Journal of Ambient Intelligence and Humanized Computing (2021) 12:7997–8008
https://doi.org/10.1007/s12652-020-02528-4

ORIGINAL RESEARCH

Language identification from multi-lingual scene text images: a CNN based classifier ensemble approach

Neelotpal Chakraborty[1] · Soumyadeep Kundu[1] · Sayantan Paul[1] · Ayatullah Faruk Mollah[2] · Subhadip Basu[1] · Ram Sarkar[1]

Received: 19 February 2020 / Accepted: 5 September 2020 / Published online: 19 September 2020

Abstract
Since the past two decades, detecting text regions in complex natural images has emerged as a problem of great interest for the research fraternity. This is because these regions of interest serve as source of information that can be utilized for various purposes. However, these regions may contain texts in multiple languages. Hence, identifying the corresponding language of a detected scene text becomes important for further information processing. Language identification of the text, captured in a wild, is an extremely challenging research field in the domain of scene text recognition. In this paper, a deep learning-based classifier combination approach is proposed to solve the problem of language identification from multi-lingual scene text images. In this work, a minimalist Convolutional Neural Network architecture is used as the base model. Five variants of an input image—three different channels of RGB color model (i.e. R for red, G for green and B for blue) along with RGB itself, and grayscale image are passed through the base model separately. The outcomes of these five models are combined using the classifier combination approaches based on sum rule and product rule. Performances of the proposed model have been evaluated on some standard datasets like KAIST and MLe2e as well as in-house multi-lingual scent text dataset. From the experimental results, it has been observed that the proposed model outperforms some state-of-the-art methods considered here for comparison.

Keywords Multi-lingual · Language identification · Scene text · CNN · Classifier combination

1 Introduction

The domain of multi-lingual scene text analysis (Zhu et al. 2016) involves detecting texts in natural scene images and identifying the language of the detected text for further processing. This research area is of great interest within the research fraternity owing to potential applications in several areas like tour guide and assistance, information retrieval, language translation, and image to text conversion (Lin et al. 2019), etc. This wide scope of applications to satiate growing human needs thus necessitates the development of some robust techniques to detect texts in complex natural images (Saidane and Garcia 2007). Furthermore, a culturally diverse country like India where languages vary from region to region throws another challenge of depicting texts in multiple languages. People of one region may not know the languages of other regions. This eventually leads to the necessity of tackling the problem of multi-lingual scene texts (Zhu et al. 2016) where the language of the detected text can be recognized.

A text in a given image is usually identified as a set of strokes with certain geometry. Textual data may not be language-wise uniform in the wild, especially in Asian countries where scene texts of varying languages may co-occur (e.g. a street hoarding with texts in English language along with texts in native/local language). To recognize a text in image format, an Optical Character Recognition (OCR) (Lin et al. 2019) system is used which is usually language specific. Hence, to distinguish these scene texts based on language, the differences in stroke patterns need to be observed, for which certain feature descriptors find great usage.

The problem of language identification is filled with challenges like stroke pattern similarities among languages like

✉ Neelotpal Chakraborty
chakraborty.neelotpal@gmail.com

[1] Computer Science and Engineering Department, Jadavpur University, Kolkata, India

[2] Computer Science and Engineering Department, Aliah University, Kolkata, India

Bangla and Hindi, which many popular techniques fall quite short of resolving to the desired levels. Language-wise discrimination among scene texts is essentially a problem of multi-class classification where some methodologies, though few, involve training of certain state-of-the-art classification models. However, these models, no matter how much robust they are, face limitations, and their misclassification rates are quite high since they are usually trained with handcrafted feature descriptors that may fail at times to identify certain hidden or unobservable details that may be potent enough to reflect the distinction between similar languages.

On the other hand, recent approaches based on Convolutional Neural Network (CNN) (Saidane and Garcia 2007) have been quite successful in language identification of texts detected from complex wild images. CNN performs automatic feature extraction by self-learning, which eliminates the need for feature engineering. However, it also involves usage of highly parameterized neural networks, training of which are time consuming. The feature map reduction is done by selecting the filter maps of a convolutional layer of a CNN model. A few existing methods have been found on language identification of scene texts that applied various neural network models on datasets consisting of some European or Asian languages.

However, very few works exist where CNN based model is applied on scene texts written in Indic languages which motivates us to apply deep learning based strategy for language identification of scene texts obtained from an in-house dataset comprising of some Indic languages like Bangla and Hindi along with English, and some publicly available standard datasets. Also, it is observed that most CNN based systems face some limitations in accurately identifying the language of a text due to lack of component information resulting from scaling down in size of the text region. Scaling down of the image is often required to reduce the training time as well as hardware resource consumption of the CNN classifier. Existing models mostly work either on grayscale or RGB channel of any text region, thereby leading to ignoring certain hidden information from either channels that could have helped the model to improve its language identification performance. Another major disadvantage of CNN based models is that they need to be trained using datasets having a large number of instances which in the domain of scene text analysis is quite low. This limitation becomes more pertinent when we consider language identification from multi-lingual scene text images having regional languages, as required number of samples is hardly available for those languages in comparison with the popular languages like English and Chinese.

In this paper, an attempt is made to ameliorate these limitations of the existing CNN based systems by passing different channels R (red), G (green), B (blue), RGB and grayscale of an input text image individually to a basic CNN model which gets trained for each of these channels to determine the language class. Then, the class confidence scores obtained from each CNN model are combined using either sum rule or product rule of classifier combination mechanism (Tulyakov et al. 2008) to determine the final language class in which the input scene text belongs to. The proposed model proves to be more advantageous than some of the recent CNN based models as maximum possible text component information is retained despite the scale down in size of the input image, and the overall model gives better accuracy scores despite the low number of training instances for a given dataset. The proposed system is experimented on standard datasets like KAIST (Jung et al. 2011; Lee et al. 2010) and MLe2e (Gomez et al. 2017) and an in-house dataset where it has been observed that the proposed model performs satisfactorily and scores better than some existing methods in terms of classification accuracy while identifying the languages of the scene text images.

2 Related work

The domain of scene text analysis is extensively explored by a large number of researchers during the last two decades and some good number of standard methods have been developed that are fairly robust to detect text in the wild. But, when it comes to analyzing the multi-lingual aspect of the detected scene texts, comparatively few works have been proposed for language identification.

Several approaches for scene text detection rely on the concept of Maximally Stable Extremal Regions (MSER), which maps stable blobs having nearly the same intensities. To make this method more dynamic, the work by Chakraborty et al. (2018) performs histogram analysis of the input image and a number of sharp peaks is obtained which is used as delta threshold for detecting MSERs. The parameters are further analyzed by Panda et al. (2020) and their relationship with image dimensions and text size is determined. An adaptive stroke-based filter is developed by Paul et al. (2015) which dynamically resolves the minute variations of stroke width of a text object. An improvement is made using Fuzzy Distance Transform (FDT) in the work by Paul et al. (2019). The technique used by Dutta et al. (2019) analyses the homogeneity of the foreground by binning process and extracts text regions on the basis of their stability and pixel density, which may be considered as an alternative for MSER. A Support Vector Machine (SVM) based One Class Classifier (OCC) is developed by Mukhopadhyay et al. (2019) where scene text detection is projected as a one class problem since non-texts do not have any definite representation. (Dhar et al. 2020) have proposed a gradient morphology-based method to detect multi-lingual scene texts in low quality images where morphological operations

are applied on the grayscale channel of an input image to extract text components while filtering out spurious components. Recent years have also witnessed the use of various CNN models in multi-lingual scene text detection. (He et al. 2017) applied deep direct regression to models like Faster-RCNN and ResNet for detecting multi-oriented scene text regions in images. To address the multi-lingual aspect of the scene texts, (He et al. 2018) applied direct regression on better models like ResNet-50.

Latest works on language identification or recognition among multi-lingual scene texts obtained from complex natural images, are comparatively few in number, as researchers worldwide have newly developed interest in this problem area. The methodology proposed by Singh et al. (2016) extracts and utilizes mid-level features based on which scene texts are classified language-wise by using a linear SVM classifier. Here, Indic languages are considered along with its English counterpart. Scene texts in English and some Indic languages are also classified in the work proposed by Jajoo et al. (2019) where a comparison of accuracies for some standard classifiers is presented. The work by Liao et al. (2015) utilizes the differences in the connected component (CC) aspects between English and Chinese languages to train a basic SVM classifier for recognition purpose. A semi-Markov model is implemented in (Weinman et al. 2008) for detecting text in the wild. The common problem among such classification models is that their misclassification rates are quite hard to manage due to being trained with handcrafted feature descriptors, which occasionally fail to observe certain hidden patterns that may vary between similar languages.

Recent years have witnessed extensive usage of deep learning architectures like CNNs for different pattern recognition purposes. (Ul-Hasan et al. 2015) applied a CNN-LSTM based model for language identification on multi-lingual texts in document images. Similar approach is also followed by Nicolaou et al. (2016) to identify languages of handwritten texts in document images. A CNN based recurrent networked is built (Shi et al. 2016a, b) to skew correct oriented scene texts. A variant of CNN called Multi-Stage Spatially-sensitive Pooling Network (MSPN) is used by Shi et al. (2015) for utilizing the power of automatic feature learning in CNNs, while making it fit to the language identification problem. In the work by Shi et al. (2016a, b), a deep network comprising spatial transformer and sequence recognition networks is implemented. Language identification of scene text is attempted by Gomez and Karatzas (2016) using CNN based text line extraction and identifying languages of these text lines by training a Naive-Bayes Nearest Neighbour classifier. A CNN based model is implemented by Shi et al. (2016a, b) to discriminate among multiple European languages of scene texts. The problem of variable aspect ratio of scene texts is countered by patch-based classification proposed by Gomez et al. (2017) for preserving discriminative parts of the image. The method proposed by Xie et al. (2019) uses both CNN and Recurrent Neural Network (RNN) where CNN encodes images and the RNN generates character sequences. The work by Deng et al. (2019) which detects and links corners for identifying possible text locations, is also independent of variable aspect ratio and orientations due to the geometric adaptive-ness of its quadrilateral proposals. To detect and map curved texts, the work by Liu et al. (2019) proposes a polygon-based curved text detector independent of any empirical combination by using an RNN based architecture. Language identification is performed by the technique proposed by Bhunia et al. (2019) which employs a CNN-LSTM framework to dynamically weigh local and global features from scene images. A re-weighted tensor decomposition technique is proposed by George (2019) to detect and remove unnecessary sparse components of the video data to finally retain the text data. A tree-based system is developed by Singh et al. (2018) where a Multi-layer Perceptron (MLP) classifier is trained using log-Gabor features of handwritten text images for recognizing their corresponding languages. To recognize English characters in ancient historical documents, (Narayanan and Kasthuri 2020) have built a system where features based on Local Binary Patterns (LBP) from individual characters are fed to a Spatial Pyramid Matching (SPM) classifier for identifying the character class. An end-to-end system is developed by Saha et al. (2020) where text candidates are extracted from scene images using a combination of MSER and Generative Adversarial Network (GAN) and trained a CNN model to identify the corresponding languages of the detected scene texts. CNN model is also used for recognizing handwritten Arabic numerals in the method proposed by Ahamed et al. (2020). Channel-based analysis is also a popular approach adopted to address certain object detection problems. An automatic road sign detector is developed by Farhat et al. (2019) where the HSV color space is explored and morphological operations are performed to filter out the spurious regions while retaining certain regular shaped components that are potential road signs. A deep residual neural network is trained by Sheng et al. (2020) in order to analyze the input scene images and identify the content in them. Another content-based image retrieval system implemented by Kavitha and Saraswathi (2020) where satellite images are analyzed and clustered on the basis of fuzzy logic. Zhu et al. (2018) proposed a deep learning-based method where the deep features are extracted from the regions of interest in any given image for understanding the local geometric invariance as well as diminishing noisy information. Zdenek et al. (2017) combined triplets of local convolutional features with the classical

bag-of-visual-words model to identify the languages of scene texts. Language identification of scene texts is also done by use of transfer learning from pre-trained CNN models which is implemented by Tounsi et al. (2017) in order to address the issue of low training data availability. A fusion of local and global CNNs is achieved by Lu et al. (2019) where local features are analyzed in collaboration with the global features of an input image to recognize the corresponding language of any scene text.

The use of CNN and other deep learning-based approaches in most of these recent methods, has yielded to a significant improvement in the accuracy in identifying the language of the corresponding scene texts. However, these methods suffer from the following limitations:

1. Information loss resulting from scaling down in size of the text image. Scaling down of an input image is often essential in reducing training time as well as hardware resource consumption of a CNN based classification model. Existing models work either on grayscale or RGB channel of any text region, thereby ignoring certain hidden information from either channels that could have been useful to the model in improving its language identification capability.
2. A CNN based classifier usually learns from a large number of instances to successfully perform the classification task at hand. In the domain of multi-lingual scene text analysis, availability of such datasets having large number of text images of certain native languages is quite low. As a result, the CNN based models mentioned above suffer from this data deficiency, thereby leading to improper learning by the model which results in their classification performance falling short of achieving the desired outcome.

To address and resolve the above limitations of the existing CNN based systems, this paper proposes a relatively simple but effective novel CNN based model for language identification of the scene texts. In this method, different channels i.e. R, G, B, RGB and grayscale of an input text image are passed individually to basic CNN model that gives a class confidence score. The scores from each of these CNN models are combined using either sum rule or product rule of classifier combination approach to decide the final language class of the scene text. The proposed model proves to be more advantageous than the latest existing CNN-based models by achieving the following:

1. Maximum possible text component information is retained despite the scale down in size of the input image.
2. The overall model gives better accuracy scores despite the low number of training instances for a given multi-lingual scene text dataset.

3 Proposed approach

The scene text language identification method proposed in this work is based on the encapsulation of deep learning (LeCun et al. 2015) and classifier combination (Tulyakov et al. 2008) approaches. Initially, information from the channels R, G, B, RGB and grayscale is obtained from an input image and individually fed to a basic CNN model. The confidence scores representing the predicted class probability from every CNN model are then fed to the classifier combination mechanism which gives the final outcome i.e. the language class to which the input scene text belongs. In this method, each of the models learns about different properties of input images and then these trained models get combined in order to enhance the performance of the overall system as shown in Fig. 1, which is beyond to the reach of individual models. The processes involved in designing the overall models are discussed individually in the following sub-sections.

3.1 Channel analysis

The input images of the dataset are color images of dimensions 48×64 pixels. Intensity maps extracted from an input image are separated into three channels R, G and B, each of

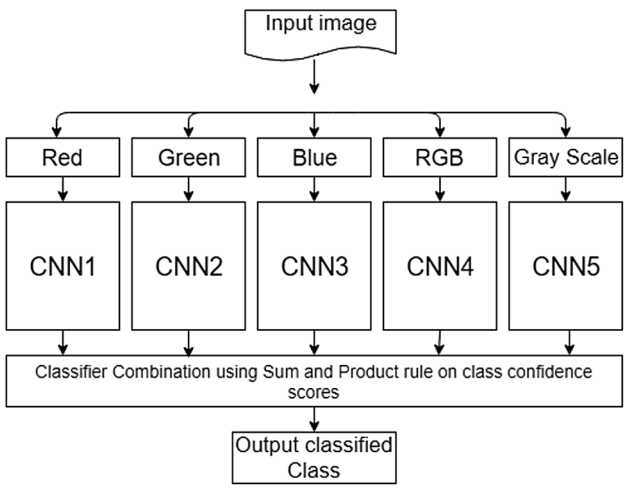

Fig. 1 Framework of the deep learning-based classifier combination model for language identification of scene texts, where information of input image from different color channels individually as well as combined along with grayscale is separately passed through a CNN classifier and gets combined using sum rule and product rule based classifier combination approaches

which is fed as input to separate CNN models along with RGB image and its grayscale version.

It is observed that information obtained from R, G and B channels separately, at times, becomes more meaningful than what is obtained from only RGB channel. This is due to possible suppression of information when either R, G and B channels remain combined or due to different lighting environment or background information. Some of the existing methods utilize grayscale information of an input image for feature extraction in order to train some classifier for language classification from scene text images. This grayscale information is the diagonal value representing intensity between the complete dark and complete bright as shown in the color model in Fig. 2.

However, certain deficiencies in accurate prediction of the language class from a complex scene image emerge in such systems. This happens because in the process of converting a color image to its grayscale version, the originally existing 3D map gets reduced to its equivalent 2D map. As a result, there remains a high chance of loss in 3D information that could have helped in better language identification. At the same time, grayscale channel of the input image may contain some useful information crucial enough for ensuring proper classification. Hence, to prevent any kind of suppression or loss in information while processing any input image, a strategy is adopted where all these five input modalities are considered in order to provide maximum information to the CNN models while training them. The distinctive information about the image from various modalities, as shown in Table 1, also helps the CNN model to find out the edge maps more soundly.

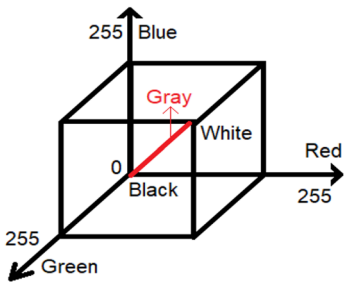

Fig. 2 A simple representation of color model. The grayscale information is represented in red (color figure online)

3.2 CNN architecture

A total of five models, all having the same architecture but with different inputs as R, G, B, RGB and grayscale, are trained. The base model on which these images are trained is a CNN based architecture (LeCunn et al. 1998) with just two convolution cum pooling layers. The main reason behind using such a simplistic base model is to make our system less resource intensive. The base model used in our work is illustrated in Fig. 3.

Each model consists of two convolutional layers. The input to the first layer being the input image or its individual channel. The first convolutional layer has a filter of size 3×3, with zero padding of size 1, stride of 1 position and 64 convolutional filters having Rectified Linear Unit (ReLU) as the activation function. It is succeeded by an average pooling layer of dimensions 2×2 with stride 2, such that the output dimensions are halved and we can extract higher order features from the respective input images, resulting in an output vector with dimensions of $24\times32\times64$. The output is fed to a second convolutional layer with same parameters as the first but with the exception of 128 convolutional filters. Another average pooling layer is added after the second convolutional layer reducing the dimensions to $12\times16\times128$. The output is multiplied by a weight matrix, and a bias matrix is added to give an output matrix containing the respective probability scores for the different classes. This last layer corresponds to an MLP with zero hidden layer. For quick reference, the summary of the CNN architecture used in this work is given in Table 2.

Methods based on deep learning approaches usually require two things: (1) adequate hardware resources, and (2) significantly large number of training instances. These are required for ensuring proper training of the CNN based classifier so that the trained model is able to perform well on the unseen data with maximum possible accuracy. While performing classification among images, the size of an image plays a significant role in training a CNN model. The more the size of input image, the more the hardware resource and time consumed during training. Hence, scaling down in size is done up to an optimal level so as to minimize this resource and time consumption. However, this scaling down usually causes significant loss of information that may be useful in maximizing the classification accuracy of the CNN. In the domain of multi-lingual scene text analysis, most of the

Table 1 Distinct information obtained from different color channels individually and combined, for a sample image. This unique information retaining capability of different channels is utilized separately

RGB	Red	Green	Blue	Grayscale
				

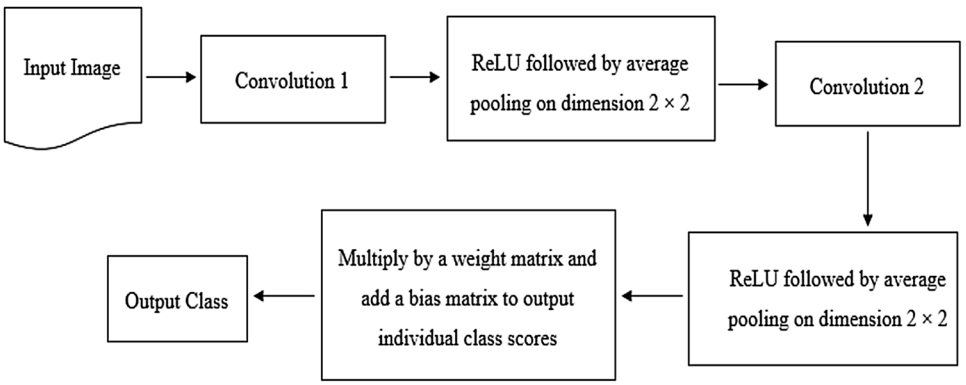

Fig. 3 A simple CNN based architecture used for language-wise classification of scene texts

Table 2 Summary of the CNN architecture used in the proposed method

Input	48 × 64 × 1 image patch for R, G, B and grayscale 48 × 64 × 3 image patch for RGB
Convolution 1	64 channels, 3 × 3 filter, zero padding of size 1, stride 1, ReLU
Average pooling 1	pool size 2 × 2, stride 2
Convolution 2	128 channels, 3 × 3 filter, zero padding of size 1, stride 1, ReLU
Average pooling 2	pool size 2 × 2, stride 2
Fully connected layer 3	Number of classes, MLP, zero hidden layers

datasets that are publicly available have a certain number of images that are usually not sufficient for adequate training of a CNN classifier. As a result, the classification performance of the CNN based models falls short of the desired outcomes. To address these adversaries, the proposed method includes training a simple CNN model for each channel such that maximum information is processed while ensuring that less time and hardware resources are consumed in order to adequately accomplish the training procedure.

3.3 Classifier combination

Classifiers combination can be carried out at three different levels: at the data level (early combination), at feature level, or at the decision level (late combination) (Mohandes et al. 2018) Data level fusion implies combination of data having different sources. Features level combination indicates simple concatenation of feature attributes having equal or varying weights. Hence, sometimes the resultant feature vector become very high dimensional, and it requires some feature dimension reduction algorithms. This mechanism is applied at the decision level in this work.

Most classifier combination methods work at decision level which follow any of the three approaches: abstract, rank, and score (Mukhopadhyay et al. 2018). In abstract approach, a single output label from each of the individual classifiers is applied as input to the grouping scheme. In case of rank-based approaches, the information used for the combination is various labels ranked from most likely to least likely produce by each classifier. Whereas in the score-based approach, each classifier gives n best labels along with their confidence scores which are then combined by using various simple to intelligent ways. In this work, for identifying the languages of the scene texts, we have also used the concept of late combination and as a classification model we have used a simple CNN architecture fed with different variants of the input images.

Some techniques are enlisted by Kittler et al. (1996) for combining classifiers using Bayes theorem such as product rule, sum rule, majority voting etc. In the proposed work, the sum rule and product role are employed. Let us consider N classifiers, K classes ($t_1, t_2, t_3, \ldots, t_K$) and a feature V that generates the feature vector a_i for classifier$i$. For a class j generated by ith classifier, posterior probability is represented as $p(t_j|a_i)$.

3.3.1 Product rule

It assigns $V \rightarrow t_j$ if the Eq. (1) holds. Product rule is very sensitive towards low probability value.

$$p^{1-N}(a_j)\prod_{i=1}^{N} p(t_j|a_i) = \max_m \left[p^{1-N}(t_m)\prod_{i=1}^{N} p(t_m|a_i) \right] \quad (1)$$

3.3.2 Sum rule

In case of sum rule, we assign $V \rightarrow t_j$ when the given relation in Eq. (2) becomes true. Sum rule provides average

performance in the cases where majority does not provide accurate result.

$$(1-N)p(t_j) + \sum_{i=1}^{N} p(t_j|a_i) = \max_m[(1-N)p(t_m) + \sum_{i=1}^{N} p(t_m|a_i)] \quad (2)$$

The outputs obtained from each CNN function is combined using classifier combination. The performance of the whole framework is evaluated for sum rule and product rule separately.

4 Experimental results and discussion

Experiments are performed on standard as well as in-house datasets to evaluate the performance of the CNN based classifier combination process for sum rule and product rule separately. The following sub-sections elaborate on the experimentation.

4.1 Datasets

The proposed method for scene text language identification is experimented on the KAIST (Jung et al. 2011; Lee et al. 2010) and MLe2e (Gomez et al. 2017) datasets. Also, we have provided detailed results on an in-house multi-lingual dataset of scene text images containing Indic languages like Bangla, Hindi and English. The overall distribution of images in the datasets are depicted in Table 3.

From the KAIST dataset, we crop 4760 word-level images. Of these, 3102 are written in English language, and 1658 are written in Korean language. Out of these images, we randomly choose 3500 word images to form the training set and the rest 1260 are considered as part of the test dataset. The MLe2e dataset is split into 450 and 261 images, representing the training and test datasets respectively. The in-house dataset consists of 200 images of words for each of the three languages: English, Bangla and Hindi, thereby leading to a total of 600 word images. For data augmentation, the images are rotated by 5, 10, −5 and −10 degrees, which increases the dataset size to 1000 images per class, totaling to 3000 images. Some sample images from all the datasets are shown in Table 4.

Table 3 Distribution of word images in different datasets

Dataset	# Images	# Word images		Languages
		Train set	Test set	
KAIST	3000	3500	1260	English, Korean
MLe2e	711	1178	643	English, Korean, Kannada, Chinese
In-house	200	2100	900	English, Bangla, Hindi

4.2 Experimental setup

The model is tested on an Intel Xeon CPU E5-1630 v4 @ 3.70 GHz with RAM of 32 GB and 8 GB Nvidia M4000 GPU. The model is built with TensorFlow (Abadi et al. 2016). The images are resized to 48 × 64 pixels and the number of features in the two convolution layers are 64 and 128 respectively. For the convolution layers, 2D convolutions with a small kernel size are considered. A max pooling strategy is considered and cross entropy is used as the loss function, along with Adam Optimizer (Kingma and Ba 2014) with an experimentally acquired learning rate. The models are trained till convergence.

4.3 Results

A CNN based architecture is used for language identification in multi-lingual scene text images and trained the model on 5 different types of input variants namely R, G, B, RGB and grayscale taken from the same image source. The sum rule and product rule techniques are applied upon the implemented using the outputs generated from the

Table 4 Sample images of scene words in various languages from different datasets

In-house			MLe2e				KAIST	
Bangla	English	Hindi	English	Chinese	Kannada	Korean	English	Korean

Table 5 Accuracies in (%)age obtained for CNNs from different channels which are combined using sum and product rules. Product rule and sum rule give almost comparable results and the classifier combination techniques improves the overall accuracies by a significant amount

Dataset	Color Channels					Classifier Combination rules	
	Red (%)	Blue (%)	Green (%)	RGB (%)	Gray scale (%)	Sum rule (%)	Product rule (%)
KAIST	92.82	91.06	92.39	93.90	93.24	96.19	**96.29**
MLe2e	91.71	92.59	92.98	94.27	93.08	96.88	**97.02**
In-house	83.11	85.66	85.44	86.00	85.22	**91.55**	91.33

The highest performance scores are marked in bold

above-mentioned models for classifier combination. The accuracies obtained by CNN models of each channel and the final accuracy scores achieved by the classifier combination techniques on the three above mentioned datasets are presented in Table 5.

The sum-rule has given the highest accuracy for in-house images, while the product-rule has given best results for KAIST and MLe2e datasets. A comparison with the present state-of-the-art techniques for each of these datasets are given in Table 6.

For in-house dataset, the test set contains 900 images (300 belonging to each language). The performance of the classifier combination with respect to each language class can be visualized from the confusion matrices corresponding to each of the results provided in Fig. 4.

Table 6 Comparison of scene text language identification performance of the proposed method with some state-of-the-art methods over different datasets. Accuracies are given in (%)age

Dataset	Method	Accuracy (%)
KAIST	Faster-RCNN (He et al. 2017)	88.00
	ResNet-50 (He et al. 2018)	87.00
	(Bhunia et al. 2019)	85.00
	(Saha et al. 2020)	95.00
	Proposed—sum rule	96.19
	Proposed—product rule	**96.29**
MLe2e	(Gomez and Karatzas 2016)	91.12
	(Gomez et al. 2017)	94.40
	(Zdenek and Nakayama 2017)	94.08
	(Tounsi et al. 2017)	94.30
	(Lu et al. 2019)	95.80
	(Bhunia et al. 2019)	96.70
	Proposed—Sum rule	96.88
	Proposed—Product rule	**97.02**
In-house	(Jajoo et al. 2019)	90.00
	(Saha et al. 2020)	**95.00**
	Proposed—Sum rule	91.55
	Proposed—Product rule	91.33

The highest performance scores are marked in bold

The change in the test set accuracy with epochs is shown as graphical representation for in-house dataset for each of the five different CNN based models in Fig. 5.

4.4 Discussion and error analysis

From the results obtained after experimentation, it is observed that CNN based classifier combination performs better than some of the state-of-the-art methods. Most of the existing methods rely on deep learning models which are trained on images with reduced map size, thereby leading to significant loss of information. Due to this deficiency, these CNN-based models suffer from inadequate training resulting in to an increased rate of misclassification. In the proposed work, although the CNNs are trained using images with reduced size, incorporation of multiple channels (one channel for each CNN classifier) of the image leads to maximization of information content that proves to be useful in enhancing the classification accuracy. The product rule applied on the confidence scores from every CNN classier achieves better accuracies than the sum rule in case of KAIST and MLe2e. However, in case of in-house dataset, the sum rule performs slightly better by achieving 0.22% more than what product rule has achieved.

Despite the high performance of the proposed model, it faces certain limitations which eventually result into scene text regions falling under wrong language classes. This is due to the fact that although information obtained from different channels are considered, the inter-channel relation remains unutilized. As a result, the response of every CNN for particular channel, varies with each other thereby giving rise to a situation where a confusion in language determination of a scene text may arise. Also stroke similarities among certain languages like Bangla and Hindi, or Korean and Chinese may lead to misclassification. Another issue that may cause trouble is the quality of the input image. Images in the datasets considered, are mostly captured using some handheld devices which may not be equally good in

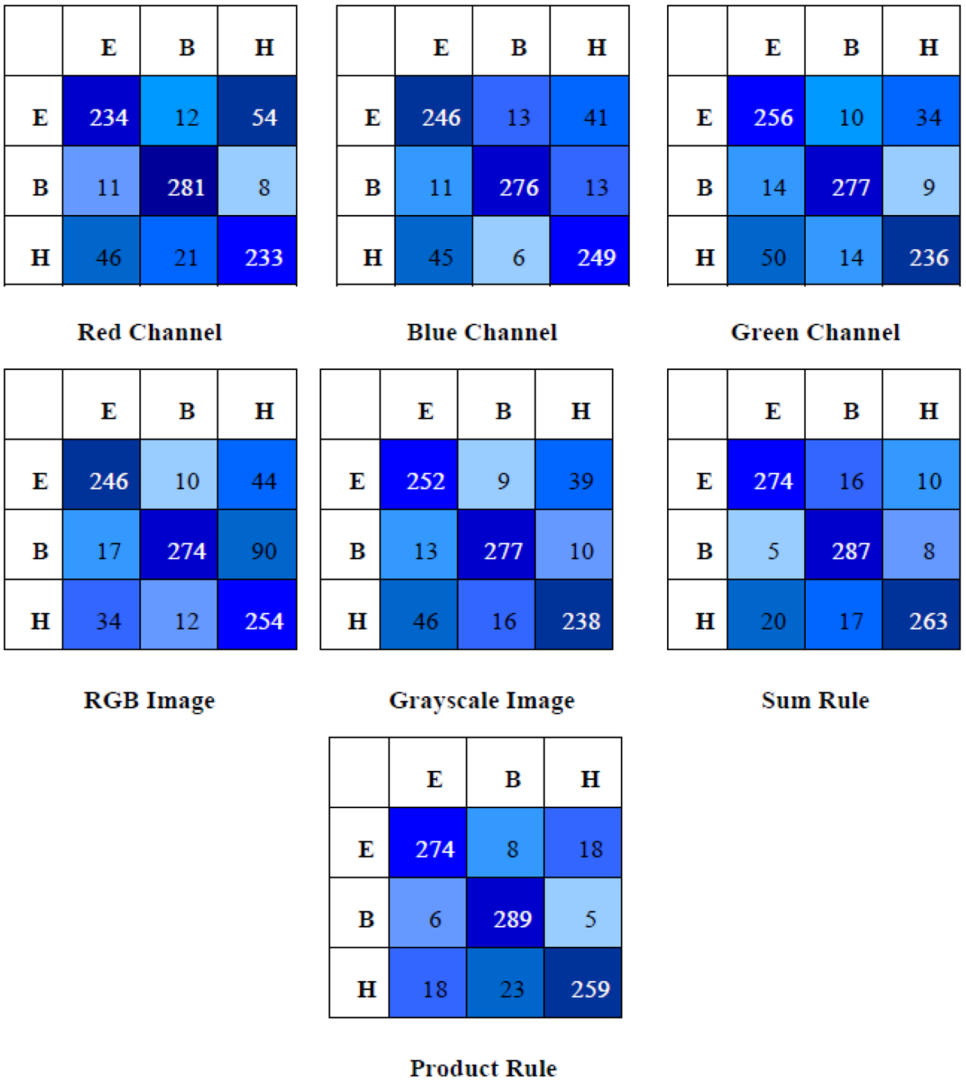

Fig. 4 The confusion matrices for the classification results on the test set of 300 images per language for in-house dataset. Here, *E*, *B*, *H* represents English, Bangla and Hindi, respectively. Performance evaluation is done for sum rule and product rule separately. The method based on sum rule performs slightly better than that using product rule

quality. This leads to capturing of images which may contain blur and noise. Also, irregular intensity distribution due to flash or shadow can diminish the system performance. Some of the images for which the method could not perform well are depicted in Fig. 6.

5 Conclusion

In this paper, multiple two-layers CNN architectures are used with different inputs taken from the same source image and then the outputs of these models are combined to address the problem of language identification in multi-lingual scene text word images. This is a necessary step in the domain of scene text analysis and recognition when the environment is multi-lingual. For this, two classifier combination approaches are used to ensemble the results of five base CNN models, each trained on different type of input images. The ensemble techniques are able to increase the performance by a significant margin. For in-house images, the highest accuracy achieved is 91.55% using sum rule, and for KAIST and MLe2e 96.29% and 97.02% respectively using product rule. In future, our aim is to amalgamate the current method in end-to-end system development for scene text understanding. Also, plans are being considered to

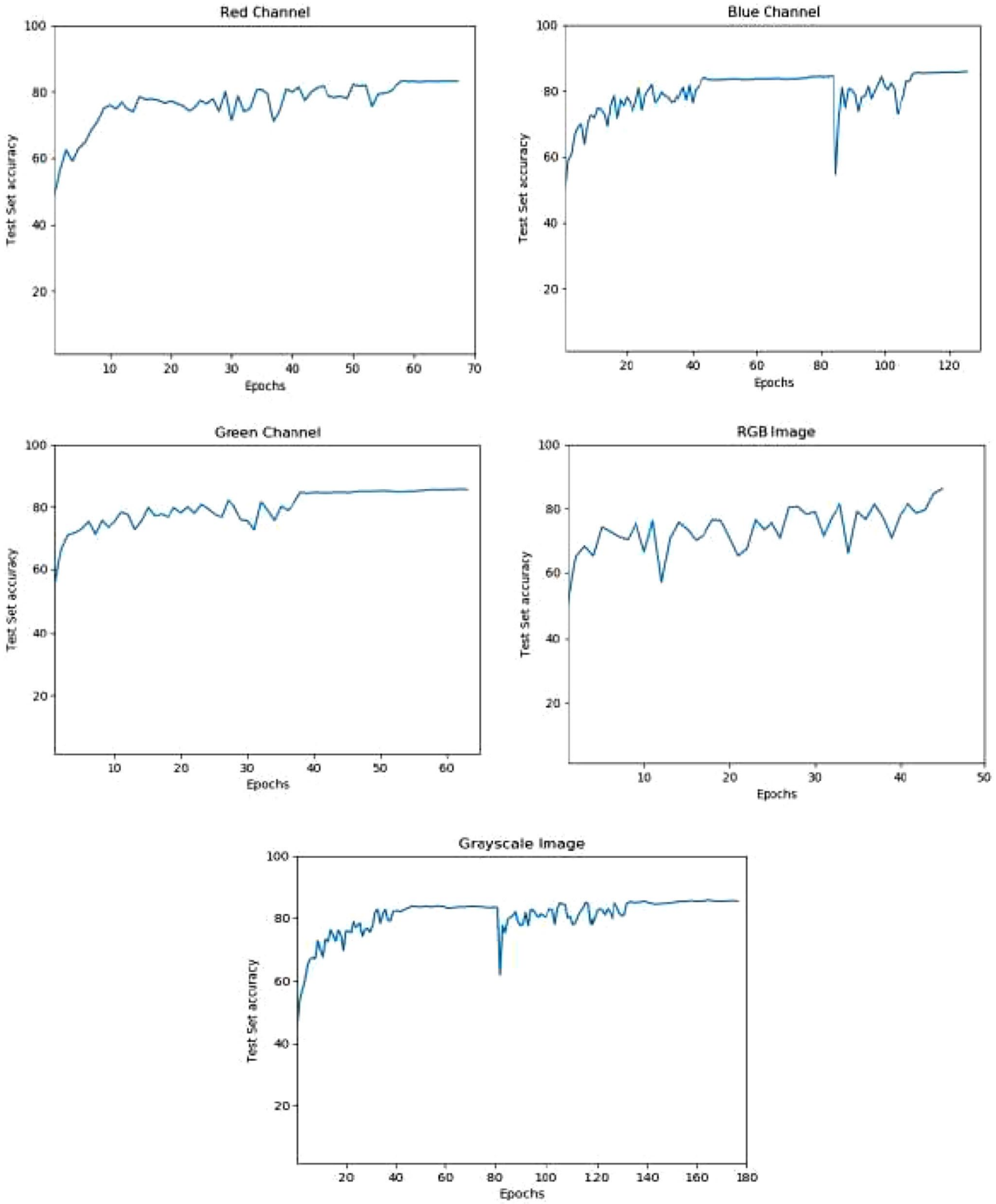

Fig. 5 Graphs showing the relation between test set accuracy and number of epochs. The classifier combination method either by sum rule or product rule perform significantly better since information from various channels individually and combined are separately processed. The CNN response for every channel considered, is combined using either sum rule or product rule to decide the class of a scene text

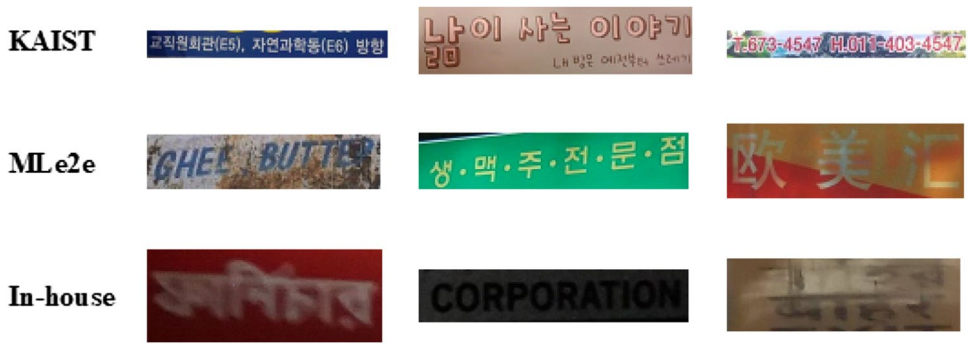

Fig. 6 Images for which the method could not perform well. Possible reasons for failure are: uncertainty over inter-channel relation, blur or noise in images, irregular intensity distribution due to flash or shadow, and stroke similarities among certain languages like Bangla and Hindi, or Korean and Chinese

incorporate other probabilistic classifier combination techniques in future.

Acknowledgement This work is partially supported by the CMATER research laboratory of the Computer Science and Engineering Department, Jadavpur University, India, PURSE-II and UPE-II, project.

Funding This work is partially funded by DBT Grant (BT/PR16356/BID/7/596/2016), UGC Research Award (F.30–31/2016(SAII)) and DST Grant (EMR/2016/007213).

Compliance with ethical standards

Conflict of Interest There is no conflict of interest.

References

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Kudlur M (2016). Tensorflow: a system for large-scale machine learning. In: 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16) (pp. 265–283).

Ahamed P, Kundu S, Khan T, Bhateja V, Sarkar R, Mollah AF (2020) Handwritten Arabic numerals recognition using convolutional neural network. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-020-01901-7

Baburaj M, George SN (2019) Tensor based approach for inpainting of video containing sparse text. Multim Tools Appl 78(2):1805–1829

Bhunia AK, Konwer A, Bhunia AK, Bhowmick A, Roy PP, Pal U (2019) Script identification in natural scene image and video frames using an attention based convolutional-LSTM network. Pattern Recogn 85:172–184. https://doi.org/10.1016/J.PATCOG.2018.07.034

Chakraborty N, Biswas S, Mollah AF, Basu S, Sarkar R (2018) Multilingual scene text detection by local histogram analysis and selection of optimal area for MSER. In: International Conference on Computational Intelligence, Communications, and Business Analytics (pp. 234–242). Springer, Singapore.

Deng L, Gong Y, Lin Y, Shuai J, Tu X, Zhang Y, Xie M (2019) Detecting multi-oriented text with corner-based region proposals. Neurocomputing 334:134–142

Dhar D, Chakraborty N, Choudhury S, Paul A, Mollah AF, Basu S, Sarkar R (2020) Multilingual scene text detection using gradient morphology. Int J Comput Vis Image Process (IJCVIP) 10(3):31–43

Dutta IN, Chakraborty N, Mollah AF, Basu S, Sarkar R (2019) Multilingual text localization from camera captured images based on foreground homogenity analysis. In: Recent Developments in Machine Learning and Data Analytics (pp. 149–158). Springer, Singapore.

Farhat W, Sghaier S, Faiedh H, Souani C (2019) Design of efficient embedded system for road sign recognition. J Ambient Intell Hum Comput 10(2):491–507

Gomez L, Karatzas D (2016) A fine-grained approach to scene text script identification. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS) (pp. 192–197). IEEE.

Gomez L, Nicolaou A, Karatzas D (2017) Improving patch-based scene text script identification with ensembles of conjoined networks. Pattern Recogn 67:85–96

He W, Zhang XY, Yin F, Liu CL (2017) Deep direct regression for multi-oriented scene text detection. In: Proceedings of the IEEE International Conference on Computer Vision (pp. 745–753).

He W, Zhang XY, Yin F, Liu CL (2018) Multi-oriented and multi-lingual scene text detection with direct regression. IEEE Trans Image Process 27(11):5406–5419

Jajoo M, Chakraborty N, Mollah AF, Basu S, Sarkar R (2019) Script identification from camera-captured multi-script scene text components. In: Recent Developments in Machine Learning and Data Analytics (pp. 159–166). Springer, Singapore.

Jung J, Lee S, Cho MS, Kim JH (2011) Touch TT: Scene text extractor using touchscreen interface. ETRI J 33(1):78–88

Kavitha PK, Saraswathi PV (2020) Content based satellite image retrieval system using fuzzy clustering. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-020-02064-1

Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv:1412.6980.

Kittler J, Hater M, Duin RP (1996) Combining classifiers. In: Proceedings of 13th international conference on pattern recognition (vol. 2, pp. 897–901). IEEE.

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436

Lee S, Cho MS, Jung K, Kim JH (2010) Scene text extraction with edge constraint and text collinearity. In: 2010 20th International Conference on Pattern Recognition (pp. 3983–3986). IEEE.

Liao WH, Liang YH, Wu YC (2015) An integrated approach for multilingual scene text detection. In: 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR) (pp. 211–217). IEEE.

Lin H, Yang P, Zhang F (2019) Review of Scene Text Detection and Recognition. Arch Comput Methods Eng 27:1–22

Liu Y, Jin L, Zhang S, Luo C, Zhang S (2019) Curved scene text detection via transverse and longitudinal sequence connection. Pattern Recogn 90:337–345

Lu L, Yi Y, Huang F, Wang K, Wang Q (2019) Integrating local CNN and global CNN for script identification in natural scene images. IEEE Access 7:52669–52679

Mohandes M, Deriche M, Aliyu SO (2018) Classifiers combination techniques: a comprehensive review. IEEE Access 6:19626–19639

Mukhopadhyay A, Singh P, Sarkar R, Nasipuri M (2018) A study of different classifier combination approaches for handwritten Indic Script Recognition. J Imag 4(2):39

Mukhopadhyay A, Kumar S, Chowdhury SR, Chakraborty N, Mollah AF, Basu S, Sarkar R (2019) Multi-Lingual scene text detection using one-class classifier. Int J Comput Vis Image Process (IJCVIP) 9(2):48–65

Narayanan VS, Kasthuri N (2020) An efficient recognition system for preserving ancient historical documents of English characters. J Ambient Intell Hum Comput 15:1–9

Nicolaou A, Bagdanov AD, Gómez L, Karatzas D (2016) Visual script and language identification. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS) (pp. 393–398). IEEE.

Panda S, Ash S, Chakraborty N, Mollah AF, Basu S, Sarka, R (2020) Parameter tuning in MSER for text localization in multi-lingual camera-captured scene text images. In: Computational Intelligence in Pattern Recognition (pp. 999–1009). Springer, Singapore.

Paul S, Saha S, Basu S, Nasipuri M (2015) Text localization in camera captured images using adaptive stroke filter. In: Information Systems Design and Intelligent Applications (pp. 217–225). Springer, New Delhi.

Paul S, Saha S, Basu S, Saha PK, Nasipuri M (2019) Text localization in camera captured images using fuzzy distance transform based adaptive stroke filter. Multim Tools Appl. https://doi.org/10.1007/s11042-019-7178-3

Saha S, Chakraborty N, Kundu S, Paul S, Mollah AF, Basu S, Sarkar R (2020) Multi-lingual scene text detection and language identification. Pattern Recog Lett. https://doi.org/10.1007/s11042-019-7178-3

Saidane Z, Garcia C (2007) Automatic scene text recognition using a convolutional neural network. In: Workshop on Camera-Based Document Analysis and Recognition (vol. 1).

Sheng F, Zhang Y, Shi C, Qiu M, Yao S (2020) Xi'an tourism destination image analysis via deep learning. J Ambient Intell Hum Comput 18:1–10

Shi B, Yao C, Zhang C, Guo X, Huang F, Bai X (2015) Automatic script identification in the wild. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR) (pp. 531–535). IEEE.

Shi B, Bai X, Yao C (2016a) Script identification in the wild via discriminative convolutional neural network. Pattern Recogn 52:448–458

Shi B, Wang X, Lyu P, Yao C, Bai X (2016b) Robust scene text recognition with automatic rectification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4168–4176).

Singh AK, Mishra A, Dabral P, Jawahar CV (2016) A simple and effective solution for script identification in the wild. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS) (pp. 428–433). IEEE.

Singh PK, Sarkar R, Bhateja V, Nasipuri M (2018) A comprehensive handwritten Indic script recognition system: a tree-based approach. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-018-1052-4

Tounsi M, Moalla I, Lebourgeois F, Alimi AM (2017) CNN based transfer learning for scene script identification. In: International Conference on Neural Information Processing (pp 702–711). Springer, Cham.

Tulyakov S, Jaeger S, Govindaraju V, Doermann D (2008) Review of classifier combination methods. In: Machine learning in document analysis and recognition (pp 361–386). Springer, Berlin, Heidelberg.

Ul-Hasan A, Afzal MZ, Shafait F, Liwicki M, Breuel TM (2015) A sequence learning approach for multiple script identification. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR) (pp. 1046–1050). IEEE.

Weinman JJ, Learned-Miller E, Hanson A (2008) A discriminative semi-Markov model for robust scene text recognition. In: 2008 19th International Conference on Pattern Recognition (pp. 1–5). IEEE. https://doi.org/10.1109/ICPR.2008.4761818

Xie H, Fang S, Zha ZJ, Yang Y, Li Y, Zhang Y (2019) Convolutional Attention Networks for Scene Text Recognition. ACM Trans Multim Comput Commun Appl (TOMM). https://doi.org/10.1145/3231737

Zdenek J, Nakayama H (2017) Bag of local convolutional triplets for script identification in scene text. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 369–375). IEEE.

Zhu Y, Yao C, Bai X (2016) Scene text detection and recognition: recent advances and future trends. Front Comput Sci 10(1):19–36

Zhu X, Wang Q, Li P, Zhang XY, Wang L (2018) Learning region-wise deep feature representation for image analysis. J Ambient Intel Hum Comput. https://doi.org/10.1007/s12652-018-0894-0