

FONT TYPE IDENTIFICATION OF HINDI PRINTED DOCUMENT

Yogendra Bagoriya¹, Nisha Sharma²

¹CDAC Noida, GGSIPU

²CDAC Noida, GGSIPU

Abstract

Optical font type identification is one of the important but often neglected problem. Due to the use of different font types OCR engine is not able to recognize the characters properly and accuracy of the system may decrease. It is not possible to reproduce old documents without knowledge of font type. Major use of different font styles in documents is to emphasis on some part of document so that any reader can notice them easily. In a document, font type changes may occur at particular points like titles, indexes, references, etc. They may be done by choosing another typeface, or changing the style or the size of the same typeface. Major applications of font type identification are document reproduction, creation of new font types, Document indexing and information retrieval, improving the recognition rate of an OCR engine. This paper presents an approach for font type identification of 5 Hindi font types as Marathi-Vakr, Shusha05, Devanagari New, Shusha02, Prem chand ki kahaniyan (Type writer written old book)

Keywords: Font identification, Typographic and Structural features, Line level Identification, Devnagari font type, classifying font type.

-----***-----

1. INTRODUCTION

In machine-printed documents [1], the OCR systems can be divided in three groups: Mono-font, Omni-font and Multi-font. Mono-font OCR systems deal with documents written with one specific font type so their accuracy is very high but they need a specific module for each font. Omni font OCR systems allow the recognition of characters of any font, and for this reason their accuracy is typically lower. Finally, Multi-font OCR systems handle a subset of the existing fonts. Their accuracy is related to the number and the similarity of the fonts under consideration. These systems achieve the best results when a single letter has very similar features in each font and it is easy to discriminate among different classes. On the other hand, the recognition is very difficult when different letters have similar features: for example the letter 'l' in one font could be very similar to the digit '1' in another font. Hence there is a need of Optical font identification (OFI) before an input can be given to OCR engine for recognition.

The OCR engine uses the information extracted through OFI(optical font type identification) system to perform character recognition and provide better efficiency .OFI is useful and necessary in[4] Document reproduction where knowledge of the font is necessary in order to reproduce (reprint) the document, improving the recognition rate of OCR systems because OCR with a known font may give better efficiency than Omni-font OCR (OCR capable of recognizing characters of any font and size),identification of logical document structures where knowledge of the font used in a word, line, or text block may be useful for defining its logical label (chapter title, section title or paragraph), .Document

indexing and information retrieval, where word indexes are generally printed in fonts different from those of the running text. Since we know that an OCR is trained for a particular type of language or some standard font type or most widely using font type, when we give input textual image of trained font type than the OCR can easily process that and can generate the output as required, but if the input image to OCR of some un-trained font type than it gives incorrect output, and the accuracy of OCR gradually decreases. This approach identifies five Hindi font types as Marathi-Vakra, Shusha05, Devanagari New, Shusha02 and Pram chand ki kahaniyan (Type writer written old book). A font type can be specified by five attributes as Typeface (Times, Courier, Helvetica), Weight (light, regular, bold, heavy), Slope (roman, italic) , Width (normal, expanded, condensed), and Size. An example of various font types identified through the given approach is shown below

Table 1 Various Hindi fonts with their effects

Shusha02	योगेन्द्र	व्यस्त भारतीय	सम्पन्न
Shusha05	योगेन्द्र	व्यस्त भारतीय	सम्पन्न
Marathi-Vakra	योगेन्द्र	व्यासत भारतीय	सम्पन्न
Devanagari New	योगेन्द्र	व्यासत भारतीय	सम्पन्न

The major steps involved in Optical font identification are shown as follow:

Pre Processing may enhance a document image preparing it for the next stage in a character recognition system.

Segmentation is a process that determines the constituents of an image.

Feature Extraction captures the essential characteristics of the font type.

Classification is the process of identifying each font type and assigning to it the correct class.

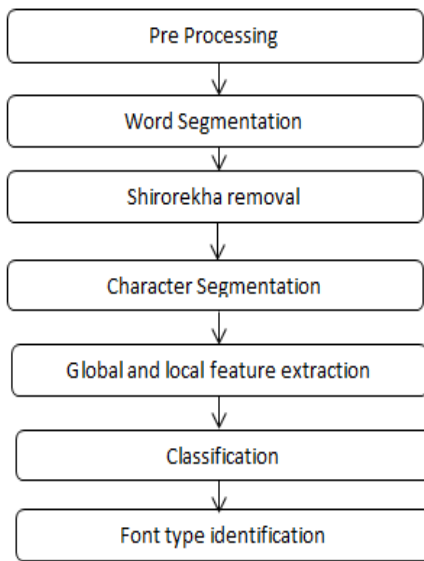


Fig1 Steps involved in Font identification

Section II of this paper discusses the various pre-processing steps performed before the input is forwarded for optical font identification. Section III gives an overview of segmentation technique used. Section IV discusses the feature extraction techniques used to extract features which can uniquely identify each font type. Section V discusses the classification of the font types. Section VI and Section VII discuss the Testing Results and the conclusion of the approach used respectively. Section VIII provides the future work of the research carried out.

Data Collection: Various Images has been collected written in different font types for the training and testing purpose. 40 images were collected for Marathi-Vakra font whereas 50 images each were collected for other four types of font i.e. Shusha05, Devanagari New, Shusha02 and PreamChand ki kahaniyan (Type writer). One sample image of each type of font is given as follow:

चहदिक रूदफल फजति स्फज्ज
दहन हप दरत जेक दफि दतह

Fig 2 Sample Image for Shusha20

बासफ कजबा नासदा कजासफ
जाहबाजा हसाफद हिबा बासद

Fig 3 Sample Image for Devnagari new

फज्जा दहभा दुज्ज तहाकज हा
दहजष्क गदक्षतह हगफ ज्जह

Fig 4 Sample Image for Shusha 05

जे नहि पायेनगे तुम सामजहति
कहबहु कजदुब जहसलर हककला

Fig 5 Sample Image for Marathi Vakra

मुलिया ने सिर से आंचल खिसकाया ;
से नहीं डरती, न बैठे-बैठे खाना चाहती हूँ;

Fig 6 Prem Chand ki Sampuran Kahaniya

2. PRE-PROCESSING

The collected font type documents are scanned using 300dpi which usually provides a low noise and good quality image. The digitized images are stored as binary images in BMP format. Pre Processing is used to enhance a document image preparing it for the next stage of system. The three steps involved in the approach are:

- Noise removal
- Binarization.
- Skew correction
- Shiro rekha removal

Noise Removal: Noise is the result of errors in the image acquisition process that result in pixel values that do not reflect the true intensities of the real scene. Connected components are used to remove the noise from the image.

Binarization: Binarization is a process of converting color or grey-scale image into binary. Otsu method is applied to convert the image into binary image. Otsu's method selects the threshold by minimizing the within-class variance of the two groups of pixels separated by the thresholding operator.

Skew Correction: Due to the possibility of rotation of the input image and the sensitivity of many document image analysis methods to rotation of the image, document skew should be corrected.

Shiro Rekha Removal: The Shiro rekha removal based on horizontal projection. Where projection start from 1st row to last row In the projection profile where we get the pixels almost equal to the width of the word or above some threshold value, than these row are considered as shiro rekha

3. SEGMENTATION

Segmentation is a process that determines the constituents of an image. Applied to text; segmentation is the isolation of characters or words. In the approach segmentation is performed by isolating each connected component that is each connected black area. This technique is easy to implement, but problems occur if characters touch or if characters are fragmented and consist of several parts. The two types of segmentation done are:

Word Segmentation: Approach is based on the white space separating the two adjacent words in the segmented lines. Compute the vertical profile for each segmented line, find the points from which the word starts and ends.

Character Segmentation: The character segmentation is based on the labeling concept. In this each connected component in shiro rekha removed image is labeled by number. The labeling is done by using Recursion.

4. FEATURE EXTRACTION

Feature extraction is the identification of appropriate measures to characterize the component images distinctly. The objective of feature extraction is to capture the essential characteristics of the symbols, and it is generally accepted that this is one of the most difficult problems of pattern recognition.

Mainly two type of feature are used as typographical feature and structural feature. The typographical are two types.

Global Typographical Feature:- The global feature are extracted from text entity like word, line, paragraph or complete page. These features are generally detected by non-experts in typography. Some global features are

- Height of line.
- Word orientation.
- Word spacing.
- Word Height.
- Word width.
- Aspect ratio.

Local Feature Extraction:- The Local features are those that need extra knowledge of previous global features .Local

feature extraction is done from individual letters. The features are based on characters:

- Slope (roman, italic)
- Width (normal, expanded, condensed)
- Pixel Density (Vertical line, Shiro rekha)
- Based on Character Spacing.
- Thickness of Shiro rekha.
- Size of vertical line.

Various feature extracted are discussed as follow:

Vertical Line Identification: The vertical line identifies after character segmentation. The vertical projection used for this. The Colum in which the numbers of black pixels are greater than threshold value or almost equal to the height of character than that row considered as vertical line.

Number of Pixels before and after the vertical line: Number of pixels after and before is calculated by vertical projection profile. This information gives us that ether the vertical line is full stop or vertical line of some character. This information differentiates between full stop or character.

Thickness of vertical line: The thickness of vertical line is one of typographical feature that is helpful to differentiate one font type to another font type.

Word Space: The word space is another typographical feature to identify the font type. The vertical projection can easily give the word space.

5. CLASSIFICATION

The global and local features of the particular font type are extracted, and then trained the system according to these features. Next time when the input image gave to the system, it again finds the feature and compare with the trained values. The maximum matching gives the result. For un-trained font type the system just gives output as non-recognized font type.

6. TESTING RESULTS

Testing is performed on different erroneous image on different font type's Images, i.e. single line containing image, paragraph containing image and complete page containing image. Accuracy achieved was 96%,for Marathi-Vakra and Prem chand ki kahaniyan,98% for Shusha05 and Devanagri New and 100% for Shusha02.The final results are given below in table:

Table 2 Testing Table

Input Font Type	Number of images tested	Output Hit	Output Miss	Accuracy
Marathi-Vakra	40	48	2	96%
Shusha05	50	49	1	98%
Devanagari New	50	49	1	98%
Shusha02	50	50	0	100%
PreamChand ki kahaniyan(Type writer)	50	48	2	96%

Unrecognized Images: Example of unrecognized images is given as follow. Image font type was unable to recognize due to lots of noise present in it.

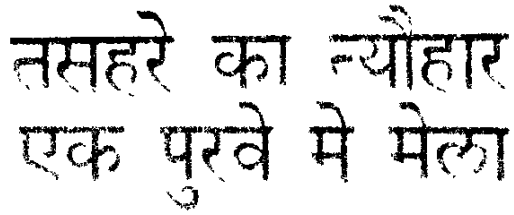


Figure 7 Unrecognized Sample Image

7. CONCLUSIONS

Font type Identification is an important aspect when analyzing document for useful information extraction. In this paper simple and fast technique is developed for detection of Font Type for improving the recognition accuracy of Hindi OCR system (without doing actual character recognition). The technique developed was applied on various font types in various different images with resolution 300 dpi. The Marathi-Vakra font type identifies with 98 % accuracy, 96% accuracy for Shusha05, 98% accuracy for Devanagari New, 100% accuracy for Shusha02, and 96% accuracy achieved in identification of Pream Chand Ki kahaniyan (Type writer written old book). The proposed approach needs more analysis and some constraints, for better result. These various methods and strategies help us in determining the font type from document images.

FUTURE WORK

The future work can be extensions up to word level font type identification. By study we found that the work on the improvement of the image can be the solution of this. As times goes the original image becomes erroneous, and during scan and binarization of image the actual feature or properties degraded like pixels from the corners removes or shape of serif or pixels lost from shiro rekha. This degradation of image can be reverted by improvement of image.

REFERENCES

- [1]. Abdelwahab zramdini and rolf ingold — optical font recognition using typographical features| IEEE transactions on pattern analysis and machine intelligence, vol. 20, no. 8, august 1998.
- [2]. B.v.dhandra, h.maliikarjun t, ravindra hegadi, and v.s.malemath — word- wose script identification based on morphological reconstruction in printed bilingual document| IEEE transactions on pattern analysis and machine intelligence, April 2000.
- [3]. Vincent, l.," morphological gray scale reconstruction in image analysis: applications and efficient algorithms," IEEE trans. On image processing vol.2 no. 2. Pp. 176- 201, 1993

- [4]. Abdelwahab Zramdini and Rolf Ingold "Optical font recognition from projection profiles" electronic publishing, vol. 6(3), 249–260 (September 1993)

BIOGRAPHIES



Mr Yogendra bagoriya received his B.tech degree from Sobhasaria engineering collage under Rajsthan technical university. He has completed his M.Tech. in Information technology from Centre for development of advance computing (CDAC), Noida. His interest area is Image processing. OOPS, Programming. Currently he is working as a software engineer in Mindfire Solutions.



Ms. .Nisha Sharma received her B.Tech. Degree in honours from Punjab Technical University .She has completed her M.Tech in computer science from Centre for development of advance computing (CDAC), Noida. Her interest area is Image processing. Database management Systems, OOPS, JAVA. Currently she is working in E-Governance team of CDAC Noida(R & D) and involved in SERB project.