www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 10 Oct 2015, Page No. 14594-14598

Recognition Of Punjabi Script Character And Number For Multiple Fonts

Guriqbal Singh[1], Vikas Mongia[2]

[1]Assistant Professor, Department of Computer Science,
Guru Nanak College Moga (Punjab), India
gurijohal21@yahoo.com

[2]Assistant Professor, Head, Department of Computer Science,
Guru Nanak College Moga (Punjab), India
vikasmongia@gmail.com

Abstract: ***In this paper there is telling about a easy and quick algorithm for recognition of Punjabi script which included both character and number. This algorithm is designed for multiple fonts because the fonts used in newspaper, magazines and books are different. We apply morphological operations on binary images. There is no need of any knowledge of phase or any kind of memory. Main advantage of this algorithm is its correctness to identify Punjabi characters and number. Only a very few work can be traced for character recognition of Indian scripts especially for the North Indian scripts like Punjabi. Input to the system is the scanned images from newspaper, magazines and old books.***

Keywords: Punjabi character recognition, number recognition, OCR, multiple font Recognition, machine printed, Templates.

1. Introduction

Optical Character Recognition is a technology used to copy and machine printed material into editable word processing file formats. This is the technology long used by libraries and government agencies to make lengthy documents quickly available electronically. According to Census of India of 2001 India has 122 major languages and 1599 other languages. For multiscript/multilingual country India, documents containing more than one Indian script are very common at distinct levels. But it is very challenging task to recognize a Punjabi script character and digits. Most of the people who speak this language live in the Punjab region of Pakistan and India. It is also widely spoken in Haryana, Himachal Pradesh and Delhi. Script identification has been discussed at paragraph level in [1] and [2] for Indian documents several algorithms have been proposed to improve recognition capabilities [3, 4]. Methods used to recognize characters inside a bitmapped image fall mainly into two categories: pattern matching, used in cheaper systems, and feature analysis, used in more sophisticated systems. Pattern matching methods have the bitmaps stored for every character of each of different font and type sizes. We create a single bitmaps for a single character and number of Punjabi script for multiple fonts. By comparing a database of stored bitmaps with the bitmap of scanned character the program tries to recognize the letters.

1.1 Character Recognition Techniques

The constant development of computer tools leads to a requirement of easier interfaces between the man and the

Computer. CR is one way of achieving this. A CR deal with the problem of reading handwritten/typewritten character offline i.e. at some point in time (in mints, sec, hrs) after it has been written. However recognition of unconstrained handwritten text can be very difficult because characters cannot be reliably isolated especially when the text is cursive handwriting. We divide the character recognition a technique into two categories first is Offline recognition and second is online recognition[10]. In case of online character recognition there is real time recognition of characters. Online systems have better information for doing recognition since they have timing information and since they avoid the initial search step of locating the character as in the case of their offline counterpart. Online systems obtain the position of the pen as a function of time directly from the interface. In offline recognition the source is either an image or a scanned form of the document whereas in online recognition the successive points are represented as a function of time and the order of strokes are also available [5]. Here in this paper only offline recognition is deal. The proposed OCR system provides the following features[6]: No more retyping, Quick Digital Searches, Edit Text and Save Space. Off-line recognition operates on pictures generated by an optical scanner. Off-line character recognition involves the automatic conversion of text in an image into letter codes which are usable within computer and text-processing applications. In case of offline character recognition the typewritten/handwritten character is typically scanned in form of a paper document and made available in the form of a binary or gray scale image to the recognition algorithm. Offline character recognition is a more challenging and difficult task as we do not have control over the medium and instrument used. The artifacts of the complex interaction between the instrument medium and subsequent operations such as scanning and binarization present additional challenges to the algorithm for the offline CR. Therefore offline character recognition is considered as a more challenging task then its online counterpart. The steps involved in character recognition after an image scanner optically captures text images to be recognized is given to the recognition algorithm.

1. Document Analysis / Preprocessing
2. Character Recognition / Classification

1.1.1 Document Analysis

The process of extraction of text from the document is called as document analysis. Recognition depends to a great extent on the original document quality and registered image quality.

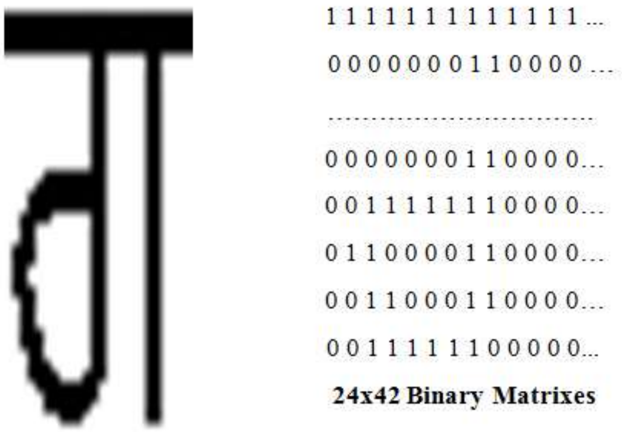
1.1.2 Character Recognition

The Punjabi character recognition algorithm has two essential components feature extractor and the classifier. Feature analysis determines the descriptors, or the feature set used to describe all characters and numbers. Given a character image, the feature extractor derives the features that the character possesses. The derived features are then used as input to the character classifier. Template matching or matrix matching, is one of the most common classification methods. Here individual image pixels are used as features. Classification is performed by comparing an input character with a set of templates (or prototypes) from each character class. Each comparison results in a similarity measure between the input characters with a set of templates. One measure increases the amount of similarity when a pixel in the observed character is identical to the same pixel in the template image. If the pixels differ the measure of similarity may be decreased. After all templates have been compared with the observed character image, the character's identity is assigned the identity of the most similar template. Template matching is a trainable process as template characters can be changed Character misclassification stem from two main sources: poor quality character images and poor discriminatory ability. Poor document quality, image scanning and preprocessing all degrade performance by yielding poor quality characters. The character recognition method may not have been trained for a proper response on the character causing the error. This type of error source is difficult to overcome because the recognition method may have its own limitations and all possible character images cannot possibly be considered in training the classifier.

2. Methodology

2.1 Creating Templates

Template matching or matrix matching, is one of the most common classification methods. Here individual image pixels are used as features. Classification is performed by comparing an input character with a set of templates (or prototypes) from each character class. So First of all create templates for alphabets and number used in Punjabi language. It is very difficult task but once you have created it then it is easy to recognize the character from any scanned Punjabi documents like book, magazine and newspaper. For template let the size of the matrix 24x42 pixels For example: in this example the Punjabi script alphabet 'r' is converted into binary bitmap.



```
1 1 1 1 1 1 1 1 1 1 1 1 1 ...
0 0 0 0 0 0 0 1 1 0 0 0 0 ...
.................................
0 0 0 0 0 0 0 1 1 0 0 0 0...
0 0 1 1 1 1 1 1 1 0 0 0 0...
0 1 1 0 0 0 0 1 1 0 0 0 0...
0 0 1 1 0 0 0 1 1 0 0 0 0...
0 0 1 1 1 1 1 1 0 0 0 0 0...
```

24x42 Binary Matrixes

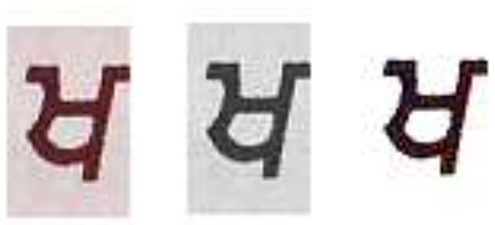
Figure 1: r alphabet of Punjabi script and its Binary Bitmap.

2.2 Document Analysis

The image is taken and is converted to gray scale image. The gray scale image is then converted to binary image. This process is called Digitization of image (Binarization). Practically any scanner is not perfect; the scanned image may have some noise. This noise may be due to some unnecessary details present in the image. By applying suitable methods the denoised image is produced. The denoised image thus obtained is saved for further processing [7].

2.2.1 Noise Removal

The process of removing noise is a pre-processing step used in OCR system to improve accuracy of the result. In generally we use scanned document images [8] for font detection, in OCR system, but the scanned images are not in good condition for processing due to noises. Mostly in old document we can see there are some spots and peaks, by which we can't get a better result, therefore the process of noise removing is a pre-processing step to be used after scanning the document. The preprocessing step for background noise cleaning is an important step after scanning images.

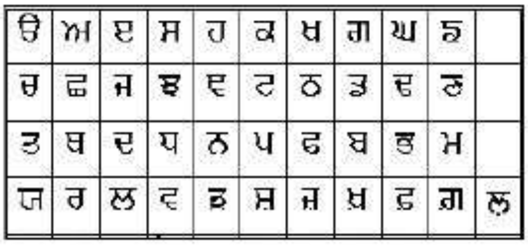


Original Image   Binary Image   After Preprocessing

Figure 2: Preprocessing of Image

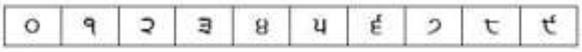
2.3 Character Extraction

The pre-processed image serves as the input to this and each single character in the image is found out [8].



ੳ	ਅ	ੲ	ਸ	ਹ	ਕ	ਖ	ਗ	ਘ	ਙ	
ਚ	ਛ	ਜ	ਝ	ਞ	ਟ	ਠ	ਡ	ਢ	ਣ	
ਤ	ਥ	ਦ	ਧ	ਨ	ਪ	ਫ	ਬ	ਭ	ਮ	
ਯ	ਰ	ਲ	ਵ	ੜ	ਸ਼	ਜ਼	ਖ਼	ਫ਼	ਗ਼	ਲ਼

Figure 3: 41 Punjabi Alphabets



੦	੧	੨	੩	੪	੫	੬	੭	੮	੯
0	1	2	3	4	5	6	7	8	9

Figure 4: Punjabi (Gurumukhi) Number from 0 to 9

2.4 Recognition

The image from the extraction stage is correlated with all the templates which are preloaded into the system. Once the correlation is completed, the template with the maximum correlated value is declared as the character present in the image. [8,11]

2.5 Post Processing

After the recognition stage, if there are some unrecognized characters found, those characters are given their meaning in the post-processing stage. Extra templates can be added to the system for providing a wide range of compatibility checking in the systems database [9].

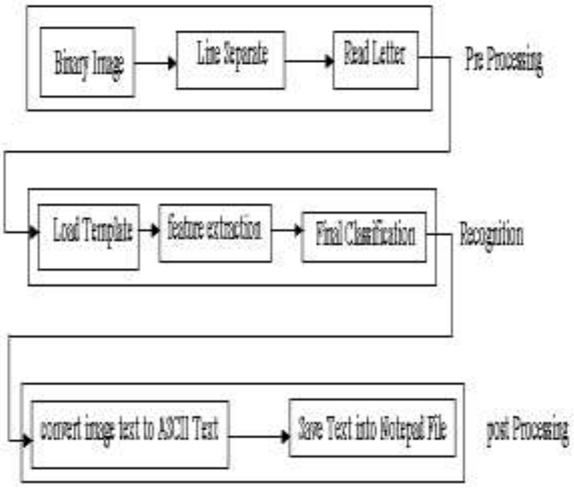
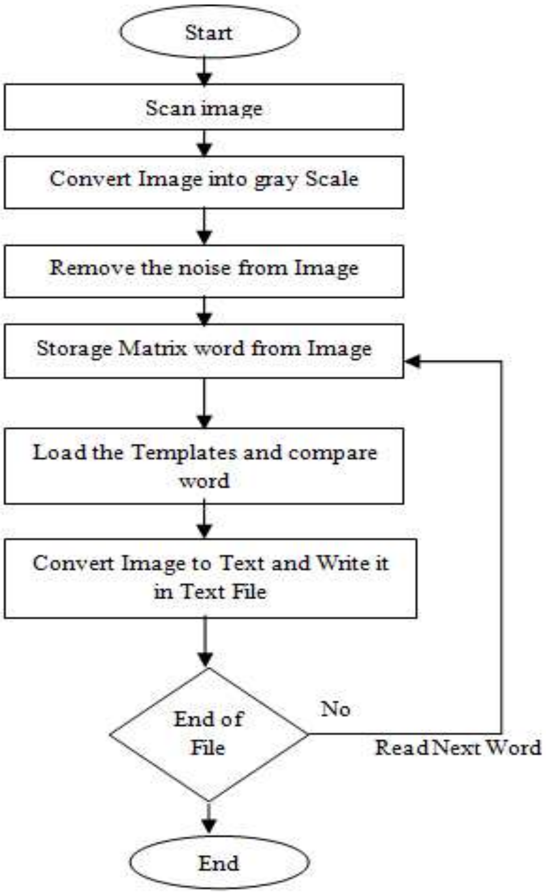


Figure 5: Construction of OCR system

3. Data Flow Diagrams


Start
Scan image
Convert Image into gray Scale
Remove the noise from Image
Storage Matrix word from Image
Load the Templates and compare word
Convert Image to Text and Write it in Text File
End of File
No
Read Next Word
End

Data flow chart show that input of proposed algorithm is scanned image. Convert it into gray scale image. Gray scale image is the black and white image which helps us to remove noise from it. Storage matrix word from image and then compares it with loaded templates. When word is matches with the template then appropriated ASCII text store in the text file.

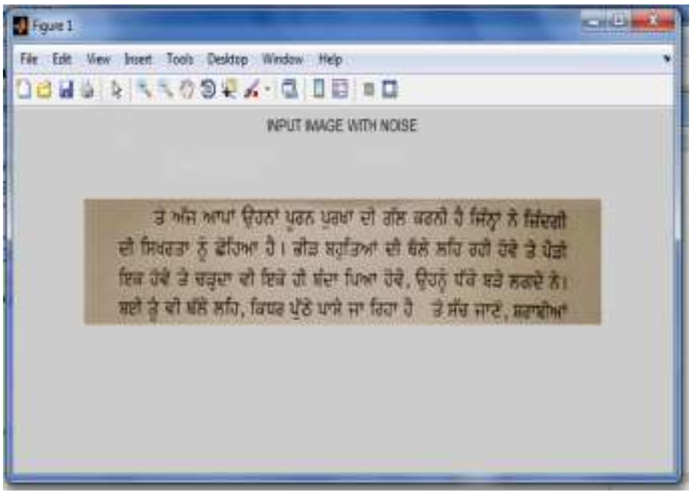
4. Results

We developed an application using Matlab R2011b and Performed test on different Punjab i fonts and characters of different size. And result is shown below. Input to this application is any scanned image magazine, news paper, old Punjabi record. The results are given below for different fonts of Punjabi character and digits.

Table 1: Margin specifications

Font	*Recognition Rate*
AnmoLlipi	100%
Asees	99.22%
Gurbani Akhar	95.28%
Joy	98.45%
AmrNeon	92.15%
Raaj	98.56%
Amrlipi	100%

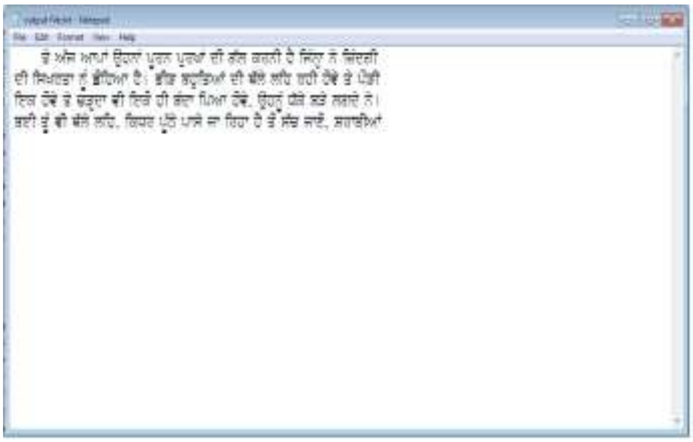
The following Screen shorts shows the input of our proposed work. The input of our algorithm is the scanned image of a Punjabi book in fig 6.



ਤੇ ਅੱਜ ਆਪਾਂ ਉਹਨਾਂ ਪੂਰਨ ਪੁਰਖਾਂ ਦੀ ਗੱਲ ਕਰਨੀ ਹੈ ਜਿਨ੍ਹਾਂ ਨੇ ਜ਼ਿੰਦਗੀ
ਦੀ ਸਿਖਰਤਾ ਨੂੰ ਛੋਹਿਆ ਹੈ। ਭੀੜ ਬਹੁਤਿਆਂ ਦੀ ਥੱਲੇ ਲਹਿ ਰਹੀ ਹੋਵੇ ਤੇ ਪੈੜਾਂ
ਇਕ ਹੋਵੇ ਤੇ ਚੜ੍ਹਦਾ ਵੀ ਇਕੋ ਹੀ ਬੰਦਾ ਪਿਆ ਹੋਵੇ, ਉਹਨੂੰ ਧੱਕੇ ਬੜੇ ਲਗਦੇ ਨੇ।
ਬਈ ਤੂੰ ਵੀ ਥੱਲੇ ਲਹਿ, ਕਿਧਰ ਪੁੱਠੇ ਪਾਸੇ ਜਾ ਰਿਹਾ ਹੈ ਤੇ ਸੱਚ ਜਾਣੋ, ਸ਼ਰਾਬੀਆਂ

Figure 6: input image with noise

The following screen short shows output Notepad file. Now it is in ASCII format. This format allow user to edit the documents.



ਤੇ ਅੱਜ ਆਪਾਂ ਉਹਨਾਂ ਪੂਰਨ ਪੁਰਖਾਂ ਦੀ ਗੱਲ ਕਰਨੀ ਹੈ ਜਿਨ੍ਹਾਂ ਨੇ ਜ਼ਿੰਦਗੀ
ਦੀ ਸਿਖਰਤਾ ਨੂੰ ਛੋਹਿਆ ਹੈ। ਭੀੜ ਬਹੁਤਿਆਂ ਦੀ ਥੱਲੇ ਲਹਿ ਰਹੀ ਹੋਵੇ ਤੇ ਪੈੜਾਂ
ਇਕ ਹੋਵੇ ਤੇ ਚੜ੍ਹਦਾ ਵੀ ਇਕੋ ਹੀ ਬੰਦਾ ਪਿਆ ਹੋਵੇ, ਉਹਨੂੰ ਧੱਕੇ ਬੜੇ ਲਗਦੇ ਨੇ।
ਬਈ ਤੂੰ ਵੀ ਥੱਲੇ ਲਹਿ, ਕਿਧਰ ਪੁੱਠੇ ਪਾਸੇ ਜਾ ਰਿਹਾ ਹੈ ਤੇ ਸੱਚ ਜਾਣੋ, ਸ਼ਰਾਬੀਆਂ

Figure 7: output Text file of Notepad

The accuracy of this system is very good as shown in table; accuracy is down only in the cases where we need to divide the image into two or four equal parts.If any other languages which do not need to divide the image.

5. Conclusions

This paper tells about Recognition of Punjabi Script Character and Number for Multiple Fonts system for offline handwritten\typed character recognition. The systems have the capability to give up brilliant results. Preprocessing techniques used in document images as an primary step in character recognition systems were described. The feature extraction step of optical character recognition is the most important. It can be used with existing OCR methods, particularly for English text. This system offers best and accurate methods of recognition of Punjabi alphabet and digits. The advantage of this method is its scalability, while it is configured to read a predefined set of document formats, currently Punjabi documents, it can be configured to recognize new types. Future scope of research is that implement this proposed work in mobile devices, extraction of text from video images, extraction of information from security documents and processing of chronological documents.

6. Acknowledgements

Guriqbal Singh, Assistant Professor, Department of Computer Science in Guru Nanak College Moga. Completed B.Tech (Computer Science and Engineering) from Punjab Technical University, Jalandhar. M-Tech (Computer Science and Engineering) from Punjab Technical University, Jalandhar Pb. (INDIA). Research Area: - Bio-Informatics, Software Engineering, and Image Processing. He works on the Punjabi character and digits recognition for multiple fonts in MAT Lab.

Vikas Mongia, Head, Department of Computer Science in Guru Nanak College Moga. Completed MCA from Punjabi University, Patiala. M.Tech(CSE) From Lovely Professional University. UGC-NET qualified. Research Area: Data Mining, Image processing, and fuzzy Logic.

References

[1] G.D Joshi, S.Garg and J.Sivaswamy, "Script Identification from Indian Documents," IAPR Intl Workshop Document analysis System, pp. 255-267(Feb. 2006).

[2] S.Chaudhury and R. Sheth "Trainable Script Identification Strategies for Indian Languages," Proc. Intl Conf. Document Analysis and Recognition pp. 657-660, (Sept. 1999).

[3] S. Kahan, T. Pavlidis, H. S. Baird, “On the recognition of printed characters of any font and size” IEEE Trans. Pattern Anal. Machine Intell., vol. 9, no. 2, pp. 274-288, March 1987.

[4] R.E. Howard, B. Boser, J.S. Denker, H.P. Graf, D.Henderson, W. Hubbard, L.D. Jackel, Y. LeCun, H.S. Baird: “Optical character recognition: a technology driver for neural networks” Circuits and Systems, 1990, IEEE International Symposium.

[5] Alon, Jonathan, “Document Analysis and Recognition”, 2005. Eighth International Conference on 29 Aug.-1 Sept. 2005.

[6] en.wikipedia.org/wiki/Optical_character_recognition

[7] Bolan Su, Shijian Lu, Tan C.L., “Combination of Document Image Binarization Techniques”, 2011 International Conference on Document Analysis and Recognition.

[8] Anand Arokia Raj, Kishore Prahallad, “Identification and Conversion of Font-Data in Indian Languages” at International Conference on Universal Digital Library (ICUDL2007) November 2007, Pittsburgh, USA.

[9] Junaid Tariq, Umar Nauman Muhammad Umair Naru, “α-Soft: An English Language OCR”, 2010 Second International Conference on Computer Engineering and Applications.

DOI: 10.18535/ijecs/v4i10.10

[10] Pranob K Charles, V.Harish, M.Swathi, "A Review on the Various Techniques used for Optical Character Recognition", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 1,Jan-Feb 2012.

[11] Udo Miletzki , "Character Recognition in practice Today and Tomorrow", 1996, , Siemens, Germany.

Author Profile

First Author

Guriqbal Singh, Assistant Professor, Department of Computer Science in Guru Nanak College Moga. Completed B.Tech (Computer Science and Engineering) from Punjab Technical University, Jalandhar. M-Tech (Computer Science and Engineering) from Punjab Technical University, Jalandhar Pb. (INDIA). Research Area: - Bio-Informatics, Software Engineering, and Image Processing. He works on the Punjabi character and digits recognition for multiple fonts in MAT Lab.

Second Author

Vikas Mongia, Head, Department of Computer Science in Guru Nanak College Moga. Completed MCA from Punjabi University, Patiala. M.Tech(CSE) From Lovely Professional University. UGC-NET qualified. Research Area: Data Mining, Image processing, and fuzzy Logic.