Machine Recognition of Printed Odiya Characters

Peeta Basa Pati* A G Ramakrishnan† and U K Arvind Rao
Biomedical Laboratory
Department of Electrical Engineering
Indian Institute of Science
Bangalore 560 012, India

Abstract

Character recognition is becoming extremely important in the present day in office automation and intelligent system development. Schemes are available for printed character analysis and recognition in different scripts worldwide. Successful attempts have been made towards the achievement of this goal in few Indian scripts. However, commercial products are still a rarity. Here an attempt to develop such a system, for recognition of printed Odiya characters, is taken and success to some extent is reported.

Various preprocessing and segmentation units are described in brief. Necessary technical details of implementation is also provided, wherever necessary. Second order geometric moments, after dividing the character in to a number of rectangular sectors, are extracted for representation of these characters. Nearest neighbor (NN) and k-NN classifiers are utilized as decision making devices. Results obtained from different feature classifier combinations are also discussed.

1 Introduction

Optical Character Recognition **(OCR)**, as the name suggests, deals with recognition of characters in an optically-scanned document image. In scanned documents the presence of pictures, graphics along with text is unavoidable. The process of OCR involves the selection of the text region in the document and subsequent processing. The text area, after being separated, is taken for certain processes which helps in isolation of individual characters in this region. Each of these characters are considered as *test* characters and are recognized. There exists a set of available characters, *reference* or *training* characters, against which these test characters are compared. A decision is taken as to which one of the library of reference patterns the test one has maximum resemblance by a chosen process of *classification*.

A group of tasks are involved in the process of extracting the test patterns from the text region. The various subtasks involved therein are:

- **Preprocessing:** Preprocessing involves noise removal, skew detection and correction, binarization (process of conversion of gray-valued to two-tone image)and page analysis (segmentation of the document page in to regions of test, graphics *etc.*).
- **Segmentation:** This process includes separating the preprocessed image into lines, words and characters in that hierarchy.
- **Feature Extraction:** The attributes of a character which makes it distinct from other characters are called the ***features***. The process of obtaining them from individual characters is called Feature Extraction.
- **Classification:** Using the extracted features a decision is made about the class to which the test pattern possibly belongs.

2 Motivation

OCR has the potential to revolutionize the life of the people, using the script, in innumerable ways. This provides the motivation for development of such a system for Odiya script. Some of its regular uses are:

1. A media for mass conversion of existing documents and literature into electronic format
2. A reading aid for blind and others as a part of text-to-speech converter
3. Automatic sorting of letters and mails in postal department and processing of bank documents

*e-mail:pati@ee.iisc.ernet.in
†Corresponding Author: email:ramkiag@ee.iisc.ernet.in

4. Machine transliteration/translation of documents and literature of other scripts

5. Automatic verification and coding of products in industries and other centers of trade and commerce

6. A recognizer for analyzing the digits of the number plates of vehicles in motion; this would help control traffic systems.

7. A reader to input text in electronic publishing.

8. Automated bill payment processing in customer-centric departments such as telephone and electricity.

9. A deciphering system that can be fit in (future) automatic vehicles for acting upon textual-signals provided on the road sides.

10. A book-keeping system, to help in examination evaluation, attendance record evaluation, mark-sheet reading, etc., for educational and law enforcement institutions.

3 Properties of Odiya Script

Odiya being an oriental script, a brief description of it is essential for clear understanding of the script and design of an OCR system for it. Following are some of the important properties:

1. The Odiya script, like other Indian scripts, is derived from ancient Brahmi script through various levels of transformations [1].

2. There are 12 vowels, 35 consonants, 10 numerals in the Odiya alphabet set (Ref. Fig. 1).

3. For almost half of the characters there exists a vertical line towards the right end.

4. Some of these characters are modifications of other basic characters.

5. Vowels can occur anywhere in the word independently, unlike other Indian scripts [1].

6. Vowels combine with consonants to modify them and special symbols get added to these consonants, called as *matras*.

7. The matras, according to the modern practice, do not touch the consonant character while modifying the later.

Figure 1: Odiya Alphabet

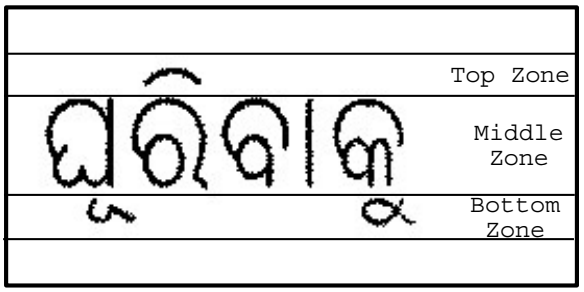

Figure 2: Three horizontal zones of Odiya word

8. Combination of a consonant with other consonants is governed by a set of rules. Sometimes they modify the base characters while other times another symbol gets added with the base character.

9. Sometimes two consonants combine to form a completely new character known as *compound* character.

10. Each modifier symbol, vowel modifier (*matra*) or consonant modifier, has a specific position with respect to the base character.

11. Compounding of two consonants is more abundant, though three consonants can also be compounded.

12. Each word can be modelled as consisting of three horizontal zones (Ref Fig.2). The zones in the center always has the base character; the zones above and below contain the modifiers.

13. Non-aspirated characters occur more frequently then aspirated ones.

14. Some of the consonant characters present in the alphabet set are never used independently but are used as modifiers only.

4 Preprocessing

The digitized document images contain inherent noise. The various sources of noise which add to the distortion of the document are: (i) inherent manuscript noise, (ii) non-transparent, dusty scanner bed, (iii) improper document placement on the scanner bed, and (iv) scanner-induced noise due to erroneous sampling, quantization and lighting. Any simple, readily available digital filter may be employed to reduce the noise present and improve the Signal to Noise Ratio (SNR) of the document image. However, from experience, the authors have noted that binarization (conversion of the gray-valued digital document to two-tone image) takes care of most of these noises.

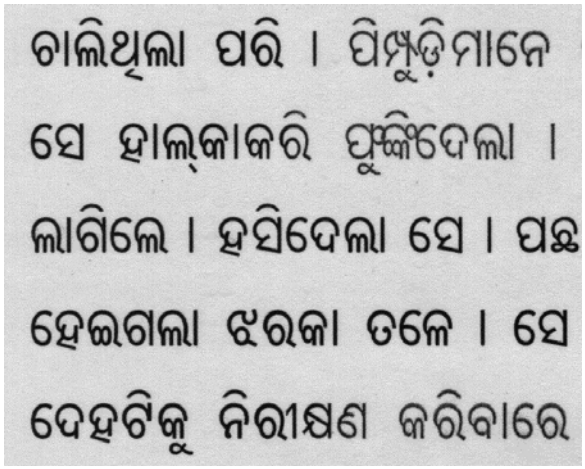

Figure 3: Original Image

The conversion of the gray-valued image to a two-tone image is known as binarization. There are various binarization techniques available in the literature [2]. A simple histogram based binarization technique is adapted. The histogram of a gray value document image contains two prominent peaks, one corresponding to the black fore-ground while the other to the background. After appropriate filtration of this function with a Gaussian filter, the valley point is chosen to be the threshold for binarization.

Due to improper placement of the document page on the scanner bed, rotation in the scanned image, *skew*, comes into effect. Detection and subsequent correction of the skew, before any further processing, is essential as skew affects the task of segmentation and hence recognition accuracy adversely. This is effectuated by a precision skew detection algorithm proposed by Mahata *et al.* [3, 4]. Correction of the skew in the image is brought about by an algorithm proposed by the above authors.

The presence of pictures and graphics along with the text in scanned documents is not unlikely. Selection and further separation of the text region from rest of the image can be accomplished by a simple texture segmentation algorithm.

Segregation of the text region in the binarized document image is followed by its segmentation at line, word and character level hierarchically. For line segmentation, the projection of the image onto the vertical axis, horizontal projection is considered. Words and Characters in a line are separated using the projection of the segmented line on to the horizontal axis, vertical

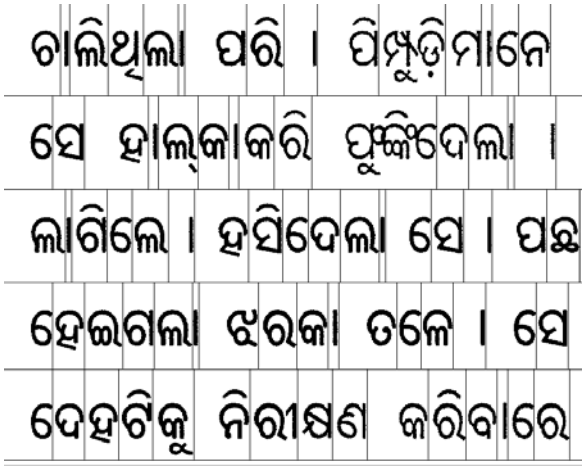

Figure 4: Preprocessed and Segmented Image

projection vector. Since the rule of the language keeps the matras (vowel modifiers for consonants) separated from base character, a connected component analysis separates the the *matras* situated either above or below the character.

5 Feature extraction and Classification

The attributes of the character which makes it different from the rest in the group are known as features. The shape information of the character should be implicit in a good feature vector. Ideally speaking, attempt to extract the features which are made use of by humans in recognizing the character should be made. But it is a real difficult task. Moreover, features used by machines need not be precisely those used by humans [5, 6].

The test character is normalized to a given size based on the aspect ratio (*character width / character height*). The normalized character is divided into a number of rectangular sectors. Second order geometric moments are extracted from each of these sectors, which contribute to the formation of the feature vector.

The moment m_n for a given function, $f(x)$ is given by:

$$m_n = E(x^n) = \int x^n f(x)\, dx$$

Extending the above equation for a 2-D function *e.g.* an image,

$$m_{l,k} = \int x^l y^k f(x,y)\, dx\, dy$$

Considering a digital image, normalizing w.r.t. mass factor, the generalized equation for second order geometric moments will be:

$$m_{l,k} = \frac{1}{\sum(i^2+j^2)} \left(\sum_{i=0}^{R} \sum_{j=0}^{C} i^l j^k I(i,j) \right)$$

where $(l,k) \in [0,1,2]$.

R, C are the number of rows and columns of the given image.

Making of a decision about the possible class value of the pattern under test based on the feature information available from it and the set of reference patterns is known as classification. Nearest Neighbor (NN) and k-Nearest Neighbor (k-NN) [7] classifiers are employed to accomplish this task of decision making. A test for the presence of the straight line (almost half the set of Odiya alphabet have a vertical bar towards its rightmost end) in the character is conducted and the derived information is compared with the available ones for confirmation. In case the results obtained from the above two methods contradict, a decision is taken to reject the character.

6 Results

The proposed algorithm is tested on a number of documents scanned from various sources. Different kinds of document are considered for this test. They are: (i) prints of computer generated fonts, (ii) prints of books edited after the year 1998, (iii) pages of books printed before 1995, and (iv) pages of a popular Odiya magazine, Sucharita.

The reference or *training* set consists of 38 classes of characters. Each class is represented by 10 templates. These patterns are chosen from the above scanned documents, with extra-care taken to ensure that the set contains characters of all shapes and sizes for a given class. Some of the reference patterns are deliberately chosen to be noisy ones, to handle the test patterns with noise. The set has size normalized patterns which touch all sides of the boundary.

After preprocessing and segmentation, around 1600 basic characters are randomly chosen to form the test set. It is ensured that all the categories of documents mentioned above, are considered for contribution to this set. Mahata *et al.* [3, 8] have reported the advantages of thinning the characters before taking them for feature extraction. Another set of test characters formed by the thinning of the above test set. Both these sets are tested for recognition separately.

The NN classifier chooses the class of the reference template with the minimum distance which it encoun-

ters last. The value of k for the k-NN classifier is selected to be 7. The following are the results of recognition:

Features	NN	k-NN
GM (Normal Chars)	82.33	72.27
GM (Thinned Chars)	41.88	39.41

7 Discussion and Future Scope

Observation of the obtained results clearly demonstrates the adverse effect of thinning on the characters in the process of recognition. Such effects could be explained by the absence of a robust thinning algorithm which could overcome the deficiencies created due to the development of breaks or improper thinning by the present algorithms.

It is also observed that the NN classifier is performing better then the k-NN classifier, the later taking more computation time. A situation where two different classes contribute in equal amount to the k-neighborhood in feature vector space is observed time and again. There exists a bias, irrespective of the method adopted to fix this dispute. Since hierchical systems are reported to be faring better then the mono-level ones in the literature, at this point the character can be taken for another level of feature extraction and classification. A different set of features or classifier or both can be utilized at the next level to resolve such ambiguities.

8 Conclusion

In the present work an OCR system for reading of printed Odiya characters is proposed for the first time. It is assumed that the document contains only one type of font in both size and style without any graphics. The preprocessing part of the task is successfully completed. Of the obtained test patterns, which consists of basic characters of the alphabet set, compound characters, matras and other special characters, only the basic characters are considered here. A comparative study of the available successful methods of feature extraction and classification is carried out and the results are demonstrated. Possible future directions of research in the area are also discussed.

References

[1] B. Chaudhuri and U. Pal, "A complete Printed *bangla* OCR System," *Pattern Recognition*, vol. 31, no. 5, pp. 531–549, 1998.

[2] B. Chaudhuri and D. Datta Majumdar, *Two Tone Image Processing and Recognition*. New Delhi: Wiley Eastern Limited, 1993.

[3] K. Mahata, "Optical Character Recognition for printed Tamil script," Master's thesis, Department of Electrical Communication Engineering, Indian Institute of Science Bangalore, 2000.

[4] K. Mahata and A. Ramakrishnan, "Precision skew detection through principal axis," in *International Conf. on Multimedia Processing and Systems*, (IIT Chennai), 2000.

[5] E. Gose, R. JohnsonBaugh, and S. Jost, *Pattern Recognition and Image Analysis*. New Delhi: Prentice Hall of India Pvt. Ltd., 1999.

[6] E. Gose, J. Bacus, and L. Ackerman, "A comparision of some computer-measured and human-measured pattern recognition properties," *Journal of Cybernetics*, vol. 1, pp. 68–74, 1971.

[7] R. Duda and P. Hart, *Pattern classification and Scene analysis*. John Wiley and Sons, 1973.

[8] A. Ramakrishnan and K. Mahata, "A complete OCR for Tamil printed text," in *Proceedings of Tamil Internet 2000*, (Singapore), 2000.