A Survey on Transformer for Optical Character Recognition

Sicheng Zhou

Department of Computer Science, University of California, Davis, Davis, United States

sizhou@ucdavis.edu

Abstract. Optical Character Recognition (OCR) transforms visual text into machine-readable form, supporting the large-scale digitization of printed, handwritten, and scene-based documents. Early approaches, such as template matching and motion analysis, relied on handcrafted patterns and were constrained to limited fonts and simple layouts. The introduction of statistical models, including Hidden Markov Models and Conditional Random Fields, expanded OCR capabilities through probabilistic sequence modeling. With the rise of deep learning, Convolutional and Recurrent Neural Networks enabled end-to-end recognition, reducing dependence on manual feature engineering and improving performance on noisy or cursive text. More recently, transformer-based models like TrOCR have redefined OCR by leveraging self-attention and large-scale pretraining, achieving state-of-the-art results across multilingual and domain-specific applications. These models excel in cross-lingual transfer, low-resource adaptation, and specialized domains such as biomedical and historical text recognition, while integrating pretrained vision–language components for greater robustness against degraded inputs. Despite these advances, challenges persist in adversarial robustness, complex document layout understanding, and fairness across underrepresented languages and scripts. Emerging research directions include zero-shot and few-shot learning, modular adapters for scalable multilingual OCR, post-OCR correction pipelines, efficiency improvements, and privacy-preserving inference. This survey outlines OCR's historical progression, highlights deep learning and transformer-based breakthroughs, and points to future work needed to address enduring challenges in this critical field of document analysis.

Keywords: Optical Character Recognition, Document Analysis, Template Matching Motion Analysis, Hidden Markov Models, Convolutional Recurrent Neural Network.

1. Introduction

Among trending Large Language Models (LLM), processing textual input is well developed and multiple useful gadgets such as Named Entity Recognition (NER) has been created to solve this issue [1]. However, most LLMs appeared to be insufficient in handling visual input. To be specific, through the preprocessing sequence of image input, manual text entry is still a common way for transforming visual data into digital form for machine learning [2]. Such a method is naturally expensive and also requires a certain amount of time for the whole processing sequence. According to the IMPACT project, one page of textual source costs around 1 Euro, which sums up to a price range of 500-1000 Euro per source [3]. In comparison, through the extraction of machine comprehensible information from physical texts, Optical Character Recognition (OCR) has seen an immense evolution in algorithms and a wide application in multiple fields, including institutional depositories or digital libraries [4]. Initially, motion analysis was the main approach to apply OCR, which was accurate but lacked significant supremacy in comparison with manual keying [5]. The transformation to neural networks introduced an enhancement on the degree of flexibility and continuous evolution of input data, increasing performance on various tasks [6]. The integration of LLM significantly transformed OCR by introducing end-to-end learning frameworks and leveraging large-scale pre-trained models capable of capturing nuanced linguistic patterns and rich contextual information.

Despite these significant advancements, the field continues to encounter persistent limitations. Contemporary systems remain susceptible to adversarial perturbations, background noise, document degradation, and challenges in distinguishing visually ambiguous characters. Song and Shmatikov demonstrated that imperceptible adjustments to text images can cause OCR engines to misclassify words in ways that diverge sharply from human perception [7]. Similarly, Hartley and Crumpton

highlighted that text degradation frequently induces OCR systems to generate spurious characters or strokes absent in the original input [8]. These findings indicate that, notwithstanding improvements in model architectures and preprocessing techniques, OCR retains structural vulnerabilities that are similar to perceptual illusions in human vision.

An important milestone in OCR research has been the development of end-to-end deep learning frameworks, most notably the Convolutional Recurrent Neural Network (CRNN). Compared to conventional feature-engineering or HMM-based methods, CRNNs integrate convolutional layers for robust spatial feature extraction with recurrent layers to capture sequential dependencies, while using Connectionist Temporal Classification (CTC) to align outputs with variable-length text [9]. This architecture enabled OCR systems to bypass explicit character segmentation, achieving state-of-the-art performance on both printed and handwritten benchmarks [10].

This paper offers a comprehensive examination of the approaches and developments in Optical Character Recognition, tracing its progression from early template-based and handcrafted feature methods to deep learning and transformer-based frameworks. We review the historical evolution of OCR techniques, emphasizing the pivotal role of CRNN in enabling end-to-end recognition, as well as the more recent impact of self-attention architectures on multilingual and document-level performance. In addition, we analyze the various applications of OCR across domains such as digitization of historical archives, automated form processing, medical documentation, and scene text recognition. Furthermore, we highlight current challenges, including robustness to degraded inputs, adversarial vulnerabilities, and the recognition of handwritten or low-resource scripts, and discuss potential research directions that seek to overcome these limitations while expanding the adaptability and accuracy of OCR applications.

2. Historical evolution of ocr techniques

2.1. Introduction of Various Models

2.1.1. Template-based recognition model

In the early stage of OCR, only a few standard fonts and a very limited number of languages could be recognized by this technology [11]. One common decision for researchers is English, as it has only 26 characters, while the recognition of each letter was instructed by a manually created template. For instance, the algorithm will perform a direct visual comparison to match the shape of the character with its stored information. This method is particularly accurate and proficient for printed fonts but has a poor performance on handwritten input or any degradation in the printed characters. Another research-oriented approach is motion analysis. Instead of treating characters as static images, these systems analyzed the temporal variation of light intensity produced as a scanning beam moved across a printed symbol. The resulting signal, effectively a trajectory of brightness changes, was then compared against reference patterns to classify characters. While this method demonstrated accuracy under controlled conditions, it proved inflexible for varied fonts and noisy documents, offering little practical advantage over manual data entry. Consequently, motion-based OCR remained largely experimental and was soon overshadowed by more robust template matching and statistical techniques.

2.1.2. Statistical model

With the rise of statistical modeling, Hidden Markov Models were introduced into OCR as a means of addressing sequential dependencies in text images. An HMM is a probabilistic system in which transitions occur between hidden states, each generating feasible observable symbols [1]. While the internal state sequence remains hidden, the model estimates the likelihood of candidate state sequences based on observed outputs. In OCR, this framework allowed characters or strokes to be treated as hidden states and image features as observable outputs, thereby providing a structured method for modeling text sequences. Although originally developed for applications such as speech recognition and later extended to areas like bioinformatics, HMMs offered OCR systems improved

flexibility compared to rigid template-based methods by learning state-transition and output probabilities directly from data.

Following the widespread use of HMM, Conditional Random Fields (CRF) were introduced into OCR as a more flexible alternative for sequence labeling. Unlike HMMs, which rely on strong independence assumptions between observations, CRFs consider the conditional dependencies of different labels in a given sequence. This allows OCR systems to incorporate rich and overlapping feature sets, such as stroke shapes, contextual cues, and positional dependencies, without being constrained by the limitations of generative models. As a result, CRF-based approaches demonstrated greater robustness in handling ambiguous or noisy character sequences, offering a clear advantage over HMM-based systems.

2.1.3. Neural Networks

Deep learning became a transformative development for OCR during the 2010s. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) were employed to process text images in an end-to-end manner. By learning direct mappings from raw pixel inputs to character sequences, these models outperformed conventional HMM- and CRF-based systems and reduced the need for handcrafted feature design. CRNN architectures, which integrate CNNs for spatial feature extraction with recurrent layers for sequence modeling, achieved significant improvements in handling cursive handwriting, complex layouts, and noisy inputs [10].

To enhance both visual and linguistic representations, pretrained models have been incorporated into OCR systems. Pretrained CNN backbones such as ResNet improve spatial feature extraction, while pretrained language models like BERT refine recognition outputs by leveraging contextual cues [7]. The integration of character-level features helps manage rare symbols and handwriting variations, and bi-directional recurrent designs enable models to capture dependencies from both preceding and succeeding characters. Together, these advances significantly improve OCR adaptability and accuracy across diverse domains.

3. Transformer-based models in OCR

3.1. Introduction of Transformers

A significant turning point in OCR was reached with the introduction of transformer-based architectures. Instead of relying on recurrent layers to model sequences, transformers employ self-attention mechanisms that capture global dependencies across the entire text image in parallel. This design enables models to handle long or complex sequences more efficiently while reducing training time. In OCR, transformers have demonstrated state-of-the-art performance. Li et al. introduced TrOCR, which integrates a pretrained Vision Transformer (ViT) as the encoder and a pretrained RoBERTa language model as the decoder [12]. By pretraining both components separately and fine-tuning them end-to-end on OCR datasets, TrOCR achieved substantial improvements on printed, handwritten, and scene text recognition benchmarks. Similarly, Nguyen et al. highlighted how pretrained transformer-based language models, such as BERT, are increasingly applied in post-OCR correction, refining noisy outputs and improving semantic consistency [13]. Together, these advances show that transformers now drive progress in OCR both at the recognition and post-processing stages.

3.2. OCR Application for Other Languages

One among their chief strengths is the ability to adapt transformer-based OCR models to function across languages. With large pretrained multilingual models such as TrOCR and vision–language transformer variants allowing OCR systems to generalize learning across high-resource languages to low-resource scripts, cross-lingual OCR has become a topical research area [14]. Such models are trained or fine-tuned across some languages or a collection of scripts such that they can embed script-agnostic visual and linguistic features that can be employed in recognition applications where little or no annotation data is provided.

Training these models on multilingual corpora has been successful in improving OCR quality across languages, especially where training data is minimal. Transfer learning and few-shot adaptation have been used to generalize pretrained OCR models to new scripts without a need for much additional supervision. Further, choosing parallel corpora and translation-based alignment schemes has made alignment of recognition consistent across languages possible and contributed to better robustness in multilingual OCR systems.

Most recent research in multilingual transformers within OCR has been equally concerned to address language-specific challenges such as script diversity, diacritics, and culturally specific text representation. By employing language-specific adapters or meta-learning methods, researchers aim to increase OCR model's flexibility across a broad spectrum of linguistic environments. Such research emphasizes the nature of transformer architecture in advancing further the bridging between high- and low-resource scripts and extending OCR to a universally global technology."

3.3. Specialized OCR

Fine-tuning pre-trained models such as TrOCR has shown remarkable advancement in specialized OCR applications, including handwritten or domain-specialized documents. Pre-trained over large-scale text–image datasets, TrOCR has shown the capability of transformer-based models to boost OCR performance in specialized domains by surpassing the performance of traditional systems in identifying complex scripts or degraded input [12]. Specialized models for OCR also go beyond the typical printed text, including biomedical documents or historical records. For instance, Drobac and Lindén applied post-correction techniques as part of combined approaches for processing biomedical and administrative texts, whereas Nguyen et al. discussed post-OCR correction techniques involving the use of domain-specialized dictionaries for legal documents or historical sources [3, 6]. Likewise, handwriting-specialized OCR systems, as discussed by Tappert et al., illustrate how specialized training boosts performance on informal or cursive scripts [5]. Transformer-based OCR must be further pretrained on domain-specific corpora (e.g., medical prescriptions or archival manuscripts) and then fine-tuned on annotated datasets relevant to that field in order to adapt to specialized recognition tasks. This approach allows the model to internalize the unique scripts, vocabularies, and contextual features that are inherent to specialized domains, thereby improving performance and reliability. Moreover, incorporating external resources, such as domain-specific lexicons and ontologies, into OCR correction pipelines has been shown to further boost accuracy by resolving ambiguities and refining recognition outputs [3]. These enhancements enable transformer-based OCR systems not only to recognize text more accurately but also to interpret its contextual meaning within specialized domains.

4. Current Challenges and Future Approaches.

4.1. Post OCR Correction

Post-OCR error correction remains an ever-enduring challenge for the OCR workflow, particularly in disciplines with specialized vocabularies. Nguyen et al. provide a comprehensive survey on post-processing strategies, with emphasis on the use of external resources such as dictionaries and ontologies specifically for the reduction of recognition errors. External resources, however, must be manually created and fail to scale efficiently, thus their scarcity across disparate disciplines [3]. Furthermore, the challenge of domain-specific ambiguity as well as the context-sensitive error is not yet addressed by generic correction strategies. Without robust post-OCR solutions, even accurate recognition systems will provide non-usable outputs for archival or professional use.

4.2. Language Support

As OCR systems aim to support a growing number of languages and scripts, adapting models efficiently remains a major challenge. Retraining large transformer architectures for every new script is computationally expensive and impractical. Language-specific adapters offer a promising solution

by enabling modular fine-tuning for individual languages without altering the full model. Building on transformer-based systems such as TrOCR, which already demonstrate strong cross-domain adaptability [12], adapters provide a lightweight mechanism to address script-specific features. Wang highlights the importance of scalable, multilingual approaches in OCR development, suggesting that adapter-based methods could play a key role in achieving inclusive and robust recognition across diverse linguistic settings [11].

4.3. Ethnic Issue

As OCR systems are increasingly deployed on sensitive materials such as medical records, ID documents, and financial statements, concerns about privacy and security have become pressing. Without adequate protections, digitized outputs risk exposing individuals to data breaches and misuse. Recent work highlights the need for privacy-preserving methods, including on-device OCR, encrypted inference, and federated learning, to limit the exposure of confidential information [14]. Beyond privacy, ethical considerations such as fairness and inclusivity are critical. OCR models often perform unevenly across languages and scripts, raising concerns about bias. Addressing these challenges will be essential for building globally responsible and trustworthy OCR technologies

4.4. Robustness

Expanding OCR capabilities to cover diverse languages and scripts presents a significant challenge, as retraining large transformer-based architectures for each new script is computationally prohibitive. Language-specific adapters have emerged as a promising approach to address this issue by enabling modular fine-tuning tailored to individual languages while preserving the shared knowledge of the base model. Transformer-based frameworks like TrOCR already demonstrate effective cross-domain adaptability, providing a foundation for such modular enhancements [12]. As Wang argues, the future of OCR lies in scalable, multilingual solutions, and adapter-based methods represent a practical pathway toward more inclusive, efficient, and script-aware recognition systems [11].

5. Conclusion

Since the adoption of deep learning and transformer-based models, OCR has progressed far beyond its early template-based and motion analysis methods. The introduction of end-to-end architectures has been especially influential in improving accuracy and adaptability. CRNN models, which combine convolutional and recurrent layers, marked a pivotal advance by enabling robust recognition of cursive handwriting and noisy inputs. Similarly, Graves et al. demonstrated the effectiveness of recurrent neural networks for unconstrained handwriting recognition, showcasing their strength in modeling sequential dependencies. Despite these milestones, OCR still faces challenges in handling degraded documents, adversarial inputs, privacy risks, and underrepresented scripts. Addressing these limitations requires methods that prioritize robustness, fairness, and efficiency, while maintaining scalability across domains. Promising directions include multilingual adapter-based models, privacy-preserving inference, and tighter integration with post-OCR correction pipelines. Furthermore, advances in zero-shot and few-shot learning offer new opportunities to extend OCR capabilities to unseen scripts and specialized domains. As the field evolves, OCR will remain a central technology for digitizing archives, administrative records, and multilingual content, thereby expanding global access to knowledge.

References

[1] Ahonen H, Heinonen O, Klemettinen M, et al. Applying data mining techniques for descriptive phrase extraction in digital document collections [C]//Proceedings of the IEEE International Forum on Research and Technology Advances in Digital Libraries (ADL'98). IEEE, 1998: 2-11. doi:10.1109/ADL.1998.670375.

[2] Singh A, Bacchuwar K, Bhasin A. A survey of OCR applications [J]. International Journal of Machine Learning and Computing, 2012, 2 (4): 314-318. doi:10.7763/IJMLC.2012.V2.137.

[3] Nguyen T T H, Jatowt A, Coustaty M, et al. Survey of post-OCR processing approaches [J]. ACM Computing Surveys, 2022, 54 (6): 124:1-124:37. doi:10.1145/3453476.

[4] Mir Asif A, Hannan S A, Perwej Y, et al. An overview and applications of optical character recognition [J]. International Journal of Computer Science and Information Technologies, 2014, 5 (4): 4587-4590.

[5] Tappert C C, Suen C Y, Wakahara T. The state of the art in online handwriting recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12 (8): 787-808. doi:10.1109/34.57669.

[6] Drobac S, Lindén K. Optical character recognition with neural networks and post-correction with finite state methods [J]. International Journal on Document Analysis and Recognition (IJDAR), 2020, 23 (3): 279-295. doi:10.1007/s10032-020-00359-9.

[7] Song C, Shmatikov V. Fooling OCR systems with adversarial text images [EB/OL]. arXiv:1802.05385, 2018. doi:10.48550/arXiv.1802.05385.

[8] Hartley R T, Crumpton K. Quality of OCR for degraded text images [C]//Proceedings of the Fourth ACM Conference on Digital Libraries (DL'99). ACM, 1999: 228-229. doi:10.1145/313238.313387.

[9] Graves A, Liwicki M, Fernández S, et al. A novel connectionist system for unconstrained handwriting recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31 (5): 855-868. doi:10.1109/TPAMI.2008.137.

[10] Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition [EB/OL]. arXiv:1507.05717, 2015. https://arxiv.org/abs/1507.05717.

[11] Wang J. A study of the OCR development history and directions of development [J]. Highlights in Science, Engineering and Technology, 2023, 72: 409-415. doi:10.54097/bm665j77.

[12] Li M, Lv T, Chen J, et al. TrOCR: Transformer-based optical character recognition with pre-trained models [EB/OL]. arXiv:2109.10282, 2021. doi:10.48550/arXiv.2109.10282.

[13] Lauar F, Laurent V. Spanish TrOCR: Leveraging transfer learning for language adaptation [EB/OL]. arXiv:2407.06950, 2024. doi:10.48550/arXiv.2407.06950.

[14] Cheema M D A, Shaiq M D, Mirza F, et al. Adapting multilingual vision–language transformers for low-resource Urdu optical character recognition (OCR) [J]. PeerJ Computer Science, 2024, 10: e1964. doi:10.7717/peerj-cs.1964.