DOI: 10.54254/2755-2721/2026.BA31810

A Review on the Application of Optical Character Recognition (OCR) in Robotics

Zhenyun Xu

American Community School Athens, Athens, Greece
mou.xuzhenyun@gmail.com

Abstract. Optical Character Recognition (OCR), once primarily associated with scanned document processing, has evolved into a pivotal perceptual competence for robots operating in human-centered environments. This review synthesizes the evolution, integration, applications, and lingering challenges of OCR technology in robotics. The paper traces the paradigm shift from classical handcrafted methodologies—including stroke-based and region-based techniques—to contemporary deep learning architectures, encompassing convolutional, transformer-based, and vision-language models. It further investigates the integration paradigms of OCR within robotic perception pipelines across heterogeneous platforms, including mobile robots, manipulators, autonomous vehicles, and aerial drones. Real-world deployment domains—such as logistics automation, service robotics, medical assistance systems, autonomous driving, and infrastructure inspection—are elaborated to elucidate the practical efficacy and deployment constraints of robotic OCR. Particular attention is paid to robotics-specific challenges, including motion blur, extreme viewpoints, environmental degradation, limited onboard computation, and safety-critical latency requirements. Despite substantial progress in recent years, robust and real-time scene text understanding in unconstrained real-world environments remains an open research frontier. Finally, the review identifies promising future research directions aimed at enabling more reliable, efficient, and context-aware reading capabilities for robots in real-world scenarios.

Keywords: Optical Character Recognition (OCR), Robotic perception, Deep-learning models, Perception pipelines

1. Introduction

The ability to read text embedded in the physical environment is one of the most distinctive human perceptual skills and, concomitantly, one of the most indispensable capabilities for autonomous robotic systems. Street signs, product labels, serial numbers, medication vials, door plates, handwritten notes, and digital displays all encode dense, explicit semantic information that is exceedingly challenging or infeasible to infer solely from geometric cues or object categorization. A warehouse robot that can read "Fragile This Side Up" mitigates costly damage; a delivery robot that recognizes "No Parking Tow Zone" avoids mission abort and potential fines; a surgical assistant that verifies "Lidocaine 2 %" on a vial prevents life-threatening medical errors. Text is compact, culturally universal, and ubiquitous in every human-built space, rendering OCR an exceptionally high-leverage perceptual modality for robotic systems.

The history of OCR stretches back many decades, initially focusing on the processing of typed or printed documents under controlled laboratory conditions. For more than half a century the field remained largely confined to the domain of flat, high-resolution scanned document processing. The watershed moment [1] arrived in the mid-2010s when fully convolutional detectors and sequence-to-sequence recognition models surpassed previous performance benchmarks on natural scene text recognition tasks. Concurrently, robots began transcending structured factories and entering homes, hospitals, offices, warehouses, and public roads , environments saturated with textual cues. The confluence of these two revolutions has propelled scene text detection and recognition to the forefront of interdisciplinary research at the intersection of computer vision and robotics.

Today, robotic platforms incorporate OCR as an integral component to navigate and interact with human-centric environments. Last-mile delivery robots recognize house numbers and traffic regulatory signs. Surgical systems verify medication identity. Agricultural robots authenticate chemical containers. Self-driving vehicles interpret speed-limit signs, construction warnings, and license plates in real time. Infrastructure-inspection drones read minute serial numbers from considerable altitudes. These examples illustrate that OCR has transitioned from a supplementary add-on to a core perceptual modality for robotic systems, on par with object detection, semantic segmentation, and depth estimation. This review synthesises the state of the art across four dimensions:

(1) the evolution of OCR algorithms from classical to modern foundation-model paradigms
(2) architectural patterns for integrating OCR into real-time robotic perception pipelines
(3) application domains of deployed (or deployable) systems
(4) the remaining technical bottlenecks and the most promising research directions.

Consequently, the ability of robots to reliably detect, recognize, and interpret text in unconstrained real-world environments is becoming a decisive factor in ensuring safe autonomy and effective human-robot interaction (HRI). A systematic review of OCR from a robotics-oriented perspective is therefore imperative to comprehensively understand not only algorithmic advancements but also the performance of these methodologies under real-world robotic constraints.

2. Evolution of OCR technology

2.1. Classical era (pre-2015)

Early OCR for natural scenes and by extension early robotic OCR were predominantly reliant on handcrafted features [2]. Among the most influential methods was the Stroke Width Transform (SWT), which computes a stroke width for each pixel and exploits the consistency of character strokes in text regions [3], thus rendering it fast, locally computed, and language-agnostic.

Other classical pipelines adopted region-based proposal methods (e.g. using MSER), HOG + sliding-window SVMs, or early document-OCR engines optimized for high-contrast, fixed-font scenarios. These "document-centric" methods, however, struggled under real-world robotic application conditions: varying lighting, arbitrary viewpoints, low resolution, cluttered backgrounds, and non-uniform fonts.

Thus, although classical OCR and detection methods established critical conceptual underpinnings, their reliability under real-world robotic constraints remained suboptimal.

2.2. The deep learning revolution (2015–2020)

The paradigm shift emerged with the application of deep learning particularly convolutional neural networks began to be applied to scene text detection and recognition. A landmark in detection was EAST (Efficient and Accurate Scene Text detector), which employs a fully convolutional network to directly predict word-level or text-line-level bounding quadrilaterals in natural scenes [4], thereby obviating numerous heuristic intermediate steps (such as candidate aggregation, word partitioning, character grouping) that had been necessary in older pipelines.

On the recognition front, models such as the convolutional-recurrent neural network (CRNN) e.g. CRNN established a robust baseline: inputting a detected text crop into a CNN + sequence model to transcribe the text [5].

Further, by combining detection and recognition into "end-to-end text spotters" (i.e. systems that directly output recognized text from raw images) this integration streamlined the inference pipeline, mitigating latency and error propagation [6]. For example, systems like TextBoxes++ embody this trend by encapsulating detection and recognition within a single forward pass.

Thanks to these advances, by 2020,real-time end-to-end systems capable of operating at tens of frames per second (FPS) on high-performance GPUs had become prevalent, significantly enhancing the viability of using OCR in dynamic, robotic settings.

2.3. Transformer and vision-language era (2020–present)

More recently, text recognition architectures have shifted from recurrent sequence models to attention-based and transformer-based paradigms. For instance, ViTSTR and the Transformer-based model from Rowel Atienza [5] "Vision Transformer for Fast and Efficient Scene Text Recognition" demonstrate that a pure transformer framework can achieve state-of-the-art performance while maintaining sufficient efficiency for practical deployment [7].

Likewise, PARSeq (autogressive transformer-based STR) exhibits enhanced robustness and context-aware decoding capabilities [8], aligning with human-like reading behaviors and achieving superior performance in handling distortions, motion blur, and font variations [9].

Moreover, as vision-language large models become more powerful and ubiquitous, OCR is increasingly being incorporated into broader scene understanding frameworks, integrating detection, recognition, context reasoning, layout analysis, and semantics, which potentially paves the way toward robots that not only read text but also reason about its contextual meaning (e.g. expiration dates, hazard warnings, instructions).

3. Integration patterns in robotic perception pipelines

OCR rarely functions as an isolated module in robotic systems; instead, three dominant integration paradigms have been developed.

In robotic systems, OCR is rarely deployed as a standalone module and is instead integrated into broader perception pipelines. One common approach is tightly coupled semantic mapping, wherein detected text instances and their decoded content are directly embedded into metric or semantic maps. This facilitates high-level symbolic reasoning and navigation, such as guiding a robot toward a door labeled "EXIT."

Another widely used approach is the loosely coupled middleware architecture. In this design, OCR functions as an independent node that disseminates recognized text, confidence scores, and spatial metadata via a communication bus. Higher-level modules, including navigation, manipulation, and human–robot interaction systems, can subscribe to this data on an on-demand basis.

A third integration paradigm is active perception. In such systems, robots proactively plan viewing angles or adjust sensor parameters to maximize text legibility prior to recognition. This may involve viewpoint optimization, zoom control, illumination adjustment, or iterative recognition with language model-aided refinement.

In practice, many robotic perception pipelines adhere to a detect crop recognizable language-model / post-processing workflow, increasingly augmented by vision-language or transformer-based models to rectify recognition errors, disambiguate context-dependent text, and enforce consistency constraints (e.g. dictionary, format constraints).

Because of recent improvements in detection and recognition speed/accuracy (notably via models like EAST + PARSeq / ViTSTR), such pipelines are increasingly feasible for deployment on mobile or power-constrained robotic platforms. Nevertheless, in most real-world systems, OCR remains a component of a comprehensive perception stack (with object detection, segmentation, depth/geometry, semantic reasoning, etc.).

4. Application domains and deployed systems

OCR has evolved into an indispensable component across diverse robotic application domains. Representative use cases include:

In warehouse and logistics environments, OCR empowers autonomous mobile robots to decipher package labels, tote identifiers, and handling instructions under adverse lighting and occlusion scenarios. Accurate text recognition mitigates sorting errors and enhances operational efficiency.

In last-mile delivery and service robotics, OCR enables robotic systems to recognize house numbers, room designations, and signage in both indoor and outdoor contexts. These systems must maintain robustness against weather perturbations, motion artifacts, and substantial variations in font styles and sign designs.

Medical and assistive robotic systems depend on OCR for verifying medication labels, patient identifiers, and dosage specifications. In such safety-critical scenarios, recognition accuracy and reliability are paramount to avert catastrophic consequences.

Autonomous driving platforms utilize OCR to interpret traffic signs, speed limit markers, construction warnings, and temporary signage. While many signs are standardized, non-standard or degraded text persists as a recalcitrant challenge.

Inspection robots and drones leverage OCR to read serial numbers, equipment labels, and traceability tags from extended distances or oblique viewing angles, often under motion blur and low resolution.

In all these domains, OCR is no longer a novel add-on but rather a core perceptual modality, complementary to geometry-based perception (depth, segmentation), semantic detection (objects), and planning/logic.

5. Persistent technical challenges

Despite remarkable laboratory performance and rapid advancements over the past decade, several fundamental challenges persist before robots can read text as reliably as humans in all real-world scenarios. The most recalcitrant difficulties are outlined below:

5.1. Environmental degradation

Motion blur (predominantly in mobile robots or drones), extreme or shallow viewing angles, low or variable illumination, specular reflections (on glossy surfaces), low spatial resolution, and tiny or stylized fonts continue to be dominant failure modes even for modern detectors and recognizers.

5.2. Resource constraints

Many robots (mobile platforms, drones, embedded systems) operate under stringent power, computational, memory, and thermal constraints. Running large transformer-based OCR models may be infeasible without aggressive model compression, quantization, or specialized hardware acceleration.

5.3. Real-time multi-modal contention

In a full robotic perception stack, OCR often competes for computational resources with concurrent tasks such as object detection, semantic segmentation, depth estimation, SLAM (localization & mapping), motion planning, sensor fusion, etc. This contention can drastically reduce OCR throughput or introduce prohibitive latency, unacceptable for safety-critical tasks (e.g. driving, medical).

5.4. Domain gap & data scarcity

Most academic datasets and benchmarks for scene text detection/recognition are curated under relatively controlled conditions (good lighting, legible text, limited distortion). By contrast, real-world robotic environments frequently present "noisy," damaged, occluded, or stylized text. There is a shortage of large-scale, high-quality, annotated datasets capturing the kinds of conditions robots actually encounter , motion blur, low light, varied surfaces, multi-lingual contexts, etc.

Because of these challenges, many published models exhibit strong performance under benchmark conditions but degrade sharply in unconstrained real-world deployments, particularly for robots operating in dynamic and uncontrolled environments.

6. Future directions

Future research endeavors should prioritize the development of ultra-lightweight OCR models enabling reliable operation on low-power robotic hardware. Multilingual and context-aware OCR systems incorporating semantic constraints will further improve robustness in real-world settings. Additionally, adapting vision-language foundation models via compression and distillation may enable robots to reason about textual content rather than merely transcribing it. Finally, continual learning frameworks and active perception strategies offer promising pathways for closing the gap between benchmark performance and real-world deployment.

7. Conclusion

This review has synthesized the role of Optical Character Recognition (OCR) as a core perceptual competence in modern robotic systems. The paper has analyzed the evolution of OCR technologies from classical handcrafted methodologies to deep learning and transformer-based paradigms, elucidating how these advances have enabled deployment in dynamic and unstructured environments. It further discussed integration architectures, real-world application domains, and the unique challenges imposed by robotic platforms, including motion-induced degradation, limited computational resources, and safety-critical latency requirements. Collectively, the analysis shows that OCR has transitioned from a document-centric tool to a fundamental component of robotic perception and decision-making pipelines.

Despite these advances, this review has several limitations. First, the study is based on a literature survey rather than empirical experimentation, and therefore does not quantitatively evaluate system performance across different robotic platforms. Second, the scope of the reviewed literature, while representative, is inherently limited and may not encompass all emerging industrial deployments. Future work could address these limitations by conducting large-scale empirical studies on real robotic systems and incorporating a broader spectrum of datasets and application scenarios. Such efforts would further clarify the practical boundaries of robotic OCR and guide the development of more robust, efficient, and deployable systems.

References

[1] Jaderberg, Max; Simonyan, Karen; Vedaldi, Andrea; Zisserman, Andrew (2016). Reading Text in the Wild with Convolutional Neural Networks. International Journal of Computer Vision (IJCV), 116(1): 1–20.

[2] Karatzas, Dimosthenis; Gomez-Bigorda, Lluis; Nicolaou, Agis; et al. (2015). ICDAR 2015 Competition on Robust Reading. In International Conference on Document Analysis and Recognition (ICDAR), pp. 1156–1160.

[3] Epshtein, Boris; Ofek, Eyal; Wexler, Yonatan (2010). Detecting Text in Natural Scenes with Stroke Width Transform. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2963–2970.

[4] Zhou, Xinyu; Yao, Cong; Wen, He; Wang, Yuzhi; Zhou, Shuchang; He, Weiran; Liang, Jiajun (2017). EAST: An Efficient and Accurate Scene Text Detector. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5551–5560.

[5] Shi, Baoguang; Bai, Xiang; Yao, Cong (2017). An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition (CRNN). IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 39(11): 2298–2304.

[6] Li, Hui; Wang, Peng; Shen, Chunhua (2017). Towards End-to-end Text Spotting with Convolutional Recurrent Neural Networks. — unified detection & recognition of scene text.

[7] Atienza, R. (2021). Vision Transformer for Fast and Efficient Scene Text Recognition. In: Lladós, J., Lopresti, D., Uchida, S. (eds) Document Analysis and Recognition – ICDAR 2021. ICDAR 2021. Lecture Notes in Computer Science(), vol 12821. Springer, Cham. https: //doi.org/10.1007/978-3-030-86549-8_21

[8] Bautista, D., Atienza, R. (2022). Scene Text Recognition with Permuted Autoregressive Sequence Models. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol 13688. Springer, Cham. https: //doi.org/10.1007/978-3-031-19815-1_11

[9] Fang, Shancheng; Xie, Hongtao; Wang, Yuxin; Mao, Zhendong; Zhang, Yongdong (2021). Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).