

OPTICAL CHARACTER RECOGNITION: A LITERATURE REVIEW OF CHALLENGES, METHODS, AND FUTURE DIRECTIONS IN THE NUSANTARA CONTEXT

Muhammad Iqbal Ripo Putra *, Muhammad Zikri Wiguna
Universitas PGRI Pontianak, Jalan Ampera no 88, Indonesia
*correspondence email: ripoputra87@gmail.com

Received: 2025-04-23

Accepted: 2025-08-15

Published: 2025-08-30

Abstract

Optical Character Recognition (OCR) has evolved from rule-based template matching to a multimodal sub-field of computer vision using deep neural networks and vision-language models. While modern OCR systems achieve near-human accuracy on printed Latin text, they struggle with low-resource scripts, degraded media, and writing systems with different visual logic. This review focuses on OCR in the Nusantara context, where over 700 languages coexist with ten traditional indigenous scripts and Jawi. It traces the methodological trajectory from feature-based classifiers to Convolutional Recurrent Neural Networks (CRNNs) with Connectionist Temporal Classification (CTC) loss, encoder-decoder transformers like TrOCR and Donut, and OCR-free vision-language paradigms. The review synthesises script-specific OCR research on Aksara Jawa, Aksara Bali, Sundanese, Batak, Lampung, Bugis-Makassar, and Jawi. Drawing on the NusaAksara benchmark and recent evaluations, the review argues that the challenge is data-related: the lack of curated datasets and the mismatch between abugida grammar and tokenisers for alphabetic languages. The paper concludes by proposing research directions that combine synthetic data generation, transfer learning from related Brahmic scripts, community-driven annotation, and culturally aware multimodal foundation models.

Keywords: optical character recognition, Nusantara scripts, vision-language models

INTRODUCTION

Optical Character Recognition (OCR) is the process by which images of typed, printed, or handwritten text are converted into machine-encoded character sequences. As one of the earliest and most enduring problems in computer vision, OCR has accompanied the digital transformation of paper-based knowledge: from postal automation in the mid-twentieth century to the searchable digitization of national libraries, the parsing of identity documents at border control, the indexing of scientific literature, and the assistive reading technologies that mediate access to the written world for visually impaired users.

The trajectory of OCR research mirrors the broader trajectory of artificial intelligence. Early systems relied on handcrafted features, deterministic segmentation, and template matching. The introduction of statistical learning broadened the range of fonts and styles that could be recognized. The deep learning revolution, beginning around 2012, replaced segmentation-then-classification

pipelines with end-to-end neural sequence models capable of jointly localizing and decoding text in natural scenes. The most recent shift, around 2022 onward, brings transformer-based vision-language models that treat OCR not as a pixel-to-character mapping but as an instance of visual reasoning, with promising consequences for layout understanding, mixed-language documents, and zero-shot transfer.

Despite these advances, OCR remains far from solved when the target script departs from the conventions of modern printed Latin. Cursive Arabic, ligature-rich Indic scripts, vertical East Asian writing, and historical manuscripts each pose distinct technical problems. Among the most underrepresented contexts in international OCR literature is Nusantara, the Indonesian archipelago, which is home to more than 700 living languages and at least ten traditional indigenous writing systems, including Aksara Jawa, Aksara Bali, Aksara Sunda, Aksara Batak, Aksara Lampung, Aksara Lontara (Bugis), Aksara Makasar, Aksara Rejang, Aksara Kerinci, Aksara Pallawa, and Aksara Kawi, alongside the Jawi script used historically for Malay in Arabic-derived characters.

The motivation for an OCR literature review with a Nusantara focus is twofold. First, the practical urgency: many of these scripts are no longer used in daily life, and the human expertise required to read them is concentrated in a shrinking population of specialists, while millions of historical documents, palm-leaf manuscripts, and epigraphic inscriptions remain inaccessible to the wider public. Without robust OCR, large-scale digitization yields image archives that cannot be searched, indexed, or analyzed at scale. Second, the scientific opportunity: Nusantara scripts present a uniquely diverse testbed for OCR research because they are predominantly abugidas of common Brahmic ancestry, which means that methods developed for one script can in principle transfer to others, while each script also exhibits idiosyncratic ligatures, diacritics, and material substrates that probe the limits of current models.

This paper aims to provide a structured synthesis of OCR research relevant to the Nusantara context. It begins with a brief history of OCR and the standard pipeline that has historically structured the field. It then reviews methodological developments grouped into three eras: classical statistical methods, deep learning models centered on convolutional and recurrent architectures, and the transformer and vision-language era. Against this technical backdrop, the paper turns to the specific challenges of OCR in Nusantara, surveys the most recent empirical work on individual scripts, and reflects on benchmarks such as NusaAksara that attempt to measure progress in a principled way. The review closes with a discussion of open problems and proposes directions for future research that align technical advancement with cultural preservation.

The Nusantara Context

Despite these advances, OCR remains far from solved when the target script departs from the conventions of modern printed Latin. Cursive Arabic, ligature-rich Indic scripts, vertical East Asian writing, and historical manuscripts each pose distinct technical problems. Among the most underrepresented contexts in the international OCR literature is Nusantara, the Indonesian archipelago, which is home to more than 700 living languages and at least ten traditional indigenous writing systems, including Aksara Jawa, Aksara Bali, Aksara Sunda, Aksara Batak, Aksara Lampung, Aksara Lontara (Bugis), Aksara Makasar, Aksara Rejang, Aksara Kerinci, Aksara Pallawa, and

Aksara Kawi, alongside the Jawi script used historically for Malay in Arabic-derived characters.

Two considerations motivate an OCR literature review with a Nusantara focus. The first is practical urgency: many of these scripts are no longer used in daily life, and the human expertise required to read them is concentrated in a shrinking population of specialists, while millions of historical documents, palm-leaf manuscripts, and epigraphic inscriptions remain inaccessible to the wider public. Without robust OCR, large-scale digitisation yields image archives that cannot be searched, indexed, or analysed at scale. The second is scientific opportunity: Nusantara scripts present a uniquely diverse testbed for OCR research because they are predominantly abugidas of common Brahmic ancestry, which means that methods developed for one script can, in principle, transfer to others, while each script also exhibits idiosyncratic ligatures, diacritics, and material substrates that probe the limits of current models.

Research Questions

This review is organised around four research questions.

1. What is the methodological trajectory of OCR research over the past three decades, and which architectural paradigms currently define the state of the art?
2. What is the present empirical state of OCR research for Nusantara scripts, and how does it compare to mainstream OCR for high-resource languages?
3. What are the cross-cutting challenges that distinguish OCR in the Nusantara context from OCR for Latin or East Asian scripts?
4. What research directions appear most promising for closing the gap between current Nusantara OCR performance and the levels achieved for high-resource scripts?

METHOD

Design

This study is a narrative literature review with scoping elements, designed to map the breadth of OCR research relevant to the Nusantara context rather than to perform a strict quantitative meta-analysis. A scoping approach is appropriate because the literature on Nusantara OCR is fragmented across journals in computer science, computational linguistics, digital humanities, and cultural heritage studies; published in both English and Indonesian; and distributed across local, regional, and international venues that do not share a common indexing scheme. A formal systematic review in the PRISMA sense would have excluded the substantial body of relevant work in Indonesian-language venues that are not consistently indexed in Web of Science or Scopus.

Data Sources

Literature was retrieved from seven categories of sources: international indexing databases including IEEE Xplore, ACM Digital Library, ScienceDirect, and SpringerLink; preprint and proceedings repositories including arXiv and the ACL Anthology; the Google Scholar discovery engine for both English and Indonesian publications; Indonesian academic journal portals including the Indonesian Journal of Computing and Cybernetics Systems (IJCCS), the JOIV International Journal on

Informatics Visualization, JURTEKSI, JINACS, and MALCOM; the Hugging Face Model Hub for openly released OCR and vision-language models; the GitHub platform for open-source implementations; and selected non-academic but authoritative sources such as the Unicode Consortium technical chapters and digital humanities project pages including the Endangered Archives Programme and the Wikimedia Wikisource Loves Manuscripts initiative.

Search Strategy

Searches were conducted using combinations of English and Indonesian keywords. English search strings included terms such as optical character recognition, OCR, text recognition, handwritten character recognition, document image analysis, scene text recognition, vision-language model OCR, low-resource script OCR, Brahmic script recognition, Javanese script OCR, Balinese palm leaf manuscript, Jawi character recognition, and Nusantara scripts. Indonesian search strings included pengenalan karakter, pengenalan aksara Jawa, pengenalan aksara Bali, OCR aksara, dan transliterasi naskah. Boolean combinations of these terms with method names (CNN, CRNN, CTC, Transformer, ViT, TrOCR, Donut, VLM, Qwen-VL) and dataset names (Hanacaraka, AMADI_LontarSet, NusaAksara) were used to refine retrieval.

For each candidate source, backward citation tracing was performed on the references it cited, and forward citation tracing was performed where citation counts were available, to identify papers that might have been missed by keyword search alone. This snowballing procedure was particularly important for retrieving older foundational papers on classical OCR methods and for locating recent papers that cite the NusaAksara benchmark.

Inclusion and Exclusion Criteria

Sources were included if they met at least one of the following criteria: they presented an OCR architecture or method that is widely cited as a methodological milestone; they reported empirical results on an Indonesian indigenous script or on Jawi; they introduced a dataset, benchmark, or evaluation protocol relevant to low-resource script OCR; or they provided technical or cultural context essential for understanding the Nusantara case, such as Unicode encoding specifications, manuscript conservation studies, or typographic analyses of the scripts.

Sources were excluded if they were purely commercial product pages without methodological transparency; if they reported OCR work on languages or scripts unrelated to the Nusantara context and did not represent methodological milestones with general applicability; or if they were duplicate publications of work already represented by a more authoritative version. Predatory journal publications were also excluded after manual inspection of editorial practices and indexing status.

Time Frame

The primary time frame for inclusion was 2015 through 2026, corresponding to the deep learning era of OCR research and the period during which the majority of OCR studies on Indonesian indigenous scripts were published. Foundational papers from earlier periods, particularly those that established methods still in active use, such as the CTC loss (2006), Hidden Markov Models for cursive script recognition (1990s and

2000s), and Hu moment invariants (1962), were included without time restriction because they remain methodologically relevant.

Analysis Approach

Sources were analysed thematically along two intersecting axes. The first axis grouped sources by methodological era: classical statistical methods, the deep learning era centred on convolutional and recurrent networks, and the transformer and vision-language model era. The second axis grouped sources by target script: Aksara Jawa, Aksara Bali and the lontar tradition, the Sumatran scripts (Sunda, Lampung, Batak, Rejang, Kerinci), the Sulawesi scripts (Lontara, Makasar), Jawi, and multi-script studies. The intersection of these axes produced the matrix that organizes the Results section. Cross-cutting challenges were identified inductively from the thematic analysis and are reported in the Discussion section. A summary of the search strategy is presented in Table 1.

Table 1. Summary of literature review methodology

Element	Specification
Review type	Narrative literature review with scoping elements
Time frame	Primarily 2015–2026; earlier foundational works included where methodologically relevant
Databases	IEEE Xplore, ACM DL, ScienceDirect, SpringerLink, arXiv, ACL Anthology, Google Scholar, Indonesian journal portals (IJCCS, JOIV, JURTEKSI, JINACS, MALCOM), Hugging Face, GitHub
Languages	English and Bahasa Indonesia
Key search terms	OCR, optical character recognition, handwritten character recognition, vision-language model OCR, Nusantara scripts, Aksara Jawa, Aksara Bali, Jawi, palm-leaf manuscript, low-resource script
Snowballing	Backward and forward citation tracing on all included papers
Analysis axes	Methodological era × target script; cross-cutting challenges synthesized inductively

FINDINGS AND DISCUSSION

The Classical OCR Pipeline

Most pre-deep-learning OCR systems decomposed the problem into a sequence of stages. The first was image acquisition and preprocessing, in which the input image was binarised, deskewed, denoised, and normalised for contrast. The choice of binarisation algorithm, whether Otsu, Niblack, Sauvola, or an adaptive variant, was often decisive for accuracy on degraded documents, because subsequent stages assumed a clean foreground-background separation. The second stage was layout analysis, which identified text regions and reading order, distinguishing body text from figures, tables, and marginalia. The third stage was segmentation, in which the page was decomposed into lines, words, and finally individual characters. Segmentation was the most error-prone stage for cursive and connected scripts

because errors here propagated irrecoverably into the recognition stage. The fourth stage featured extraction, in which each character image was converted into a vector of measurements: geometric moments, gradient histograms, projection profiles, or zoning features. The fifth was classification, typically using k-nearest neighbours, decision trees, or support vector machines trained on labelled character samples. The sixth and final stage was post-processing, in which a language model or dictionary corrected obvious errors and resolved ambiguities such as the visual confusability of letters and digits.

This pipeline had two deep liabilities. The first was error propagation: a poor binarisation made segmentation harder, a poor segmentation made classification harder, and a poor classification made post-processing more aggressive and therefore more likely to introduce its own errors. The second was inflexibility: each stage had to be retuned for each new script, document type, and capture condition. The deep learning era addressed both liabilities by collapsing the pipeline.

Classical Statistical Methods

Before deep learning, the most accurate OCR systems were built around carefully engineered features and shallow classifiers. Hu moment invariants, introduced in the 1960s, provided rotation- and scale-invariant descriptors that proved useful for isolated character recognition; they have been applied as recently as the late 2000s to Jawi character classification through what some authors call a tree-root algorithm. Histograms of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) descriptors, ported from general computer vision, also achieved competitive results on isolated character benchmarks. For sequential text, the dominant paradigm before 2014 was the Hidden Markov Model (HMM), imported from speech recognition. In an HMM-based OCR system, the image of a line of text is sliced into thin vertical frames, each frame is converted into a feature vector, and the sequence of feature vectors is modelled as the output of a hidden Markov chain over character states. HMMs were used extensively for Arabic and Jawi handwriting recognition because they handle the cursive nature of these scripts gracefully. Support Vector Machines (SVMs) and Multi-Layer Perceptrons (MLPs) also featured prominently, often in hybrid systems where an HMM provided the sequence model and an SVM or MLP provided the per-frame classifier. Indonesian studies on Aksara Jawa and Aksara Lampung from the 2010s frequently relied on such hybrid pipelines, achieving acceptable accuracy on small, balanced datasets of isolated characters but degrading sharply on continuous text or handwritten samples.

The Deep Learning Era

The publication of AlexNet in 2012 and the subsequent dominance of convolutional neural networks (CNNs) on the ImageNet benchmark transformed visual recognition. For OCR, CNNs first replaced handcrafted feature extractors in the classification stage. A CNN trained end-to-end on character images learns hierarchical visual features that consistently outperform engineered descriptors, particularly on noisy or stylistically varied inputs. Architectures originally designed for natural image classification, including LeNet variants, VGG, ResNet, DenseNet, EfficientNet, and MobileNet, were quickly adapted for character classification, with ResNet and its descendants becoming the de facto default for isolated character tasks.

The breakthrough that made end-to-end text-line recognition practical was the Convolutional Recurrent Neural Network (CRNN), introduced by Shi and colleagues in 2015. A CRNN consists of a convolutional feature extractor that produces a sequence of feature vectors along the horizontal axis of a text line, followed by a bidirectional Long Short-Term Memory (BiLSTM) recurrent network that contextualises each position, and finally a Connectionist Temporal Classification (CTC) loss that aligns the variable-length output sequence to the ground-truth character sequence without explicit segmentation. CRNN with CTC eliminated the most fragile step of the classical pipeline, character segmentation, and became the architectural backbone of PaddleOCR, EasyOCR, and the later versions of Tesseract.

Subsequent refinements replaced CTC with attention-based decoders, producing sequence-to-sequence OCR models analogous to the encoder-decoder machine translation architectures of the late 2010s. Attention decoders relax the monotonic alignment assumption of CTC, which benefits scripts with complex ligature behaviour. Hybrid models that combine CTC and attention objectives, sometimes called dual-head or joint CTC-attention models, became standard practice in the late 2010s. The principal limitations of CRNN-based systems are well known: they struggle with long-range dependencies, with sparsely supported characters, and with scripts whose vertical extent or diacritic complexity violates the assumption that text can be represented as a horizontal one-dimensional sequence of fixed-height slices.

Transformers and Vision-Language Models

TrOCR, released by Microsoft in 2021, was the first widely adopted pure-transformer OCR model. It uses a Vision Transformer encoder to convert a line image into a sequence of patch embeddings and a text Transformer decoder to autoregressively generate the character sequence. By eliminating both the convolutional backbone and the recurrent layer, TrOCR demonstrated that the transformer architecture alone, when pre-trained on large amounts of synthetic and real text-line data, can match or surpass CRNN-based systems on standard benchmarks. Donut, introduced the same year, took the transformer paradigm one step further by proposing an OCR-free document understanding model: instead of first running an OCR engine and then extracting structured information from the recognised text, Donut takes the document image and the desired output schema directly and emits structured JSON.

The most consequential development of the past three years has been the emergence of large multimodal foundation models such as GPT-4o, Claude, Gemini, Qwen-VL, LLaVA, and BLIP-2, which combine a vision encoder with a large language model decoder and are trained on web-scale image-text data. These vision-language models (VLMs) can perform OCR as one capability among many, without any task-specific architecture. They handle layout, mixed languages, handwriting, and partially occluded text in a single prompt-response interaction. Empirical evaluations across enterprise document types of report VLM accuracies in the range of 42 to 67 per cent on complex multi-structured documents, which surpasses traditional OCR pipelines while dramatically simplifying deployment.

For Indonesian indigenous scripts, however, off-the-shelf VLM performance is dismal. The NusaAksara benchmark, presented at ACL 2025 by Adilazuarda and colleagues, evaluates GPT-4o, Llama 3.2 Vision, Aya 23, PP-OCR, and other systems on a curated dataset spanning eight scripts and seven languages, including Balinese,

Batak, Javanese, Lontara, Makasar, Rejang, Sundanese, and the un-Unicode Lampung script. The authors find that most contemporary NLP technologies cannot handle Indonesia's local scripts, with several systems achieving near-zero performance on basic OCR and transliteration tasks. A promising response is targeted fine-tuning: the Jawi-OCR-Qwen-v1 model, released on Hugging Face in 2024, is a fine-tuned version of Qwen2-VL-2B-Instruct specialized for historical Malay texts in Jawi script, achieving usable accuracy on historical manuscripts that off-the-shelf models could not handle.

Empirical Findings by Script

The cross-script empirical literature reveals significant heterogeneity in research maturity. Aksara Jawa is by a wide margin the most studied Nusantara script, reflecting both the size of the Javanese-speaking population and the existence of the publicly available Hanacaraka handwritten character dataset. Studies published between 2018 and 2025 have explored architectures ranging from vanilla CNNs to ResNet-18, Xception with transfer learning, and Vision Transformers. Reported isolated-character accuracies routinely exceed 95 per cent, with the best recent results approaching 99 per cent on the Hanacaraka test split. Less common but more practically valuable are studies that move from isolated characters to full transliteration, such as the 2024 JOIV pipeline that segments Javanese script images, recognises characters with a transfer-learned CNN, and post-processes the predictions with a Levenshtein-distance dictionary lookup to produce Latin transliterations.

Aksara Bali OCR research is dominated by the challenges of the palm-leaf substrate. The AMADI_LontarSet, released in 2016 as the first handwritten Balinese palm-leaf manuscripts dataset, has been the central benchmark. Subsequent work has explored binarisation, text-line segmentation, and end-to-end character recognition. The Wikisource Loves Manuscripts initiative, launched by the Wikimedia Foundation in 2023, and the Balinese WikiPustaka Basa Bali project, launched in 2021, have together digitised over twenty thousand lontar manuscripts and made them available with original-script and transliterated versions. Recent papers report that U-Net-based deep binarisation outperforms classical thresholding on degraded lontar images by substantial margins on intersection-over-union and F-measure metrics.

Aksara Sunda has been the subject of CNN-based handwritten character recognition studies that follow the same template as the Aksara Jawa literature: small custom datasets, transfer learning from ImageNet weights, and reported accuracies above 90 per cent on isolated characters. Aksara Lampung, despite remaining outside Unicode, has attracted similar work, motivated by its inclusion in primary school curricula in Lampung province. Aksara Batak, used in northern Sumatra, has been less studied; available work has focused on basic character classification and on the philological question of how to standardise a script with regional variants across the Toba, Karo, Mandailing, Simalungun, and Pakpak Batak communities. Aksara Rejang, Aksara Kerinci (also called Aksara Incung), and Aksara Pallawa have each been addressed in isolated case studies, often as part of student theses, with results that are not always comparable due to inconsistent datasets and metrics.

Aksara Lontara, also known as Buginese script, and the related Makasar script are used to write the Bugis, Makassar, Mandar, and several smaller languages of South Sulawesi. Both are abugidas of Brahmic descent. OCR research on these scripts is

sparse but growing, with recent work integrating them into the multi-script Nusantara benchmark studies. Jawi, the Arabic-derived script used historically for Malay across the Indonesian-Malaysian region from roughly the fourteenth to the early twentieth century, occupies a special place in Nusantara OCR research because it differs typologically from the Brahmic family. Jawi is a cursive abjad with positional letter forms, like Arabic, but with additional letters and orthographic conventions for representing Malay phonemes. Early OCR work on Jawi used Hu moment invariants, hidden Markov models, and multi-classifier hybrid systems; more recent work has applied CNN-based character recognition to mobile applications. The Jawi-OCR-Qwen-v1 model represents the current state of the art, demonstrating that small VLM fine-tunes can outperform earlier feature-based approaches by a wide margin.

The most comprehensive cross-script empirical effort is the 2023 Indonesian Journal of Computing and Cybernetics Systems study, which evaluated CNN, ConvMixer, and Vision Transformer backbones for both script-type classification and per-character recognition across ten traditional scripts. The authors report 96 per cent accuracy on script-type classification and per-character accuracies ranging from 93 per cent to nearly 100 per cent across the ten target scripts. The 2025 NusaAksara benchmark complements this with the most comprehensive multimodal and multilingual evaluation suite currently available, covering eight scripts across seven languages and defining tasks including image segmentation, OCR, transliteration, translation, and language identification. Its results, however, are sobering. State-of-the-art models, including GPT-4o, Llama 3.2, and Aya 23, perform at or near zero accuracy on several scripts, indicating that contemporary NLP and OCR technology is structurally biased toward romanised and high-resource scripts.

Discussion

Why Nusantara OCR Is Hard

The findings reported in Section 3 reveal a striking gap between OCR performance on high-resource scripts and on Nusantara scripts. Modern architectures achieve near-human accuracy on printed Latin text and impressive results on isolated Aksara Jawa characters, yet frontier vision-language models register near-zero scores on continuous Nusantara text. This gap is not explained by any single factor; it emerges from the interaction of four structural characteristics of the Nusantara context.

The first factor is script diversity and abugida structure. Most Nusantara indigenous scripts are abugidas, in which each consonant carries an inherent vowel and additional vowels are indicated by diacritic marks that attach above, below, before, or after the consonant. A single base consonant in Javanese, such as the letter *ha*, can combine with *swara*, *panyigeg*, and *wyanjana* diacritics to produce well over one hundred visually distinct glyph forms. Extrapolated across the full inventory of base consonants, the number of distinct shapes that a Javanese OCR system must recognise runs into the tens of thousands. This is fundamentally different from the alphabetic case, where the system needs to learn perhaps a hundred symbols. Approaches that treat each composite glyph as an atomic class quickly run into combinatorial explosion; approaches that decompose composite glyphs into base-plus-diacritic require segmentation logic that is sensitive to script-specific conventions.

The second factor is material substrate and document degradation. Nusantara manuscripts are written on materials whose physical properties differ markedly from

European paper. Balinese and Javanese palm-leaf manuscripts, known generically as lontar, are produced from the cured leaves of the lontar palm (*Borassus flabellifer*). Characters are incised into the leaf with a metal stylus and then rubbed with soot or oil to make them legible. The leaves are stored in tropical conditions that promote insect damage, fungal growth, and ink fading. Many surviving lontars are several centuries old and have suffered cumulative degradation that produces low-contrast images with broken strokes, missing characters, and intrusive backgrounds. Classical binarisation methods perform poorly on these substrates; deep-learning-based binarisation with U-Net architectures substantially improves the separation of textual content from degraded background, suggesting that for palm-leaf OCR, preprocessing is not a solved precursor to recognition but a research frontier in its own right.

The third factor is data scarcity and annotation cost. The single most binding constraint on OCR for Nusantara scripts is the scarcity of high-quality annotated data. Where Latin OCR can draw on billions of synthetic and real text-image pairs, Aksara Jawa has perhaps a few thousand publicly available labelled character images, Aksara Bali a few hundred annotated palm-leaf pages, and several other scripts essentially none. Annotation is expensive because it requires expert readers who can both interpret the script and operate digital annotation tools, and the pool of such experts is small and shrinking. This scarcity biases the literature toward isolated-character classification, because curating a balanced dataset of single characters is much cheaper than annotating continuous lines. It also biases the field toward transfer learning from ImageNet or from Latin OCR models, which may not transfer well because the visual statistics of Brahmic scripts differ systematically from those of Latin text.

The fourth factor is tokenisation and evaluation metrics. Standard OCR evaluation metrics, Character Error Rate (CER) and Word Error Rate (WER), assume that the notions of character and word are well defined in the target script. For abugidas, this is contested. A composite glyph, such as a consonant with two stacked diacritics, may be considered one character (one visual unit) or three characters (one base plus two combining marks at the Unicode level). Different choices yield different CER values for identical predictions. Word boundaries are similarly fraught: many Nusantara scripts traditionally use minimal or no inter-word spacing, so tokenisation must be performed by the OCR system itself rather than read off from whitespace. For evaluation to be meaningful and comparable across studies, the community needs shared conventions on Unicode normalization, on the granularity of CER, and on the handling of unsupported characters such as the Lampung block.

Gaps in the Literature

Three gaps in the current literature are particularly conspicuous. First, the empirical literature on most Nusantara scripts is dominated by isolated-character classification, while real-world applications demand continuous-text recognition. Studies that report 99 percent accuracy on isolated handwritten characters give a misleading impression of progress because they do not measure the harder problem of segmenting and decoding running text in authentic documents. Second, there is no widely accepted benchmark for Nusantara OCR beyond NusaAksara, which itself is recent and not yet broadly adopted. Without shared benchmarks, claims of methodological progress cannot be compared across papers in a principled way. Third, the literature is methodologically conservative: most Indonesian-language papers continue to apply

CNN architectures and report accuracy on small custom datasets, even as the international literature has moved toward vision-language models, OCR-free document understanding, and large-scale benchmarks.

Applications and Implications

The most consequential application of Nusantara-focused OCR is the large-scale digitisation of cultural heritage. National and regional libraries, museums, religious institutions, and private collectors across Indonesia hold an unquantified but very large body of manuscripts, inscriptions, and printed material in indigenous scripts. Photographic digitisation alone transforms physical fragility into digital permanence but does not by itself enable search, indexing, or computational analysis. OCR is the bridge from image archive to searchable corpus, and it is the prerequisite for the kind of computational philology that has transformed the study of Latin, Greek, Sanskrit, and Chinese texts over the past two decades.

A second class of applications is educational. Several Indonesian provincial curricula include indigenous scripts in primary and secondary education, particularly in Java, Bali, and Lampung. OCR systems embedded in mobile applications can support student learning by recognising handwritten characters, providing immediate feedback, and offering Latin transliterations of unfamiliar text. Real-time camera-based OCR can also support visitors to heritage sites who may encounter inscriptions and signage in indigenous scripts that they cannot read. A third class is accessibility: OCR is a core component of assistive technology for blind and low-vision users. In a multilingual and multi-script environment such as Indonesia, accessibility tools that cover only the Latin script offer second-class service to communities that interact with indigenous scripts in daily life or in heritage contexts. Extending OCR coverage to Nusantara scripts is, therefore, an equity issue as well as a technical one.

Future Research Directions

Several research directions appear especially promising. First, synthetic data generation at scale, conditioned on font diversity, substrate simulation, and noise modelling derived from authentic degraded manuscripts, can address the data scarcity that bottlenecks supervised OCR. Diffusion models and generative adversarial networks for document image synthesis are particularly attractive because they can be conditioned on text content while producing visually realistic degradation. Second, cross-script transfer learning that exploits the shared Brahmic ancestry of Javanese, Balinese, Sundanese, Rejang, Batak, Lontara, Makasar, and Kawi can amortise the cost of annotation across scripts that share visual and structural properties.

Third, parameter-efficient fine-tuning of strong open-source vision-language models such as Qwen-VL, MiniCPM-V, and InternVL, using techniques such as LoRA and prefix tuning, offers a path to script-specific OCR systems that inherit the broad visual competence of frontier models while requiring only modest amounts of script-specific training data. The Jawi-OCR-Qwen-v1 model is an early demonstration of this approach. Fourth, community-driven annotation platforms, modelled on Common Voice for speech and on the Wikimedia ecosystem for manuscripts, can produce the labelled corpora that no individual research group can fund alone. Fifth, evaluation methodology itself deserves attention: the community needs reference Unicode normalisation conventions, agreed CER granularities, and shared test sets that are not

contaminated by being incorporated into the pre-training of foundation models. Sixth, the integration of OCR with downstream tasks, including transliteration, machine translation, named entity recognition, and historical linguistic analysis, should be pursued as an end-to-end research agenda rather than as a sequence of disconnected technical milestones.

Limitations of This Review

This review has three limitations that should be acknowledged. First, as a narrative scoping review rather than a systematic review, it does not claim exhaustive coverage of the literature; some relevant student theses and Indonesian-language proceedings papers may be omitted. Second, the boundary between OCR research and adjacent fields such as document image analysis, historical paleography, and document understanding is porous, and the inclusion decisions reflect judgments about relevance that other reviewers might draw differently. Third, the field is moving rapidly: the recent dominance of vision-language models means that the empirical landscape can shift substantially within a single year, and conclusions about the state of the art should be treated as snapshots rather than as durable claims.

CONCLUSION

OCR has undergone three major methodological transitions in roughly a century: from template matching to statistical pattern recognition, from statistical pattern recognition to deep learning with convolutional and recurrent networks, and from deep learning pipelines to transformer-based vision-language models. Each transition has expanded the range of documents that can be reliably digitised and has narrowed the gap between human and machine reading performance for high-resource scripts. In the Nusantara context, this progress is real but uneven. Modern architectures achieve high accuracy on isolated handwritten characters of Aksara Jawa, Sundanese, Lampung, and similar scripts when curated datasets are available. They struggle with continuous text, on cursive scripts such as Jawi, and with degraded substrates such as palm-leaf manuscripts. Frontier vision-language models, despite their remarkable general competence, achieve near-zero accuracy on most Indonesian indigenous scripts when evaluated rigorously, as the NusaAksara benchmark demonstrates. The bottleneck is not architectural sophistication but training data coverage, evaluation conventions, and integration with the cultural and linguistic communities that own the scripts in question.

A productive path forward combines four ingredients: large-scale curated annotation through community partnerships, synthetic data generation grounded in authentic degradation patterns, parameter-efficient fine-tuning of strong open-source multimodal models, and shared evaluation protocols that respect the typological properties of abugidas and cursive abjads. If the OCR research community treats Nusantara scripts not as exotic edge cases but as central test cases for general-purpose document intelligence, the result will be both better OCR for Indonesia and better OCR systems overall. The long history of OCR research suggests that the scripts that pose the hardest problems are the ones that drive the most enduring methodological advances.

REFERENCES

- Adilazuarda, M. F., Wijanarko, M. I., Susanto, L., Nur'aini, K., Wijaya, D. T., & Aji, A. F. (2025). NusaAksara: A multimodal and multilingual benchmark for preserving Indonesian indigenous scripts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 28371–28401).
- Aji, A. F., et al. (2022). One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (ACL).
- Burie, J.-C., Coustaty, M., Hadi, S., et al. (2016). ICFHR2016 competition on the analysis of handwritten text in images of Balinese palm leaf manuscripts. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition* (ICFHR).
- Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 1800–1807).
- Dewa, C. K., Fadhilah, A. L., & Afiahayati. (2018). Convolutional neural networks for handwritten Javanese character recognition. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 12(1), 83–94.
- Du, Y., et al. (2020). PP-OCR: A practical ultra lightweight OCR system. arXiv. <https://arxiv.org/abs/2009.09941>
- Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the International Conference on Machine Learning* (ICML).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 770–778).
- Heryanto, A., Nasrudin, M. F., & Omar, K. (2017). Multi-classifier Jawi handwritten sub-word recognition. *International Journal on Advanced Science, Engineering and Information Technology*.
- Indrawan, G. (2020). Balinese character recognition on mobile application based on Tesseract open source OCR engine. *Journal of Physics: Conference Series*.
- Khare, S., & Sharma, M. (2024). A survey of deep learning in optical character recognition. *MATEC Web of Conferences*, 392, Article 01126. <https://doi.org/10.1051/mateconf/202439201126>
- Kim, G., Hong, T., Yim, M., et al. (2021). OCR-free document understanding transformer (Donut). arXiv. <https://arxiv.org/abs/2111.15664>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (NeurIPS).
- Li, M., et al. (2022). TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Mahdi, M. G., Sleem, A., & Elhenawy, I. (2024). Deep learning algorithms for Arabic optical character recognition: A survey. *Multicriteria Algorithms with Applications*, 2, 65–79.
- Mahastama, A. W., & Krisnawati, L. D. (2019). Improving projection profile for segmenting characters from Javanese manuscripts. In *Proceedings of the 1st International Conference on Intermedia Arts and Creative Technology*.

- Mohsenzadegan, K., Tavakkoli, V., & Kyamakya, K. (2022). A smart visual sensing concept involving deep learning for a robust optical character recognition under hard real-world conditions. *Sensors*, 22(16), 6025. <https://doi.org/10.3390/s22166025>
- Mulyanto, A., Susanti, E., Rossi, F., Wajiran, W., & Borman, R. I. (2021). Penerapan convolutional neural network (CNN) pada pengenalan aksara Lampung berbasis optical character recognition (OCR). *Jurnal Edukasi dan Penelitian Informatika*, 7(1), 52–59.
- Naufal, M., et al. (2024). Transliterating Javanese script images to Roman script using convolutional neural network with transfer learning. *JOIV: International Journal on Informatics Visualization*.
- Perdana, A. B. (2024). Diverse scripts, same mistakes: Identifying common pitfalls in the typographic designs of Indonesian traditional scripts. *Jurnal Desain*.
- Ramadhanti, I. (2023). Deep learning approaches for Nusantara scripts optical character recognition. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 17(3), 325–336.
- Shi, B., Bai, X., & Yao, C. (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*, 39(11), 2298–2304. <https://doi.org/10.1109/TPAMI.2016.2646371>
- Sudewo, E. D. B., Biddinika, M. K., & Fadlil, A. (2023). Javanese script Hanacaraka character prediction with ResNet-18 architecture. *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*.
- Unicode Consortium. (2024). *The Unicode standard* (Version 16.0, Chapter 17: Indonesia and Philippines). <https://www.unicode.org>
- Varela, M. E. (2024). *Qwen for Jawi v1: A model for Jawi OCR* [Model repository]. Hugging Face. <https://huggingface.co/>
- Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *In Advances in Neural Information Processing Systems* (NeurIPS).
- Widiarti, A. R., Pulungan, R., Harjoko, A., Marsono, & Hartati, S. (2018). A proposed model for Javanese manuscript images transliteration. *Journal of Physics: Conference Series*, 1098, Article 012014.