Received: 6 June 2024 | Revised: 12 November 2024 | Accepted: 31 December 2024 | Published online: 24 January 2025

RESEARCH ARTICLE

Artificial Intelligence and Applications
2025, Vol. 00(00) 1–8
DOI: 10.47852/bonviewAIA52023569

BON VIEW PUBLISHING

A New Hybrid Wavelet Decomposition-based Networks for Script Identification in Scene Images

Shivakumara Palaiahnakote[1,*], **Umapada Pal**[2] **and Taha Mansouri**[1]

[1] *School of Science, Engineering and Environment, University of Salford, UK*

[2] *Computer Vision and Pattern Recognition, Indian Statistical Institute, India*

Abstract: Script identification is challenging because of the unpredictable nature of the scene text. This paper presents a new model for achieving accurate script identification irrespective of intra and inter-class variations. The distinct features that represent the scene text of different scripts uniquely are extracted by fusing inception, which captures multi-scale features, and dense network, which captures fine-grained features. To strengthen the feature extraction, the proposed work uses wavelet decomposition, which enhances the fine details like edges in the images. Furthermore, for extracting text style, we propose a soft style attention module, which captures the unique style of scene text. The above modules are integrated as a hybrid model for accurate script identification. To evaluate the proposed model, we conducted comprehensive experiments on benchmark datasets, namely CVSI2015, SIW-13, and MLe2e, and combined datasets (combining distinct classes of all three benchmark datasets). The results of the proposed model on different datasets show that the performance is superior to the state-of-the-art methods in terms of accuracy.

Keywords: deep learning, dense network, inception architecture, attention model, scene text, script identification

1. Introduction

Script identification for improving the performance of optical character recognition is important for several applications because developing successful universal OCR is not feasible [1–3]. The reason is that the nature of the scene text of different scripts is unpredictable. In addition, the number of scripts is not finite. Furthermore, the number of applications of OCR increases drastically due to digitization and its importance to local languages. There are methods for script identification as it is not a new challenge for document image analysis [2, 3]. However, the problem of script identification is considered an open challenge. The key reason is that the existing methods are limited to several scripts, and hence, the methods work well for particular datasets. Therefore, the methods are not consistent for different datasets. Thus, there is a scope for developing a new model, which can be consistent for different datasets and achieve accurate results.

To address the challenges of script identification, inspired by the accomplishments of the recent deep learning approaches, we explore fusing different deep learning models, namely dense and inception networks for extracting distinct features. This is because the dense network is good for extracting fine-tuned features while the inception network is good for extracting multi-scale features [4, 5]. Since the problem is complex and the work aims to achieve the best results, the features extracted using the above network combination are insufficient. Therefore, to strengthen the discriminative ability of the features, motivated by the edge enhancement ability of the wavelet decomposition, we propose to integrate wavelet decomposition with the network for feature extraction [6]. In the same way, inspired by the attention models that focus on dominant information in the images, we explore soft style attention module or extracting text style features [7]. Overall, the proposed work fuses the strengths of the above modules to achieve the best results for script identification.

The key contributions are as follows. (i) Proposing the combination of dense and inception networks in a new way for feature extraction, (ii) exploring wavelet decomposition to enhance the fine details in the images to improve the feature extraction's strength, (iii) proposing a soft style attention module for text style features to enhance the performance of script identification, and (iv) proposing lightweight models for achieving efficiency without losing accuracy.

The structure of the paper is as follows. The critical analysis of existing methods is presented in Section 2. In Section 3, a description of dense, inception, wavelet decomposition, and soft style attention modules are discussed. Experimental results and analysis of different experiments are presented in Section 4. Section 5 summarizes the findings of the proposed work.

2. Related Work

The analysis of existing script identification methods is discussed here. Gomez et al. [3] proposed patch-based features for

***Corresponding author:** Shivakumara Palaiahnakote, School of Science, Engineering and Environment, University of Salford, UK. Email: s.palaiahnakote@salford.ac.uk

scene script identification. Cheikhrouhou et al. [8] developed an end-to-end multi-task deep neural network for script identification for spotting document images. However, the work is confined to handwritten text but not scene text. Dutta et al. [9] used the inception network to extract global and local features for script identification in the scene images. Khalil et al. [10] developed an end-to-end model for text detection and script identification in scene images. However, the success of the method depends on the success of text detection.

Guo et al. [11] introduced a model that combines deep convolutional neural networks and spatial pyramid pooling for script identification in ancient books. The approach performs well for the document images, and hence, it may not be suitable for scene images. Li et al. [7] used a self-attention network for script identification in scene images. The technique works well for particular datasets but not for all the benchmark datasets. Udupa et al. [12] aimed to explore YOLOv5 for text detection and script identification. The authors used their dataset for the classification of scripts. However, the performance of the method has been tested using their own dataset. Shi et al. [13] and Lu et al. [14] proposed deep learning models for script identification based on attention and patch-mining approaches. The method is said to be computationally expensive as it involves heavy models. Mahajan et al. [15] focused on Indian script identification based on CNN enhancement. This method requires a large number of samples to achieve the best results. In the same way, Ma et al. [16] proposed hierarchical feature fusion blocks and attention mechanisms for scene script identification. The features are not robust and flexible to achieve better results. Zhang et al. [2] used Res2net for identifying scene scripts. Since these methods follow conventional features, the methods may not be effective for complex datasets. Yang et al. [17] dealt end-to-end model for detection and identification by combining visual and textual features. The method works well for specific datasets, and the method lacks generalization ability. Roy et al. [18] introduced the combination of Xception and log-polar transformed of the original images for script identification in scene images. The approach focuses on extracting global features and text style features and proposes a style-enhanced network for fusing two different features to improve the performance of the script identification. However, the method is sensitive to blurred images. Khan et al. [19] developed a multi-scale deep neural network-based model for script identification. The approach performs multiple CNNs simultaneously for feature extraction at different scales. At the same time, the weight computation model assigns suitable weight to widen the gap between different scripts.

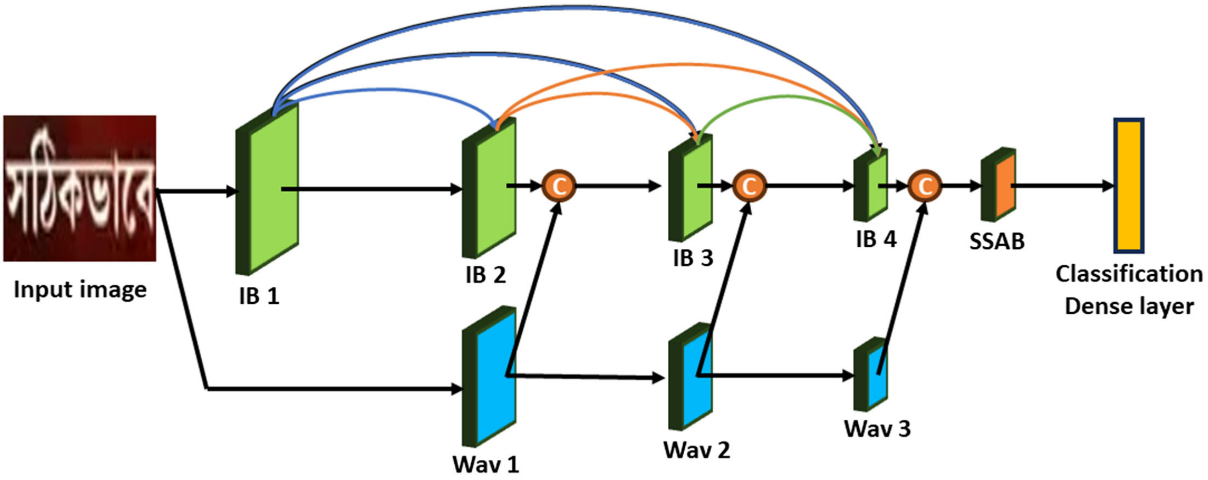
As noted from the above review, some methods used conventional features and the recent deep learning approaches for script identification. Most methods focus on extracting local and global features for improving script identification performance. However, none of the methods reported consistent performance for different benchmark datasets of script identification. In addition, most methods use heavy models for achieving the best results, and therefore, the models are expensive in terms of the number of computations. These observations motivated us to propose a novel method for script identification.

3. Proposed Methodology

As discussed in the related work section, achieving consistent results for different benchmark datasets is challenging. Therefore, we aimed to propose a model that integrates the strengths of multiple networks so that a hybrid model can achieve the expected results. Inspired by the special properties of the dense network which can extract fine-tuned features and the Inception network which can extract spatial multi-level features [4, 5], we adapt the same models for classification of scripts. Similarly, to enhance the fine details in the images, motivated by the wavelet decomposition which can extract different oriented edge information, we use the same wavelet decomposition to improve the image quality [6]. This adds more richness to the above feature extraction. For successful script identification, the text style and shape of characters are most important; to extract such observation, the proposed work explores the soft style attention module which extracts the text style and shape of characters [7].

Overall, the proposed work used lightweight models to achieve efficiency without losing accuracy. The proposed work introduces wavelet transform to enhance the fine details in the images irrespective of adverse situations. Due to this, the style of stroke information is preserved which provides important clues for differentiating the scripts. As a result, the combination of inception-dense and soft style attention modules extracts elegant features for successful script identification. The structure of the hybrid model can be seen in Figure 1, where one can see how the

Figure 1
The overall architecture. In this hybrid CNN architecture, IB indicates the Inception blocks used along with the dense connections shown. Wav are the wavelet decomposed feature maps concatenated with the output features of IBs



different modules are integrated into one model. This is where the proposed model is different from the state-of-the-art methods for script identification.

3.1. Inception-DenseNet

DenseNet, which is a Densely Connected Convolutional Network, has garnered widespread acclaim in the computer vision community for its unique structural attributes and numerous advantages. Specifically, DenseNet excels in effective feature extraction for script identification tasks, which often require discerning subtle distinctions in writing styles. Its dense connectivity empowers the model to capture fine-grained features, enabling precise discrimination between various scripts. For example, the script can be Latin, Cyrillic, Arabic, or others. Additionally, DenseNet demonstrated its efficiency due to a small number of parameters. Its adaptability further shines through, allowing researchers to tailor network architecture to suit specific script identification challenges, such as accommodating varying stroke patterns, character shapes, and writing styles. Notably, DenseNet’s ability to handle limited script samples with aplomb, thanks to feature reuse and efficient training, makes it an ideal choice for script recognition tasks constrained by small datasets.

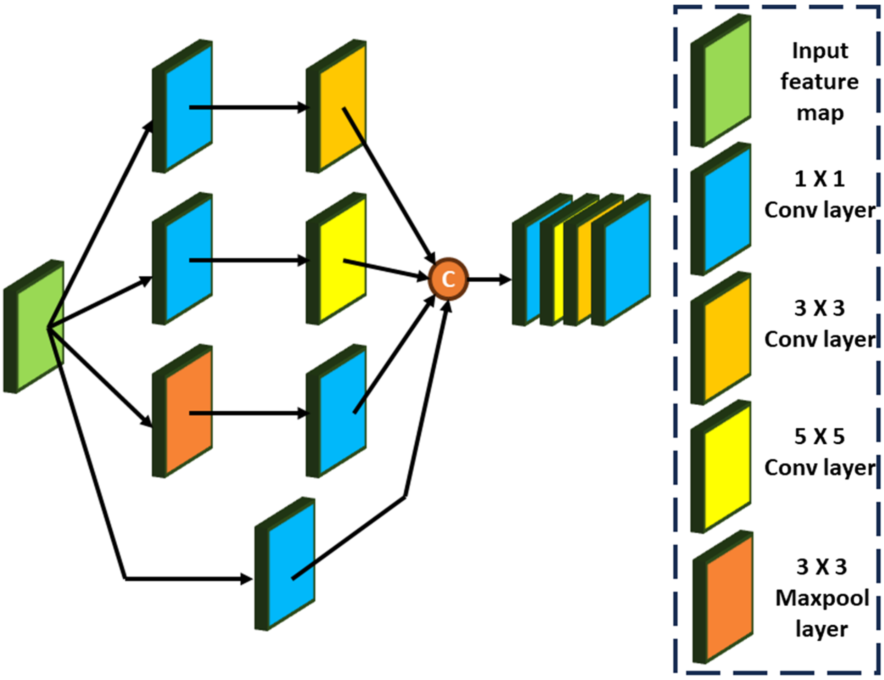
Inception blocks, on the other hand, are indispensable for their role in enabling script identification models to capture multi-scale features, expand receptive fields, and process information in parallel. This pivotal feature extraction mechanism caters to the diverse nature of scripts, accommodating variations in stroke thickness, character size, and complexity. By ensuring the capture of both global and local dependencies through different receptive fields, Inception blocks enhance the model’s ability to discern subtle script variations accurately. Consequently, Inception blocks contribute significantly to the model’s precision, adaptability, and overall performance in script identification tasks. A visual demonstration of the ability of Inception blocks to capture spatial features of various receptive fields can be seen in Figure 2. In summary, one can argue that the combination of DenseNet and Inception network is essential for extracting distinct features for script identification.

3.2. Discrete wavelet transform

The wavelet transform is a combination of low- and high-pass filters, and it helps in retaining high-frequency coefficients which represent edge pixels in the image. In the same way, the wavelet decomposition makes the enhancement step robust to extract minute changes in the images. DWT operates by representing the signal in terms of wavelet basis functions, allowing it to capture both high and low-frequency information effectively. Edge detection is important for extracting fine details in the images, and it helps to identify the boundaries and transitions between objects or regions in an image. The following special properties of DWT help us to detect the fine details, namely edges in the images.

Multi-scale Analysis: DWT decomposes an image into different scales or resolutions, each containing information about specific frequency components. Edges in an image exist at multiple scales, from fine details to larger contours. DWT captures these edges at different scales, enabling a multi-scale analysis that helps detect edges of various sizes and orientations. **Localization:** It not only identifies edges but also provides information about their precise locations within the image. **Feature Extraction:** Edges often carry essential information about object shapes and structures. DWT can be used to extract edge-related features that are useful for subsequent script identification steps.

Figure 2
Illustration of the Inception block



The DWT decomposition of a 2D image into approximation (low-frequency) and detail (high-frequency) components is given by:

$$A_{j+1}[n,m] = \sum\nolimits_{k1}^{n} \sum\nolimits_{k2}^{m} \alpha_{j,k1,k2} * A_j[n-k1,\ m-k2] \quad (1)$$

$$D_{j+1}[n,m] = \sum\nolimits_{k1}^{n} \sum\nolimits_{k2}^{m} \sigma_{j,k1,k2} * A_j[n-k1,\ m-k2] \quad (2)$$

In Equations (1) and (2), $A_{j+1}[n,m]$ represents the approximation coefficients at scale j for the 2D image, $D_{j+1}[n,m]$ represents the detail coefficients at scale j for the 2D image, $\alpha_{j,k1,k2}$ and $\sigma_{j,k1,k2}$ are the Morlet wavelet and scaling function coefficients, respectively, at scale j.

3.3. Soft style attention module

Motivated by the successful deployment of Soft Attention [7] in medical imaging for segmentation, it is explored to extract spatial attention to boost the value of important features and suppress the noise-inducing features. In this, K different attention maps are generated and are unified to produce a single attention map based on a learnable weight. We use the Soft Attention block to provide the initial spatial enrichment to the input tensor X of dimension H × W × C to produce Xs of dimension H × W × C. Xs is then treated by the squeeze and excitation channel attention mechanism to generate Xsc of dimension H × W × C. Xsc consists of information regarding the relevant channels while having lower weightage assigned to the noisy channels. It is then max pooled to extract the dominant spatial features that correspond to the semantic boundaries of the script, i.e., the strokes and edges. The script-aware attention weights are generated by applying sigmoid activation to the pooled feature and then multiplied with Xsc to produce Xattn of dimension H × W × C. A detailed block diagram representation of the proposed soft style attention module is shown in Figure 3. The heatmaps shown in Figure 4 demonstrate the ability of soft style attention module to highlight spatial features relevant to the script to streamline the focus of the model and boost performance.

Figure 4
Heatmaps of the soft style attention module (SSAM) showcasing the ability to highlight spatial structural features of the scripts

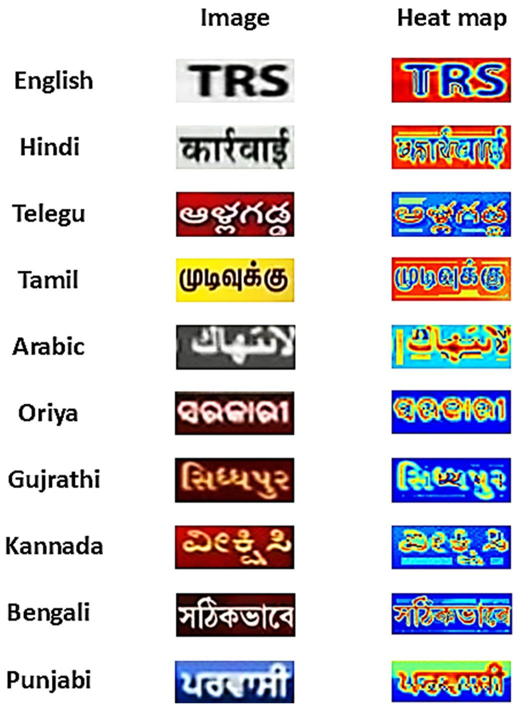
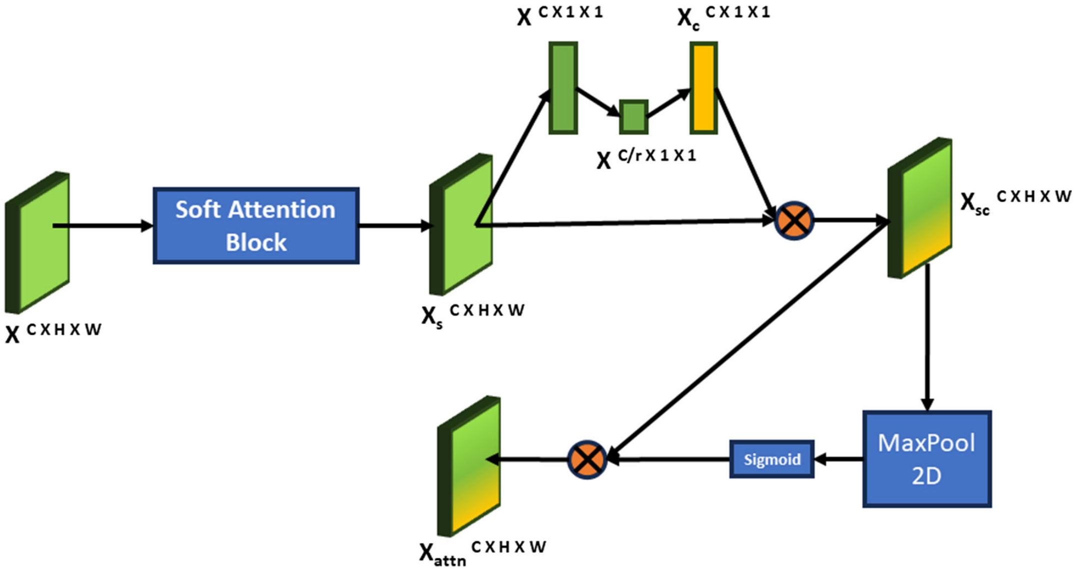


Figure 3
The soft style attention module (SSAM)



4. Experimental Results

4.1. Dataset and evaluation

To evaluate the proposed method, we conducted experiments on the three standard datasets, namely CVSI2015, SIW-13, and MLe2e. **CVSI2015 dataset [18]:** This one consists of word images extracted from various video sources, encompassing a diverse range of ten distinct scripts: English, Hindi, Bengali, Oriya, Gujarati, Punjabi, Kannada, Tamil, Telugu, and Arabic. **SIW-13 dataset [20]:** This one is another benchmark dataset that provides word-level images representing 13 different scripts, including Arabic, Cambodian, Chinese, English, Greek, Hebrew, Japanese, Kannada, Korean, Mongolian, Russian, Thai, and Tibetan. **MLe2e dataset [2]:** This is one more highly challenging dataset that supports script identification for four different script types, including Latin, Chinese, Kannada, and Korean. **Combined dataset**: Additionally, to show the robustness of the proposed model, we combined the distinct classes of all the above three benchmark datasets, which resulted in 21 class classification problems. This combined dataset is much more complex than individual datasets. This is because as the number of classes increases, the complexity of the problem increases. Overall, the above four datasets are considered for experimentation and evaluation of the proposed and existing methods. The details of all three datasets mentioned above are given in Table 1.

Table 1
A summary of the number of images used for training, testing, and validation for the three datasets and the combined dataset

Dataset	Class	Training	Testing	Validation
CVSI2015	10	6412	3234	1069
SIW-13	13	9103	3299	3889
MLe2e	4	826	642	351
Combined	21	15164	7175	6483

For measuring the performance of the proposed and existing methods on different benchmark datasets, we use standard measures of Recall, Precision, F-measure, and Accuracy. Furthermore, a comparative study has been done in terms of accuracy because most existing methods use accuracy for reporting the results. To ensure a fair comparison, the proposed work follows the same.

Implementation Details: We conducted our experiments using TensorFlow on a single NVIDIA P100 GPU. We configured the batch size to 24 to ensure optimal training conditions when training with all three datasets with a learning rate of 0.01. We used the TensorFlow-keras library to code the model. To optimize our model, we employed the Adam optimizer, incorporating a momentum of 0.9 and a weight decay of 1×10^{-4}. In our experimentation, we resized the text lines from the input images to a uniform size of 256 × 256 and normalized it to bring the value of pixels down to values between 0 and 1 for a uniform gradient flow. To train the model, we have utilized categorical cross-entropy loss. We employed a train-test split of 80–20%, and from the training data, we used 30% for validation. The same experimental setup has been used for all the experiments including running the existing methods.

4.2. Ablation study

It is noted from the proposed methodology section that the DenseNet, Inception blocks, soft style attention module, and wavelet decomposition are the key elements of the proposed hybrid method. To assess the contribution and effectiveness of each key component, we conducted the following experiments on the CVSI2015 dataset. The results are recorded in Table 2. It is observed from Experiments (i)–(v) that as the key components are added to the proposed model, the performance of script identification improves. This shows that the proposed key components contribute equally to achieving the best results for script identification. It is also noted that individual component performance does not achieve the best as the proposed model. This indicates that the individual elements are not capable, and hence, it is necessary to fuse the strength of each component to solve the complex problem with high accuracy.

4.3. Comparative study with the state-of-the-art

To evaluate the proposed method for script identification, the accuracy is calculated for the three benchmark datasets as reported in Table 3. It is noted from Table 3 that the proposed method is the best at accuracy compared to the existing methods on CVSI2015 and MLe2e datasets. This shows that the proposed method is superior to the existing methods for two benchmark datasets. Compared to CVSI2015 and MLe2e datasets, the SIW-13 is a bit complex due to more number of scripts. Therefore, the proposed method is effective. Roy et al. [18] report better performance than all the methods listed in Table 3 including the proposed method on the SIW-13 dataset. This is because the method [18] elegantly fuses global and test-style features. Therefore, it performs well for the SIW-13 dataset. However, the same method [18] does not achieve the best results for CVSI2015 and MLe2e datasets compared to the proposed method. This is the limitation of the method [18]. The reason for the poor results of the other existing methods is that the scope of the methods is limited, and the features used in the existing methods are not robust in contrast to the features used in the

Table 2
Validating the key steps of the proposed method on CVSI2015 (in %)

Exp.	Steps	Accuracy	Precision	Recall	$F1$-score
(i)	Baseline DenseNet (4 layers)	90.26	92.83	89.12	90.93
(ii)	Baseline DenseNet with Inception Block	93.32	94.39	92.87	93.62
(iii)	Baseline DenseNet + Inception + Soft Attention module	97.65	98.47	96.10	97.27
(iv)	Baseline DenseNet + Inception + SSAM	98.14	98.95	97.99	98.07
(v)	Proposed model (Adapted DenseNet + Inception +SSAM + Wavelet Decomposition)	**98.99**	**99.38**	**98.25**	**98.81**

Table 3
Accuracy of the proposed and existing methods on CVSI2015. SIW-13 and MLe2e (in %)

Methods	CVSI2015	SIW-13	MLe2e
Shi [13]	94.30	89.40	–
Gomez [3]	97.20	94.80	94.40
Ma [16]	98.78	97.30	97.20
Lu [14]	97.90	96.11	89.42
Dutta [8]	98.97	95.70	95.01
Mahajan [15]	97.40	–	–
Guo [11]	93.50	–	98.74
FAS-Res2net [2]	96.00	94.70	–
SANet [7]	**99.03**	96.18	–
Roy et al. [18]	98.99	**97.83**	98.75
Ours	**99.03**	97.53	**98.82**

Table 4
Comparison in terms of accuracy on the combined dataset

FAS-Res2net [2]	SANet [7]	Ours
77.09	81.21	81.75

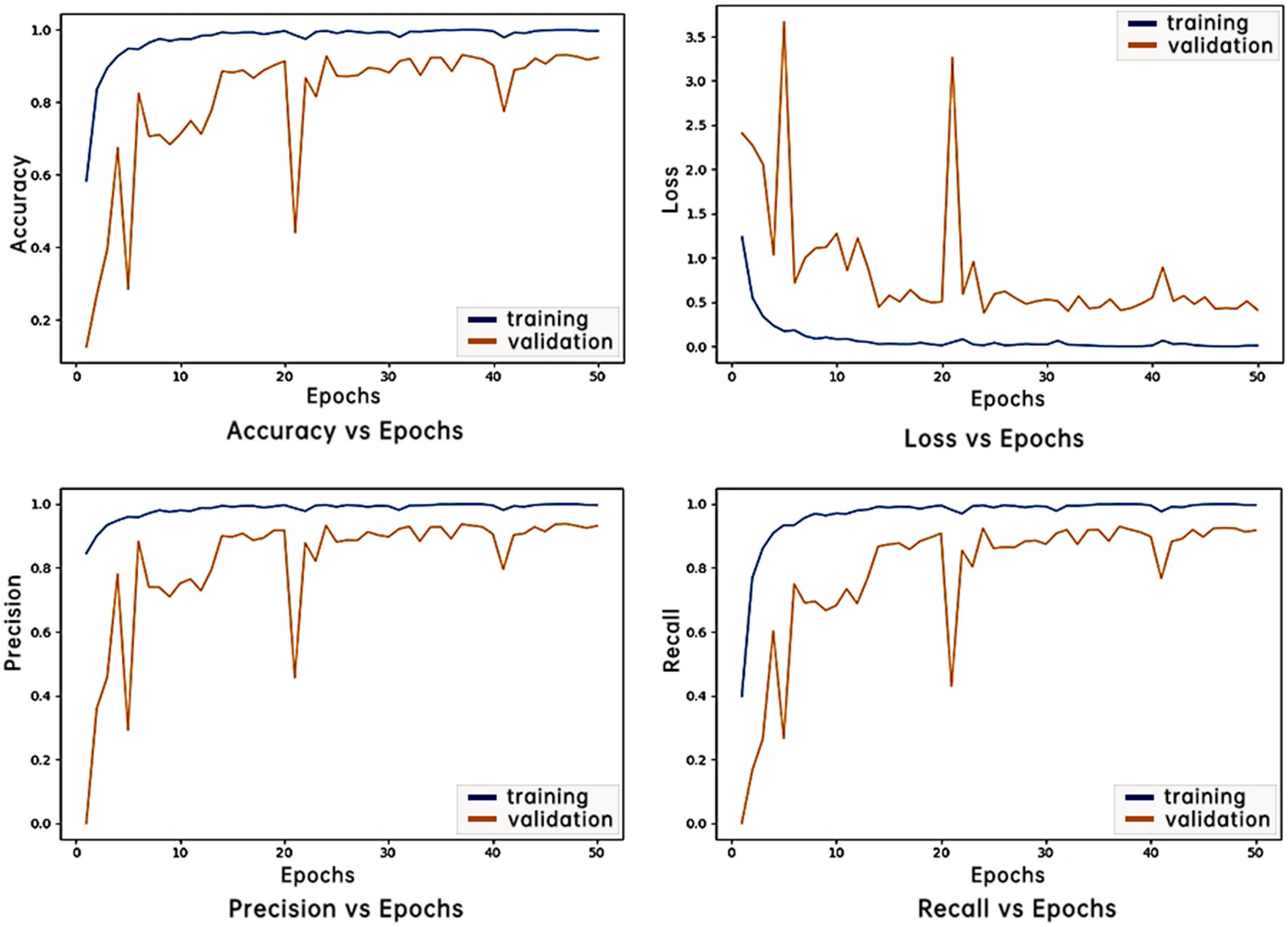
proposed method. In addition, most of the existing methods use heavy models, which require a large number of computations to achieve good performance. On the other hand, the proposed model is simple and combines light modules to achieve the best results. When we consider the overall performance of the proposed and existing methods on all three datasets, the proposed method is promising and impressive.

To demonstrate the consistency and effectiveness of the proposed method as the number of classes increases beyond those in individual datasets, accuracy was calculated for the combined dataset. The results in Table 4 confirm that the proposed method outperforms existing methods. Thus, based on the performance of both individual and combined datasets, we conclude that the proposed method is robust and effective regardless of class count or dataset complexity. Conversely, the results highlight that existing methods struggle with handling a larger number of classes, as previously discussed.

4.4. Discussion

To show the proposed model is efficient, the experiments of training and validations are presented in Figure 5. Figure 5 shows that all the experiments used 50 Epochs to achieve high performance. Therefore, one can infer that the proposed model

Figure 5
The training and validation curves of the proposed model on the CVSI2015 dataset



Accuracy vs Epochs

Loss vs Epochs

Precision vs Epochs

Recall vs Epochs

does not require more number of computations for successful script identification with high performance.

5. Conclusion and Future Work

We have proposed a simple and effective model for script identification. For all three benchmark datasets and one more new combined dataset, the proposed work fuses the DenseNet, Inception blocks, wavelet decomposition, and soft style attention module. The key advantage of the method is that it integrates the merit of each key component mentioned above, which results in a hybrid model for script identification. To show the robustness of the proposed model, it is tested on three benchmark datasets and one more combined dataset which has 21 classes. It is evident from the results that the proposed method outperforms the existing methods in terms of accuracy. However, when we include more low-resource languages for script identification, the performance of the proposed method may be degraded. This is because, in the case of low-resource languages, the characteristics such as shape and text style may be shared to a large extent. This can be solved by adding a language model to the proposed model. The reason is that the language model fuses both image and textual information in contrast to the existing method which uses only image information.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

Palaiahnakote Shivakumara is the Editor-in-Chief and Umapada Pal is an Advisory Board Member for *Artificial Intelligence and Applications*, and were not involved in the editorial review or the decision to publish this article. The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in CVSI 2025 at https://www.ict.griffith.edu.au/cvsi2015/Dataset.php, in GitHub at https://github.com/lluisgomez/script_identification, and in Kaggle at https://www.kaggle.com/datasets/ayush02102001/cvsi-script-identification-dataset.

Author Contribution Statement

Shivakumara Palaiahankote: Methodology, Writing – original draft. **Umapada Pal:** Writing – review & editing, Visualization, Supervision. **Taha Mansouri:** Validation, Resources.

References

[1] Ke, W., Hou, Q., Liu, Y., Song, X., & Wei, J. (2024). SARN: Script-aware recognition network for scene multilingual text recognition. *Expert Systems with Applications*, *250*, 123753. https://doi.org/10.1016/j.eswa.2024.123753

[2] Zhang, Z., Mamat, H., Xu, X., Aysa, A., & Ubul, K. (2023). FAS-Res2net: An improved res2net-based script identification method for natural scenes. *Applied Sciences*, *13*(7), 4434. https://doi.org/10.3390/app13074434

[3] Gomez, L., Nicolaou, A., & Karatzas, D. (2017). Improving patch-based scene text script identification with ensembles of conjoined networks. *Pattern Recognition*, *67*, 85–96. https://doi.org/10.1016/j.patcog.2017.01.032

[4] Nabi, S. T., Singh, P., & Kumar, M. (2023). Writer identification from offline handwriting images in Urdu script with dense-net: A deep learning approach. In *14th International Conference on Computing Communication and Networking Technologies*, 1–6. https://doi.org/10.1109/ICCCNT56998.2023.10307034

[5] Chen, L., Peng, L., Yao, G., Liu, C., & Zhang, X. (2019). A modified inception-ResNet network with discriminant weighting loss for handwritten Chinese character recognition. In *International Conference on Document Analysis and Recognition*, 1220–1225. https://doi.org/10.1109/ICDAR.2019.00197

[6] Han, X. K., Aysa, A., Yadikar, N., Mamat, H., & Ubul, K. (2017). Script identification based on nonsubsampled contourlet transform. In *14th IAPR International Conference on Document Analysis and Recognition, 1*, 697–702. https://doi.org/10.1109/ICDAR.2017.119

[7] Li, X., Zhan, H., Shivakumara, P., Pal, U., & Lu, Y. (2023). SANet-SI: A new self-attention-network for script identification in scene images. *Pattern Recognition Letters*, *171*, 45–52. https://doi.org/10.1016/j.patrec.2023.04.015

[8] Cheikhrouhou, A., Kessentini, Y., & Kanoun, S. (2021). Multi-task learning for simultaneous script identification and keyword spotting in document images. *Pattern Recognition*, *113*, 107832. https://doi.org/10.1016/j.patcog.2021.107832

[9] Dutta, K., Dastidar, S. G., Das, N., Kundu, M., & Nasipuri, M. (2021). Script identification in natural scene text images by learning local and global features on inception net. *Computer Vision and Image Processing*, *1567*, 458–467. https://doi.org/10.1007/978-3-031-11346-8_40

[10] Khalil, A., Jarrah, M., Al-Ayyoub, M., & Jararweh, Y. (2021). Text detection and script identification in natural scene images using deep learning. *Computers & Electrical Engineering*, *91*, 107043. https://doi.org/10.1016/j.compeleceng.2021.107043

[11] Guo, H., Yang, D., Liu, Y., & Zhao, J. (2023). Script identification of ancient books by Chinese ethnic minorities using multi-branch DCNN and SPP. *Pattern Analysis and Applications*, *26*(2), 809–821. https://doi.org/10.1007/s10044-023-01146-y

[12] Udupa, C., Upadhyaya, A., Patil, B. S., Seeri, S. V., Patil, P., & Hiremath, P. S. (2022). Text localization and script identification in natural scene images and videos. In *International Conference on Connected Systems & Intelligence*, 1–7. https://doi.org/10.1109/CSI54720.2022.9924044

[13] Shi, B., Bai, X., & Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2298–2304. https://doi.org/10.1109/TPAMI.2016.2646371

[14] Lu, L., Wu, D., Tang, Z., Yi, Y., & Huang, F. (2021). Mining discriminative patches for script identification in natural scene images. *Journal of Intelligent & Fuzzy Systems*, *40*(1), 551–563. https://doi.org/10.3233/JIFS-200260

[15] Mahajan, S., & Rani, R. (2022). Word level script identification using convolutional neural network enhancement for scenic images. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(4), 1–29. https://doi.org/10.1145/3506699

[16] Ma, M., Wang, Q. F., Huang, S., Huang, S., Goulermas, Y., & Huang, K. (2021). Residual attention-based multi-scale script

identification in scene text images. *Neurocomputing*, *421*, 222–233. https://doi.org/10.1016/j.neucom.2020.09.015

[17] Yang, K., Yi, J., Chen, A., Liu, J., Chen, W., & Jin, Z. (2022). ConvPatchTrans: A script identification network with global and local semantics deeply integrated. *Engineering Applications of Artificial Intelligence*, *113*, 104916. https://doi.org/10.1016/j.engappai.2022.104916

[18] Roy, A., Palaiahnakote, S., Pal, U., Antonacopoulos, A., & Blumenstein, M. (2025). XLSI: A new Xception and log polar transform based approach for scene text script identification. *Pattern Recognition*, *15319*, 183–198. https://doi.org/10.1007/978-3-031-78495-8_12

[19] Khan, T., Saif, M., & Mollah, A. F. (2024). MuSIC: A novel multi-scale deep neural framework for script identification in the wild. *IEEE Access*, *12*, 166955–166976. https://doi.org/10.1109/ACCESS.2024.3494023

[20] Shi, B., Bai, X., & Yao, C. (2016). Script identification in the wild via discriminative convolutional neural network. *Pattern Recognition*, *52*, 448–458. https://doi.org/10.1016/j.patcog.2015.11.005

How to Cite: Palaiahnakote, S., Pal, U., & Mansouri, T. (2025). A New Hybrid Wavelet Decomposition-based Networks for Script Identification in Scene Images. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA52023569