Chapter 1
Character Recognition System Using CNN for Sanskrit Text

R. Dinesh Kumar, M. Kalimuthu, and B. Jayaram

Introduction

Documents are in the form of papers that are understood by the human, but it is not possible for the computer system. So these documents are to be converted into a computer system readable form. OCR is the procedure of converting the scanned document into handwritten or printed text, symbols, letters and numerals and also can convert it into a possible format, for example, ASCII code [1]. Typically OCR can be used for processing the handwritten character and pattern recognition and is motivated greatly by a desire to enhance machine-to-man communication. Presently, few products are available for handwritten character recognition process, though different kinds of methods and procedures have been proposed [2].

Sanskrit Language

Sanskrit language is no longer spoken but still contains written text.

There are a few beliefs about the name 'Sanskrit':

R. Dinesh Kumar (✉)
Department of Computer Science and Engineering, Siddhartha Institute of Technology and Science, Hyderabad, India
e-mail: me.dineshkumar@gmail.com

M. Kalimuthu
Department of Computer Science and Engineering, Malla Reddy Institute of Technolgy, Hyderabad, India

B. Jayaram
Department of Computer Science and Engineering, RMK Engineering College, Chennai, Tamilnadu, India

Ch. Satyanarayana et al. (eds.), *Proceedings of the International Conference on Computer Vision, High Performance Computing, Smart Devices and Networks*,
Advanced Technologies and Societal Change,
https://doi.org/10.1007/978-981-19-4044-6_1

1. Sanskrit is termed as the voice of Devas.
2. Sanskrit is also termed as Devanagari because of its excessive utilization in Brahmins of Gujarat.
3. An additional viewpoint is in that Devnagar area of Kashi, therefore it was termed as Devanagari (Sanskrit).

Sanskrit is the most precise scientific basis language. For a long period, it has been Indo Aryan's script language. It is also utilized by Marathi, Hindi and Nepali languages. Widely spoken language is Hindi since its script is in Sanskrit and Sanskrit has got the dialect status. In the initial stage, Hindi was stated as the state language and Sanskrit as the start script of the few states like Haryana, Madhya Pradesh, Himachal, Uttaranchal, Bihar and so on [3]. Currently, it is found that Sanskrit is connected with every other script. In this script, all letters are equal which means there is no concept of small or capital letters and is half syllabic in nature [4, 5].

Problems in Sanskrit Text

- All separate characters are fused by a headline named 'ShiroRekha' in the case of Sanskrit script. This fusion system creates it hard to isolate separate characters from the single words.
- There are different kinds of isolated dots, which are vowel modifiers, for example, 'Chandra Bindu', 'Visarga', and 'Anuswar', which add up to the confusion.
- Descender and Ascenders recognition are also difficult, attributed to the difficult nature of language.
- It comprises composite characters.
- Minor differences in same characters.
- It comprises a huge volume of stroke and character classes.

Vowels and Consonants

Sanskrit script comprises 18 vowels ('svar') and 34 consonants ('vyanjan'). In addition, vowels and consonants are also comprised of vowel modifiers named matra's which are located at right or left part of the Sanskrit script.

Literature Survey

The framework of Sanskrit Character Recognition system is shown in Fig. 1.1.

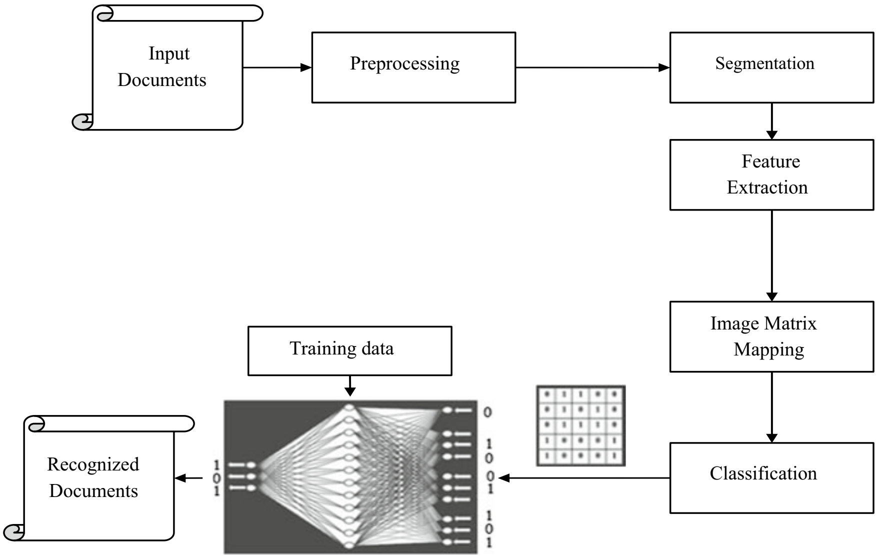

Fig. 1.1 Character recognition system

Input Documents

The documents are scanned with the proper scanner to provide the digital image for recognition. After this process, input image size is specified by the user which means the length and width of the document [6].

Preprocessing

After scanning, the next step is to remove noise from the scanned image [7]. The noise-free image is checked for skewness. Skewness is defined as the tilt (angle) in the bitmapped image of the scanned document. It is normally caused if the document is not straightly inserted into the scanner. But, most of the character recognition procedure is sensitive to the skew (orientation) of the input document, cropping is essential to implement the algorithms which can correct and detect the skew automatically.

Segmentation

After preprocessing, the noise-free image is passed to the segmentation phase, where the image is decomposed into individual characters. Sanskrit text character recognition process is carried out by applying the line and character segmentation techniques

[8]. The Sanskrit text images are segmented into the lines and each line is segmented into the words in terms of the upper modifier, consonant and lower process. Then the segmented words are converted into the straight lines which are used in Sanskrit text recognition process.

Feature Extraction

The feature extraction process examines a character segment and chooses a set of features that can be utilized uniquely to recognize the segmented text or character. The selection of a representative and stable set of a feature is the heart of the character recognition system. Different kinds of features are extracted to perform classification [9].

Classification

Classification is an important stage that is used for main decision-making process which is done by using the extracted features from the previous stages. The classification process identifies the character according to their preset rules [10]. The classification process is making decisions by using the class membership pattern. But this task is difficult because of the decision rule. Thus the feature extraction scheme is applied for reducing the misclassification probability. After this process, the classification process is done, but classification becomes an issue where characters fall into an unknown pattern.

Naïve Bayesian Classifiers

The Bayesian classifier is known to be capable of universal approximation and the output of a Bayesian network can be related to Bayesian properties. The Bayesian network has three input layers, namely input layer, hidden layer and the output layer in which each layer consumes the non-linear inputs and produces the linear output.

SVM Classifiers

The optimized features are applied to the Support Vector Machine which chooses the maximum fitness value to recognize the handwritten characters with better results. The extracted zones are considered as the features and are classified by applying the Support Vector Machine.

In this type, error function is minimized.

$$\frac{1}{2}w^r w + C\sum_{i=1}^{N}\xi_i \tag{1.1}$$

subject to the constraints

$$y_i\left(w^r \emptyset(x_i) + b\right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, i = 1, \ldots., N \tag{1.2}$$

Note that $y \in \pm 1$ is the class label and x_i is the independent variables.

$$\frac{1}{2}w^r w - vp + \frac{1}{N}\sum_{i=1}^{N}\xi_i \tag{1.3}$$

subject to the constraints

$$y_i\left(w^r \phi(x_i) + b\right) \geq \rho - \xi_i, \xi_i \geq 0, i = 1, \ldots .N \text{ and } \rho \geq 0 \tag{1.4}$$

CNN Classifier

CNN is a text classifier using feed-forward artificial neural networks and uses a different multilayer perceptron model to involve minimal preprocessing. The outcome of each convolution wills flames when a particular pattern is identified. As a result of varying the size of the kernels and concatenating their outcome, its permits to detect patterns of various sizes containing two, three, or five adjacent words. The output patterns could be in terms like ‘I hate’, ‘very good’ and consequently CNNs can categorize them in the sentence in spite of their position.

Performance Analysis

The performance analysis of CNN with Naive Bayesian classifiers, SVM classifiers is carried out. Then the evaluated performance metrics are listed as follows (Table 1.1).

a. Sensitivity

Table 1.1 Efficiency of character recognition methods

Metrics	Naive Bayesian classifiers	SVM classifiers	CNN classifiers
Sensitivity	83.66	89.13	91.13
Specificity	84.42	90.63	92.63
Accuracy	86.56	91.05	93.45

b. Specificity
c. Accuracy.

Mean Square Error

MSE calculates the difference between the pixel values of the original image and the recognized image. So, the MSE is calculated by using following Equation.

$$\mathrm{MSE} = \frac{1}{\mathrm{MN}} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(f(i, j) - f'(i, j) \right)^2 \tag{1.5}$$

In Eq. (1.5) $f(i, j)$ is represented as the original image and $f^{'}(i, j)$ is denoted as the recognized character image. M is the height of the image and N is the width of the image.

Unicode Mapping

The Unicode standard reflects the basic principle which emphasizes that each character code has a width of 16 bits. Unicode text is simple to parse and process Unicode characters have well-defined semantics. Hence, Unicode is chosen as the encoding scheme for the current work. After classification, the characters are recognized and a mapping table is created in which the Unicode for the corresponding characters are mapped. Based on the mapping, the Sanskrit characters are recognized with minimum error rate also enhances the recognition accuracy.

Results and Discussion

The efficiency of the CNN network is analyzed using the sensitivity and specificity metrics Table 1.2 clearly explains that the CNN method consumes minimum error rate while classifying the Sanskrit characters when compared to the existing method. The minimum error rate leads to increase the accuracy of the CNN system. From the above discussions, the proposed system classifies the Sanskrit characters with efficient manner when compared to the other traditional methods.

Table 1.2 Mean square error of CNN

Different classifiers	Mean square error
Naïve Bayesian classifier	0.9
SVM classifier	0.789
CNN classifier	0.645

Conclusion

The system explains the various processes of Sanskrit character recognition process namely preprocessing, segmentation, feature extraction and classification. The character recognition system methods are analyzed and the several quality measures such as sensitivity, specificity and accuracy have been used to analyze the effectiveness of the proposed techniques. Thus the CNN-based recognition system recognizes the exact handwritten characters in offline with minimum error rate and high accuracy when compared to the traditional classifiers. Thus, the CNN classifier overcomes the above-discussed classifiers' drawback with minimum error and high efficiency.

References

1. Agarwal, P.: Hand-written character recognition using Kohonen network. IJCST **2**(3), 10–18 (2011)
2. Alirezaee, S., Aghaeinia, H., Faez, K., Fard, A.S.: An efficient feature extraction method for the middle-age character recognition. Lect. Notes Comput. Sci. **3645**(PART II), 998–1006 (2005)
3. Ameri, M.R., Haji, M., Fischer, A., Ponson, D., Bui, T.D.: A feature extraction method for cursive character recognition using higher-order singular value decomposition. In: 2014 14th International Conference on Frontiers in Handwriting Recognition, pp. 512–516 (2014)
4. Dwivedi, N., Srivastava, K., Arya, N.: Sanskrit Word Recognition Using Prewitt's Operator and Support Vector Classification. ICECCN, pp. 265–269 (2013)
5. El Qacimy, B., Kerroum, M.A., Hammouch, A.: Feature extraction based on DCT for handwritten digit recognition. Int. J. Comput. Sci. Issues (IJCSI) **11**(6), 27–27 (2014)
6. Joshi, M.R., Sabale, M.V.V.: Offline Character Recognition for Printed Text in Devanagari Using Neural Network and Genetic Algorithm, pp. 1–8 (2013)
7. Kale, K.V., Deshmukh, P.D., Chavan, S.V., Kazi, M.M., Rode, Y.S.: Zernike moment feature extraction for handwritten Devanagari compound character recognition. Sci. Inf. Conf. **3**(1), 459–466 (2013)
8. Liu, X.Y., Blumenstein, M.: Experimental analysis of the modified direction feature for cursive character recognition. In: Proceedings—International Workshop on Frontiers in Handwriting Recognition, IWFHR, pp. 353–358 (2004)
9. Dineshkumar, R., Suganthi, J.: Sanskrit character recognition system using neural network. Indian J. Sci. Technol. **8**(1), 65–69. E-ISSN: 0974-5645 (2015)
10. Dineshkumar, R.: Offline Sanskrit handwritten character recognition framework based on multilayer feed forward network with intelligent character recognition. Asian J. Inf. Technol. **15**(11), 1678–1685. ISSN: 1682-3915 (2016)