International Journal of Information & Computation Technology.
ISSN 0974-2239 Volume 4, Number 17 (2014), pp. 1839-1845
© International Research Publications House
http://www. irphouse.com

Recognition of Gurmukhi Text from Sign Board Images Captured from Mobile Camera

Shilpa Arora [1], Dharamveer Sharma[2], Silky Arora [3]

[1] Department of Computer Science, Punjabi University, Patiala (Punjab)
[2] Assistant Professor, Department of Computer Science,
Punjabi University, Patiala (Punjab)
[3]Department of Computer Science, Punjabi University, Patiala (Punjab)
[1]arorashilpa69@yahoo.com, [2] dveer72@hotmail.com, [3]arorasilky08@gmail.com

Abstract

This paper presents recognition of Gurmukhi text from sign board images which are captured through mobile phone camera. The images are binarized and noise free. This consists of three stages. In first step the extracted text is segmented into characters. In second step the features which uniquely classify the characters are extracted using Zoning. In third stage the classifier SVM is used to recognize the text.

Keywords: SVM, classifier, recognition, Gurmukhi, images

1. Introduction

Automatic text recognition from images receives a growing attention because of potential applications in image retrieval, robotics and intelligent transport system, etc. In addition, extraction and recognition of texts in images is useful to blind and foreigners with language barrier as well. However, developing a robust scheme for extraction and recognition of texts from camera captured image is a great challenge due to several factors which include variations of style, color, spacing, distribution and alignment of texts, background complexity, influence of luminance, and so on. A large number of algorithms have been proposed in the literature to cope with these issues. Work for the development of complete OCR systems for Indian language scripts is major field of research. Research in the field of recognition of Gurmukhi script faces major problem mainly related to the unique characteristics of the script like connectivity of characters on the headline, characters in a word present in both horizontal and vertical directions, two or more characters in a word having intersecting minimum bounding rectangles along horizontal direction, existence of a large set of visually similar character pairs, multi-component characters, touching characters which are present even in clean documents and horizontally overlapping text segments. In this paper software to recognize text from sign board images has been developed. Before segmentation step the image is checked. If it is skewed it is de skewed by rotating the image.The rest of the paper is organized like section 2 covers review of literature, section 3 elaborates objectives, assumptions and the proposed technique, section 4 covers results, section 5 concludes the paper and references are given at end.

2. Review of Literature

A lot of research has been done on the recognition of isolated Gurmukhi characters, resulting in number of different techniques. Chen et al. [1], presents an edge based method to recognize text from images. There are quite some relevant works on text detection reported in the literature, including overlay text location in both images and videos. Edge (gradient) and edge layout are often used. Lehal and Chandan [2], presents a Complete Machine-Printed Gurmukhi OCR System. This is the first time that a complete multi-font and multi-size OCR system for Gurmukhi script has been developed. It has been tested on good quality images from books and laser print outs and has recognition accuracy of more than 97%. Research in the field of recognition of Gurmukhi script faces major problems mainly related to the unique characteristics of the script like connectivity of characters on the headline, characters in a word present in both horizontal and vertical directions, two or more characters in a word having intersecting minimum bounding rectangles along horizontal direction, existence of a large set of visually similar character pairs, multi-component characters, touching characters which are present even in clean documents and horizontally overlapping text segments. This paper addresses the problems in the various stages of the development of a complete OCR for Gurmukhi script and discusses potential solutions. The overall system design of the Gurmukhi OCR system developed and implemented. As with most of the OCR systems, there are five main processing stages: Digitization, Pre-processing, Segmentation, Recognition and Post-processing. Sharma and Arora [3], presents a software module to recognize isolated machine printed characters of Gurmukhi script. KNN and SVM performances are analysed in this paper. Ohya et al. [4], presents a method for recognizing characters in scene images. An image segmentation method based on local thresholding is applied and an evaluation of the gray-level differences between regions and their back ground is taken. Lim et al. [5] made a simple assumption, while detecting text, that text usually has a higher intensity than the background. They counted the number of pixels that are lighter than a predefined threshold value and exhibited a significant color difference relative to their neighborhood, and regarded a frame with a large number of such pixels as a text frame. This method is extremely simple and fast. Jhajj and Sharma [6], presents a system to recognize isolated handwritten Gurmukhi characters. Zoning and two classification methods, k-nearest neighbor, SVM (support vector machines) have been used and compared. Paper work is done by classifying image area in two classes, text and non text using SVM (support vector machine). Singh and

Budhiraja [7], This paper presents an overview of the various O.C.R. systems for Gurmukhi which are developed for handwritten isolated Gurmukhi text. Many researchers have proposed various techniques for handwritten Gurmukhi script. Sharma and Singh [8], work was carried out to detect lines present in scanned document in handwritten Gurmukhi script. So firstly we are to find out the lines present in the document then to find words present in each line detected at the first step. Using the detected words it is to segment characters present in each word. Therefore using line detection algorithm (the first approach) lines were detected. Kaur and Josan [9], identified the features and train a model based on the feature vector which is then used to classify text and non text area in an image. This algorithm is in sensitive to text orientation the output of the text extraction algorithm is can be fed to an OCR system to recognize the contained information. The results obtained on varied set of images are compared with respect to precision and recall rates. This work is based on feature extraction and classification using SVM. Assumption is made that images are of good quality and free from noise. Wu. and Manmatha [10], The system uses a text segmentation procedure to focus attention to regions where text may occur, and then a Chip Generation module to and actual text strings within these regions. Reasonable heuristics on text strings, Such as height similarity, spacing and alignment are used in this module. Multi-scale processing is used to account for significant font size variations..

3. Proposed Technique

In this paper the system to recognize Gurmukhi text from sign board images has been developed. The code has been implemented in Microsoft Visual Studio 2010.LIBSVM tool has been used for classification of extracted text.

3.1. Objectives of Research

The objectives of the proposed study are outlined as follows:

- To develop a software module to recognize text of Gurmukhi script from sign board images.
- To analyze different kernels of SVM.

3.2. Assumptions

There are following assumptions while developing the system

- The images are noise free.
- Images are binarized images.
- The images are collected using mobile phone cameras of different resolutions.
- Skewed images are also collected.
- Only middle zones characters are considered.
- Sihari and bihari of Gurmukhi script are recognized as kanna.
- There is no benchmark database for Gurmukhi script. Therefore own database has been used for training the software module.

3.3. Segmentation

Segmentation is a technique, which partitions text into individual lines, words and characters Segmentation is of three types which are Line segmentation, Word segmentation and Character segmentation. Horizontal profiles are created for line segmentation. Vertical profiles are created for word segmentation and then characters are extracted into individual upper, middle and lower zone characters from words.

3.4. Classification:

The character is classified using SVM technique. Support vector machine is a machine learning technique. An SVM classifier discriminates two classes of feature vectors by generating hyper-surfaces in the feature space, which are optimal in a specific sense that is the hyper-surface obtained by the SVM optimization is guaranteed to have the maximum distance to the nearest training examples, the support vectors. The model is trained using 7257 different images of Gurmukhi characters. In testing phase the images taken from different sign board images are used as input. The results are shown in next section.

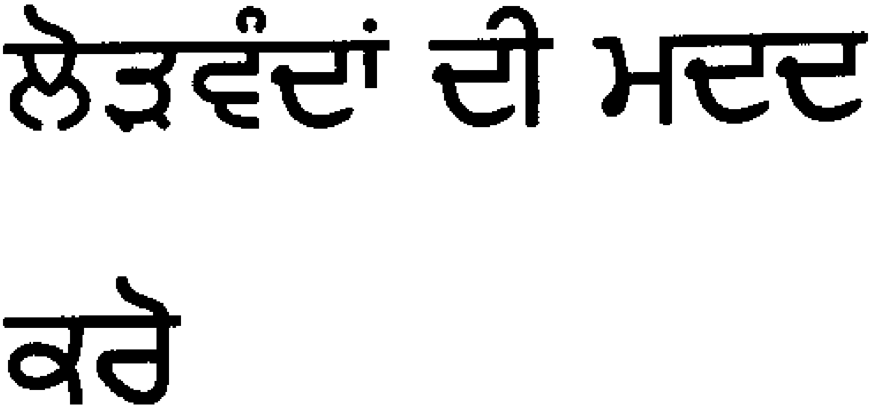
4. Results

Images are collected using mobile cameras of different resolutions. The word accuracy is found by dividing correctly recognized words with total number of words. The total numbers of words which are tested using this software are 657.Table below shows the results using different kernels of SVM.

Table 1. Recognition accuracy of Gurmukhi text using SVM kernels

Feature extraction	SVM kernel	Correctly recognized	Recognition accuracy
zoning	**linear**	**607**	**92.38%**
zoning	**polynomial**	**561**	**85.38%**

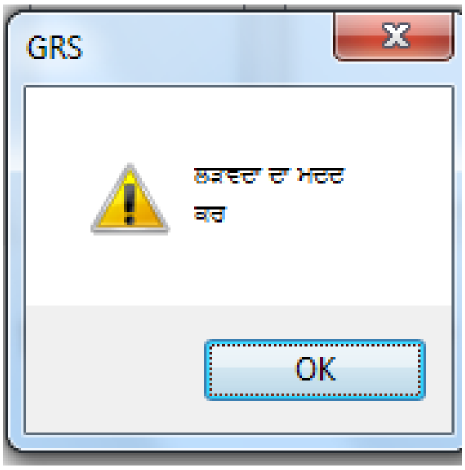
Some samples of the images which are collected using mobile phone cameras and their corresponding results are shown below. As we have assumed the characters which lies in middle zone are recognized and bihari is recognized as kanna of Gurumukhi script.



ਲੋੜਵੰਦਾਂ ਦੀ ਮਦਦ

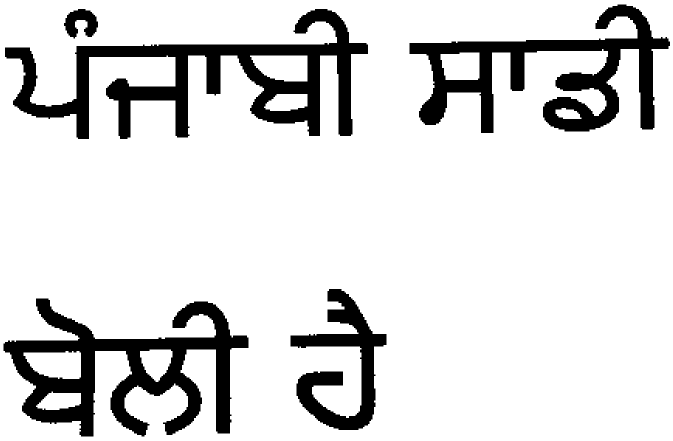
ਕਰੋ

Fig 1. Test sample 1



ਲੜਵਦਾ ਦਾ ਮਦਦ
ਕਰ

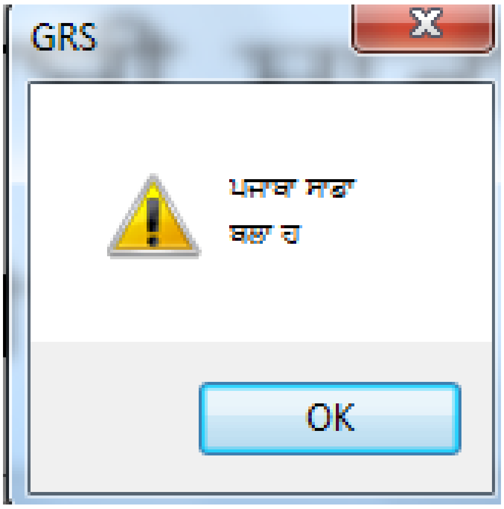
Fig 2. Result using svm linear kernel



ਪੰਜਾਬੀ ਸਾਡੀ

ਬੋਲੀ ਹੈ

Fig. 3 Test sample 2



ਪਜਾਬ ਸਤਾ
ਬਲਾ ਹ

Fig. 4 Result using SVM Polynomial kernel

5. Conclusion and Future Scope

In this paper the software module to recognize Gurmukhi text from sign board images has been developed.Tthe extracted text is segmented into characters using projection profiles. The zoning methods is used to extract features. The SVM classifier with linear and polynomial kernels is used to recognize text. Higher accuracy is achieved with linear kernel. The work presented in this paper can be further extended by considering all upper, middle and lower zones characters of Gurmukhi script. The results can be refined in post processing.

References

[1] Chen, X., Yang, J., Zhang, J. and Waibel, A., “Automatic Text Detection and Recognition of Signs in Natural Scene Images”, IEEE Transactions on Image Processing, vol. 13, no.1, pp. 87–99, 2004.

[2] G. S. Lehal and Chandan Singh, "A Complete Machine Printed Gurmukhi OCR System", Vivek, vol. 16, pp. 10-17, 2006.

[3] D.Sharma and S. Arora, ” Recognition of Isolated Machine Printed Characters of Gurmukhi Script Captured through Mobile Camera” in proceedings of the International Conference on Sciences, Engineering and Technical Innovations(IMTC), vol. 1, pp. 380-384, 2014.

[4] J.Ohya, A. Shio, S. Akamatsu, ” Recognizing Characters in Scene Images”, IEEE transactions on pattern analysis and machine intelligence. vol. 16, no. 2, pp.214-220, 1994.

[5] Lim, Y.K., Choi, S.H., and Lee, S.W., “Text Extraction in MPEG Compressed Video for Content-based Indexing”, Proc. of Int. Conf. on Pattern Recognition, vol. 4, pp. 409-412, 2000.

[6] P. Jhajj, D. Sharma, “Recognition of Isolated Handwritten Characters in Gurmukhi Script”, International Journal of Computer Applications, vol. 4, no. 8, pp. 9-17, 2010.

[7] P. Singh and S. Budhiraja, “Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script – A Survey”, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 vol. 1, Issue 4, pp. 1736-1739, 2012.

[8] R. K. Sharma and A. Singh, ” Segmentation of Handwritten Text in Gurmukhi Script”, International Journal of Image Processing, vol. 2, pp.1793-8201, 1995.

[9] S. Kaur and G.S Josan, “Gurmukhi Text Extraction From Image Using Support Vector Machine (SVM) ”, International Journal of Engineering Research and Applications (IJERA) ISSN : 0975-5462 vol. 3, no. 4, pp. 9-15, 2011

[10] V. Wu. and R. Manmatha, “Text Finder An automatic System to detect and recognize text in images”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, pp. 1224-1229, 1999.

[11] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm