Electronic Letters on Computer Vision and Image Analysis 23(1):15-31; 2024

Off-line identifying Script Writers by Swin Transformers and ResNeSt-50

Afef Kacem Echi* and Takwa Ben Aïcha Gader*

* *University of Tunis, ENSIT-LaTICE, 05 Avenue Taha Hussein, Tunis, 1008, Tunisia*

Received 17th of October, 2023; accepted 29th of May 2024

Abstract

This work proposes two novel deep learning models for writer identification, achieving state-of-the-art performance. The first model leverages the Swin Transformer, known for its ability to capture long-range dependencies and handle variations in handwritten text. It operates on sequences of image patches, learning robust representations of a writer's style through extensive training on large datasets. The second model utilizes ResNeSt-50, a deep convolutional neural network with modules designed for feature attention and efficient learning. ResNeSt-50 excels at extracting complex, distinctive features from handwritten samples, enabling highly accurate writer distinction. Experimental results demonstrate exceptional accuracy: the Swin Transformer achieves 98.50% accuracy (patch-level) on the CVL database containing cursive German and English text, while ResNeSt-50 attains 96.61% accuracy (page-level) on the same dataset. These results showcase the effectiveness of both models in writer identification and their potential for handling complex handwriting styles.

Key Words: Writer identification, Deep learning, Swin Transformer, ResNeSt-50, Handwriting analysis.

1 Introduction

In the current digital era, where communication occurs mainly online, and document analysis is automated, identifying the text's author has become crucial in many fields. It is especially important in forensic analysis, historical text analysis, and automated document processing. The precise determination of the authorship of handwritten texts holds significant implications for law enforcement, historical research, and even the validation of digital signatures.

Although publicly available datasets such as IAM, CVL, and Firemaker have achieved high recognition rates, identifying a writer based on small regions of handwritten text is still challenging. This is because limited information is available to model the writer's handwriting style. This results in low recognition rates on the IAM and CVL datasets when using handcrafted features for word-based writer identification. Instead of analyzing an entire document or handwriting sample, this study focuses on identifying individual patches. This approach is particularly useful when dealing with degraded or fragmented documents, where only small handwriting sections are available.

Correspondence to: takwa.ben.aichaa@gmail.com

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.1787
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

Over time, several methods have been devised to tackle this complex task, with recent progress in deep learning techniques exhibiting encouraging outcomes. Two main types of methods are used in this context, namely holistic and feature-based approaches. Holistic approaches focus on capturing the overall writing style of an individual, considering global features such as the handwriting's shape, slant, and spacing. These methods often employ statistical and machine learning algorithms to extract and compare high-level characteristics. Feature-based approaches analyze individual strokes and curves in handwriting. These techniques leverage image processing and pattern recognition algorithms to extract discriminating features that differentiate one writer from another. Combining global and local characteristics, hybrid methods emerged to improve writer identification systems' accuracy and robustness.

This paper explores the efficacy of deep learning using the new vision Swin Transformer [1] and the ResNeSt-50 [2] within the feature-based approach. Specifically, we investigate their ability to capture and differentiate unique writing styles in the context of the CVL database, which consists of images with cursively handwritten German and English texts chosen from literary works. The Swim Transformer's ability to capture long-range dependencies, model temporal dynamics, and adapt to different tasks makes it a promising choice for handwriting analysis and writer identification. It can offer improved accuracy, robustness, and interpretability compared to traditional methods, especially when dealing with complex handwriting patterns and variations across writers. The ResNeSt-50 (Residual Neural Network with Squeeze-and-Excitation (SE) and Next Stage modules) model, built on the ResNeSt architecture, is a highly effective deep-learning model for identifying writers. Its advanced architecture and efficient image processing capabilities make it capable of recognizing distinct handwriting styles.

The paper is organized into the following sections: Section 2 presents a comprehensive analysis of the related literature and existing approaches to writer identification. Section 3 will discuss the state of the art and introduce the proposed systems. In Section 4, the significance of deep learning in writer identification is emphasized. Sections 5 and 6 detail the methodology, including the Swin Transformer and the ResNeSt models architecture, training procedure, and evaluation metrics. Section 7 accounts for the experimental results and analysis, discussing the findings, limitations, and potential improvements. Section 8 concludes the paper by summarizing key contributions and highlighting avenues for future research.

2 Related Works

Identifying a document's or written text's author involves analyzing various handwriting characteristics. One common approach is to extract specific features from the handwritten text, such as individual letter shapes, stroke patterns, word spacing, slant, baseline alignment, and other relevant attributes. These features can then be input to machine learning algorithms for writer identification, like k-nearest neighbors (KNN), support vector machines (SVM), and random forests. These algorithms are trained on a dataset of known handwritten samples and learn to differentiate between different writers based on recognized patterns. The training data's quality and diversity heavily influence these algorithms' efficacy.

Deep learning approaches have recently gained attention and proven effective in writer identification, achieving state-of-the-art results. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are the most widely used in this field. CNNs are effective in capturing spatial patterns and have been used to extract features from handwriting images. Meanwhile, RNNs, especially the Long Short-Term Memory (LSTM) networks, are suitable for modeling sequential data, such as the temporal ordering of strokes in handwriting. Considering the time dimension of data, they are more effective in identifying online writers [6].

A new study [7] has introduced a promising end-to-end system that utilizes a Global Context Residual RNN (GR-RNN) to identify writers based on limited handwritten data. The GR-RNN system employs global average pooling to gather information and RNNs to model the connection between local and fragment-based features. Results from tests on several datasets, including IAM, CVL, Firemaker, and CERUG-EN, reveal that this technique performs the best on Firemaker. The authors suggest that this approach can extract intricate details related to writing style.

The FEM-WI (Funneling Ensemble Method for Writer Identification) is a two-level ensemble system designed to identify writers [8]. This method employs various feature-dependent base classifiers and a meta-classifier. The authors of the method introduced four new feature descriptors. The approach was able to successfully identify over 90% of the writers in the IAM and Firemaker datasets.

In their study, [9] utilized CNN AlexNet with ImageNet transfer deep learning for feature extraction. They worked with text-line images of handwriting in English and Arabic. Eight input patches were created, including the original, contoured, sharpened, and sharpened contours, along with their negatives. Deep features were obtained from 227×227 image patches. Before data augmentation, skew detection, correction, normalization, segmentation, and patch sliding window approach were performed. SVMs were employed for classification. The QUWI collection comprised 1017 writers, four digitized pages, and 60 words per writer. The authors extracted features from several layers, including Conv3, Conv4, Conv5, Fc6, Fc7, and AlexNet Fc6 and Fc7 freeze layers. The freeze Conv5 layer yielded the highest accuracy of 92.78% for English, 92.2% for Arabic, and 88.11% for both languages.

In a study by [14], it was suggested to use multi-task learning that combines word recognition and writer identification techniques. One-word image writer identification was studied to capture the writing style. The information loss is minimized by considering both implicit and explicit features, creating a more generalized model. The study used a CNN based on the AlexNet architecture with two pathways to transfer features from the secondary to the primary (writer identification) task in an end-to-end system to enhance the writer identification performance metrics. The study also explored word recognition, word-length estimation, character attribute recognition, and their combinations to improve writer identification. The approaches were tested using the CVL and IAM datasets, and the results showed that deep adaptive learning can improve writer identification by capturing complex linkages.

In a recent study by [14] and another by [15], a CNN architecture was utilized with feature pyramid and fragment branches, along with a deep learning strategy that used multi-task learning. The feature pyramid branch was responsible for extracting feature maps, while the fragment branch trained writer identification using fragments from the word image and feature maps. [11] proposed a framework for identifying writers using deep learning, which utilized ResNet and a new handwriting thickness descriptor. This framework was tested on various datasets and achieved high accuracies, with results of 97.50%, 99.61%, 96.16%, and 88.95% on the IAM, Firemaker, CVL, and CERUG-EN datasets, respectively. These findings confirm the usefulness of the framework for recognizing handwritten characters.

A study in [12] utilized a CNN to distinguish writers by excluding the classification layer and utilizing the second last fully connected layer output as a feature vector. Another method for writer identification, DeepWriter, was introduced in [13]. This approach involves using local handwritten patches in pairs and can identify writers independently of the text content. By expanding the training data and improving the system, this approach demonstrated good results for both Chinese and English characters.

Generally, these proposed systems follow three approaches: end-to-end modeling, deep learning modeling, and transfer learning modeling, as explained below. Each approach has strengths and can be applied depending on available resources, dataset size, and specific requirements.

- **End-to-end Modeling**: The end-to-end modeling process involves training a single model that can analyze raw input data (like handwritten text images) and predict the writer's identity without any manual feature extraction or preprocessing. In writer identification, this method utilizes deep neural networks like convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to learn the direct mapping between input data and the writer's identity. End-to-end models can automatically detect important features from data, capturing complex patterns and relationships that are hard to identify manually. However, these models often require a large amount of labeled data for training and can be computationally intensive.

- **Deep Learning Modeling** : Deep learning models like CNNs and RNNs are often utilized when identifying writers. CNNs are adept at capturing local patterns and features in images, making them ideal

for analyzing individual characters or handwritten text patches. On the other hand, RNNs are better suited for modeling text sequences, enabling the analysis of longer portions of handwriting. These deep learning models can be trained using supervised learning techniques, where known writers' identities are labeled as data to optimize the model parameters. They can learn complex representations of input data and provide accurate predictions. However, they typically require significant labeled training data and can be computationally demanding.

– **Transfer Learning Modeling**: To better understand transfer learning, it involves using knowledge acquired from a similar yet different task and applying it to the target task. In writer identification, transfer learning is used by taking a deep learning model pre-trained on a large dataset (such as general image classification) and then fine-tuning it on a smaller dataset of labeled handwritten samples. With transfer learning, the pre-trained model can capture general features and representations useful for many tasks, including writer identification. Fine-tuning allows the model to specialize and adapt to the specific characteristics of the writer identification task, even with limited training data. This method is especially useful when there is a shortage of labeled data for writer identification.

Writer recognition in handwriting analysis often focuses on words rather than pages or patches. Depending on available data and task requirements, identification can occur at various levels: word, patch, or page. At the word level, individual words or short text segments are analyzed for unique handwriting features like letter shapes, stroke formations, and spacing. Algorithms and ML techniques extract and compare these features for identification. Patch-level analysis expands to larger text portions, capturing broader handwriting characteristics for increased accuracy. Page-level identification treats entire documents as units, considering layout and global features. Deep learning methods like CNNs or RNNs can be employed for comprehensive page-level analysis and recognition.

Table 1: Related works summaries.

References	Writer identification level	Appraoch
[6]	Word	End-to-end, Deep learning, Transfer learning
[7]	Word	End-to-end, Deep learning
[8]	Page	Deep learning, Transfer learning
[9]	Word	Deep learning, Transfer learning
[14]	Word	End-to-end, Deep learning, Transfer learning
[15]	Word	Deep learning
[11]	Patch	End-to-end, Deep learning
[12]	Word	Deep learning
[13]	Patch	End-to-end, Deep learning

As shown in table 1, these approaches are generally accurate even when dealing with datasets containing many authors, as seen in current literature. The most common architectures used are CNNs and transfer learning, which are already trained and perform well on short datasets. Deep learning is also frequently utilized and has proven useful for writer identification. It is important to understand that identifying a writer can be difficult, requiring many handwriting samples from potential writers for accurate analysis. Furthermore, variations in writing styles, intentional disguises, or limitations in the quality of the available samples can all impact the accuracy of the identification process. Nevertheless, advancements in machine learning and pattern recognition techniques have improved the reliability of writer identification. This has made it a valuable tool in forensic analysis and documentation. We investigated the use of new artificial intelligence technologies: the Swin transformer and ResNeSt-50 models which offer advantages for writer identification tasks. Swin Transformer's hierarchical structure and efficient computation enable capturing information at multiple scales, making it suitable

for large-scale datasets. ResNeSt-50's "Split-Attention" mechanism enhances feature representation, helping in capturing subtle details crucial for identification. Additionally, ResNeSt-50's scalability and availability of pre-trained models facilitate transfer learning on smaller datasets.

3 Deep Learning for Writer Identification

Deep learning models are a subset of machine learning that uses artificial neural networks to solve complex tasks. These models are successful due to their ability to learn hierarchical features from data automatically. Several types of deep learning models, including Convolutional Neural Networks (CNNs), are used for tasks involving structured grid data like images and videos. Recurrent Neural Networks (RNNs) are designed for sequential data where the order of input elements is important, and they use recurrent connections to capture temporal dependencies. U-Net is a specific architecture used for image segmentation tasks, while Transformers are designed to handle sequential data efficiently by employing a self-attention mechanism. Transformers are particularly useful in natural language processing tasks like language translation and text generation. These deep learning models have a range of applications, including image classification, object detection, facial recognition, time series prediction, sentiment analysis, medical image segmentation, and image-to-image translation.

Convolutional neural networks (CNNs) are best suited for data in grid-like structures, while recurrent neural networks (RNNs) and Transformers are more effective for dealing with sequential data. However, RNNs can suffer from vanishing and exploding gradient problems due to their recurrent connections. Transformers, on the other hand, use self-attention mechanisms that overcome these issues and allow for parallelization. U-Net specializes in image segmentation tasks and has a contracting and expanding path. Transformers have achieved impressive results in various natural language processing (NLP) tasks and have also been adapted for vision tasks.

Different deep learning models have specific applications. CNNs are useful for tasks that involve images and videos, while RNNs are employed for NLP, speech recognition, and time series analysis. Encoder-decoder architectures are used for machine translation, text summarization, and image captioning. U-Net is primarily designed for medical image segmentation and image-to-image translation. Transformers are versatile and can be used for various tasks, including NLP, computer vision, and reinforcement learning.

These models represent only a subset of the vast field of deep learning, and many other specialized architectures and variations are developed for specific applications. Deep learning continues to evolve, with researchers constantly pushing the boundaries of what can be achieved using neural networks. It is important to note that the success of deep learning for writer identification depends on the availability of a sufficiently large and diverse dataset and the choice of the appropriate model architecture and training techniques. Distinguishing subtle writing style traits that change over time or in different settings makes writer identification difficult. However, deep learning can automate this process. The next section outlines the proposed system based on the Swim Transformer deep model, which can effectively capture individuals' unique writing styles and characteristics for writer identification.

4 The Proposed Swin Transformer-based System

It is important to mention that the Swin Transformer is a widely utilized backbone architecture for various vision tasks, including image classification and object detection. It is an extension of the Vision Transformer (ViT) architecture, which was initially proposed for natural language processing tasks. In contrast to the Transformer, the Swin Transformer is highly efficient and has superior accuracy. The Swin Transformer has demonstrated exceptional proficiency in the following areas:

– **Capturing long-range dependencies**: The Swim Transformer's self-attention mechanism can detect long-range dependencies in input sequences, including strokes and pen movements in handwriting. This

allows the model to comprehend the context and relationships between various handwriting parts, which could result in more precise analysis and identification.

– **Modeling temporal dynamics**: Writing by hand involves a step-by-step process, and the Swim Transformer can accurately represent the changing patterns of handwriting by focusing on important details at various points in time. This technology can accurately capture the subtleties of stroke sequence, velocity, and pen pressure, which are all key factors in analyzing handwriting and identifying its author.

– **End-to-end learning**: The Swim Transformer can learn task-specific features from raw handwriting sequences without needing hand-engineered features. This approach simplifies the analysis pipeline and may improve performance by allowing the model to learn relevant features specific to the task. Additionally, the Swim Transformer is adaptable to different tasks related to handwriting analysis, such as handwriting recognition, signature verification, and writer identification. By adjusting the output heads or incorporating additional task-specific modules, this model can be customized to specific requirements, providing flexibility and versatility.

– **Robustness to variations**: Identifying individual writers based on their handwriting can be difficult due to the significant variations in handwriting styles. However, the Swim Transformer's capability to model long-range dependencies and capture context can aid in handling these variations, thus increasing the accuracy and robustness of the identification process. Additionally, the Swim Transformer's attention mechanism provides insights into the model's decision-making process, making it possible to understand which parts of the handwriting the model focuses on when making predictions. This feature interpretability is valuable for analysis and verification, as it can explain the model's decisions.

Below is a detailed explanation of the Swin Transformer architecture and its key concepts. It is important to note that the Swin Transformer was specifically designed to overcome the challenges faced by the original ViT. It introduced two crucial ideas - hierarchical feature maps and shifted window attention. The name 'Swin Transformer' is derived from the Shifted Window Transformer. The overall structure of the Swin Transformer is depicted in Figure 1. The most significant components of the Swin Transformer are the 'Patch Merging' and 'Swin Transformer' blocks, which we will explore further in the following sections.

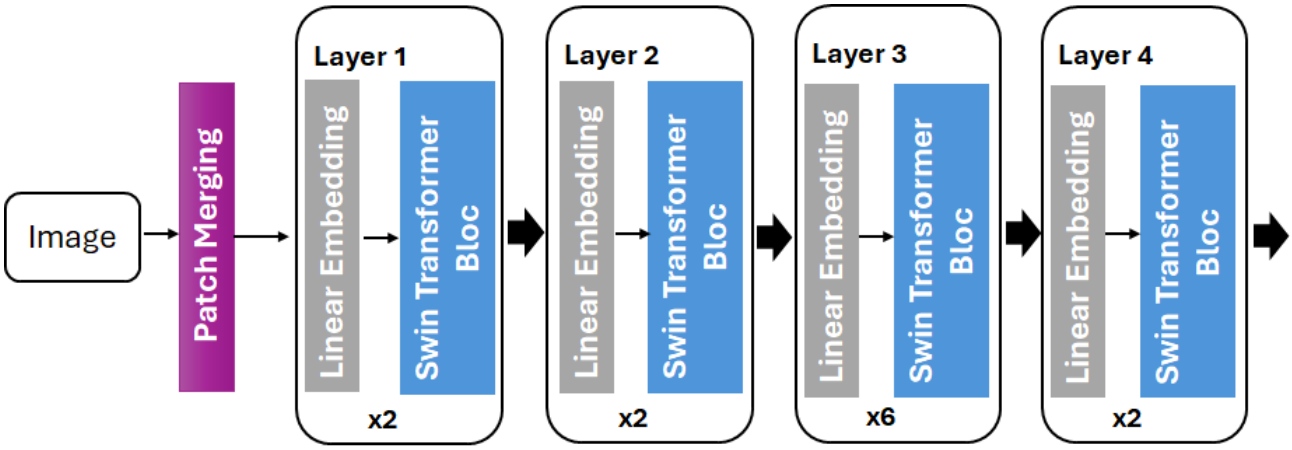

Figure 1: Overall Swin Transformer architecture, Adapted from [1].

– **Hierarchical Feature Maps**: The Swin Transformer approach is distinct from ViT as it generates hierarchical feature maps. These maps are essentially the tensors that result from each layer. The term "hierarchical feature maps" implies that these maps are amalgamated from one layer to another, reducing the spatial dimension (downsampling) as we progress through each layer.

– **Patch Merging**: Convolutional Neural Networks (CNNs) like ResNet utilize convolution for downsampling feature maps. On the other hand, the Swin Transformer uses a patch merging technique to achieve downsampling without convolution. A patch refers to the smallest element in a feature map. Through the patch merging operation, the input is downscaled by a factor of n by combining $n \times n$ patches and concatenating them depth-wise.

– **Swin Transformer Block**: The system consists of two sub-units (as shown in Figure 2). Each sub-unit contains a normalization layer, an attention module, another normalization layer, and a Multilayer Perceptron (MLP) layer. The first sub-unit utilizes a Window MSA (W-MSA) module, while the second sub-unit utilizes a Shifted Window MSA (SW-MSA) module.

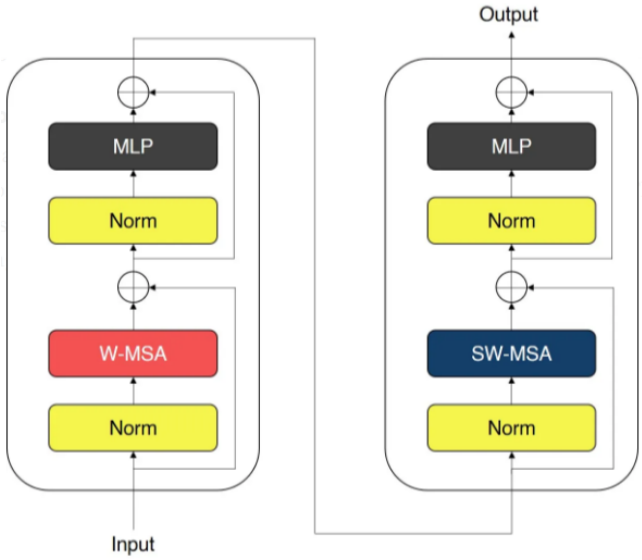

Figure 2: The Swin Transformer block, with 2 sub-units. The first sub-unit applies W-MSA, and the second applies SW-MSA, Adapted from [1].

– **Window-based Self-Attention**: The ViT uses global self-attention to determine the relationship between each patch and all other patches in the image, but this approach becomes impractical for high-resolution images due to its quadratic complexity. The Swin Transformer solves this problem using a window-based MSA approach, where the attention is calculated only within each window, a group of patches. This significantly improves complexity over the standard MSA, as the window size remains fixed throughout the network. The complexity of window-based MSA is linearly proportional to the number of patches, equivalent to the image size, making it a much better option than the quadratic complexity of standard MSA.

– **Shifted Window Self-Attention**: Although window-based MSA is useful, it is limited. The network's modeling capability is limited because self-attention is confined to each window. To address this challenge, the Swin Transformer includes a Shifted Window MSA (SW-MSA) module after the W-MSA module. Significant cross-connections are created between them by shifting windows, improving network performance.

5 The Proposed ResNeSt-50-based System

Note that ResNeSt-50 is a deep-learning architecture that was originally developed for image classification tasks. However, it can also be utilized for writer identification and provides a range of benefits, including:

– **High Accuracy**: ResNeSt-50 is a state-of-the-art image classification model known for its exceptional accuracy in recognizing various image patterns. In the context of writer identification, the model can distinguish between different writing styles with high precision.

– **Transfer Learning**: Pre-trained ResNeSt-50 models can be used for large-scale image datasets like ImageNet. This transfer learning technique enables the model to be fine-tuned on a smaller dataset for writer identification, taking advantage of the general features it learned from diverse images.

– **Deep Architecture**: ResNeSt-50 is a deep convolutional neural network (CNN) with multiple layers, making it capable of capturing intricate patterns and details in handwritten text. This depth helps the model learn complex representations of a writer's unique style.

– **SE Module**: The SE module within ResNeSt emphasizes important features while suppressing less relevant ones, which can benefit writer identification by highlighting distinctive handwriting characteristics and reducing noise.

– **Next Stage Module**: The Next Stage module modifies the standard bottleneck building block and enhances feature extraction capabilities, improving performance in identifying subtle differences in handwriting.

– **Scalability**: ResNeSt-50 is scalable, meaning its depth and width can be adjusted to meet specific requirements. This flexibility allows the model's capacity to be fine-tuned based on the size and complexity of a writer identification dataset.

The ResNeSt-50 model, similar to other ResNets, follows a hierarchical method of processing visual information (see Figure 3). It breaks down images into smaller patches or blocks, allowing it to identify broad global features and specific local details. This hierarchical approach is vital in distinguishing between different writing styles. The model has a distinctive structure that incorporates STM units, which efficiently process images. These units save computational resources while capturing complex details, making them ideal for identifying writers across a large dataset.

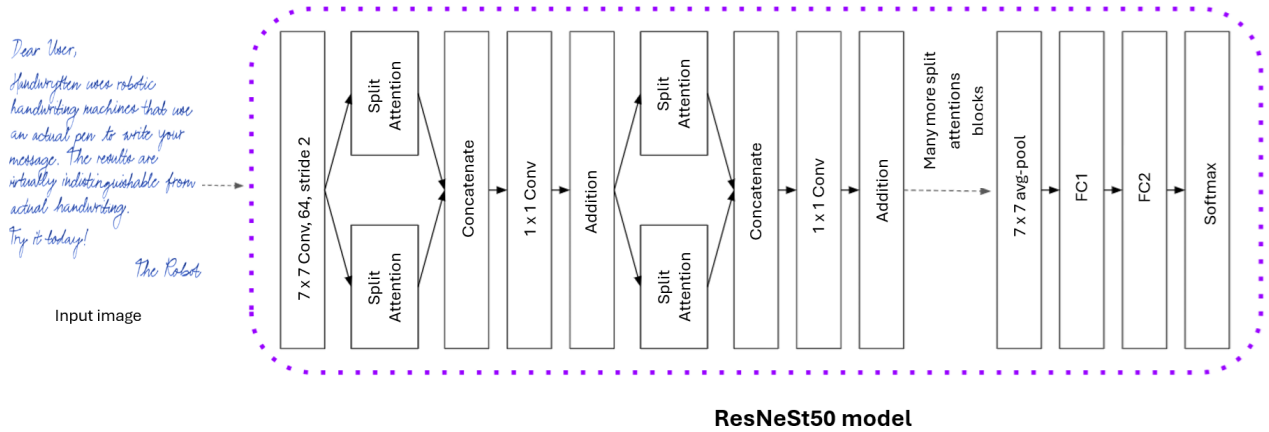

Figure 3: ResNest50 General Architecture.

– **Input Layer:** Typically, RGB images with dimensions of 1x224x224 pixels are used as input for ResNeSt-50.

– **Convolutional Blocks:** ResNeSt-50 consists of convolutional blocks with convolutional layers, batch normalization, and ReLU activations.

– **Residual Connections:** ResNeSt-50, like ResNet, utilizes skip connections to mitigate the vanishing gradient problem during training.

– **Nested Residual Blocks:** ResNeSt-50 stands out for its use of nested residual blocks. These blocks split the input feature maps into several groups, each processed simultaneously using its convolutional path. The resulting output from each group is then combined to create a more comprehensive representation. This process of nesting and aggregation enhances the model's ability to capture multi-scale information effectively.

– **Split-Attention Mechanism:**ResNeSt-50 has a split-attention mechanism crucial to its performance. It helps the network concentrate on separate parts of the feature maps within each nested residual block, resulting in better precision when capturing intricate patterns and details.

– **Output Layer:** Typically, a neural network's last output layer comprises a global average pooling layer, followed by a fully connected layer. The number of neurons in this layer varies depending on the task. For instance, in image classification, the number of neurons is usually equivalent to the number of classes in the dataset.

– **Efficiency**: ResNeSt-50 achieves excellent performance while being computationally efficient.

6 Experimental Results and Discussion

6.1 Used Database

We utilized the CVL database [16], a public writer retrieval, identity, and word-spotting database (see Figure 4 for samples). The database contains 7 handwritten texts (1 German, 6 English). The dataset included 310 writers. 283 writers had to write 5 pieces, and 27 wrote 7. A cropped version (just handwritten) and a 300 dpi RGB color image of the handwritten and printed text are available for each text. A unique ID identifies the writer, while an XML file stores each word's Bounding Boxes.

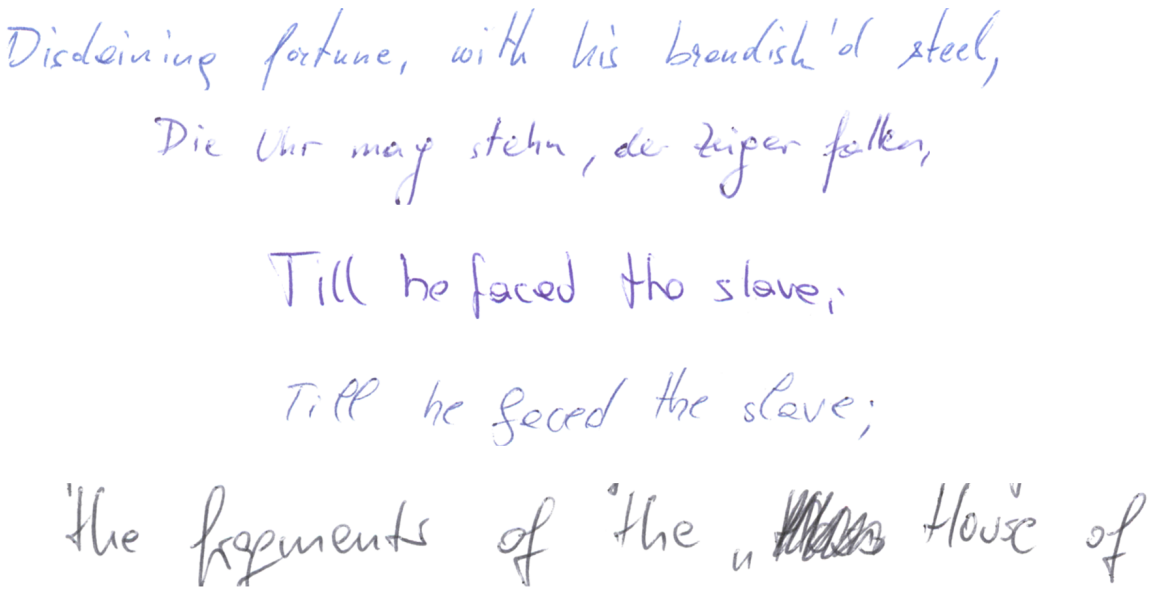

Figure 4: Samples from the CVL database for the same writer with different pens [5].

6.2 Proposed Swin Transformer-based System

6.2.1 Training and Validation

We evaluated the model's training on the CVL dataset using accuracy and loss metrics. The parameter settings used during the training process are shown in Table 2. The training was stopped after 600 epochs, and the resulting curves demonstrated significant progress. The training loss curve and validation loss plot decreased until they reached a stable point with a small gap between them, as shown in Figures 5 and 6. We concluded that the model loss on the validation dataset was lower than on the training dataset. In summary, the accuracy and loss curves indicate a good fit. We achieved promising results during training by configuring the model architecture and selecting hyperparameters thoroughly.

Table 2: Image pre-processing and Training Setting on the CVL database.

Image pre-processing	Segmenting the CVL database into two sets: 8924 images for training and 3648 for testing. All images are resized to ($30 \times 30 \times 3$). No other processing.
Training setting	$patch\_size = (2, 2)$, $dropout\_rate = 0.03$, $num\_heads = 8$, $embed\_dim = 64$, $num\_mlp = 256$, $qkv\_bias = True$, $window\_size = 2$, $shift\_size = 1$, $image\_dimension = 32$, $num\_classes = 310$, $learning\_rate = 10^3$ $batch\_size = 128$, $num\_epochs = 1000$, $validation\_split = 0.1$, $weight\_decay = 0.0001$, $label\_smoothing = 0.1$

Figure 5: Training and Validation Accuracy.

6.2.2 Testing

We conducted a test using the CVL database, which consists of 3648 images with dimensions of $30 \times 30 \times 3$, to evaluate our writer identification method. Our evaluation metric was accuracy (refer to equation 1), which

Figure 6: Training and Validation Loss.

measures the percentage of correctly identified writers among all the instances tested. The Swin Transformer proved effective, achieving an impressive accuracy of 98.50%, surpassing the state-of-the-art methods.

$$\text{Accuracy} = \frac{\sum_{i=1}^{n}(P_i = \text{True}_i)}{n} \times 100\ [\%] \tag{1}$$

where P_i is the string of characters that the model recognizes for the i^{th} input image, True_i is the true transcription of the i^{th} image, and n is the size of the test database.

6.3 Proposed ResNeSt-50-based System

6.3.1 Training and Validation

We used the ResNeSt-50 model for training and validation when identifying writers on the CVL database. The parameter settings used during the training process are shown in Table 3. Our learning curves (see Fig. 7) demonstrate that the model performed well, with a balanced performance. We noticed a consistent decrease in the loss on the training curve, which indicates that the model learned effectively from the dataset. In addition, the validation curve also showed a descending trend that converged to a stable point. It is important to note that the gap between the training and validation loss was minimal, highlighting the model's strong generalization ability. This 'good fit' scenario indicates that the ResNeSt-50 model learned effectively from the data without overfitting or underfitting, making it a reliable choice for identifying writers on the CVL database.

Table 3: Image pre-processing and Training Setting.

Image pre_processing	Segmenting the CVL database into three sets: 70% images for training, 30% for validation. All images are resized to (224 × 224 × 3).
Training setting	batch_size = 32, lr = $1e^{-5}$, optimizer = Adam, Loss = CrossEntropy, num_epochs = 8

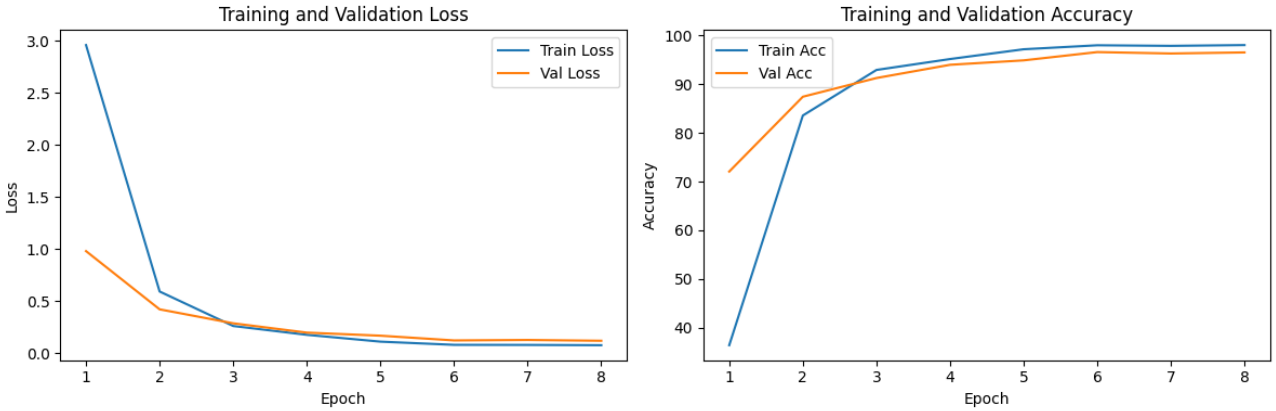

Figure 7: Train and validation Loss and Accuracy.

6.3.2 Testing

Using ResNeSt-50, we achieved 96.61% test accuracy, slightly lower than the Swin Transformer. Figure 8 displays a writer identification test example.

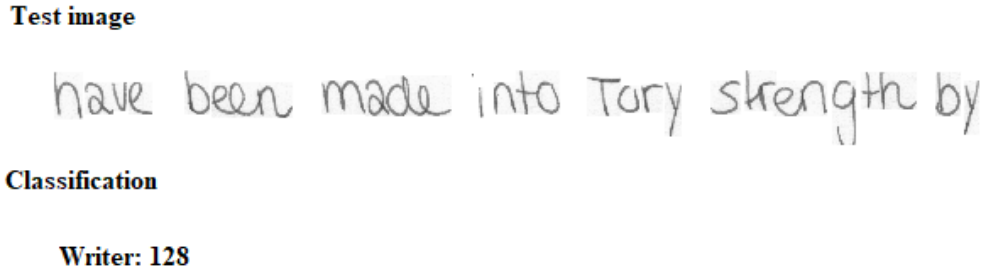

Figure 8: Test example.

6.3.3 Comparison of the proposed Models

Both the Swin Transformer and ResNeSt-50 are deep learning architectures applied here to the task of writer identification. Here are some differences between the two:

1. Architectural Differences:
   - Swin Transformer: It uses a "shifted window" approach for hierarchical visual data processing. The input image is divided into non-overlapping windows, and attention mechanisms are applied hierarchically.
   - ResNeSt-50: It is an extension of the widely used ResNet architecture. It incorporates a unique split-attention mechanism that improves the network's feature detection capability by grouping channels and enabling them to attend to distinct input areas separately. ResNeSt-50 is a variation of ResNeSt that has 50 layers.
2. Parameter Efficiency:
   - Swin Transformer: IT can achieve competitive performance with fewer parameters than other architectures, making them advantageous in scenarios with limited computational resources.
   - ResNeSt-50: It can achieve competitive performance with fewer parameters than other architectures, making it advantageous in scenarios with limited computational resources.
3. Performance:
   - Regarding the writer identification task, the performance of Swin Transformer and ResNeSt-50 can differ based on the dataset and task. Ultimately, deciding between the two options may require empirical evaluation to determine which is best suited for the dataset.
4. Computational Requirements:
   - Swin Transformer's hierarchical approach might make it more efficient regarding memory and computation compared to deeper networks like ResNeSt-50.

Swin Transformer and ResNeSt-50 are deep-learning architectures with different characteristics and trade-offs. Choosing between them should be based on empirical evaluation of specific datasets and computational resources.

6.4 Comparison with Related Works

To provide a thorough analysis, we compared our systems' performances with existing methods, highlighting the superior performance of the proposed models (see Table 4). The results show a significant improvement in the accuracy of writer identification, demonstrating the potential of our method to advance the field and contribute to the progress of writer identification research.

Tested on the CVL dataset, and compared to the ResNeSt-50, the swin transformer provides high accuracy. This is due to its hierarchical structure and computational efficiency. The Swin Transformer's architecture allows for capturing information at multiple scales, which is beneficial for analyzing handwriting patterns across different levels of granularity, such as individual words or letters. Additionally, its shifted window mechanism reduces computational complexity, making it more feasible to apply to large-scale datasets commonly encountered in writer identification tasks. These features enable Swin Transformer to potentially achieve higher accuracy and efficiency compared to ResNeSt-50, particularly when dealing with complex and varied handwriting styles.

Note that most related works are done at the word level, but identifying writers at patch and page levels, as we proposed in this work, poses greater difficulty compared to the word level due to the complexity and

Table 4: Methods and best performance for writer recognition.

Ref.	Method	Mode, level	Dataset	Accuracy
[6]	RNN with bidirectional long short-term memory	on-line, word	BIT (133 writers), Chinese (186 writers)	100% for English, 99.46% for Chinese
[7]	GR-RNN	word	IAM (657 writers), CVL (310 writers), Firemaker (250 writers), CERUG-EN (105 writers)	96.4%, 99.3%, 98.8%, 99.1%
[8]	FEM-WI	Page	IAM (657 writers), Firemaker (250 writers)	above 90%
[9]	CNN AlexNet with transfer learning for feature extraction and SVM for classification	word	QUWI (1017 writers with four digitized pages and approximately 60 words written by each writer)	92.78% for English, 92.2% for Arabic, and 88.11% for a combination of both languages
[14]	CNN (AlexNet adaptation)	word	CVL (310 writers), IAM (657 writers)	only Top1 and Top5 are presented
[15]	FragNet	word, page	IAM (657 writers), CVL (310 writers), Firemaker (250 writers), CERUG-EN (105 writers)	only Top1 and Top5 are presented
[11]	ResNet	off-line, patch	IAM (657 writers), Firemaker (250 writers), CVL (310 writers), CERUG-EN (105 writers)	97.50%, 99.61%, 96.16%, 88.95%
[12]	CNN	word	ICDAR 2011 and 2013 Writer Identification contest, CVL (310 writers)	Top1, Top2, Top5 and Top10 for ICDAR contests, Top2, Top3, Top4 for CVL
[13]	Deep multi-stream CNN	Patch	IAM and HWDB	99.01% (301 writers), 97.03% (657 writers with one English sentence input), 93.85% (300 writers with one Chinese character input)
Ours	**Swin Transformer**	**off-line, patch**	**CVL**	**98.50%**
Ours	**ResNeSt-50**	**off-line, page**	**CVL**	**96.61%**

variability of handwriting features. Analyzing larger text segments requires capturing diverse features such as stroke variations and spatial relationships, increasing computational complexity. However, it also offers richer insights into writing styles, demanding advanced algorithms for accurate identification.

7 Conclusion and Future Work

In this work, we aimed to showcase the efficacy of the Swin Transformer and the ResNeSt-50 models, two advanced deep-learning techniques, for identifying multi-script writers on the CVL database at the patch level and the IAM database at the page level.

ResNeSt-50 is an effective tool for identifying writers due to its high accuracy in image classification tasks. Its ability to learn and capture intricate patterns and features within images is particularly helpful in identifying and distinguishing between different handwriting styles. ResNeSt-50 can leverage pre-trained models on large image datasets like ImageNet, allowing for transfer learning and fine-tuning on smaller datasets. The architecture is scalable and adaptable to input image sizes and resolutions, making it suitable for various handwriting samples. ResNeSt-50's deep layers extract hierarchical and abstract features from input images, capturing unique characteristics of each writer's handwriting regardless of writing style and quality variations. The model can benefit from regularization techniques to prevent overfitting and improve generalization capabilities, a crucial aspect when dealing with limited training data. ResNeSt-50 can also be efficiently parallelized for GPU acceleration, leading to faster training and inference times, making it essential for real-time or large-scale writer identification systems.

On the other hand, the Swin Transformer is a promising option for identifying writers, with several advantages that make it stand out. Its hierarchical architecture allows it to capture local and global features, making it ideal for detecting subtle nuances in handwriting styles. Its scalability makes it suitable for handling large datasets and high-resolution images. Swin Transformer's Shifted Window mechanism reduces computational complexity, leading to faster training times and more efficient performance. Its state-of-the-art results in various computer vision tasks suggest that it has the potential to excel in writer identification. Finally, the Swin Transformer's modular architecture offers flexibility for adapting to different tasks and datasets, making it a compelling choice for various writer identification scenarios.

Due to its hierarchical representation and scalability, the Swin Transformer is a hopeful framework for writer identification assignments. Nevertheless, it presents the usual difficulties linked with deep learning models, such as data prerequisites, computational resources, and interpretability. Its triumph in a specific application will rely on the amount and quality of data accessible, as well as the specific needs of the task.

Through rigorous training on a sizable dataset of handwritten text samples, we harnessed the models's capacity to capture distinctive features and establish robust representations of each writer's unique style. The experimental results revealed exceptional performance, with an accuracy of 98.5% achieved by the Swin Transformer on the test database comprising 3648 images.

Research on identifying writers mainly focused on analyzing handwritten texts in a single script. However, it is widely recognized that in most cultures, a significant portion of the population speaks and writes in at least two languages. By examining writing samples from the same individuals in multiple scripts, we can investigate the fascinating challenge of identifying writers who use multiple scripts. This approach can reveal recurring writing patterns among various scripts, aiding in author identification.

Moreover, identifying multi-script writers based on small regions of text images still needs improvement compared to other deep-learning challenges. Building upon this research, future work could explore other deep-learning architectures and datasets to enhance the performance of writer identification systems further. Investigating the influence of different types of data augmentation, incorporating contextual information, and considering temporal dependencies in the analysis could be potential avenues for improving accuracy and expanding the applicability of writer identification techniques.

References

[1] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, *Swin transformer: Hierarchical vision transformer using shifted windows*, Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012, 2021.

[2] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, and others, *Resnest: Split-attention networks*, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2736–2746, 2022.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and others, *An image is worth 16x16 words: Transformers for image recognition at scale*, arXiv preprint arXiv:2010.11929, 2020.

[4] Florian Kleber, Stefan Fiel, Markus Diem, Robert Sablatnig, *CVL-database: An off-line database for writer retrieval, writer identification, and word spotting*, Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, pp. 560–564, 2013.

[5] Faraz Ahmad Khan, Muhammad Atif Tahir, Fouad Khelifi, Ahmed Bouridane, Resheed Almotaeryi, *Robust off-line text-independent writer identification using bagged discrete cosine transform features*, Expert Systems with Applications, volume 71, pages 404–415, 2017.

[6] Xu-Yao Zhang, Guo-Sen Xie, Cheng-Lin Liu, Yoshua Bengio, *End-to-end online writer identification with recurrent neural network*, IEEE Transactions on Human-Machine Systems, volume 47, number 2, pages 285–292, 2016.

[7] Sheng He, Lambert Schomaker, *GR-RNN: Global-context residual recurrent neural networks for writer identification*, Pattern Recognition, volume 117, page 107975, 2021.

[8] Enock Osoro Omayio, Indu Sreedevi, Jeebananda Panda, *Funnelling Ensemble Method for Writer Identification (Fem-Wi)*, Available at SSRN 4022914, 2022.

[9] Arshia Rehman, Saeeda Naz, Muhammad Imran Razzak, Ibrahim A Hameed, *Automatic visual features for writer identification: a deep learning approach*, IEEE Access, volume 7, pages 17149–17157, 2019.

[10] Sheng He, Lambert Schomaker, *Fragnet: Writer identification using deep fragment networks*, IEEE Transactions on Information Forensics and Security, volume 15, pages 3013–3022, 2020.

[11] Malihe Javidi, Mahdi Jampour, *A deep learning framework for text-independent writer identification*, Engineering Applications of Artificial Intelligence, volume 95, page 103912, 2020.

[12] Stefan Fiel, Robert Sablatnig, *Writer identification and retrieval using a convolutional neural network*, Proceedings of the 16th International Conference on Computer Analysis of Images and Patterns, pages 26–37, 2015.

[13] Linjie Xing, Yu Qiao, *Deepwriter: A multi-stream deep CNN for text-independent writer identification*, 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pages 584–589, 2016.

[14] Sheng He, Lambert Schomaker, *Deep adaptive learning for writer identification based on single handwritten word images*, Pattern Recognition, volume 88, pages 64–74, 2019.

[15] Sheng He, Lambert Schomaker, *Fragnet: Writer identification using deep fragment networks*, IEEE Transactions on Information Forensics and Security, volume 15, pages 3013–3022, 2020.

[16] Kleber, Florian, et al. ”Cvl-database: An off-line database for writer retrieval, writer identification and word spotting.” 2013 12th international conference on document analysis and recognition. IEEE, 2013.