Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ31030

A Review of the Development of Text Recognition Techniques

Bo Zhang

Polytechnic Institute, Purdue University, West Lafayette, USA
zhan5055@purdue.edu

Abstract. Document Understanding plays an essential role in the interaction of Artificial Intelligence (AI) systems with real-world information, and Optical Character Recognition (OCR) serves as its foundation. The traditional design of an OCR system relies on several separate modules for text detection, recognition, and understanding, which have been used for a long time. The traditional design has been widely used, while it suffers from several weaknesses: errors can occur during propagation, and computation between modules can be inefficient. To simplify this process, an end-to-end approach was proposed, which has become a trend in the field of document understanding. This review compares two end-to-end approaches: TrOCR refines the traditional OCR architecture by introducing a transformer and unifying separated modules, while the Document Understanding Transformer (Donut) bypasses text recognition and understands the document directly. By comparing two approaches, the paper finds that end-to-end approaches offer significant advantages in accuracy and efficiency, representing a promising direction for future document understanding systems.

Keywords: End-to-End, Transformer, Document Understanding, OCR

1. Introduction

Document understanding plays an essential role in AI development, as it enables machines to extract information from images, such as screenshots or photos of documents, empowering AI systems to interact with the real world. Optical Character Recognition (OCR) plays a key role in this process, and this technology has been widely used across industries, from automated data entry, such as document digitalization, to accessibility tools for visually impaired users. Traditional OCR systems are typically based on Convolutional Neural Networks (CNN) and involve several steps to process the input image, including preprocessing, segmentation, feature extraction, classification, and post-processing [1]. However, these architectures have significant limitations, such as errors that can propagate between modules, low efficiency, and a lack of the ability to understand text [1].

To overcome this issue, researchers have adopted end-to-end learning frameworks that integrate all modules into a unified system. And currently, researchers have achieved significant results with the transformer approach in an end-to-end framework, as its self-attention mechanisms enable models to capture long-range dependencies and contextual relations more effectively than previous architectures. The current research in end-to-end text document understanding has diverged into two paths: one is to use a transformer to refine the traditional OCR architecture and reduce multiple

DOI: 10.54254/2755-2721/2026.TJ31030

modules into an integrated module; the other approach is independent of traditional OCR methodologies, which bypasses the text recognition step, but processes visual information directly, and it's known as the OCR-free approach.

This review examines two representative models of these two paths: TrOCR, which refines traditional OCR through a unified transformer architecture [2], and Document Understanding Transformer (Donut), which bypasses text recognition to directly extract structured information [3]. This paper will focus on analyzing the architectural innovations and methodologies of both approaches, comparing their performance with traditional methods, and discussing how these approaches can inspire future research on document understanding. By reviewing the history of text recognition and document understanding techniques, this paper presents a clear path of how the techniques evolved from pattern matching to transformer-based models, focusing on the characteristics of each era, and predicting the future trend of such models.

2. Literature review

2.1. Early OCR and rule-based approaches

The development of OCR can be traced back to the 1930s, even earlier than the invention of computers. During this period, text recognition was usually finished by applying multiple "rules" to compare the given text with the patterns stored in the system's database. According to Mori's literature review in 1992, there were two major approaches to OCR in the early times [4]. In 1929, Austrian scientist Gustav Tauschek submitted his "Reading Machine" patent in German, which is considered one of the earliest OCR-related concepts in the world [5]. In Tauschek's design, the core mechanisms are optical scanning, photoelectric conversion, and comparison with preset letter templates. Such a design was later called the template-matching approach, and it has deeply inspired later OCR research and has multiple branches based on it. Template-matching approaches were accepted by the business, which requires scanning standardized documents, such as bank documents. Although the original patent was purely mechanical, its idea remained the basis of text recognition research.

The other approach to OCR in the early phase is structural analysis, which recognizes text by analyzing its characteristics, such as thinning the text to make it easier to recognize [6] or analyzing strokes, intersections, and character components. Such a concept was also proposed in the early phase; however, because this approach requires heavy computing, it was limited by cost and hardware constraints at the time. Even after the 1950s, when computers were invented, due to the high cost of computing devices, such an approach was not widely adopted in the commercial sector. However, this approach still plays an essential role in the development of OCR, especially after computers became more affordable. A representative product based on this technique is the IBM 1287, a text scanner that shows greater tolerance for handwritten texts.

These early methods were groundbreaking for their time. However, due to hardware limitations and the limitations of this field of study, early systems are considered rigid, so their recognition results are sensitive to text font or document condition. The advent of machine learning would later enable more flexible and data-driven approaches.

2.2. From statistical learning to transformer

As computers and large datasets become more widely available, OCR research is boosted, marking a fundamental change in how text recognition systems are designed and trained. The 1990s marked a

DOI: 10.54254/2755-2721/2026.TJ31030

paradigm shift in OCR research, as the field had transitioned to a more data-driven statistical approach. Unlike in pattern matching and structural analysis, thanks largely to increased computational power, researchers in this era didn't need to manually input characteristics to the system; they could instead train a machine learning model on labeled images to achieve the same or even better performance.

In the 1990s, based on the theoretical foundations of statistical pattern recognition, two major approaches dominated the field: Hidden Markov Models (HMMs) and Support Vector Machines (SVMs). HMMs are based on the original Markov Chain and have been proposed for decades, and they perform well at modeling the sequential nature of text, especially in handwriting recognition, where character temporal dependencies are important [7]. SVMs were proposed in 1995 and offer superior classification performance by finding optimal decision boundaries in high-dimensional feature spaces [8]. However, both approaches have their limitations: they require hand-crafted features for training and limit their flexibility across different OCR tasks.

The next revolution in text recognition is neural networks, which can extract features directly from the data [9]. Although Conventional Neural Networks (CNNs) still struggle with variable-length sequences, they still make text recognition more flexible than previous approaches. In 2012, AlexNet achieved unprecedented accuracy with a deep CNN, and this success quickly inspired other researchers to spur rapid architectural innovation [10]. One outcome is the popularity of hybrid architectures, such as the Convolutional Recurrent Neural Network, which leverages the strengths of CNNs and Recurrent Neural Networks (RNNs) to achieve outstanding performance and quickly become the baseline for document and scene text recognition [11].

Attention Is All You Need marks a new era of machine learning [12]. Unlike RNNs, transformers' self-attention mechanisms process entire sequences in parallel, significantly increasing the training efficiency. The self-attention mechanisms allow each position to directly attend to all other positions, enabling better capture of long-range dependencies, making this architecture highly promising for document recognition and understanding tasks. Transformers can also learn more effectively with data and parameters, allowing researchers to leverage large pre-trained models through transfer learning. These advantages make the transformer the ideal choice for developing an end-to-end document understanding system, which is a current trend in the field.

3. Case study

The advent of transformer architecture enabled two distinct paradigms for end-to-end document understanding. The first approach refines the traditional OCR pipelines by integrating detection, recognition, and understanding modules into one unified transformer-based architecture. The other approach bypasses text recognition, processing visual information directly to extract structured outputs, a technique known as the OCR-free approach. This section introduces the representatives of these two approaches: Transformer-based Optical Character Recognition (TrOCR) and Donut.

3.1. TrOCR

TrOCR was proposed by Microsoft researchers in 2021, which optimizes the traditional OCR pipeline using a unified transformer architecture [3]. This architecture combines detection, recognition, and language modeling into a single end-to-end trainable framework, which addresses the problems of error propagation and inefficiency in multi-stage systems. TrOCR uses an encoder-decoder architecture that combines pretrained models from both visual and linguistic domains. Visual encoders are based on pre-trained models such as DeiT or BEiT, which split images into

DOI: 10.54254/2755-2721/2026.TJ31030

smaller parts and use self-attention mechanisms to capture both local features and global context. Text decoders are based on language models such as RoBERTa and MiniLM, which generate texts using autoregression and predict morphemes based on visual encoding and prior predictions. The transformer architecture enables TrOCR to perform better across different tasks and reduces its reliance on task-specific data.

The training strategy of TrOCR includes two stages: In the pre-training stage, the model is trained on a large-scale dataset of general printed text images, followed by task-specific pre-training on a handwritten, printed, or scene text dataset. During the fine-tuning stage, the model is trained on labeled datasets according to the specific tasks this model should handle. This strategy allows the model to adapt to specific recognition tasks while still retaining its general OCR capabilities.

Compared with traditional OCR models, TrOCR has a few advantages: Firstly, TrOCR does not require an external language model for image understanding, as it employs the pre-trained image Transformer and a text Transformer that leverage large-scale unlabeled data for image understanding and language modeling. Secondly, TrOCR is easy to implement and maintain since it does not require convolutional networks as the backbone and does not introduce image-specific inductive biases. Also, TrOCR can be easily extended to multilingual recognition by leveraging multilingual pre-trained models on the decoder side and expanding the dictionary. And finally, of course, TrOCR leads the text recognition benchmark in contemporary models, showing its potential in the field.

3.2. Donut

Donut is the model introduced in 2022. The purpose of this model challenges a fundamental hypothesis of document processing: that text recognition is necessary for document understanding. Unlike TrOCR and other OCR models, Donut is an OCR-free model that bypasses the text detection step and maps document images to structured outputs directly. For tasks like form parsing or information extraction, intermediate text conversion might be unnecessary or even harmful for document understanding, since the structure of texts might also include information, discarding it might introduce errors that propagate downstream. While the methodology of Donut can properly avoid this issue.

Donut remains the encoder-decoder design. Swin Transformer is used as the visual encoder, a hierarchical visual transformer that processes images via sliding windows to simultaneously capture local details and global structure while preserving spatial hierarchies. As the Swin Transformer retains the two-dimensional document layout, it keeps the original layout of the document, which is critical for understanding forms, tables, and complex documents. The decoder generated structured output directly from visual features, producing task-specific formats, such as JSON or Key-Value pairs, rather than plain text. Under this architecture, the model can extract information directly from visual patterns, keep the original layout, and avoid the error propagation between modules.

Donut's training strategy emphasizes task-specific optimization. The model is pre-trained on large-scale document datasets, such as IIT-CDIP, to acquire general document comprehension capabilities and to learn to associate visual patterns with semantic meanings without character recognition. Then, for specific tasks, researchers can train it to produce corresponding structured outputs during fine-tuning. This direct optimization toward the final objective, rather than relying on intermediate text recognition steps, achieves specialized performance.

This OCR-free approach offers significant advantages. Firstly, it simplified the pipeline for converting an image to structured text output, eliminating the need for separate detection and recognition modules. Also, by bypassing the text recognition stage, it prevents the error propagation caused by OCR inaccuracies. And since this architecture retains two-dimensional layout

DOI: 10.54254/2755-2721/2026.TJ31030

information, it achieves outstanding performance in recognizing complex documents. And finally, since this architecture eliminates multi-stage processing and consolidates it into a single model, it significantly enhances computational efficiency.

4. Discussion

Transformer has produced two different paradigms for document understanding, exemplified by TrOCR and Donut. Although both leverage self-attention mechanisms and pre-trained models, they differ in their views on the role of text recognition in document understanding tasks. As a successor to traditional OCR, TrOCR retains the text recognition stage, allowing texts to be inspected by humans and, in real applications, fed to downstream NLP models for correction when necessary. Such an approach builds on decades of OCR research and modernizes it with transformers. While Donut discards the text recognition step, it still recognizes the entire document. For many document understanding tasks, structured output matters more than character-level transcription. By optimizing directly from end goals, Donut avoids multi-stage complexity and error propagation.

In the architectural aspect, despite both models sharing an encoder-decoder design, there are still major differences between them. TrOCR treats images as flat patch sequences, while Donut's Swin Transformer maintains hierarchical spatial structure, preserving two-dimensional layouts. TrOCR's decoder, based on language models, generates natural texts benefiting from linguistic priors; Donut's decoder produces task-specific structured tokens, learning semantic representations directly from visual patterns. This fundamental contrast underlies their practical tradeoffs.

These two models also demonstrate their advantages in specific tasks. TrOCR is ideal for traditional OCR tasks, such as digitizing books, historical documents, and scene texts. Its interpretable outputs enable human verification, and it achieves outstanding results on benchmarks. Donut excels when structuring information extraction, such as parsing forms, extracting receipt data, or document classification. Its direct task optimization and spatial layout awareness yield superior performance, particularly for documents where meaning emerges from two-dimensional structures.

While each approach has its limitations, TrOCR still requires separate text detection and struggles with complex layouts. Donut lacks transparency, which makes debugging and verification difficult for researchers. Future research might focus on both methods, aiming to develop hybrid architectures that can either generate explicit text or support direct structured extraction, combining TrOCR's linguistic grounding with Donut's spatial awareness.

5. Conclusion

This paper examines the evolution of document understanding techniques, ranging from pattern matching to contemporary Transformer-based methodologies, and analyzes two representative studies that illustrate divergent paradigms of document understanding. The literature review indicated that the current trend in this field is to adopt end-to-end architectures that minimize redundant modules in classic structures. This study adds to the field by systematically comparing methods that use OCR and those that don't. The results help professionals select the appropriate method for a specific situation. This study is constrained by its emphasis on two representative models and predominantly English-centric benchmarks, which may limit its applicability to multilingual contexts. In the future, the author will focus on more models in multiple languages, emphasizing their efficiency across languages and formats. This review provides a general overview of how document recognition techniques have evolved and highlights the current trend in this area.

DOI: 10.54254/2755-2721/2026.TJ31030

This review analyzes two exemplary papers, TrOCR and Donut, offering insights that may guide future advancements in document understanding systems.

References

[1] Mittal, R., & Garg, A. (2020, July). Text extraction using OCR: a systematic review. In 2020 second international conference on inventive research in computing applications (ICIRCA) (pp. 357-362). IEEE.
[2] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., & Park, S. (2021). OCR-free Document Understanding Transformer. ArXiv. https: //arxiv.org/abs/2111.15664
[3] Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., & Wei, F. (2021). TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. ArXiv. https: //arxiv.org/abs/2109.10282
[4] Mori, S., Suen, C. Y., & Yamamoto, K. (1992). Historical review of OCR research and development. Proceedings of the IEEE, 80(7), 1029-1058.
[5] G. Tauschek, “Reading machine, ” U.S. Patent 2 026 329, Dec. 1935.
[6] Beun, M. (1973). A flexible method for automatic reading of handwritten numerals. Philips Tech. Rev, 33(4 Part I), 89-101.
[7] Bose, C. B., & Kuo, S. S. (1994). Connected and degraded text recognition using hidden Markov model. Pattern Recognition, 27(10), 1345-1363.
[8] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121-167.
[9] LeCun, Y., Haffner, P., Bottou, L., & Bengio, Y. (1999). Object recognition with gradient-based learning. In Shape, contour and grouping in computer vision (pp. 319-345). Berlin, Heidelberg: Springer Berlin Heidelberg.
[10] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121-167.
[11] Shi, B., Bai, X., & Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence, 39(11), 2298-2304.
[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.