Sindhi OCR and Word Spotting in Document Images: Recent Advances, Challenges, and Future Research Directions

Surendra Ramteke
SSBTs College of Engineering & Technology Bambhori,
Jalgaon (MS) INDIA

ABSTRACT

Optical character recognition (OCR) and word spotting are two key functions of document intelligence vital for digitizing, indexing and retrieving handwritten, printed and historical Sindhi documents. It is still challenging due to the extended Perso-Arabic script, the cursive connectivity of script, the contextually formed shapes of letters, the use of ligatures, the similar visual appearance of various letters and symbols with distinction using dots, and significant font and handwriting differences. In this review, the authors present a unified overview of the area of Sindhi document recognition and provide special emphasis on the recent publications from 2020-2026. Recent works include the development of segmentation independent/isolated character systems, application-oriented ligature segmentation, handling of large datasets, Transformer-based contextual recognition, and sequence-learning architectures. The 2026 SOCR is an implementation of a Vision Transformer that is integrated with a text Transformer and is tested on large synthetic and real Sindhi word and text-line datasets. The value of the CNN–BiLSTM–CTC, attention, Transformer encoders, multimodal learning and adaptive augmentation for related cursive scripts is shown by complementary research on Arabic handwritten recognition for the research period 2025–2026. In contrast, there is only limited word-spotting research in Sindhi, and there are a number of unexplored areas, including query-by-example and query-by-string retrieval, metric learning, self-supervised pretraining, and unified OCR–retrieval systems. The review defines key research gaps, such as benchmark scarcity, writing-independent evaluation, historical-document processing, open-vocabulary recognition, and standardized retrieval protocols; and suggests a future research agenda for building a strong document intelligence for Sindhi.

Keywords

Sindhi OCR, Sindhi word spotting, document image analysis, Transformer, handwritten text recognition, Arabic-derived scripts, deep learning, document retrieval.

1. INTRODUCTION

OCR translates into machine-readable sequences the text that is embedded within document images, and word spotting identifies cases where a query word has appeared in a document without having to transcribe the entire document. They are both critical to languages with low resources where the handwritten and historical collections (and print) have not yet been fully digitised. Sindhi is an especially difficult language for document analysis, due to the many context dependent character forms, the many dots and secondary components, the many ligatures, and the cursive nature of the writing system of its commonly used characters.

Previous OCR research on Sindhi has confirmed the significance of glyph analysis, segmentation and isolated character recognition, and deep convolutional feature learning [1]–[2]. These contributions laid the technical foundation of present-day Sindhi OCR, but numerous systems were based on either character or ligature segmentation and were tested on sets of isolated or synthetically generated samples. A study of ligature recognition on a large scale later proved that recognition could be constructed at the ligature level, while the ambiguity of explicit character boundaries was decreased [3]. The OHSCR benchmark also enhanced the availability of writer-diverse handwritten character data [4].

The most significant achievement in the past few years is the end-to-end contextual-aware Sindhi Optical Character Recognizer (SOCR) reported by Ali et al. [5]. The hybrid model is a Vision Transformer that is integrated with a text Transformer, and synthetic word and text-line datasets have been created on this scale as well as a real-world printed Sindhi set. This is a change from the context-free visual classification to contextual sequence recognition. Meanwhile, studies in Arabic script have demonstrated that attention-enhanced CNN–BiLSTM–CTC systems, hybrid CNN–Transformer models, deep CNN models and multimodal learning can be used to improve the recognition of cursive handwritten words and characters [6]–[7]. These related-script developments are very much pertinent in the methodological context, but the reported findings of these could not be directly considered as Sindhi performance.

A second major observation is that there is a discrepancy between OCR and word spotting. The segmentation and recognition of Sindhi literature has been done well, while word spotting in direct Sindhi query-by-example (QbE) and query-by-string (QbS) is not well developed. The present review, therefore, focuses upon the recent advances made in OCR of Sindhi script and draws upon the related studies in Arabic script to propose directions for word retrieval in Sindhi that are technically valid. The major goal is to establish a short-term research agenda for comprehensive, context- and retrieval-aware document intelligence for Sindhi in a segmentation-less environment.

2. REVIEW SCOPE AND METHODOLOGY

The review examines peer-reviewed studies published between 2020 and 2026, while retaining selected earlier foundational Sindhi research needed to explain script-specific segmentation and recognition. Sources are classified as direct Sindhi evidence, related-script transfer evidence, or general document-analysis methods. Reported accuracies are not pooled because datasets, tasks, splits and evaluation units differ. The review is a structured narrative synthesis, not a systematic meta-analysis.

Analysis of literature is carried out based on five aspects: (i) Recognition unit, (ii) Segmentation strategy, (iii) Feature

representation and architecture, (iv) Dataset and evaluation protocol, (v) Relevance to future Sindhi OCR or word spotting. The evidence in Sindhi and Arabic script is clearly differentiated. This clarification is to avoid overstating the maturity of word spotting for Sindhi and not to have the results of the Arabic, Urdu or Persian word spotting interpreted as Sindhi results.

3. SCRIPT CHARACTERISTICS AND TECHNICAL CHALLENGES

Document recognition in Sindhi is a complex problem as it is a challenging task for cursive documents and even more so for the language specific problem of recognition. Characters can be different shapes depending on the context in which they appear in a word, and some symbols differ in the number or the position of the dots. Ligatures are formed to make long, connected components and dots and diacritics can be detached. Line, ligature and character segmentation are further complicated by font variation, degradation of the scan, skew, unequal spacing and handwriting variation. These properties inspire a shift from the rigid segmentation pipeline to sequence models and contextual visual-language architectures.

Table 1. Script Characteristics and Technical Challenges

Challenge	Impact on OCR	Impact on Word Spotting	Promising Direction
Cursive connectivity	Ambiguous character boundaries	Unstable word/component boundaries	Segmentation-free sequence learning
Dots and secondary components	Confusion among similar symbols	False visual matches	Multi-scale attention
Context-dependent shapes	Position-sensitive recognition	Large intra-word variation	Contextual Transformer decoding
Ligature diversity	Large recognition vocabulary	Complex visual word structure	Subword/ligature embeddings
Font and writer variability	Domain shift	Query-document mismatch	Transfer and metric learning
Limited labeled data	Overfitting	Weak retrieval benchmarks	Synthetic/self-supervised learning
Historical degradation	Missing/noisy strokes	Poor embedding correspondence	Restoration-aware encoders

4. EVOLUTION OF SINDHI OCR

4.1 Foundations: Segmentation and Character Recognition

Initially, the research on OCR for Sindhi documents was based on explicit decomposition of the document images. The image structure of the script was studied by glyph identification and the connected-component analysis [1]. Handwritten recognition systems were then created, which involved a combination of preprocessing, zoning features and neural classifiers [8] and dedicated segmentation studies were undertaken to isolate handwritten text into lines, words and characters [9]. These studies showed that segmentation is increasingly difficult and if a character is connected, its boundaries are not reliable.

CNN-based methods then decreased the reliance on hand crafted features. Under controlled isolated character conditions, high accuracy of multi-font and multi-size printed Sindhi character recognition system has been achieved [2]. While this yielded positive findings on the importance of deep visual features, it did not reveal the much greater complexity of natural pages, where segmentation and layout, degradation, language context, and unseen fonts all play a role.

4.2 Ligature-Level Modeling and Benchmark Development

Ali et al. created a very large font-diverse system for Sindhi ligature recognition system with 22,597 font diverse unique ligatures and obtained a baseline recognition accuracy of 91.85% [3]. Ligature-level modeling fits the cursive Perso-Arabic scripts well, as there is no need to enforce the boundaries between characters. Vocabulary growth is its dilemma: if a ligature is not seen or rarely used, it can be challenging for the system to recognize it.

The OHSCR dataset contains 7,800 hand-writings of character images from 150 writers [4]. While writer diversity in this data is helpful for recognition research, writer-level benchmarks are not enough for end-to-end OCR or word spotting. Future datasets should maintain full word and line context, should contain writer-disjoint splits, offers Unicode transcription, offer bounding boxes, and should offer retrieval annotations.

4.3 Recent 2025–2026 Sindhi OCR Advances

Recent evidence extends beyond isolated recognition. The SOCR study [5] combines a Vision Transformer with a text Transformer for contextual Sindhi word and text-line recognition, reporting 88.04% accuracy on its real-word evaluation. It has synthetic word and line resources for large-scale training, with an emphasis on domain-specific evaluation in real-world performance. The printed-text segmentation work [10] indicates that the line segmentation accuracy is 98.3% for 400 images and the ligature segmentation research work [11] is on a different level of processing. These results cannot be compared against each other based on the accuracy scale, as the different prediction sets and test sets are used.

4.4 Transition to Contextual and Pretrained Recognition

The evolution from isolated symbols to sequence recognition is supported by the Sindhi SOCR study [5] and by pretrained Transformer OCR research. Li et al. [12] developed TrOCR using pretrained image and text Transformers, with synthetic pretraining and supervised fine-tuning. This is relevant as a transferable architecture rather than as direct evidence of Sindhi performance. Future Sindhi evaluations should test unseen ligatures, fonts, and document domains, and report CER and WER alongside exact-match accuracy.

5. RECENT RELATED-SCRIPT METHODS RELEVANT TO SINDHI

The recently conducted research on OCR of Arabic language gives significant methodological evidence for architectures which can be adapted to Sindhi. In Arabic handwritten word recognition, Imane et al. introduced an attention enhanced CNN–BiLSTM–CTC method with Word Beam Search and adaptive augmentation [6]. The architecture is particularly important for Sindhi, since, there is no need to segment characters in the sequence for CTC, and attention can highlight discriminative local features like dots and shape changes.

Al-Maamari et al. used CNN features transfer learning to integrate the Transformer Encoder for Arabic printed and handwritten character recognition, achieving 99.51% and 98.19% accuracy on the test sets of the print and handwritten Arabic character sets, respectively [13]. Finally, the methodological implication for the Sindhi case is that CNN

features and self-attention are complementary, as local stroke patterns can be captured by the CNN, and the Transformer captures the broader dependencies.

Attar tested the deep CNN for handwritten character recognition of isolated Arabic characters, and obtained 96.8% accuracy, which is better than the results obtained by the classical SVM and KNN baselines tested in the same study [14]. While isolated character classification is not the same as a word or line of Sindhi handwriting recognition, the study shows that strong convolutional visual encoders are still useful as a building block of larger sequence systems.

Alkhateeb has introduced a multimodal learning model for handwritten word recognition in Arabic in 2026 [7]. It is applicable to Sindhi as future document intelligence could also be a problem of visual appearance and linguistic or semantic representation of the document, not only a problem of image classification using OCR. These are also multimodal learning objectives that are suitable for joint OCR and word-spotting.

Related-script evidence should also be distinguished from transferable, language-independent OCR design. TrOCR [12] provides a visual-to-text sequence baseline, whereas its reported benchmark performance cannot be compared numerically with the Sindhi ligature or word results in Table 2. For low-resource collections, synthetic pretraining and carefully audited fine-tuning are promising but require testing on held-out real Sindhi pages.

6. WORD SPOTTING: CURRENT GAP AND RESEARCH OPPORTUNITY

Word spotting: Identifying document image instances of a query. In QbE, the query is in the form of images while in QbS it is typed string and must match the image and textual representations. Word spotting may be recognition-based or recognition-free, and can work on segmented images of words or on documents.

Until now, no available literature has been found that can be used as a word-spotting benchmark for Sindhi language, which can be comparable to the growing number of OCR resources for Sindhi. This gap is significant as word spotting can offer valuable access to archives in situations where full OCR is not reliable. A practical Sindhi system must, therefore, allow for both QbE and QbS retrieval and should be evaluated on the basis of mean average precision, Precision, recall and query-level robustness with respect to fonts, writers, and degradation of documents.

One of the interesting methods is to learn a common embedding space for the word images in Sindhi and the Unicode strings. Open-vocabulary retrieval can be facilitated by Siamese networks, contrastive objectives, PHOC-style attribute representations, and/or Transformer-based cross-modal encoders. A large word corpus provided by SOCR [5] provides a chance to pre-train such embeddings and real world data or handwritten data can be used for domain adaptation.

6.1 Recent Deep Word-Spotting and Self-Training Methods

Boudraa et al. [15] suggested PUNet, which is a U-Net-based architecture with PHOC-like attribute representations for word spotting in handwritten-document images. Wolf and Fink [16] investigated self-training for both handwritten word recognition and retrieval with synthetic initialization and confidence-filtered pseudo-labels. What they do offer are experiments that demonstrate a methodology for reducing manual annotation needs – not results specific to Sindhi. In the case of Sindhi, two aspects of QbE and QbS need to be done separately, namely mAP and Precision and query coverage that includes unseen words and cross-writer tests.

7. COMPARATIVE ANALYSIS OF RECENT LITERATURE

Table 2. Comparative Analysis of Recent Literature

Ref.	Year	Language/Task	Method	Key Result/Resource	Relevance
[7]	2026	Sindhi word/line OCR	Vision Transformer + text Transformer	SWD 10.3M; STLD 47.2M; RW-Sindhi; 88.04% real-word accuracy	Direct state-of-the-art Sindhi OCR
[13]	2026	Sindhi ligature segmentation	Ligature segmentation	Dedicated Sindhi segmentation study	Direct preprocessing/segmentation
[11]	2026	Arabic handwritten words	Multimodal learning	Recent multimodal recognition framework	Transferable architecture
[12]	2025	Printed Sindhi segmentation	Adaptive thresholding + projections	98.3% line segmentation on 400 images	Direct Sindhi preprocessing
[8]	2025	Arabic handwritten words	CNN-BiLSTM-CTC + attention	End-to-end sequence recognition with adaptive augmentation	Strong model for Sindhi adaptation
[9]	2025	Arabic printed/handwritten chars	CNN features + Transformer	99.51%/98.19% reported accuracy	Hybrid local-global representation
[10]	2025	Arabic handwritten chars	Deep CNN	96.8% reported accuracy	Visual encoder evidence
[5]	2023	Sindhi ligatures	Deep recognition pipeline	22,597 ligatures; 91.85% baseline	Foundation for ligature OCR
[6]	2024	Sindhi handwriting dataset	Benchmark dataset	7,800 images; 150 writers	Writer-diverse Sindhi benchmark

7.1 Evaluation Criteria and Evidence Synthesis

Since this article is based on literature review, the evaluation is a synthesis of results reported by the studies cited and not an experiment performed new. The comparison differentiates between task type (character, ligature, word or line recognition), language, data source and evaluation metric. The reported values for accuracy are qualitative only, since there are differences of corpora, train/test protocols, fonts, writers and units of prediction that make it impossible to control an architecture ranking.

7.2 Quantitative Comparison of Reported Results

For the ligature recognition setting, the Sindhicore ligature recognizer achieves 91.85% real word accuracy, as reported by the study on the SOCR; and the study on the SOCR reports that the accuracy for the real-word is equal to 88.04%. These are not

numbers for a performance gain or loss, but are numbers of different prediction systems. According to Arabic-script studies, the accuracy of the printed characters is 99.51% and 98.19% for handwritten characters [13] and 96.8% for isolated handwritten characters [14]. The related-script results are only for architectural feasibility, not to provide evidence of equivalent Sindhi performance.

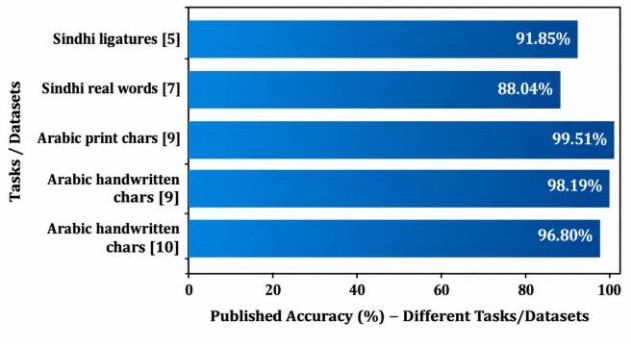


Fig. 1. Reported recognition accuracies from different studies

7.3 Dataset Scale and Evaluation Limitations

In the OHSCR resource, there are 7,800 isolated handwritten character images of 150 writers [4]. The numbers of font-diverse ligatures [3] is a vocabulary count and is not the image count to be compared to the number of unique images in an equivalent vocabulary. SOCR are in charge of creating synthetic word (10.3 million) and text line (47.2 million) resources and a real printed Sindhi evaluation set [5]. While this type of resource scale is promising, the gap between synthetic and real domain is still fairly large, and the absence of shared writer-disjoint word-spotting benchmarks is a large drawback.

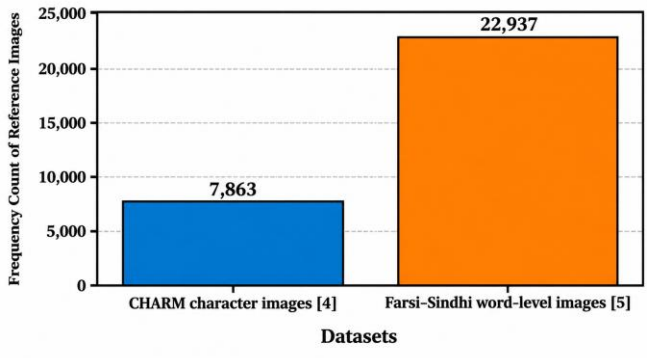


Fig. 2. Selected resource counts with explicitly different units [3], [4].

7.4 Critical Discussion and Reproducibility

Chosen character error rate (CER) word error rate (WER), exact word accuracy, line-level accuracy and inference time should be reported for a Sindhi OCR system, using a common held-out set. For word spotting, mean average precision (mAP), Precision and recall under query-by-example and query-by-string are computed. The document-disjoint, font-disjoint, and writer-disjoint splits should be written, and confidence intervals across repeated trials. These are suggested evaluation indices, but the papers surveyed offer no common scales on which to compute all these.

One of the most critical challenges across the studies surveyed – isolated character recognition versus word/line recognition – introduces a significant confounder: in one mode, one must not have to deal with the layout problems and decoding difficulties of Chinese characters in connected writing, while in the other mode, one must face the problems of contextual ambiguity within the word/line. A reproducible future evaluation should report document-level provenance and train/validation/test partitions (when applicable) separated by writer, source document and font; character error rate (CER), word error rate (WER), exact-match accuracy for OCR; and mean average precision (mAP), Precision, recall, and query-level confidence intervals for QbE/QbS retrieval. The results should be broken down by the following: print text vs. handwritten text; synthetic vs. real samples; document degradation; and vocabulary seen vs. unseen. If there are matched experimental runs, then uncertainty intervals and paired significance tests should be reported. In this review, no new experimental scores or claims of statistical significance are made.

7.4.1 Cross-Script Contextual OCR Evidence

Complementary evidence is provided by recent visual-text recognition methods, which is not necessarily comparable. In ABINet visual and linguistic corrections are separated through iterative bidirectional modeling [17] and in SVTR local/global patch mixing in a visual only recognizer [18]. Donut [19] shows OCR-free understanding of documents, but works on a different downstream task. A framework for selecting pretraining objectives is given in the 2025 critical survey on self-supervised text recognition [20]. No one of these studies is a assessed Sindhi benchmark; transfer has to be experimented.

7.5 Critical Synthesis of 2020–2026 Methods

The literature reviewed identifies three complementary, but different, lines of evidence. Several studies that focus on the script characteristics, datasets and reported recognition results for Sindhi have been done [3]–[5], [10] and [11]. The choices for the design of the sequence model are general Transformer OCR [12]. Word-spotting and self-training studies show retrieval representations and adapting strategies in other document collections [15] and [16]. The most outstanding doubt remains whether a common Sindhi encoder can be used to accurately transcribe and to enable good QbE/QbS retrieval, for both Perso-Arabic and Devanagari scripts. No claim of improvement is made without a shared data set and experimental protocol.

7.6 Word-Spotting and Retrieval Synthesis (2020–2026)

Different design choices are transformer based word spotting [21] and historical-document retrieval without learned representations [22]. Beyond exact image similarity, there are retrieval methods that extend to semantic image retrieval, such as semantic re-ranking [23], unsupervised segmentation-free indexing [24] and semantic document-image analysis [25]. The surveys of deep word spotting [26] and uncertainty-aware keyword spotting [27] underscore the desire to provide information about the query type, supervision, indexing assumptions and retrieval metrics. There are research directions that are promising, but not tested in Sindhi, such as word-length-aware spotting [28], self-supervised Transformer retrieval [29] and one-pass query-by-string region ranking [30]. As a result, comparisons of cross-study accuracy are not performed unless the datasets, tasks and evaluation protocols are identical.

8. RESEARCH GAPS

Table 3. Research Gaps in Existing Research Work

Gap	Evidence/Problem	Recommended Direction
Standardized word-spotting benchmark	No established recent Sindhi QbE/QbS benchmark was identified.	Create public word/page datasets with Unicode queries and standard retrieval metrics.
Real-world generalization	SOCR performance is lower on real-world printed data than on synthetic data [7].	Use domain adaptation, degradation modeling, and real-world pretraining.
Handwritten word/line resources	Existing Sindhi benchmark work is still strongly character-oriented [6].	Build writer-disjoint word, line, and page datasets.
Open-vocabulary recognition	Ligature vocabularies can become extremely large [5].	Use CTC, subword decoding, byte/character tokenization, or Transformer decoders.
Historical documents	Recent work focuses mainly on contemporary printed/handwritten samples.	Develop restoration-aware OCR and retrieval for archival Sindhi.
Unified OCR and retrieval	OCR and word spotting are usually treated separately.	Train a shared encoder with transcription and metric-learning branches.
Confidence and explainability	Recent Sindhi OCR emphasizes accuracy more than calibrated trust.	Add uncertainty calibration, error localization, and explainable attention.

8.1 Evidence-based Priorities for Future Sindhi Studies

The expanded 2020–2026 evidence suggests four priorities: (i) reproducible printed and handwritten Sindhi test splits; (ii) explicit separation of OCR transcription and word-retrieval outcomes; (iii) comparison of segmentation-based and segmentation-free QbE/QbS retrieval [22], [24], [30]; and (iv) domain-shift, annotation-efficiency and uncertainty analysis [27], [20]. The proposed bilingual Perso-Arabic/Devanagari extension remains a hypothesis until both scripts are represented in training and external test collections.

9. PROPOSED FUTURE FRAMEWORK

The following unified future architecture can embed document enhancement and layout analysis, CNN/ViT-based visual encoding, multi-scale contextual feature fusion, Transformer- or BiLSTM-based sequence encoding and shared visual-semantic representation learning into one framework. The learned representation can then be used as an input to two complementary branches: an OCR branch that can incorporate CTC- or attention-based decoding, and language-context refinement; and a word-spotting branch that can include discriminative embeddings for Query-by-Example (QbE) and Query-by-String (QbS) retrieval. Before the recognized and retrieved information can be stored in a searchable digital archive, confidence calibration and explainability mechanisms can also be added to estimate prediction reliability and help with the interpretation of predictions. This approach combines contextual modeling capabilities of the SOCR [5] with recent CNN–BiLSTM–CTC work [6] which provides sequence learning without segmentation with CNN–Transformer models [13] providing complementary local and global feature representation. To minimize reliance on large and fully annotated Sindhi datasets, self-supervised pretraining on top of large collections of unlabeled document images can be used prior to supervised fine-tuning. Additionally, font and ligature diversity, hand-writing and vocabulary diversity can be further enhanced by synthetic data, but realistic degradation of documents, scanning artifacts and writer variability should be added to reduce the domain gap between synthetic and real world documents. Training and testing procedures are especially critical for handwritten Sindhi documents in order to evaluate the genuine generalization. In the word-spotting branch, contrastive or triplet-learning objectives can be used to optimize the visual-semantic embedding space, and both seen and unseen query words should be used to evaluate it. This integrated architecture can evolve the document analysis of Sindhi to a single platform for OCR, word retrieval, confidence estimation and digital document search.

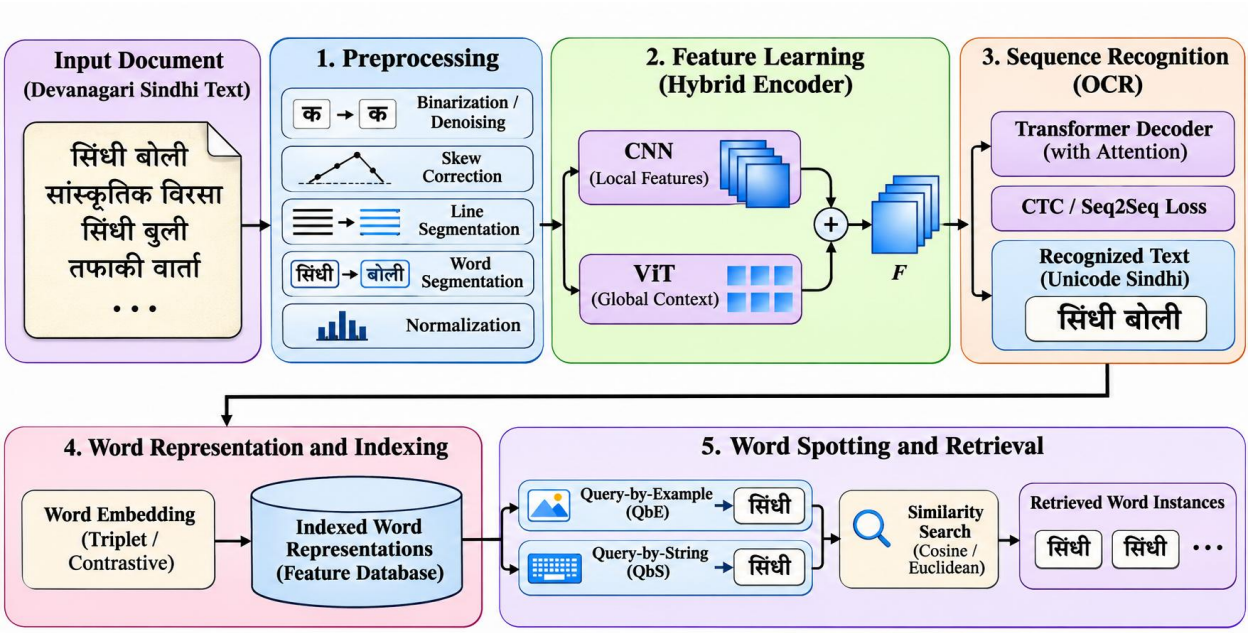


Fig. 3. Proposed unified architecture for Sindhi OCR and word spotting with Devanagari input examples

9.1 Proposed Architecture and Data Flow

Document images are initially processed for denoising, skew removal, layout analysis and line / word extraction as shown in Fig. 3. Script identification enables character inventories for Perso-Arabic and Devanagari Sindhi to be kept separately. A CNN–Vision Transformer encoder model extracts both local stroke and dot patterns, as well as the global spatial context. Shared features drive sequence modelling and visual-semantic representation learning. The OCR branch is used for learning the Unicode text from the CTC or attention-based decoding while the retrieval branch is a branch to obtain the word discriminative word embeddings for retrieval by query-by-example (QbE) and query-by-string (QbS). A confidence calibration method and explanation techniques are suggested to assist in the inspection of uncertain predictions.

9.1.1 Literature-Guided component Justification

The visual-text recognition pathway is inspired by pre-trained Transformer OCR [12] and pre-trained contextual Sindhi sequence recognition [5]. The retrieval branch is inspired by the spotting based on words and self-training to recognize words for retrieval (PHOC-oriented word spotting [15] and self-training for recognition and retrieval [16]). A script-aware tokenizer and Unicode normalization must be used for both scripts as well as independent error reporting. Shared encoder, dual heads and confidence estimation are proposed to be included; it is not known if the three could work together. Ablation studies should be conducted between fusing CNN only, ViT only, and fusing CNN and ViT (with and without self-training and retrieval supervision).

9.1.2 Extended Module-Level Evidence

Visual encoding can be compared with the SVTR [18] and context-aware principles of recognition in ABINet [17] using CNN–ViT fusion. The candidate designs for retrieval comprise of Transformer/PBCS embeddings [21], PHOC-enhanced PUNet [15], unsupervised indexing [24] and self-supervised keyword representations [29]. Also, query-by-string search may be assessed in comparison to region-ranking methods [30] and semantic re-ranking should be used as an ablation, not assumed to be a benefit [23]. Alongside CER/WER and mAP, there should be an Inference time, memory, a calibration and an annotation cost, without experiments gains are not claimed.

9.2 Validation Plan and Limitations

The framework is a future research direction, not a model trained or evaluated in this review. A subsequent study should report character error rate, word error rate, exact word accuracy, mean average precision, Precision, Recall, runtime, and confidence calibration. Script-specific test sets, writer-disjoint splits, unseen-font evaluation, and historical-document degradation tests are necessary. Performance should be reported separately for Perso-Arabic and Devanagari Sindhi, with cross-script transfer assessed only where paired annotations exist. No performance gains are claimed here.

9.2.1 Recommended Experimental Protocol

Establish document-disjoint and writer-disjoint partitions, reserve a real-world external test set, and stratify by script, print/handwriting and degradation. Evaluate OCR using CER, WER and exact word accuracy; evaluate QbE and QbS with mAP and Precision 1/5/10. Report confidence calibration, inference latency, parameter count, and paired bootstrap confidence intervals. Compare all ablations on identical splits, and avoid treating published cross-dataset accuracies as direct baselines.

10. CONCLUSION

In this review, the recent advancements in Optical Character Recognition (OCR) and Document Image Analysis (DIA) for Sindhi scripts in the respective period of 2020 to 2026 were discussed, and the foundational studies and related developments in other scripts were also discussed. The literature surveyed shows a gradual shift of the recognition methods from character-based to deep learning architectures using convolutional neural networks, sequence modelling, attention, and Vision Transformers. The recent Sindhi recognition studies showcase the potential of contextual learning and large scale synthetic datasets, but the recognition performance is still subject to the characteristics of the datasets, the document quality, the script variation, and the assessment framework.

There are several issues that are identified in the comparative analysis that needs further investigation such as lack of publicly

available benchmark datasets, limited evaluation of the degraded historical documents, handwriting variations and typography, and lack of standard word spotting benchmarks for Sindhi. In addition, most current OCR and word-retrieval systems are studied separately, and it is not easy to create a single system for transcription and content-based retrieval.

To meet these research challenges, a conceptual hybrid CNN–Vision Transformer framework has been proposed for integrated Sindhi OCR and word spotting. The proposed architecture features document preprocessing, extraction of visual features at local and global granularity, recognition of visual context sequences, discriminative representation of words. The proposed architecture incorporates the following features: local and global visual feature extraction, preprocessing of the documents, recognition of visual context sequences, and discriminative representation of words, which are necessary for query by example and query by string retrieval. The extension plan towards Sindhi documents written in Perso-Arabic and Devanagari scripts adds new demands for script specific pre-processing, script annotation and recognition.

Future work should involve building representative databases, designing script-aware and self-supervised learning algorithms and testing combined OCR and retrieval systems on standard test sets. Character Error Rate (CER), Word Error Rate (WER), Word Level Accuracy (WLA), mean average precision and Precision are also important criteria for recognition performance, whereas writer-independent testing, cross-domain evaluation, computational efficiency, and robustness of the model are important criteria for assessing retrieval effectiveness.

In conclusion, the reviewed evidences indicated that the contextual recognition, strong visual representation and word-level retrieval is an effective approach to construct searchable digital Sindhi document collections. However, the architecture proposed needs to be systematically validated in the lab before any improvements in its effectiveness can be confirmed.

11. ACKNOWLEDGMENTS

The authors gratefully acknowledge **Kavayitri Bahinabai Chaudhari North Maharashtra University (KBCNMU), Jalgaon, Maharashtra, India,** for providing financial support under the VCRMS Scheme for carrying out this research work. The funding and institutional support provided by the University are sincerely acknowledged.

12. REFERENCES

[1] N. A. Memon, F. Abbasi, and S. Zardari, “Glyph identification and character recognition for Sindhi OCR,” *Mehran University Research Journal of Engineering and Technology*, vol. 36, no. 4, pp. 933–940, 2017, DOI: 10.22581/muet1982.1704.18.

[2] A. A. Chandio, M. Leghari, M. Leghari, and A. H. Jalbani, “Multi-font and multi-size printed Sindhi character recognition using convolutional neural networks,” *Pakistan Journal of Engineering and Applied Sciences*, vol. 25, 2019, DOI: 10.67635/pjeas.2019.25.1.801.

[3] Z. Ali, S. A. Khan, M. K. Shahzad, and H. S. M. Bilal, “A large-scale font-diverse Sindhi ligature recognition system,” in *Proc. 2023 International Conference on Frontiers of Information Technology (FIT)*, 2023, pp. 132–137, DOI: 10.1109/FIT60620.2023.00033.

[4] J. A. Naveed, M. A. Soomro, L. Saleem, and M. K. Shaikh, “OHSCR: Benchmarks dataset for offline handwritten Sindhi character recognition,” *Sir Syed University Research Journal of Engineering & Technology*, vol. 14, no. 1, pp. 55–61, 2024, DOI: 10.33317/ssurj.618.

[5] S. Ali, T. Gao, Z. Zhu, and M. Ahmed, “End-to-end contextual-aware deep learning pipeline for Sindhi text recognition,” *Expert Systems with Applications*, vol. 315, Art. no. 130960, 2026, DOI: 10.1016/j.eswa.2025.130960.

[6] B. Imane, A. Alae, K. Ghizlane, and M. Mrabti, “Enhancing Arabic handwritten word recognition: A CNN-BiLSTM-CTC architecture with attention mechanism and adaptive augmentation,” *Discover Applied Sciences*, vol. 7, Art. no. 460, 2025, DOI: 10.1007/s42452-025-06952-z.

[7] J. H. Alkhateeb, “A multimodal learning framework for Arabic handwritten word recognition and future research directions,” *Scientific Reports*, published online Jul. 1, 2026, DOI: 10.1038/s41598-026-59980-z.

[8] S. A. Awan, Z. H. Abro, A. H. Jalbani, D. N. Hakro, and M. Hameed, “Handwritten Sindhi character recognition using neural networks,” *Mehran University Research Journal of Engineering and Technology*, vol. 37, no. 1, pp. 191–196, 2018, DOI: 10.22581/muet1982.1801.17.

[9] S. A. Awan, D. N. Hakro, Z. H. Abro, and A. H. Jalbani, “Segmentation of Sindhi handwritten text,” *Sindh University Research Journal (Science Series)*, vol. 50, no. 2, pp. 205–208, 2018, DOI: 10.26692/surj.v50i2.1333.

[10] P. B. Khokhar, S. Talpur, M. Ismail, H. Abbas, and M. A. Khan, “Application-oriented segmentation of printed Sindhi text for document recognition and natural language processing systems,” *Quaid-e-Awam University Research Journal of Engineering Science and Technology*, vol. 23, no. 2, pp. 15–26, 2025, DOI: 10.52584/QRJ.2302.03.

[11] S. Goel, G. S. Lehal, and V. Lamba, “Ligature segmentation in Sindhi optical character recognition,” in *Communications in Computer and Information Science*, vol. 2611, pp. 100–113, 2026, DOI: 10.1007/978-3-032-11491-4_10.

[12] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, “TrOCR: Transformer-based optical character recognition with pre-trained models,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 13094–13102, 2023, doi: 10.1609/aaai.v37i11.26538.

[13] M. R. Al-Maamari, R. Ramteke, A. M. Al-Hejri, *et al*., “Integrating CNN and transformer architectures for superior Arabic printed and handwriting characters classification,” *Scientific Reports*, vol. 15, Art. no. 29936, 2025, DOI: 10.1038/s41598-025-12045-z.

[14] E. T. Attar, “Deep convolutional neural network for isolated Arabic handwritten character recognition: Design, evaluation, and comparative study,” *Scientific Reports*, vol. 15, Art. no. 42467, 2025, doi: 10.1038/s41598-025-26658-x.

[15] O. Boudraa, D. Michelucci, and W. K. Hidouci, “PUNet: Novel and efficient deep neural network architecture for handwritten documents word spotting,” *Pattern Recognition Letters*, vol. 155, pp. 19–26, 2022, DOI: 10.1016/j.patrec.2022.01.019.

[16] F. Wolf and G. A. Fink, “Self-training for handwritten word recognition and retrieval,” *International Journal on Document Analysis and Recognition*, vol. 27, pp. 225–244, 2024, DOI: 10.1007/s10032-024-00484-9.

[17] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7098–7107.

[18] Y. Du *et al*., "SVTR: Scene text recognition with a single visual model," in *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 2022, pp. 884–890, DOI: 10.24963/ijcai.2022/124.

[19] J. Kim *et al*., "OCR-free document understanding transformer," in *Computer Vision—ECCV 2022*, pp. 498–517, 2022, DOI: 10.1007/978-3-031-19815-1_29.

[20] C. Penarrubia, J. J. Valero-Mas, and J. Calvo-Zaragoza, "Self-supervised learning for text recognition: A critical survey," *International Journal of Computer Vision*, vol. 133, pp. 6221–6250, 2025, DOI: 10.1007/s11263-025-02487-3.

[21] M. Mhiri, M. Hamdan, and M. Cheriet, "Handwriting word spotting in the space of difference between representations using vision transformers," *Pattern Recognition Letters*, vol. 174, pp. 39–45, 2023, DOI: 10.1016/j.patrec.2023.08.008.

[22] H. H. Mohammed, N. Subramanian, and S. Al-Madeed, "Learning-free handwritten word spotting method for historical handwritten documents," *IET Image Processing*, 2021, DOI: 10.1049/ipr2.12216.

[23] S. Papazis, A. P. Giotis, and C. Nikou, "Enhancing keyword spotting via NLP-based re-ranking: Leveraging semantic relevance feedback in the handwritten domain," *Electronics*, vol. 14, no. 14, Art. no. 2900, 2025, DOI: 10.3390/electronics14142900.

[24] K. Zagoris, A. Amanatiadis, and I. Pratikakis, "Word spotting as a service: An unsupervised and segmentation-free framework for handwritten documents," *Journal of Imaging*, vol. 7, no. 12, Art. no. 278, 2021, DOI: 10.3390/jimaging7120278.

[25] O. Tüselmann and G. A. Fink, "Neural models for semantic analysis of handwritten document images," *International Journal on Document Analysis and Recognition*, vol. 27, pp. 245–263, 2024, DOI: 10.1007/s10032-024-00477-8.

[26] L. Kumari and A. Sharma, "A review of deep learning techniques in document image word spotting," *Archives of Computational Methods in Engineering*, vol. 29, pp. 1085–1106, 2022, DOI: 10.1007/s11831-021-09605-7.

[27] F. Daraee, S. Mozaffari, and S. M. Razavi, "Handwritten keyword spotting using deep neural networks and certainty prediction," *Computers and Electrical Engineering*, vol. 92, Art. no. 107111, 2021, DOI: 10.1016/j.compeleceng.2021.107111.

[28] H. Wang *et al*., "Word length-aware text spotting: Enhancing dense text detection and recognition for camera-captured document image," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, Art. no. 5026915, 2025, DOI: 10.1109/TIM.2025.3560748.

[29] S. Khamekhem Jemni, S. Ammar, M. A. Souibgui, Y. Kessentini, and A. Cheddad, "ST-KeyS: Self-supervised Transformer for keyword spotting in historical handwritten documents," *Pattern Recognition*, vol. 170, Art. no. 112036, 2026, DOI: 10.1016/j.patcog.2025.112036.

[30] P. Zhao, W. Xue, Q. Li, and S. Cai, "Query by strings and return ranking word regions with only one look," in *Computer Vision—ACCV 2020*, pp. 3–18, 2021, DOI: 10.1007/978-3-030-69544-6_1