

Enhanced scene text recognition using deep learning based hybrid attention recognition network

Ratnamala S. Patil, Geeta Hanji, Rakesh Huded

¹Department of Electronics and Communication, PDA College of Engineering, Kalaburagi, India

Article Info

Article history:

Received Nov 11, 2023

Revised Mar 22, 2024

Accepted Apr 17, 2024

Keywords:

Alignment-free sequence-to-sequence

Attention mechanisms

Convolutional neural network

Hybrid attention recognition network

Scene text recognition

ABSTRACT

The technique of automatically recognizing and transforming text that is present in pictures or scenes into machine-readable text is known as scene text recognition. It facilitates applications like content extraction, translation, and text analysis in real-world visual data by enabling computers to comprehend and extract textual information from images, videos, or documents. Scene text recognition is essential for many applications, such as language translation and content extraction from photographs. The hybrid attention recognition network (HARN), unique technology presented in this research, is intended to greatly improve efficiency and accuracy of text recognition in complicated scene situations. HARN makes use of cutting-edge elements including alignment-free sequence-to-sequence (AFS) module, creative attention mechanisms, and hybrid architecture that blends attention models with convolutional neural networks (CNNs). Thanks to its novel attention processes, HARN is capable of comprehending wide range of scene text components by capturing both local and global context information. Through faster network convergence, shorter training times, and better utilization of computing resources, the suggested technique raises bar for state-of-the-art. HARN's versatility makes it a good choice for range of scene text recognition applications, including multilingual text analysis and data extraction. Extensive tests are conducted to assess the effectiveness of HARN approach and demonstrate its ability to greatly influence real-world applications where accurate and efficient text recognition is essential.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Ratnamala S Patil

Department of Electronics and Communication, PDA College of Engineering

Kalaburagi, India

Email: ratnamala_12@rediifmail.com

1. INTRODUCTION

The language seen in images of the natural world contains important and practical information. The applications indicated above are regularly used in a variety of real-world situations. Computer vision applications may be found in many different fields, including geolocation, content-based picture or video retrieval, robotic navigation, automatic license plate identification, helping people who are blind or visually impaired, and image interpretation [1]. Despite significant progress made in the field of natural scene text recognition, a number of issues still need to be addressed and resolved. There are many difficulties and complexity in the realm of image processing. Numerous elements, including as background noise, blurriness, blockage, low resolution, intricate backdrops, and differences in text properties like size, color, and orientation, provide obstacles in this specific environment [2]. Deep learning methods come with a high resource cost, including memory, computing power, and energy. Deep learning techniques or the foundational ideas of computer vision inform the algorithms used for scene text recognition and

identification. Because of their resource needs, these approaches can be challenging to apply in real-time embedded systems, particularly those running on hardware that only supports integer operations. Many methods have been developed to recognize text in naturalistic settings [3], [4].

The use of deep learning algorithms has led to significant advancements in the field of scene text recognition in recent years. Significant advancements have been achieved in the field of sequence learning-based methods. In contemporary approaches [5], [6], the employment of an encoder-decoder architecture is a commonly used strategy for word detection in pictures of natural situations. One popular method for encoding the input picture is to employ convolutional neural networks (CNNs) and recurrent neural networks (RNNs), respectively. RNNs with attention processes [7] or connectionist temporal classification (CTC) [8] are the methods used to decode the encoded input into the target text. The present paradigm for scene text recognition mostly relies on the attention-based encoder-decoder paradigm.

A typical system designed to deal with text seen in real-world settings must have text detection and identification as basic features. Text detection's primary objective is to locate and isolate the exact region of a picture containing text, with a high level of accuracy being the major focus. Text may be retrieved from the original image after the region has been identified. Text recognition's primary objective [8] is to extract and accurately recognize text from a well-defined picture. Figure 1 depicts the text recognition procedure.

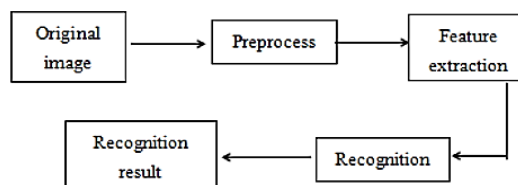


Figure 1. Text recognition process

Scene text detection is the first and most important step in the scene text recognition process. Conventional approaches such as linked component-based strategies and sliding window-based techniques have been used for scene text identification. To identify possible locations for scene text recognition, sliding window techniques are utilized. The method makes use of techniques that entail using a moving window to scan the input image of a natural scene. Connected component-based techniques are an excellent way to extract characters from scene text taken in actual environments. To complete this goal, these algorithms use extreme area extraction and color clustering techniques. Improving scene text identification performance is the aim of the hybrid technique. The advantages of linked component-based and sliding window-based techniques are combined in the suggested methodology. In their investigation, they used the contrast-enhancement maximally stable extremal regions (CE-MSERs) detector. The detector uses a wide variety of hybrid approaches to optimize and improve the contrast of the image. The maximally stable extremal areas algorithm is a special technique designed to enhance the distinction between areas with and without text [9]. It serves as the basis for this notion.

Scene text recognition is a crucial stage in the process that has to be implemented in order to convert scene text regions into string texts. Various methods have been developed and refined to facilitate the recognition and categorization of text in various contexts. The word classification-based method, the sequence-based method, and the character classification-based approach are some of the techniques available for text recognition in scenes. Character candidates are identified and recognized individually through character classification-based processes, akin to those utilized by [10]. Word classification techniques rely on word training and a preexisting vocabulary to augment their efficacy. Four primary classes may be distinguished among the typical irregular short tandem repeat (STR) techniques: the attention-based method, the multi-direction encoding-based approach, the text shape rectification-based technique, and the CTC-based approach. Shape rectification is the act of eliminating distortion from text, which produces a more consistent visual representation of words and enhances irregular text recognition. Before using traditional text recognizers, irregular text is first transformed into regular text using shape rectification techniques [11]. The first network to be presented for text normalization was the spatial transformer network (STN). The system's main goal was to correct individual letters as well as word representations in their whole.

To improve its performance, Wang *et al.* [12] included the thin-plate spline (TPS) transformation method into its text normalization module. Improving the module's capacity to manage complex text distortions was the aim of this inclusion. The growing ubiquity of sophisticated rectification modules may be

ascribed to their pivotal function in effectively handling an extensive array of distortions. Numerous factors affect the recognition algorithms' performance and memory use. The use of CTC has substantially assisted the development of numerous applications, such as online handwritten character identification and speech recognition. In the STR domain, the CTC technique is a well-known and often used prediction tool. By applying a prediction method based on CTC during the model's training phase, the early algorithms that used CTC in STR showed better results. This method made use of developments in voice recognition. Although STR has performed satisfactorily, CTC is not without its limits. Working with narrow distributions might provide difficulties due to the intricate underpinning mechanism of the CTC method. Prediction issues with irregular two-dimensional data, like STR, might not be a good fit for it.

In the realm of scene text recognition, the need for innovative solutions to address the complexities of real-world scenarios is ever-increasing. The motivation to advance the state-of-the-art in this domain stems from the recognition of the challenges faced in accurately understanding text within images. These are driven by the opportunity to develop and refine techniques that can revolutionize scene text recognition, enabling applications ranging from image-based translation to content extraction from visual data. With a commitment to improving accuracy and efficiency, the aim is to make a meaningful impact on the field, providing tools and insights that can benefit various industries and applications.

- A novel hybrid attention recognition network (HARN) is developed which improves text recognition accuracy.
- The hybrid approach combines CNNs with attention models, providing an efficient solution for local context modeling and long-term relationship modeling in text recognition tasks.
- The AFS module is developed that introduces advanced alignment techniques, enhancing recognition accuracy, especially in complex and lengthy sequences, addressing a critical need in scene text recognition.
- The integration of attention mechanisms within the model architecture enables the capture of both local and global context information, improving the accuracy of scene text recognition.

The research is organized into four sections: the first section introduces; first section of the research starts with background of scene text recognition along with problem associated with the same. Second section of the research focuses on various state-of-art techniques for scene text recognition. In third section of the research, HARN is developed and its architecture is designed. At last, HARN is evaluated in fourth section considering different dataset to prove the proposed model efficiency.

2. RELATED WORK

Scene text recognition is widely recognized as a challenging and long-standing research area in computer vision [13]. Numerous academics have dedicated their efforts to enhancing the capability of scene text recognition in real-world scenarios. Moreover, scene text recognition serves a multitude of applications, including but not limited to picture retrieval, visual navigation systems, and text reading for individuals who are blind or visually impaired. The task of scene text identification is highly challenging due to the presence of numerous scene text orientations in our surroundings. The different scene orientations include horizontal, random, curved, and vertical orientations [14]. Moreover, photos that include text can be categorized into two distinct groups: images of natural landscapes and scanned documents. The presence of text in photographs of natural scenes often exhibits variations in appearance and is often found in intricate surroundings. In contrast, text observed in scanned papers typically possesses consistent and uniform backgrounds. The process of recognizing text can be accomplished using a conventional technique known as optical character recognition (OCR). OCR can achieve a high level of accuracy when processing text that is displayed in a legible font with a clear and uncluttered backdrop. OCR is not suitable for text recognition in natural environments due to various issues. The challenges encompass congested and intricate environments, low-quality and blurred images, diverse text orientations, and a multitude of fonts. The performance of text recognition algorithms can be hindered by challenges in differentiating between text and non-text elements, particularly when certain items closely resemble text to human readers.

The integration of scene text detection and scene text recognition has garnered considerable interest in recent years, with several proposed models being the focus of attention. Various techniques have been developed for the detection of text orientation in scenes, including curved, arbitrary, and horizontally oriented text. The current state-of-the-art solutions primarily prioritize scene texts that are horizontally oriented due to their superior readability. The approach proposed by [15] is centered around the recognition of horizontally oriented scene text. The task is achieved through the utilization of a patch-based classification system, effectively preserving the distinctive areas of an image. A methodology was presented for information retrieval that utilizes the hidden markov model (HMM) to locate horizontally oriented date information.

Various methodologies [16] have been devised in recent scholarly investigations to identify text within oriented scenes. The techniques can be categorized into two main groups: 2D based techniques and

rectification-based techniques. Rectification networks are utilized by rectification-based methodologies to transform the oriented scene text into a standardized format. The corrected texts are then sent into recognizers. The character-based approaches employ a two-step procedure in which potential characters are initially localized and subsequently recognized. Several illustrative techniques have been proposed for the candidate character localization process. The utilization of maximum stable extremal regions (MSERs) is employed for the purpose of detecting character components. Strokes are employed by [15] for character localization. In addition, a generative shape model has been developed to effectively extract character traits from a limited set of distinct photographs. The production of word-level recognition results typically involves the use of heuristic algorithms or language models during the candidate character recognition stage. Nevertheless, character-based techniques exhibit certain limitations in terms of accurately determining the exact positions of characters. The primary factors contributing to this issue include a cluttered background and insufficient letter spacing.

An OCR task refers to the process of recognizing and extracting text from scenes. The OCR technology performs effectively in controlled environments. However, it encounters difficulties when dealing with complex typefaces and backgrounds, such as characters written in cursive. Image deterioration can occur due to various environmental factors such as blur, light, and orientation. These factors can lead to significant issues. Subsequent iterations in the development process encompass the extraction of pertinent elements from textual sections and the management of variations in size, color, and layout. The accuracy of text recognition can be influenced by various factors such as font sizes, styles, colors, textures, and orientation [16]. The proposed approach involves utilizing appropriate text localization techniques to effectively handle these complexities. Various methodologies have been suggested to achieve the tasks of extraction, recognition, and location.

The convolutional recurrent neural network (CRNN) system utilizes convolutional and recurrent neural networks to generate character sequence features. This system is designed as an end-to-end solution. The vanilla CTC algorithm is designed for 1D probability distributions. The conditional probability of labels from 2D distributions can be computed using the 2DCTC approach, as described in [17]. Therefore, it is ideally suited for analyzing unpredictable or inconsistent text. The organizations robust text recognizer with automatic rectification (RARE), attentional scene text recognizer (ASTER), MORAN, and residual colorectal neoplasia (RCN) successfully transformed the inconsistent text images into accurate versions. Subsequently, text recognition was accomplished by employing an attention-based recognition network within an encoder-decoder framework. The TPS transformation method can be employed to accurately rectify inconsistent scene text [9]. The rectification of irregular scene text can be enhanced by incorporating a symmetry constraint into the rectification network, as suggested by [11]. Guo *et al.* [10] developed a multi-object rectification network that operates by modifying the offset of each pixel. A potential solution for correcting individual characters in an input picture is to employ a character-aware neural network instead of correcting the entire picture. For the purpose of enhancing spatial information retention, 2D-based approaches employ encoding techniques to convert the input picture into 2D representations.

A dual relation module [18], [19] has been developed specifically for scene text recognition in order to extract complementary characteristics simultaneously. The module consists of two branches: a long-range contextual branch and a local visual branch. The local visual branch utilizes a topologically-aware operation to effectively capture the unique characteristics within individual characters and identify distinguishing elements among multiple characters. The long-range contextual branch integrates inter-character associations into feature maps through a straightforward yet efficient approach. The dual relation module is a plug-and-play element that can be seamlessly integrated into contemporary deep architectures. The user has provided the text [20]. The following text presents a model specifically designed for attention, which aims to enhance attention alignment by acquiring more refined visual characteristics. The initial stage involves the application of deformable convolution to extract high-quality visual features. The subsequent stage employs a layered self-attention architecture with intermediate supervision to enhance the quality of visual feature representation. In addition, the attention alignment process is facilitated by employing spatial location encoding.

3. PROPOSED METHODOLOGY

The proposed model combines an attention-based encoder, an AFS module, and an attention-based decoder, named HARN consists of an attention-based encoder, an AFS and an attention-based decoder. Attention-based decoder consists of two modules known attention decoder+transformer and the other is the AFS module. The model for training known as the attention decoder+transformer model ensures proper guidance for better understanding of and representation of features. Upon assumption it is considered that the attention-based encoder and the AFS module that predicts the character sequences, that minimize the parameters in comparison with the recognizer. Figure 2 shows the proposed architecture.

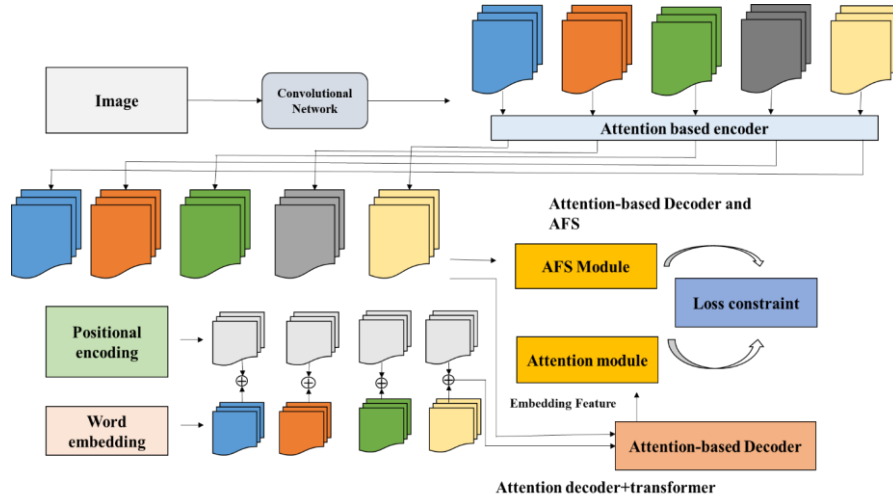


Figure 2. Proposed architecture

3.1. Attention-based encoder

Content-based attention has been emerged as a main part of the sequence model and relevant models for different tasks which accommodate long interactions without considering the distance. However, they are not capable of extracting fine-grained local information. The CNN utilizes a commonly employed position-based kernel to perform a scan on the two-dimensional input data. The main limitation of the system is the lack of global relationship modeling. However, the system is able to capture local feature patterns through the established links. The creation of a recognition model involves the combination of convolutional and attention models. The convolutional module is primarily dedicated to local context modeling, whereas this module focuses on establishing long-term relationships.

3.2. Extended attention

The two-layer sub-sample module enhances memory utilization by eliminating discriminative features, as opposed to the conventional feature selection stage that focuses on extracting visual features using deeper CNNs through the utilization of two convolutional layers. A linear operation is applicable for resizing the encoder-length vector A through the shape of (V, f_s) wherein V denotes the sequence length and f_s is the size. Once this is done the input is fed to the encoder which consists of a collection of P_g encoder blocks which consists of a forward pass and attention-based convolutional module, the attention mechanism consists of a PS module. The elementwise addition consisting of word-embedding with temporal encoding throughout the attention score within each layer for generalization on various input lengths. Figure 3 shows the attention-based encoder decoder framework. The output is computed by (1):

$$H_j = \text{softmax} \left(\frac{SM^V + ST^V}{\sqrt{f_m^j}} \right) \tag{1}$$

Wherein S depicts the queries and $S = AY_s, Y_s \in T^{f_s, f_m^j}$, M depicts the keys as $M = AY_m, Y_m \in T^{f_s, f_m^j}$. T denotes positional encoding and $T = RY_t, Y_t \in T^{f_s, f_m^j}$, $R \in T^{V * f_s}$. The output is integrated and projected as depicted as (2):

$$PS(Z) = \text{integrate}(H_1, H_2, \dots, H_j) Y_q \tag{2}$$

Wherein $Y_q \in T^{f_x * f_x}$ consists of a linear transformation as $f_x = j * f_x^j$ and $f_m^j = f_x^j$.

3.3. Short-range context modelling

Here, the model integrates a convolutional module with the self-attention module by capturing both global and local contexts. Point-wise convolution inside the self-attention module makes up the Conv module. The gated linear activation layer, sometimes referred to as pointwise convolution via factor expansion, makes up the convolution module. Context modelling is accomplished and processing is reduced

with a 1-D (dimension) depth convolutional module. An activation layer that trains the regularization of deep models receives the batch-wise output. The local and global perspectives of the picture are successfully combined by the self-attention model and the convolutional module. The two feed-forward levels include segments that make up the forward pass module. The encoder's output is represented by (3). Here, however FP refers to the forward pass, Conv denotes the convolutional module and norm_layer refers to normalization. For input A_k to the encoder block k , $Output_k$ refers to the output of the block, for $k \in [1, P_g]$.

$$\begin{aligned} A_k &= A_k + 0.5 \text{FP}(A_k), A_k = A_k + \text{PS}(A_k) \\ A_k &= A_k + \text{Conv}(A_k), \text{Output}_k = \text{norm_layer}(A_k + 0.5\text{FP}(A_k)) \end{aligned} \quad (3)$$

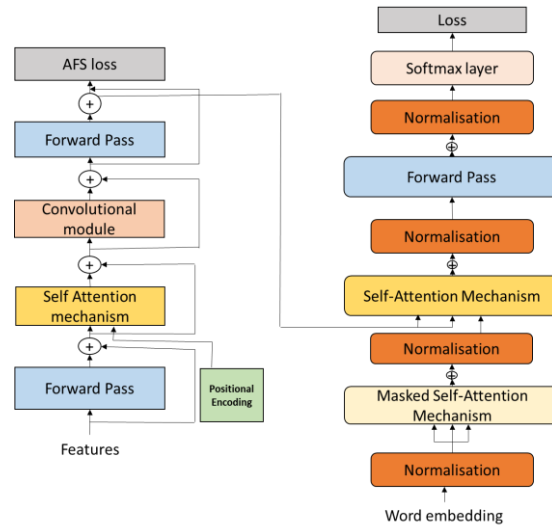


Figure 3. Attention-based encoder decoder framework

3.4. Attention decoder+transformer model

A picture is the input used by the two-layer subsampling module. To change the encoder vector's length, the module performs a linear operation. The length of the encoder vector is modified by combining the attention-based encoder with a feature-based representation. The AFS module and the decoder-based transformer model are the two modules that implement the features. Effective assumption is the main emphasis of the decoder-based transformer model, and the AFS module aims to achieve the same thing. The decoder block, which makes up the system, is separated into three parts: forward pass, ensemble attention, and forward self-attention. Using the function, a character-level embedding approach is first used to turn the input character into a vector. Every embedding gets an additional copy of the location encoding with the same dimension. By keeping positions from engaging in concurrent positions, the forward self-attention method ensures precise prediction and concentrates on data that comes before the present position. The value for ensemble attention is obtained from the encoder's output, which contains the requests for ensemble attention. The final stage of computing, the forward pass, is separated into two layers: a linear layer with the linear activation function and a linear layer with the rectified linear unit (ReLU) activation. The connections that are still present in the decoder block's sub-layer make up the normalization layer.

3.5. Alignment-free sequence-to-sequence and decoder block

The AFS method is widely recognized as a robust and effective approach for addressing sequential problems. This is accomplished by employing dynamic programming techniques and leveraging. It is important to recognize that the approach relies on several conditional independence assumptions.

The inclusion of these additional presumptions is deemed essential for attention decoders. The training process of the attention decoder can be challenging when dealing with noisy data and long sequences, as it requires achieving precise alignment. In addition, the attention-based model generates predictions at each time step by leveraging the previous characters' activities. The frame-by-frame decoding process is a significant contributing factor to the difficulties encountered when utilizing text recognition in real-world scenarios. The recommended approach for addressing the aforementioned issues is to utilize an attention decoder and an AFS. The decoder consists of two main components: the attention module and the AFS module. The attention module is built upon the transformer-decoder architecture. The primary function

of the AFS module is to transform the output of the encoder into an AFS form. This transformation is achieved using a single linear layer. The attention module based on the transformer-decoder is utilized in the computation of attention-based probabilities, as demonstrated in previous research. The attention mechanism is a regularization approach that is specifically utilized in the training phase. It is implemented into the AFS module. In order for the AFS forward-backward algorithm to operate correctly, it is crucial that the input and output sequences are aligned precisely.

In addition, the inclusion of the AFS module can accelerate the convergence rate of the network. The utilization of the attention module does not require any conditional independence assumptions. The inclusion of this feature enhances the ease of configuring a reliable and efficient monitoring system during the training process of the recognizer. The integration of the attention module and the AFS module is achieved by utilizing a decoder that incorporates both AFS and attention methods. The utilization of the shared encoder network by the decoder module is done in an efficient manner. The attention architecture employs encoding and decoding methods to effectively utilize feature information preceding the current position. Moreover, through the integration of the global probability calculation approach into the AFS design, the system can utilize feature information that goes beyond the current position. The combination of attention decoder and AFS accelerates network convergence and enhances recognition performance. The objective that has been declared is as (4):

$$N = \vartheta \text{loss}_{\text{AFS}} + (1 - \vartheta) N_{\text{AL}} \quad (4)$$

ϑ is the parameter that is tuned in the range $0 \leq \vartheta \leq 1$. N_{AL} shows the attention loss, consists of loss function. loss_{AFS} is the AFS loss. The objective is to minimize the negative log-likelihood when compared to the ground truth.

$$\text{loss}_{\text{AFS}} = -\log R\left(\frac{I}{U}\right) \quad (5)$$

Here R depicts the ground-truth label sequence wherein $U = u_1, \dots, \dots, u_V$ is the sequence that the encoder produces using the training image. K . Each u_v is the probability distribution for the character set E , comprising the task's whole label set as well as the blank label. A mapping function O is defined through the sequence $\rho \in E^V$. O maps ρ by eliminating the duplicate labels first, followed by the blank label. The conditional probability is summed as the probabilities of ρ that is mapped by O with I .

$$R\left(\frac{I}{U}\right) = \sum_{\rho: O(\rho)=I} R\left(\frac{\rho}{U}\right) \quad (6)$$

The probability of ρ is denoted by the $R\left(\frac{\rho}{U}\right) = \sum_{v=1}^V u_v^{\rho_v}$, wherein $u_v^{\rho_v}$ is the probability associated with label ρ_v at time-stamp v .

4. PERFORMANCE EVALUATION

In the performance evaluation of text recognition methodologies across three datasets, including street view text-perspective (SVTP), IC13, and CUTE80, varying levels of accuracy is observed. The evaluation of this proposed framework is conducted using both irregular and regular benchmark datasets. The proposed mHARN utilizes deep learning libraries and is implemented on a system configuration featuring a 4GB CUDA-enabled graphics card with 16 GB of RAM, operating on the Windows platform, the results are evaluated in the form of graph and tables.

4.1. Dataset details

4.1.1. street view text dataset

The street view text (SVT) [21] dataset is a collection of 647 outdoor street images sourced from Google Street View, designed for text recognition research. Each image is meticulously annotated with information about the location, size, orientation, and transcribed text present within it. The dataset's notable complexity stems from the wide array of text instances it contains, showcasing various languages, fonts, sizes, orientations, and appearances. This rich diversity makes it an invaluable resource for evaluating text detection and recognition algorithms, especially in the context of real-world challenges such as variable lighting conditions, occlusions, and image quality. The street view text (SVT) dataset is often split into training and test sets, enabling researchers to develop and assess text recognition models for handling the complexities of text in outdoor scenes.

4.1.2. ICDAR 2013 (IC13)

The ICDAR 2013 (IC13) [22] dataset is a collection of 1,015 scene text images, many of which are inherited from the ICDAR 2003 (IC03) dataset. These images are crucial for research and evaluation in the field of scene text recognition. Each image is meticulously annotated, providing information about text region location, size, orientation, and the transcribed text. The dataset's complexity arises from the diverse range of text instances it contains, including variations in languages, fonts, sizes, orientations, and backgrounds. Researchers use IC13 to develop and assess text recognition algorithms, facing challenges like changing lighting conditions, occlusions, and variable image quality. The dataset is often divided into training and test sets for model development and evaluation, and proper citation is essential when utilizing it for research to acknowledge its creators and source.

4.1.3. CUTE-80

The CUTE80 [23] dataset comprises 288-word images primarily emphasizing non-linear, curved text layouts. This dataset is particularly valuable for research in OCR and computer vision, offering a specialized focus on the recognition of text following curved or irregular paths, as opposed to traditional horizontal or linear text. CUTE80 enables researchers to develop and evaluate algorithms specifically generated to address the particular difficulties in identifying and interpreting text in unconventional layouts, including text shown inside creative creations or on curved surfaces.

4.2. Experimental analysis

The experimental analysis evaluates the performance of various text recognition methodologies across three datasets: SVTP, IC13, and CUTE80. The proposed HARN methodology is evaluated against the existing state-of-art techniques. Table 1 shows the results of SVTP dataset. For SVTP dataset the accuracy comparison graph is plotted, as shown in Figure 4 the RARE method depicts an accuracy of 71.8%, demonstrates moderate text recognition capabilities. ADCD surpasses RARE with an accuracy of 75.8%, indicating improved performance. Arbitrary orientation network (AON) reports an accuracy of 73%, while ACE achieves 73.9%, both falling within a similar performance range. Decoupled attention network (DAN) stands out with an accuracy of 80%, demonstrating strong text recognition capabilities. Char-net reports an accuracy of 78.9%, similar to end-to-end trainable scene text recognition system (ESIR) and ASTER, which achieve accuracies of 79.6 and 78.5, respectively. ScRN exhibits strong performance with an accuracy of 80.8. CASR-DRNet leads the list with a high accuracy of 85, indicating robust text recognition performance, while PS significantly outperforms the rest with an accuracy score of 95.04%.

Table 1. Results for SVTP dataset

Method	Value
RARE [24]	71.8
ADCD [25]	75.8
AON [26]	73
ACE [27]	73.9
DAN [28]	80
Char-net [27]	78.9
ESIR [29]	79.6
ASTER [30]	78.5
ScRN [30]	80.8
CASR-DRNet [5]	85
PS	95.04

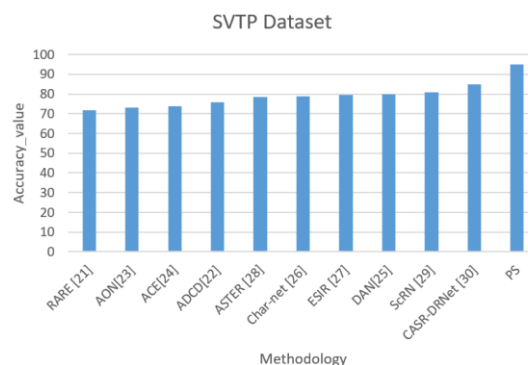


Figure 4. Accuracy comparison for SVTP dataset

4.2.1. IC13 dataset

For IC13 dataset the accuracy comparison graph is plotted, as shown in Table 2 and Figure 5 includes various text recognition methodologies, each accompanied by its respective accuracy value. IBSR shows a strong accuracy of 89.6%, demonstrating proficient text recognition. CA-FCN excels with an accuracy of 91.5%, indicating robust text recognition capabilities. CRF achieves an accuracy of 81.8%, while RARE performs well with an accuracy of 88.6. R2AM maintains a solid performance with an accuracy of 90. EDG stands out with an impressive accuracy of 92.9. ESIR demonstrates a high accuracy of 91.3, while ASTER excels with an accuracy of 91.8. ScRN performs exceptionally well with a high accuracy of 93.9. CASR-DRNet leads the list with a remarkable accuracy of 95.3, showcasing superior text recognition performance. PS significantly outperforms the others with an accuracy score of 98.42.

Table 2. Results for IC13 dataset

Method	Accuracy
IBSR [31]	89.6
CA-FCN [32]	91.5
CRF [33]	81.8
RARE [24]	88.6
R2AM [34]	90
EDG [27]	92.9
ESIR [29]	91.3
ASTER [30]	91.8
ScRN [30]	93.9
CASR-DRNet [5]	95.3
PS	98.42

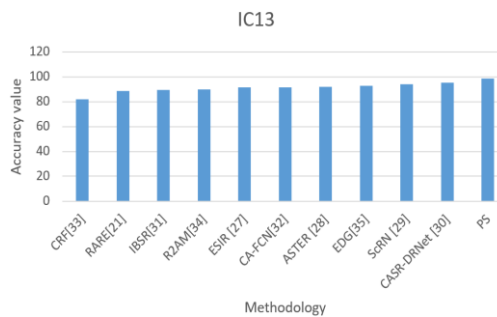


Figure 5. Accuracy comparison for IC13 dataset

4.2.2. CUTE-80

For CUTE80 dataset the accuracy comparison graph is plotted, as shown in Table 3 and Figure 6 the list encompasses various text recognition methodologies, character attention fully convolutional network (CA-FCN) reports an accuracy of 79.9%, showcasing commendable text recognition capabilities. ADCD achieves 69.3%, indicating moderate performance in comparison. RARE exhibits an accuracy of 59.2, implying more modest text recognition performance. PRN excels with a strong accuracy of 88.2, indicating proficient text recognition. SEED demonstrates solid text recognition with an accuracy of 83.6, while ESIR reports a competitive accuracy of 83.3. ASTER exhibits an accuracy of 79.5, marking a noteworthy performance. ScRN achieves a high accuracy of 87.5, demonstrating strong text recognition capabilities. CASR-DRNet leads the list with a remarkable accuracy of 89.2, showcasing superior text recognition. PS significantly outperforms the others with an exceptional accuracy score of 98.96.

Table 3. Results for CUTE80 dataset

Method	Accuracy
CA-FCN [32]	79.9
ADCD [25]	69.3
RARE [24]	59.2
PRN [35]	88.2
SEED [36]	83.6
ESIR [29]	83.3
ASTER [30]	79.5
ScRN [30]	87.5
CASR-DRNet [ES] [5]	89.2
PS	98.96

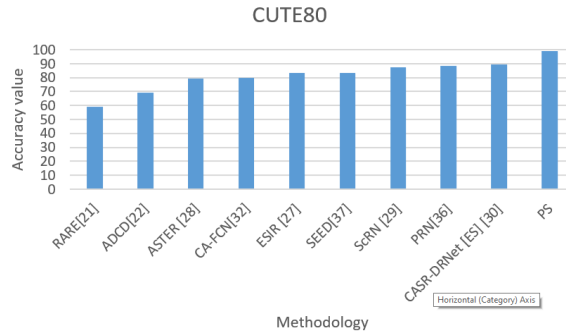


Figure 6. Accuracy comparison for CUTE-80 dataset

4.3. Comparative analysis

In the context of text recognition across diverse datasets, a comparative analysis showcases significant variations in system performance and the degree of improvement achieved by proposed solutions. For the SVTP dataset, the existing system reports an accuracy of 85%, while the proposed system significantly enhances this to 95.04%, resulting in an impressive 11.81% improvement. Moving to the IC13 dataset, the existing system starts at 95.3%, and the proposed system pushes this to 98.42%, signifying a noteworthy 3.26% increase in accuracy. In the case of the CUTE-80 dataset, the existing system stands at 89.2%, and the proposed system exhibits a more modest yet valuable improvement of 0.85%, reaching an accuracy of 89.96%. These results illustrate the efficacy of the proposed systems in enhancing text recognition across various datasets, with the SVTP dataset demonstrating the most substantial advancement, followed by IC13 and CUTE-80, each showcasing varying degrees of progress in text recognition capabilities. Table 4 shows the comparison analysis.

Table 4. Comparison analysis

Dataset	Existing system	Proposed system	Improvisation in %
SVTP	85	95.04	11.81
IC13	95.3	98.42	3.26
CUTE80	89.2	89.96	0.85

5. CONCLUSION




In conclusion, this paper introduces the HARN, a pioneering methodology in scene text recognition. HARN's innovative design, featuring the AFS module, advanced attention mechanisms, and a hybrid architecture, significantly enhances text recognition accuracy and efficiency in complex scene contexts. By addressing alignment challenges, capturing both local and global context information, and efficiently modeling local context and long-term relationships, HARN presents a versatile solution for a wide range of text recognition challenges. Through comprehensive experiments, we have demonstrated HARN's effectiveness in improving accuracy, reducing training time, and optimizing computational resources. The proposed methodology holds potential for numerous real-world applications, where precise and efficient text recognition is crucial, from data extraction in images to multilingual translation and content analysis. As we look ahead, HARN represents a noteworthy step forward in the field of scene text recognition, offering an efficient and adaptable tool to address the evolving demands of text understanding in diverse visual data.

REFERENCES




- [1] P. Dai, H. Zhang, and X. Cao, "SLOAN: scale-adaptive orientation attention network for scene text recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 1687–1701, 2021, doi: 10.1109/TIP.2020.3045602.
- [2] A. A. Chandio, M. Asikuzzaman, M. R. Pickering, and M. Leghari, "Cursive text recognition in natural scene images using deep convolutional recurrent neural network," *IEEE Access*, vol. 10, pp. 10062–10078, 2022, doi: 10.1109/ACCESS.2022.3144844.
- [3] O. Y. Ling, L. B. Theng, A. C. Weiyen, and C. McCarthy, "Development of vertical text interpreter for natural scene images," *IEEE Access*, vol. 9, pp. 144341–144351, 2021, doi: 10.1109/ACCESS.2021.3121608.
- [4] R. Bagi, T. Dutta, and H. P. Gupta, "Cluttered TextSpotter: an end-to-end trainable light-weight scene text spotter for cluttered environment," *IEEE Access*, vol. 8, pp. 111433–111447, 2020, doi: 10.1109/ACCESS.2020.3002808.
- [5] M. Li, B. Fu, Z. Zhang, and Y. Qiao, "Character-aware sampling and rectification for scene text recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 649–661, 2023, doi: 10.1109/TMM.2021.3129651.

- [6] L. Wu, Y. Xu, J. Hou, C. L. P. Chen, and C.-L. Liu, "A two-level rectification attention network for scene text recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 2404–2414, 2023, doi: 10.1109/TMM.2022.3146779.
- [7] O. A. Ademola, E. Petlenkov, and M. Leier, "Resource-aware scene text recognition using learned features, quantization, and contour-based character extraction," *IEEE Access*, vol. 11, pp. 56865–56874, 2023, doi: 10.1109/ACCESS.2023.3283931.
- [8] C. Xue, J. Huang, W. Zhang, S. Lu, C. Wang, and S. Bai, "Image-to-character-to-word transformers for accurate scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2023, doi: 10.1109/TPAMI.2022.3230962.
- [9] P. Lu, H. Wang, S. Zhu, J. Wang, X. Bai, and W. Liu, "Boundary TextSpotter: toward arbitrary-shaped scene text spotting," *IEEE Transactions on Image Processing*, vol. 31, pp. 6200–6212, 2022, doi: 10.1109/TIP.2022.3206615.
- [10] J. Guo, R. You, and L. Huang, "Mixed vertical-and-horizontal-text traffic sign detection and recognition for street-level scene," *IEEE Access*, vol. 8, pp. 69413–69425, 2020, doi: 10.1109/ACCESS.2020.2986500.
- [11] R. Mahadshetti, G.-S. Lee, and D.-J. Choi, "RMFPN: end-to-end scene text recognition using multi-feature pyramid network," *IEEE Access*, vol. 11, pp. 61892–61900, 2023, doi: 10.1109/ACCESS.2023.3280547.
- [12] Y. Wang *et al.*, "PETR: rethinking the capability of transformer-based language model in scene text recognition," *IEEE Transactions on Image Processing*, vol. 31, pp. 5585–5598, 2022, doi: 10.1109/TIP.2022.3197981.
- [13] B. Li, X. Tang, X. Qi, Y. Chen, C.-G. Li, and R. Xiao, "EMU: effective multi-hot encoding net for lightweight scene text recognition with a large character set," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5374–5385, Aug. 2022, doi: 10.1109/TCSVT.2022.3146240.
- [14] Y. Xu, P. Dai, Z. Li, H. Wang, and X. Cao, "The best protection is attack: fooling scene text recognition with minimal pixels," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1580–1595, 2023, doi: 10.1109/TIFS.2023.3245984.
- [15] K. Narwani, H. Lin, S. Pirbhulal, and M. Hassan, "Towards AI-enabled approach for urdu text recognition: a legacy for urdu image apprehension," *IEEE Access*, pp. 1–1, 2024, doi: 10.1109/ACCESS.2022.3203426.
- [16] Y. Gao, Y. Chen, J. Wang, and H. Lu, "Semi-supervised scene text recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 3005–3016, 2021, doi: 10.1109/TIP.2021.3051485.
- [17] S. Anbukkarasi, V. E. Sathishkumar, C. R. Dhivyaa, and J. Cho, "Enhanced feature model-based hybrid neural network for text detection on signboard, billboard and news tickers," *IEEE Access*, vol. 11, pp. 41524–41534, 2023, doi: 10.1109/ACCESS.2023.3264569.
- [18] J. Xu, W. Ding, and H. Zhao, "Based on improved edge detection algorithm for english text extraction and restoration from color images," *IEEE Sensors Journal*, vol. 20, no. 20, pp. 11951–11958, Oct. 2020, doi: 10.1109/JSEN.2020.2964939.
- [19] S. Anbukkarasi and S. Varadhaganapathy, "A novel approach for handwritten tamil character recognition system," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 12, pp. 1489–1495, 2020, doi: 10.5373/JARDCS/V12SP3/20201401.
- [20] M. A. Panhwar, K. A. Memon, A. Abro, D. Zhongliang, S. A. Khuuro, and S. Memon, "Signboard detection and text recognition using artificial neural networks," in *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, IEEE, Jul. 2019, pp. 16–19, doi: 10.1109/ICEIEC.2019.8784625.
- [21] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, "Recognizing text with perspective distortion in natural scenes," in *2013 IEEE International Conference on Computer Vision*, IEEE, Dec. 2013, pp. 569–576, doi: 10.1109/ICCV.2013.76.
- [22] D. Karatzas *et al.*, "ICDAR 2013 robust reading competition," in *2013 12th International Conference on Document Analysis and Recognition*, IEEE, Aug. 2013, pp. 1484–1493, doi: 10.1109/ICDAR.2013.221.
- [23] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Systems with Applications*, vol. 41, no. 18, pp. 8027–8048, Dec. 2014, doi: 10.1016/j.eswa.2014.07.008.
- [24] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 4168–4176, doi: 10.1109/CVPR.2016.452.
- [25] X. Yang, D. He, Z. Zhou, D. Kifer, and C. L. Giles, "Learning to read irregular text with attention mechanisms," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, California: Aug. 2017, pp. 3280–3286, doi: 10.24963/ijcai.2017/458.
- [26] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "AON: towards arbitrarily-oriented text recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 5571–5579, doi: 10.1109/CVPR.2018.00584.
- [27] Y. Liu, Z. Wang, H. Jin, and I. Wassell, "Synthetically supervised feature learning for scene text recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 449–465, doi: 10.1007/978-3-030-01228-1_27.
- [28] T. Wang *et al.*, "Decoupled attention network for text recognition," in *AAAI 2020-34th AAAI Conference on Artificial Intelligence*, 2020, pp. 12216–12224, doi: 10.1609/aaai.v34i07.6903.
- [29] F. Zhan and S. Lu, "ESIR: end-to-end scene text recognition via iterative image rectification," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2019, pp. 2054–2063, doi: 10.1109/CVPR.2019.00216.
- [30] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: an attentional scene text recognizer with flexible rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035–2048, 2019, doi: 10.1109/TPAMI.2018.2848939.
- [31] B. Shi, X. Bai, and C. Yao, "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*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017, doi: 10.1109/TPAMI.2016.2646371.
- [32] M. Liao *et al.*, "Scene text recognition from two-dimensional perspective," in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 2019, pp. 8714–8721, doi: 10.1609/aaai.v33i01.33018714.
- [33] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep structured output learning for unconstrained text recognition," in *3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings*, 2015.
- [34] C.-Y. Lee and S. Osindero, "Recursive recurrent nets with attention modeling for OCR in the wild," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 2231–2239, doi: 10.1109/CVPR.2016.245.
- [35] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 2315–2324, doi: 10.1109/CVPR.2016.254.
- [36] Z. Qiao, Y. Zhou, D. Yang, Y. Zhou, and W. Wang, "SEED: semantics enhanced encoder-decoder framework for scene text recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 13525–13534, doi: 10.1109/CVPR42600.2020.01354.




BIOGRAPHIES OF AUTHOR

Ratnamala S. Patil    received her Bachelor's Degree in Electronics and Communication Engineering from the Visvesvaraya Technological University, Belgaum, India in 2014 and Master Degree in Digital Communication and Networking from same University in 2016. She is currently pursuing her Ph.D. degree from the same university. She is presently working as Assistant Professor in Electronics and Communication Engineering Department Shambasva University Kalaburagi, Karnataka, India. Her primary area of interest is image processing, machine learning, and pattern recognition. She can be contacted at email: ratnamala_12@rediffmail.com.



Geeta Hanji    working presently as Professor in Department of Electronics and Communication Engineering, Poojya Doddappa Appa College of Engineering, Kalaburagi. She has 18 years of Teaching and 10 years of Research Experience, and completed her B.E., M.Tech., and Ph.D. in Electronics and Communication Engineering. Her research area includes digital image processing and pattern recognition. She published more than 55 research papers in above mentioned areas. She has 30 years of teaching experience and 18 years of research experience. She can be contacted at email: geetanjaliapatil123@gmail.com.



Rakesh Huded    have earned an engineering degree in Electronics and Communication from SDM College of Engineering in Dharwad, affiliated with Visvesvaraya Technological University, Belagavi, in 2011. Followed by M.Tech. degree from PDA College of Engineering in Kalaburagi, also affiliated with Visvesvaraya Technological University, Belagavi, in 2013, and culminating with a Ph.D. in Image Processing from Sri Satya Sai University of Technology and Medical Sciences in Sehore in 2019, currently serves as an Assistant Professor at PDA College of Engineering in Kalaburagi, Karnataka. With over five years of experience, and currently guiding research scholar under Visvesvaraya Technological University in Belagavi. He can be contacted at email: rhuded@gmail.com.