sabo, Punjab, India

**Abstract: Optical character recognition is the essential area of research in the world. Text segmentation and extraction is an important task under the OCR System. OCR system depends on the quality of the input document and type of Text Font. The researcher had segmented the text from images in many languages like English, Hindi, Urdu etc but the main problem is to extract the Punjabi word. In this paper we extract the Punjabi word with edge detector from machine printed document images detector using MATLAB.**

**Keywords-** OCR, Segment, Extract, Edge, Noise

## I. INTRODUCTION

OCR is the translation of scanned images of handwritten, typewritten or printed document into machine encoded form. This machine encoded form is editable text and solid in size [1]. Multimedia documents contain texts, graphics and pictures. Texts within an image play vital role in retrieval systems as they contain valuable information and can be easily extracted. Text extraction from document images is one of the major problems of handwritten text. Segmentation is an important issue in document analysis. There are two types of segmentation algorithms namely, region based and pixel based algorithms [2]. Text data is particularly interesting, because text can be used easily and clearly to describe the contents of an image. Character extraction is done in order to speed up the data entry. The current automatic data retrieval system existing in the market uses indirect method of reading data from the document. For e.g. Watermark based system to read confidential data and bar code system to read the price of goods in the retail commercial outlets. A system of extracting of text from text documents is introduced here which can replace the existing technology. There are a lot of possible uses of character extraction in Banking, Healthcare. Character extraction from images finds many useful applications in document processing, analysis of technical papers with tables, maps, charts, and electric circuits, identification of parts in industrial automation, and content-based image/video retrieval from image/video databases, educational and training video and TV programs such as news contain mixed text-picture-graphics regions [3]. The most familiar example is the ability to scan a paper document into a computer where it can then be edited in popular word processors such as Microsoft Word.

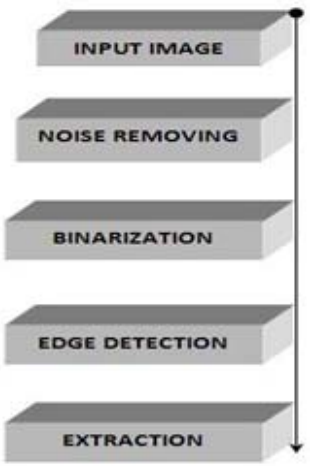


Fig. 1 The Proposed Model

## II. LITERATURE REVIEW

According to Rajiv Kumar et al. [5] OCR is the process of recognizing a segmented part of the scanned image as a character. OCR process consists of three major sub processes - pre processing, segmentation and then recognition. The segmentation process is the most important phase of the overall OCR process. It is the most significant process because if the output of segmentation phase is incorrect then we cannot expect the correct results; it is just like garbage in and garbage out. But on the same time, segmentation is complex too. If the document is handwritten then the situation becomes more cumbersome, because in that case only few points are there which can be used to make segmentation.

According to Dr. Shivaprakash Koliwad et al. [2] the detection and extraction of text regions in an image is a well known problem in the computer vision research area. Text extraction is a critical and essential step as it sets up the quality of the final recognition result. It aims at segmenting text from background, i.e isolating text pixels from those of background. Since readymade mixed mode image data is not available, it is necessary to create our own database. The database plays an important role as segmentation is to be done in an image. In educational videos and in presentation of lectures, graphic play an important role. In television industry text and images are simultaneously transmitted. In such similar application compression of data and bandwidth play an important role. To achieve better compression and bandwidth utilization properly, an efficient segmentation technique is necessary.

According to Rehna V. J et al. [3] Optical Character Recognition is the process of automatic conversion of machine-printed text into computer processable codes. In project they are emphasizing on extracting uppercase, lowercase letters and numerals from document images using segmentation and feature extraction in MATLAB image processing tool box. The proposed method can extract characters from document image (which may be scanned or camera captured) of any font size, colour, space and can be rewritten in an editable window like Notepad, WordPad where the characters can even be edited; thus, improving accuracy and hence, saves time.

According to Er. Balwinder Singh et al. [6] a character recognition system by using morphological operators on binary images. As a consequence, they will deal with the Punjabi language characters. This recognition system is merely feature-based, with no need of a learning phase or any kind of memory.

## III. NEED & SCOPE OF THE STUDY

Recently there is growing trend among worldwide researchers to extract word of many languages and scripts. Much of research work is done in English, French and Hindi languages. However, on Punjabi script, the research work is comparatively lagging. The work on Punjabi script is in beginning stage. Sometimes there is Punjabi script on the image that is difficult to recognize. Main scope of the study is to solve the recognition related problem by using recognition and extracting techniques. This work extends the technique used for Document word Extraction from Images. The research work is related with the Punjabi script.

## IV. OBJECTIVE OF THE STUDY

- Multimedia document contains graphs, pictures, charts and scripts. So the scan documents must segment before other document processing tasks.
- The objective of this study to extract the Punjabi word using edge detector from Machine printed document Images.

## V. RESEARCH METHODOLOGY

In this work the machine printed document image is considered as the input. First step is to read the input image. The machine printed document image is loaded into the system as the input. This is image in RGB format. Then the image is converted from RGB to binary in the pre-processing stage and the background is removed. Two fundamental functions namely line crop and letter crop are followed by feature extraction and segmentation stage. For analyzing lines, to treat it as start of the line it checks for continuous array of white pixels and if there are Black pixels it will consider as end of line. This procedure is continued till the last line of the image is displayed. In letter crop function, each character is cropped from the individual line, resized to the required dimensions and is displayed. This process is continued till the last character of the first line is displayed. This course of action is applied to the remaining lines of the input image. After letter crop, the extracted characters are compared with those in the database, if they are found matching; the character is interpreted, formatted and displayed. If not matching, the maximum matched character is displayed.

## VI. WORD EXTRACTION WITH EDGE DETECTOR

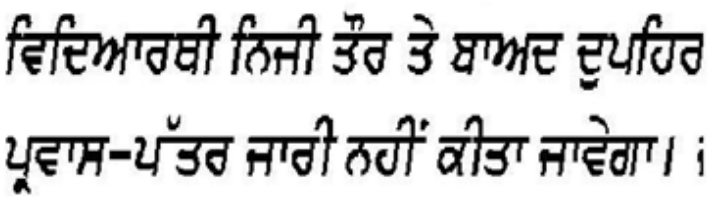

Fig. 2 Original Image

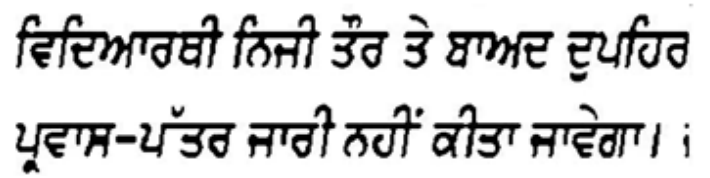

Fig. 3 Remove Noise

Fig. 4 Detect Edges

Fig. 5 Extract Word

## VII. STEPS FOR WORD EXTRACTION

- First step is to read the Original Image.
- Remove noise from the original image.
- Next step is to find the edges of Punjabi word using Edge Detector.
- Last step is to extract the word.

## VIII. CONCLUSION

In this paper we extract the Punjabi word using edge detector sobel from machine printed document image using MATLAB. In first step detect the edges of a whole text and then extract the single word from image. This work can improve by using some other techniques and improve the efficiency of the algorithm.

## REFERENCE


[1] Kartar Singh Siddharth , Mahesh Jangid, Renu Dhir, Rajneesh Rani, "Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features", International Journal on Computer Science and Engineering, Jalandhar , Vol. 3 No. 6, 2011

[2] Dr. Shivaprakash Koliwad, Jayanth.J," A Comparative Study of Segmentation in Mixed-Mode Images", International journal of Computer Application, Karnataka, Volume 31– No.3, October 2011

[3] Rehna, V. J, R. Neha, Sampada. H. K, "Character extraction and recognition from document images using segmentation and feature extraction (2012)", IRNet Transactions on Electrical and Electronics Engineering, Bangalore, Volume-1, Issue-2, 2012.

[4] Mandeep Kaur, Sanjeev Kumar," A Recognition System for Handwritten Gurmukhi Characters, International Journal of Engineering Research & Technology, Amritsar, Vol. 1 Issue 6, August 2012

[5] Rajiv Kumar and Amardeep Singh, "Character Segmentation in Gurumukhi Handwritten Text using Hybrid Approach ", International Journal of Computer Theory and Engineering, Vol. 3, No. 4, August 2011

[6] Usha Rani, Er. Balwinder Singh, Er. Ravinder Singh, "Machine Printed Punjabi Character Recognition Using Morphological Operators on Binary Images", International Journal of Engineering Research & Technology, Patiala , Vol. 1 Issue 3, May 2012.

[7] Rafael C. Gonzalez, Richard E. Woods, Digital Image Processing, 2nd ed., Prentice Hall, 2001.