DOI: 10.21917/ijsc.2023.0453

SCRIPT IDENTIFICATION FROM CAMERA CAPTURED INDIAN DOCUMENT IMAGES WITH CNN MODEL

Satishkumar Mallappa[1], B.V. Dhandra[2] and Gururaj Mukarambi[3]

[1]*Department of Mathematical and Computational Sciences, Sri Sathya Sai University for Human Excellence Kalaburagi Campus, India*
[2]*Department of Computer Science, Garden City University, India*
[3]*Department of Computer Science, Central University of Karnataka, India*

Abstract

Compared to typical scanners, handheld cameras offer convenient, flexible, portable, and noncontact image capture, which enables many new applications and breathes new life into existing ones, but camera-captured documents may suffer from distortions caused by a nonplanar document shape and perspective projection, which lead to the failure of current optical character recognition (OCR) technologies. This paper presents a new CNN model for script identification from camera-captured Indian multilingual document images. To evaluate the performance of the proposed model 9 regional languages, one national language and one international Roman languages are considered. Two languages, Hindi national language, and Roman English language are taken as the common languages with regional language for the study. The proposed method is applied on Bi-script, Tri-script, and Multi-script combinations. The average recognition accuracy for three script combinations is 92.92%, for bi-script 91.33%, and for tri-script 87.33%. is achieved. The proposed method is the unified approach used for identifying the script from bi-script, tri-script and multi- script camera-captured document images and is the novelty of this paper. The proposed model is compared with the Alexnet pretrained CNN model, and it achieved the highest recognition accuracy.

Keywords:

OCR, Deep Neural Network, Alexnet, CNN, Script Identification

1. INTRODUCTION

Traditionally digitized textual content from books, newspapers, and articles has been processed using scanners and interpreted with the help of optical character recognition (OCR). The technical advances in digital cameras in recent days have led the OCR community to consider the usage of advanced cameras in place of scanners for document image capturing. Compared to scanners, the digital cameras are portable devices and are quite easier to use, capturing images from any viewpoint. These reasons have recently led to increased interest in camera-based document analysis. On the other hand, digital cameras are accompanied by image quality problems. While one captures an image of a document page, one may find that the resulting image is warped or has unnecessary geometric distortions. This is particularly true when one controls taking images of an opened, thick, bound book. In addition to challenges like uneven lighting and lens distortion. These distortions has inspired to do work in this area and explore many avenues. The objective of this paper is to identify the scripts from the different combinations of the scripts viz., bi-script, tri-script and multi-script from the document images are captured from the Camera.

This proposed work flows as follows, in the section 2 the brief details about the previous work carried out in the related area is given. Creation of dataset is presented in the section 3. Section 4 is contains the proposed method's experimental setup information. The experimental results along with discussion are shown in section 5. At last the conclusion and future work are followed in section 6.

2. LITERATURE REVIEW

A summary of the challenges to Camera based documents can be found in [1]. Script identification from the camera-based document image can be found in [2] [3] [4] [5] [6] [7]. Still there is no unified approach that addresses the script identification from the Camera captured document images with respect to time complexity, accuracy, space complexity etc. Hence, the problem requires addressing this problem in different angles. In this paper a new CNN model is developed to identify scripts from the camera-captured document images. In the literature, one can find very negligible amount of work related to script identification from Camera captured document images, scene text, natural scenes, and video scripts. In [2], they have addressed the script identification from the camera-based document images using the generation of templates and signatures and computed statistical methods to extract the features. The classification uses the hamming distance metric to calculate the distance between the train and test sets. From this, they could achieve 91.00% recognition accuracy. Camera-captured block-wise script identification is presented in [6]. They have considered LBP features extracted from three scripts; English, Hindi, and Kannada. KNN and SVM classifiers obtained 99.70% and 98.00% recognition accuracy. In [8] authors have reported the attention-based mechanism of a convolutional-LSTM network to identify the scripts from the natural scene images and video frames. The local and global CNN features are extracted from the input image, and the fusion method was applied. Using the CNN-LSTM network, they could achieve the recognition accuracy from four standard databases, namely, SIW-13, CVSI2015, ICDAR-17, and MLe2e, as 90.23% for ICDAR-17, for MLe2e 96.70%, for SIW-13 96.50%, and for CVSI-15, 97.75%. [9] proposed the HOG, GLCM Textures, and shape features for script identification from Camera captured multi-script scene Text components. On each word image, the HOG, and GLCM features are extracted and combined to construct a new feature vector. This combined feature vector was submitted to five popular classifiers namely; SVM, Multi-class, Naïve Bays, MLP, and Multi-class. The MLP classifier has obtained 90.00%, with the highest recognition accuracy amongst other classifiers. The bi-script identification from camera-based document images at block-level is reported in [5], English and Gujarathi combination of scripts have given the maximum accuracy of 86.45% by using KNN classifier. The camera-based bus sign-board script identification is presented in [10], from the input image they have extracted the

Gabor, Wavelet, and Log-Gabor features and obtained the 97.05% recognition accuracy from Kannada and Malayalam scripts.

From the literature, it is evident that very insignificant work has been carried out on script identification based on camera-captured document images. Few papers reported in the literature on camera-based document images have used hand-crafted texture features and CNN are features but no work is reported on bi-script, tri-script, and multi-script identification from Indian camera-based document images using CNN. This has motivated to address above proposed problem.

3. DATASET DESCRIPTION

To carry out the proposed work, the own dataset has been constructed, due to the unavailability of a standard dataset based on camera-captured Indian document images.

The dataset contains 33000 (each script have 3000) camera captured printed document images belongs to 11 scripts as shown in below Fig.1. The document is captured by using a digital camera. Following Fig.1 shows the sample dataset used for the proposed experimental paragraphs must be indented and justified.

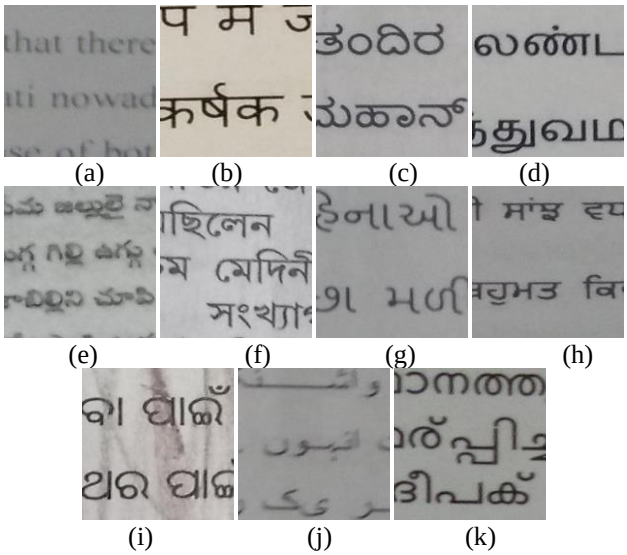

Fig.1. Sample input images (a) English (b) Hindi (c) Kannada (d) Tamil (e) Telugu (f) Bangla (g) Gujarathi (h) Punjabi (i) Oriya (j) Urdu (k) Malayalam

The proposed CNN model for identification of the scripts from the document images is explained through the following Fig.2.

4. PROPOSED METHOD

The entire manuscript should bein Times New Roman. Other font types may be used if required only for special purposes. Refer to Table.1 for font sizes.

In the CNN model, we have considered 5 convolution layers (3x3 filter size, 2,4,8, 16,32 filters with11stride) batch normalization layers5, Relu5 layers, 2 max pooling layers,1 fully connected, 1 softmax and lastly, classification layer as the class output.

Convolutional layers [13] [14] [15] uses the filter to generate a feature vector, that summarizes the existence of potential features in an input image. The filter should be smaller than the input image of size 3x3 in the experiment, and the dot product is used to multiply a filter-sized patch of the input with the filter. A dot product is the element- wise multiplication carried out in dot product of the input and filter's filter-sized patch, then the summation is obtained to yield a single value.

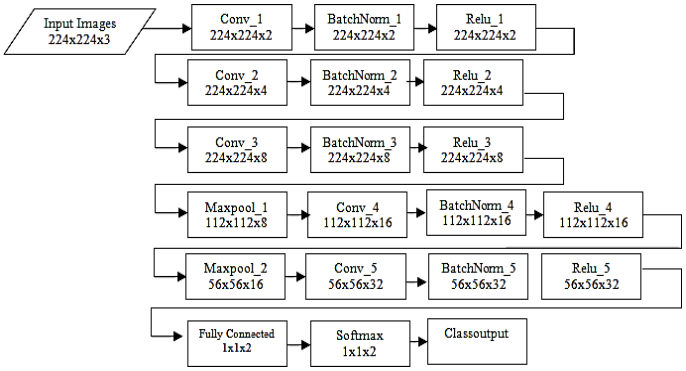

Fig.2. Proposed CNN Architecture

Batch normalization [16] [17] is a technique for standardizing the inputs to a network, which can be applied to the activations of a previous layer or directly to the inputs. Batch normalization reduces the generalization error by speeding up training (in some situations by halving or bettering the epochs) and providing some regularization.

In deep learning models, the Rectified Linear Unit (Relu) [18] [19] [20] is the most broadly used activation function. If the function receives any negative input, it returns 0; otherwise it returns the input value. When creating multilayer Perceptron and convolutional neural networks, the Relu is the default activation. It is written as $f(x)=\max(0,x)$.

The maximum element from the region of the feature map covered by the filter is selected using the max pooling [21] technique. As a result, the output from the max- pooling layer would be a feature map containing the most prominent features from the former feature map.

Feed forward neural networks are the Fully Connected Layer [22] [23]. Fully Connected Layer is the network's final layers. The output of the final Pooling or Convolutional Layer is flattened and sent to the fully connected layer as input of the fully connected layer. The final Pooling and Convolutional Layer produces a 3- dimensional matrix, which may be flattened by partitioning all of its values into a vector.

Softmax layer [24] [25] [26] is used for multi-classification issues. Before using softmax, certain vector components may be negative or higher than one, and they may not sum to one. The softmax layer produces a probability distribution, and the values of the output sum to 1. It is primarily used to normalize the output of neural networks such that it falls between zero and one. It is used to express the network output's certainty.

For classification and weighted classification problems with mutually exclusive classes, a classification layer computes the cross-entropy loss. The layer derives the number of classes from the previous layer's output size. Include a fully connected layer with output size K and a softmax layer before the classification layer, for example, to set the number of classes K of the network.

In this proposed experiment we have also considered the widely used model known as the alexnet pretrained CNN model [27] [28] . AlexNet is the name of a convolutional neural network that had a significant influence on machine learning, particularly in the application of deep learning to machine vision.

With these layers the model is designed and it has given a encouraging recognition accuracy with few layers and it is an light weight model to train the data of any size. It is suitable to train small dataset.

5. EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate the performance of the proposed method, an experiment is carried out on 33000 camera-based document images for 11 scripts (Kannada, English, Telugu, Tamil, Malayalam, Gujarathi, Bangla, Oriya, Hindi, Gujarathi, Punjabi, and Urdu). From the dataset 23100 images are used for training and 9900 used for testing. The input images were captured by digital Camera from various documents viz., Textbooks, Newspapers, Magazines, Novels, and computer-printed documents. To capture the hard document the 4920x3264 megapixel mobile camera is used. The 227x227 image block has been segmented from the camera-captured images. To check the performance of the proposed model and Alexnet a pretrained model the experiments are performed on the said dataset, and the recognition accuracies are obtained. In Table.2 contains the average recognition accuracy obtained for bi-script using Alexnet and proposed CNN models.

5.1 BI-SCRIPT IDENTIFICATION

The experimental results of the proposed CNN for script identification has shown 91.63% and 92.92% average recognition accuracies for Alexnet model and proposed CNN model respectively. The bilingual scripts combinations are formed from English with other 10 Indian scripts. The reason for keeping English script as constant is that, this script is used with almost all regional scripts. Following Table.1 presents the average recognition accuracy achieved from Alexnet and proposed models.

Table.1. Average recognition accuracy of Bi-script for SFTA, Alexnet and proposed CNN model

Script Combinations	SFTA	Alexnet	Proposed CNN Model
English-Hindi	71.38%	90.55%	93.80%
English-Kannada	78.67%	91.00%	93.40%
English-Telugu	70.22%	95.52%	97.40%
English-Tamil	75.18%	84.22%	83.80%
English-Malayalam	83.68%	85.96%	87.80%
English-Bangla	76.30%	94.55%	93.20%
English-Gujarathi	73.63%	92.65%	93.80%
English-Punjabi	91.62%	95.21%	97.80%
English-Oriya	84.50%	94.96%	93.80%
English-Urdu	82.30%	91.68%	94.40%
Total Average Recognition	78.75%	91.63%	92.92%

From the Table.1, it is clearly seen that the proposed model is better as compared to SFTA and Alexnet model with respect to recognition accuracy. In the SFTA, English with Punjabi scripts are obtained maximum of 91.62% accuracy. The English with Telugu scripts have gained the highest recognition accuracy of 95.52% from Alexnet model and 97.80% is obtained from English with Punjabi scripts from proposed model. In the forthcoming section the experiment is carried out on Tri-script identification to check the performance of the proposed model over SFTA and Alexnet model.

5.2 TRI-SCRIPT IDENTIFICATION

Extensive experiments are carried out to test the performance of the proposed model CNN model. For the experiment, the combinations of three scripts are formed such that the first two scripts viz., English and Hindi are kept fixed and third regional script keeps on changing. An average recognition accuracy of 72.40%, 90.08% and 91.33% is achieved from SFTA, Alexnet and proposed CNN models respectively. Following Table 2.presents the average recognition accuracy of all models.

Table.2. Average recognition of Tri-script by using SFTA, Alexnet and proposed CNN models.

Script Combinations	SFTA	Alexnet	Proposed CNN Model
Eng-Hin-Kan	66.96%	82.41%	85.33%
Eng - Hin - Tel	70.09%	92.61%	93.60%
Eng - Hin - Tam	67.77%	87.21%	89.60%
Eng - Hin - Mal	70.11%	89.99%	91.73%
Eng - Hin - Ban	67.65%	84.21%	85.47%
Eng - Hin - Guj	71.61%	90.54%	91.73%
Eng - Hin - Pun	82.28%	95.21%	95.33%
Eng - Hin - Ory	77.76%	94.00%	95.60%
Eng - Hin - Urd	77.34%	94.52%	93.60%
Total Average Recognition	72.40%	90.08%	91.33%

The Table.2 is shown the Tri-script identification using SFTA, Alexnet and proposed CNN models. The English, Hindi and Punjabi scripts have obtained highest recognition as 82.28% and 95.21% recognition accuracy from SFTA and Alexnet models. In proposed CNN model the English, Hindi and Oriya scripts are performed good by acquiring the highest of 95.60% recognition accuracy. From this it is clearly seen that the proposed CNN model has given highest recognition accuracy in the Tri-script environment.

5.3 MULTI-SCRIPT IDENTIFICATION

In this section an experiment carried out on multi-scripts. The 11 scripts are submitted to the alexnet and proposed models and obtained the average recognition of 85.98% and 87.33% from alexnet and proposed CNN models respectively. The Table.3 below presents the recognition accuracy achieved from the proposed models.

Table.3. Average recognition accuracy of Multi-scripts from SFTA, Alexnet and CNN models

From 11 Combined Scripts	SFTA	Alexnet	Proposed CNN Model
Eng-Kan-Tel-Tam-Mal-Ban-Hin-Ory-Pun-Guj-Urd	39.85%	85.98%	87.33%

The Table.3 presented the multi-script identification using SFTA, Alexnet and Proposed CNN model. In this multi-script environment the SFTA has shown poor performance by gaining only 39.85% recognition accuracy. The Alexnet has able to get 85.98% recognition accuracy, in this situation the proposed CNN model has performed better as compare to other two methods by obtaining the highest of 87.33% recognition accuracy

5.4 COMPARATIVE ANALYSIS

In the Table.4 below the comparison of the proposed CNN model with SFTA features Dhandra et al [4] and Alexnet is presented.

Table.4.Comparative analysis of proposed method with other methods

Script combinations	SFTA	Alexnet	Proposed (CNN model)
Bi-Script	78.75%	91.63%	92.92%
Tri-Scrpt	72.40%	90.08%	91.33%
Multi-Script	39.85%	85.9%	87.33%

From the Table.4 is observed that the proposed CNN model has has shown improved performance as compare to other two methods (SFTA and Alexnet).

Table.5. Comparative analysis of proposed CNN model performance on various standard datasets

Dataset	Type of Dataset	Scripts	Dataset Size	Recognition Accuracy
SIW-13	Natural Scene images	Ara, Cam, Chi, Eng, Gre, Heb, Jap, Kan, Kor, Mon, Rus, Thai, and Tib (13).	16291	71.23%
CVSI-2015	VideoScripts	Eng, Hin, Ben, Ori, Guj, Pun, Kan, Tam, Tel, and Ara (10)	10655	91.74%
OwnDataset	Camera-captured documents	Eng,Hin, Urd,Tel, Kan,Tam, Ori,Guj, Ban,Mal, And Punj (11)	33000	87.33%

Note: Ara->Arabic, Cam->Cambodian, Chi>Chinese, Eng-> English, Gre-.Greek, Heb->Hebrew, Jap->Japanese, Kan->Kannada, Kor->Korean, Mon->Mongolian, Rus->Russian, Thai, Tib-> Tibetan, Hin->Hindi, Ben->Bengali, Ory->Oriya, Guj-> Gujrathi, Pun->Punjabi, Kan-Kannada, Tam->Tam, Tel->Telugu, Mal-Malayalam

In this pipeline we have applied our proposed CNN model on other two standard datasets (SIW-13 and CVSI-2015). Hence, our model is performed good on CVSI-2015dataset by obtaining the 91.74% recognition accuracy, in the same way with SIW-13dataset our model has given 71.23% recognition accuracy. At last on our own dataset proposed CNN model has given 87.33% recognition accuracy.

In the ending of this discussion, we can say our developed CNN model have achieved good results, particularly in the multilingual script our model shown its best part by giving more than 85%. If it is properly fine tuned to the model then we can expect the increased recognition accuracy.

6. CONCLUSION AND FUTURE WORK

This article has presented a camera-based script identification at the word level by developing a new CNN model. A single unified model has been created and can identify scripts from bi-script, tri-script, and multi-script combinations. The different standard and own datasets were used to evaluate the performance of the proposed CNN model. To assess the performance of the proposed CNN model, the three different script combinations, bi-script, tri-script, and multi-script, are submitted to the proposed model. The results obtained from the three script combinations are 92.92%, 91.33%, and 87.33%, respectively, which are quite encouraging. It should be required to improve in recognition accuracy particularly multi-script combinations.

For future work, the following points are to be considered to improve the proposed model's overall performance.

- It requires an effort to develop a lightweight, robust model that can give higher recognition accuracy.
- Add more standard datasets to check the robustness of the proposed model.
- To make the proposed model stronger, the camera-based multilingual handwritten datasets must be created and applied.
- The proposed model should compare with other research work and different pre-trained CNN models..

REFERENCES

[1] D. Doermann and H. Li, "Progress in Camera-Based Document Images Analysis", *Proceedings of International Conference on Document Analysis and Recognition*, pp. 606-616, 2013.

[2] L. Li and C.L. Tan, "Script Identification of Camera-Based Images", *Proceedings of International Conference on Pattern Recognition*, pp. 1-4, 2008.

[3] G. Mukarambi, B.V. Dhandra and S. Mallappa. "Camera-based Bi-Lingual Script Identification at Word Level using SFTA Features", *International Journal of Recent Technology and Engineering*, Vol. 8, pp. 2988-2994, 2019.

[4] B.V. Dhandra, Satishkumar Mallappa and Gururaj Mukarambi, "Script Identification at Line-level using SFTA and LBP Features from Bilingual and Trilingual Documents Captured from the Camera", *International Journal of Computer Applications*, Vol. 13, No. 4, pp. 975-980, 2020.

[5] B.V. Dhandra, Satishkumar Mallappa and Gururaj Mukarambi, "Script Identification of Camera based Bilingual Document Images using SFTA Features", *International Journal of Human-Computer Interaction*, Vol. 15, pp. 1-12, 2019.

[6] B.V. Dhandra, Satishkumar Mallappa and Gururaj Mukarambi, "Script Identification from Camera based Tri-Lingual Document", *Proceedings of International Conference on Sensing, Signal Processing Security*, pp. 214-217, 2017.

[7] . B.V. Dhandra, Satishkumar Mallappa and Gururaj Mukarambi, "Camera-based Tri-Lingual Script Identification at Word Level using a Combination of SFTA and LBP Features", *International Journal of Advanced Science and Technology*, Vol. 29, No. 3, pp. 6609-6617, 2020.

[8] A.K. Bhunia, A. Konwer, A. Bhowmick, P.P. Roy, and U. Pal, "Script Identification In Natural Scene Image and Video Frames using an Attention based Convolutional-LSTM Network", *Pattern Recognition*, Vol. 85, pp. 172-184, 2019.

[9] M. Jajoo and R. Sarkar, "Script Identification from Camera-Captured Multi-script Scene Text Components", *Proceedings of International Conference on Recent Developments in Machine Learning and Data Analytics*, pp. 740-746, 2019.

[10] O.K. Fasil, S. Manjunath and V.N. Manjunath Aradhya, "Word-Level Script Identification from Scene Images", *Advances in Intelligent Systems and Computing*, Vol. 516, pp. 417-426, 2019.

[11] Xin Zhang, Yongcheng Wang, Ning Zhang, Dongdong Xu and Bo Chen, "Research on Scene Classification Method of High-Resolution Remote Sensing Images based on RFPNet", *Applied Sciences*, Vol. 67, No. 1, pp. 1-10, 2019.

[12] A.F. Costa, G. Humpire Mamani and A.J.M.H. Traina, "An Efficient Algorithm for Fractal Analysis of Textures", *Proceedings of International Conference on Computer Graphics and Image Processing*, pp. 39-46, 2012.

[13] Rikiya Yamashita, Mizuho Nishio, Richard Kinh, Gian Do and Kaori Togashi, "Convolutional Neural Networks: An Overview and Application in Radiology", *Insights Imaging*, Vol. 9, pp. 611-629, 2018.

[14] S. Albawi, T.A. Mohammed and S. Al-Zawi, "Understanding of a Convolutional Neural Network", *Proceedings of International Conference on Engineering and Technology*, pp. 1-6, 2017.

[15] X. Lu and Tzyy Chyang, "CNN Convolutional Layer Optimisation based on Quantum Evolutionary Algorithm", *Connection Science*, Vol. 33, No. 3, pp. 482-494, 2021.

[16] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", *Proceedings of International Conference on Machine Learning*, pp. 448-456, 2015.

[17] A. Fagbohungbe, and Lijun Qian, "Effect of Batch Normalization on Noise Resistant Property of Deep Learning Models", *Proceedings of International Conference on Machine Learning and Deep Learning*, pp. 1-12, 2022.

[18] S. Brownlee and A. Jason, "Gentle Introduction to the Rectified Linear Unit (ReLU)", *Proceedings of International Conference on Machine Learning*, pp. 1-6, 2019.

[19] Chaity Banerjee, Tathagata Mukherjee and Eduardo Pasiliao, "The Multi-Phase ReLU Activation Function", *Proceedings of International Conference on Computing*, pp. 1-7, 2020.