Off-line Handwritten Script Identification from Eastern Indian Document Images Using Logistic Model Tree

Sk Md Obaidullah, Nibaran Das and Kaushik Roy

Abstract Script identification from document images is a complex real-life problem for a multi-script country like India where 13 official scripts are present. To develop an optical character recognizer for a specific language, it is necessary to identify the script first by which the document is written. In this paper, scripts from the off-line handwritten document images written by any one of the four popular scripts in eastern India, namely Bangla, Roman, Devanagari, and Oriya, are identified. A document-level approach is followed for the same. Using some mathematical, structural, and script-dependent feature, a multi-dimensional feature set is constructed. Finally, logistic model tree (LMT) is applied for classification and an average accuracy rate of 95.5 % is obtained with a fivefold cross-validation.

Keywords Document image analysis · Handwritten script identification · Off-line documents · Classification · Optical character recognizer

S.M. Obaidullah (✉)
Department of Computer Science and Engineering, Aliah University, Kolkata, West Bengal, India
e-mail: sk.obaidullah@gmail.com

N. Das
Department of Computer Science and Engineering, Jadavpur University, Kolkata, West Bengal, India
e-mail: nibaran@gmail.com

K. Roy
Department of Computer Science, West Bengal State University, Barasat, West Bengal, India
e-mail: kaushik.mrg@gmail.com

© Springer India 2015

L.C. Jain et al. (eds.), *Intelligent Computing, Communication and Devices*,
Advances in Intelligent Systems and Computing 308,
DOI 10.1007/978-81-322-2012-1_72

1 Introduction

Optical character recognition is an active area of research since many years. It is useful for converting the physical document into digital form for making a paperless world in future. Document digitization also helps for better indexing and retrieval of huge volume of data available in modern society. The work is more relevant for a multilingual and multi-script country like India where 13 different scripts including Roman and 23 different languages including English [4] are present. There are many languages which use same script for writing. As an example, Bangla is a popular script in the eastern part of India which is used to write Bangla, Assamese, and Manipuri languages, whereas Devnagari is a popular script which is used to write different languages such as Hindi, Marathi, Nepali, and Konkani. So, here, it is not possible to develop a general purpose optical character recognizer targeting a particular language. Before feeding the particular language to the optical character recognizer, script needs to be identified first. That is why development of a script identification system is an essential requirement. Another problem arises when a single document is written using multiple scripts. Postal documents, filled up preprinted application forms, commercial advertisement documents, etc., are example of such multi-script documents. In these cases, word-level, line-level, or block-level script identification is must before choosing language-specific optical character recognizer.

Script identification can be classified into two broad categories, namely printed script identification and handwritten script identification. Handwritten script identification can be classified into two categories, namely off-line script identification and online script identification. Few works are reported in literature on script identification based on Indic scripts and non-Indic scripts. Among the pieces of work, Zhou et al. [14] identified Bangla and English printed and handwritten scripts using connected component profile-based features. Singhal et al. [12] identified Roman, Devanagari, Bangla, and Telugu scripts from handwritten document images with the help of rotation invariant texture features using multi-channel Gabor filter and graylevel co-occurrence matrix. Hochberg et al. [3] identified six scripts, namely Arabic, Chinese, Cyrillic, Devnagari, Japanese, and Latin, using some features such as horizontal and vertical centroid, sphericity, aspect ratio, and white holes. They performed the work at document level. In another work, Roy et al. [10] identified six popular Indian scripts, namely Bangla, Devnagari, Malayalam, Urdu, Oriya, and Roman, using features such as component-based features, fractal dimension-based features, and circularity-based features. This is a first kind of work involving six Indian scripts altogether. In a block-level script identification technique, Basu et al. [1] identified Latin, Devnagari, Bangla, and Urdu handwritten numeral scripts using similar-shaped digit pattern-based features. Using fractal-based features, Moussa et al. [8] identified Arabic and Latin scripts from line-level handwritten document.

Figure 1 shows block diagram of a multi-script document processing system. In Fig. 2, different multi-script documents are shown. The paper is organized as

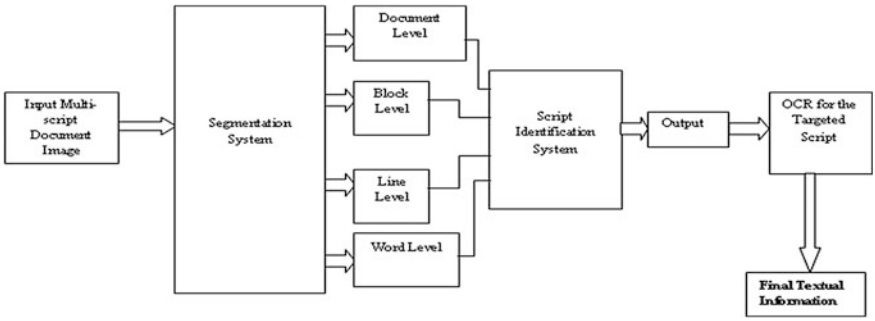

Fig. 1 Block diagram of multi-script document processing system

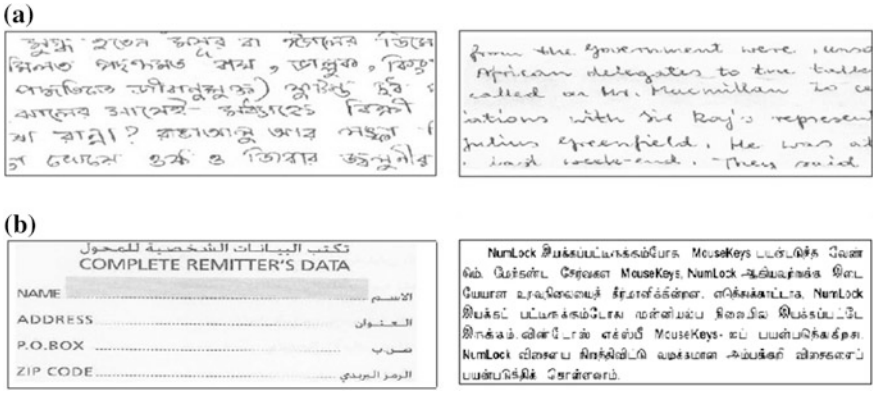

Fig. 2 Different multi-script documents. **a** Different document written by different scripts. **b** Same document written by different scripts

follows: In Sect. 2, data collection and preprocessing are described. In Sect. 3, feature extraction techniques are discussed, and the classification procedure with experimental result is described in Sect. 4. Finally, conclusion and scope of future works are described in Sect. 5. References are available in the last section.

2 Data Collection and Preprocessing

One of the major challenges in language and script identification work is absence of standard database. For this work, data are collected from different sources such as university and post office. From outside states, some data are collected through friends and different connections of the authors. Altogether, 32 Bangla, 32 Roman, 30 Devnagari, and 32 Oriya handwritten document pages are considered. Originally, the images are in gray tone and digitized at 300 dpi. A two-stage-based

approach is used to convert the images into two-tone image (0 and 1). In the first stage, a pre-binarization [11] is done using a local window-based algorithm in order to get an idea of different regions of interest. On the pre-binarized image, run length smoothing approach (RLSA) is applied to overcome the limitations of the local binarized method used earlier. After this, using component labeling, each component is selected and mapped them in the original gray image to get respective zones of the original image and the final binarized image is obtained using histogram-based global binarization algorithm [11] on these regions of the original image.

3 Feature Extraction and Selection

Feature extraction and selection is the most important task in any language or script identification work. Good features mean which are robust and easy to compute. The major features used for this work are component-based feature, shape-based feature, fractal-based feature, and freeman chain-code-based feature. Some abstract or mathematical features are also computed. Altogether, a 41-dimensional feature set consisting of features from all the above-mentioned categories is computed. Some of the important features from the feature set that we have applied are discussed below.

3.1 Component-Based Feature

Component analysis is one of the most useful and widely used tools in image processing. Here, using component analysis, we have classified all the components into three categories, namely (i) large component, (ii) medium component, and (iii) small component. An experimental threshold value is assumed for categorizing the components. For example, to calculate small component, the threshold value is assumed to be five pixels. Dots and comma characters fall under this category.

3.2 Shape-Based Feature

Under shape-based feature, occurrence of circularity at component level is calculated in a particular script. Following are the steps followed:

- Minimum enclosing circle is drawn which will enclose the component minimally, and the radius (r_1) of the circle is being stored.

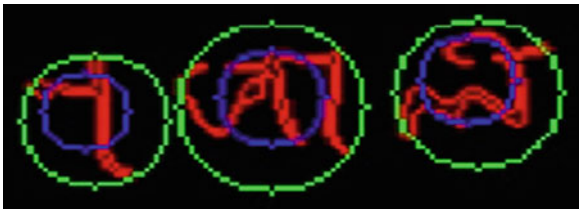

Fig. 3 Computation of circularity of component on Bangla script using fitted circles (*blue* minimum encapsulating and *red* best fitted)

- Circle fitting is done. Circle fitting refers to the fitting of a circle in the component in as minimum manner as possible. Its radius (r_2) is also stored.
- The difference of the two radii is stored to indicate the circularity of the component. The more the circularity of the component, the lesser will be the difference between the two radii (Fig. 3).

In fact, the circular components will have zero difference between the two radii or will have a difference tending to zero.

3.3 Fractal Dimension-Based Feature

Among the structural features, fractal dimension is one of the most important features. A fractal [7] is defined as a set for which the Hausdorff–Besikovich dimension is strictly larger than the topological dimension. The fractal dimension is a useful method to quantify the complexity of feature details present in an image. The fractal dimension is an important characteristic of the fractals because it contains information about their geometric structures. By employing fractal analysis, researchers typically estimate the dimension from an image. The fractal dimension of continuous object is an entity specified in terms of well-defined mathematical limiting processes. A fractal is an irregular geometric object with an infinite nesting of structure at all scales (self-similarity) (Fig. 4).

The upper part and the lower part play a significant role in feature extraction from the document image. In case of Devnagari script or Bangla script, the upper part will mainly contain matra or shirorekha pixels, whereas the lower part will contain the base pixels of the component. In case of Roman script or Urdu script, there will be no matra or shirorekha. So if pixel density is calculated, there will be difference in pixel density of upper part and lower part of the components of different scripts.

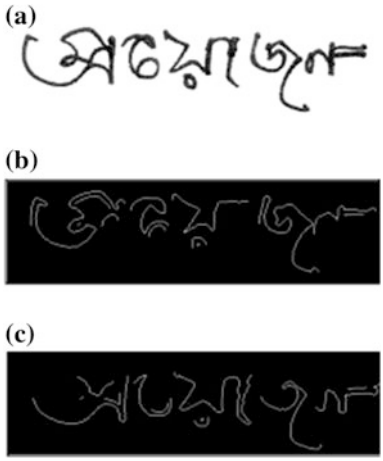

Fig. 4 Fractal dimension-based feature. **a** Original component. **b** *Upper* part of the contour. **c** *Lower* part of the contour

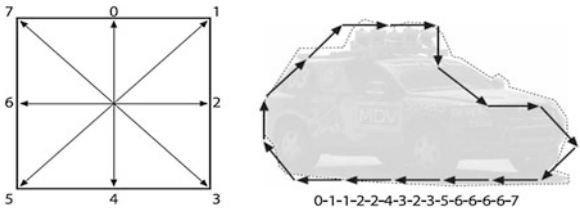

Fig. 5 Freeman chain code [2]

3.4 Feature Based on Freeman Chain Code

In Bangla and Devnagari scripts, horizontal lines present on the upper part of the writing are called 'Matra' or 'Shirorekha.' This is a unique distinguishing feature of these two scripts from the rest. We use cvFindContours() function in OpenCV [5] in CV_CHAIN_CODE mode for identifying these lines as a sequence of integers as shown in Fig. 5. Some slanting line presents in other scripts is also identified by the technique.

4 Classification Using Logistic Model Tree

Based on the above-normalized features, we employed logistic model tree (LMT) classifier under WEKA tool [2] for identification of handwritten Bangla, Roman, Devnagari, and Oriya scripts. WEKA is one of the widely used tools in the area of

Table 1 Confusion matrix

Script name	Bangla	Roman	Devnagari	Oriya	Average accuracy rate (%)
Bangla	96.8	0	3.2	0	95.5
Roman	4.2	95.8	0	0	
Devnagari	8.4	0	91.6	0	
Oriya	3.2	0	0	96.8	

machine learning. It contains tools for various applications such as data preprocessing, classification, clustering, regression, association rules, and visualization.

4.1 Logistic Model Tree Classifier

For present work LMT classifier is used. The model is build using a classification tree with logistic regression function at the leaves. The algorithm can deal with binary and multi-class target variables, numeric and nominal attributes, and missing values. For more detail, refer [6, 9, 13].

5 Result and Discussion

In the experiment a total of 126 document images are used consisting of 32 Bangla, 32 Roman, 30 Devnagari and 32 Oriya scripts. Table 1 shows confusion matrix where Bengali and Oriya obtain highest accuracy rate where as Devnagari obtained lowest among the four. Overall, 95.5 % average accuracy is obtained using LMT classifier with a fivefold cross-validation. In this result, observation is that Devnagari script gives lowest accuracy because of its similarity with Bengali script in some features such as presence of 'matra.' That is why 8.4 % Devnagari scripts are misclassified as Bengali script (Fig. 6).

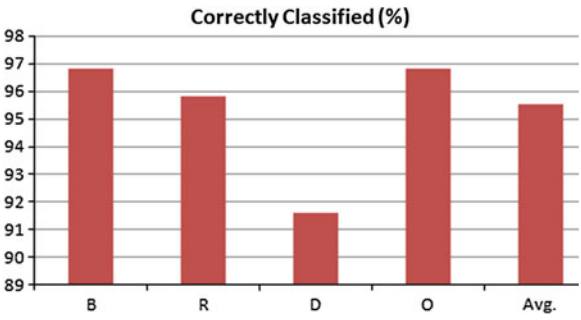

Fig. 6 Correctly classified percentage of all the scripts

Table 2 Comparative study

Name of algorithm	Scripts considered	Average accuracy rate (%)
Hochberg	Arabic, Chinese, Cyrillic, Devnagari, Roman, Japanese	88
M. Hangarge	Roman, Devnagari, Urdu	88.6
L. Zhou	Roman and Bangla	95
Proposed method	Bangla, Roman, Devnagari, Oriya	95.5

Table 2 provides a comparative study with other result available so far in handwritten script identification problems. The proposed method considering four scripts performs considerably well compared to other three available methods.

6 Conclusion

Script identification from four popular eastern Indian scripts in handwritten document images is proposed. Many works are available on printed script identification problem but attention is very less on handwritten script identification category. That is why emphasis needs to be given on the problem of handwritten script identification. So far, all the discussions were restricted to off-line script identification area. Future plan of the authors includes extending the work considering all 13 official Indian scripts and working in the online and video environment for real-life automatic script identification problem.

References

1. Basu, S., Das, N., Sarkar, R., Kundu, M., Nasipuri, M., Basu, D.K.: A novel framework for automatic sorting of postal documents with multi-script address blocks. Pattern Recogn. **43**(10), 3507–3521 (2010)
2. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**, 10–18 (2009)
3. Hochberg, J., Bowers, K., Cannon, M., Kelly, P.: Script and language identification for handwritten document images. Int. J. Doc. Anal. Recogn. **2**(2/3), 45–52 (1999)
4. http://www.rajbhasha.gov.in/8thschedulehin.pdf
5. http://www.opencv.org
6. Landwehr, N., Hall, M., Frank, E.: Logistic model trees. Mach. Learn. **59**(1–2), 161–205 (2005)
7. Mandelbrot, B.B.: The fractal geometry of nature. Freeman, NY (1982)
8. Moussa, S.B., Zahour, A., Benabdelhafid, A., Alimi, A.M.: Fractal-based system for Arabic/Latin, printed/handwritten script identification. In: Proceedings of International Conference on Pattern Recognition, pp. 1–4 (2008)

9. Obaidullah, S.M., Roy, K., Das, N.: Comparison of different classifier for script identification from handwritten document. In: Proceedings of ISPCC 2013 at Shimla (2013)
10. Roy, K., Das, S.K., Obaidullah, S.M.: Script identification from handwritten document. In: Proceedings of the Third National Conference on Computer Vision Pattern Recognition, Image Processing and Graphics, pp. 66–69. Hubli, Karnataka, Dec 2011
11. Roy, K.: On the development of an optical character recognition system for Indian postal automation. PhD thesis, Jadavpur University (2008)
12. Singhal, V., Navin, N., Ghosh, D.: Script-based classification of hand-written text document in a multilingual environment. In: Research Issues in Data Engineering, p. 47 (2003)
13. Sumner, M., Frank, E., Hall, M.: Speeding up logistic model tree induction. In: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 675–683 (2005)
14. Zhou, L., Lu, Y., Tan, C.L.: Bangla/english script identification based on analysis of connected component profiles. In: Lecture Notes in Computer Science, 2006, vol. 3872/2006, 24354, doi:10.1007/11669487_22