MultiFOLD: A Multimodal Framework to correct OCR Lapses in cluttered Documents

Anonymous Submission

Abstract. Optical Character Recognition (OCR) systems achieve high accuracy on clean, printed text but struggle with cluttered or degraded documents, leading to frequent recognition errors. Existing correction methods either rely on labor-intensive manual proofreading or fail to incorporate contextual cues effectively. To address this, we introduce MultiFOLD, a multimodal framework that integrates Automatic Speech Recognition (ASR) and context-aware refinement for post-OCR correction. Unlike prior approaches that treat OCR and ASR outputs independently, MultiFOLD fuses textual and spoken corrections in a confidence-weighted alignment strategy and employs a fine-tuned ByT5 sequence-to-sequence model for enhanced text prediction. Our framework reduces the manual annotation burden by allowing users to verbally correct OCR errors, leveraging both speech-based and text-based correction mechanisms. Through extensive experiments performed on 800 cluttered documents, we demonstrate a 52.39% reduction in annotation time and a 36.04% reduction in Character Error Rate (CER) compared to manual OCR correction. MultiFOLD outperforms OCR-only and ASR-only baselines by a reduction in CER of 29.20% and 6.81%, respectively. These results establish MultiFOLD as an efficient, scalable, and open-source solution for document digitization.

Keywords: Optical Character Recognition · MultiFOLD · Cluttered Documents

1 Introduction

Optical Character Recognition (OCR) has significantly advanced document digitization by converting printed and handwritten text into machine-readable formats. However, OCR systems remain highly error-prone when processing cluttered, degraded, or low-quality documents, where ink smudges, overlapping text, and background noise interfere with character recognition [14]. These errors lead to incorrect transcriptions, missing words, and hallucinated characters, requiring extensive human intervention for correction. Traditional post-OCR refinement relies on manual proofreading or rule-based heuristics, both of which are labor-intensive and fail to generalize across different document types. Recent deep-learning-based approaches leverage sequence-to-sequence models for text correction, but they remain ineffective when words are contextually ambiguous or partially occluded.

A promising direction to mitigate these errors is multimodal learning, where multiple input modalities such as text, speech, and visual cues are leveraged

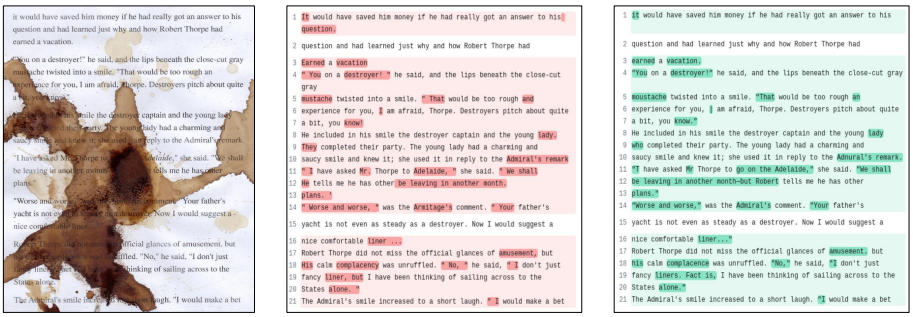

Fig. 1. Qualitative example: MultiFOLD effectively corrects OCR errors by integrating ASR with context aware refinement.

for enhanced OCR refinement. Humans frequently rely on auditory and contextual information to comprehend distorted text, motivating the use of Automatic Speech Recognition (ASR) as an auxiliary correction modality. By allowing annotators to verbally correct OCR errors, ASR provides an additional signal for ambiguous words. However, ASR alone suffers from homophone ambiguities and lacks contextual grounding in the original text, requiring a mechanism that can fuse multimodal inputs effectively.

Unlike prior multimodal correction methods, Figure 1 shows an example of MultiFOLD, introducing a confidence-weighted OCR-ASR fusion strategy, which demonstrates better performance than manual OCR correction. MultiFOLD is particularly useful in digitizing legal records, handwritten manuscripts, and historical texts.

To address these challenges, we introduce MultiFOLD, a novel multimodal OCR correction framework that integrates ASR-based speech corrections with OCR refinement to enhance text recovery in cluttered documents. Unlike prior approaches that treat OCR and ASR independently, MultiFOLD employs a confidence-weighted OCR-ASR fusion strategy to dynamically align corrections and a fine-tuned ByT5 sequence-to-sequence model to refine transcriptions using contextual information. Our approach enables users to efficiently correct OCR errors through speech input, reducing manual annotation effort while maintaining high accuracy. Overall, in this work, we present MultiFOLD and demonstrate its effectiveness in reducing OCR errors. Specifically, we make the following contributions:

- We propose MultiFOLD, a multimodal OCR correction framework that integrates speech-assisted refinement with textual correction, significantly reducing manual effort.
- We develop a confidence-weighted OCR-ASR fusion strategy that dynamically aligns speech corrections with OCR outputs, reducing error propagation. We fine-tune ByT5 for context-aware text correction, leveraging both OCR predictions and ASR transcriptions to refine noisy document text.

– We conduct extensive experiments demonstrating a 52.39% reduction in annotation time and a 36.04% reduction in Character Error Rate (CER), establishing MultiFOLD as an efficient and scalable document digitization solution.

Through these contributions, MultiFOLD bridges the gap between manual post-OCR correction and fully automated refinement, leveraging multimodal learning to improve efficiency and accuracy. The remainder of this paper is structured as follows: Section 2 reviews existing approaches for OCR correction and annotation tools. Section 3 details the MultiFOLD framework and its implementation. Section 4 presents our experimental setup, evaluation results, and ablation studies. Finally, Section 5 concludes with future research directions.

2 Related Work

OCR correction has been extensively studied, with approaches ranging from rule-based heuristics to deep learning-driven refinements. Despite advancements, existing methods struggle with cluttered and degraded documents, necessitating a multimodal approach to enhance text recovery. In this section, we review prior research in three key areas: post-OCR text correction, human-in-the-loop annotation tools, and multimodal learning for document processing.

2.1 Post-OCR Text Correction

Traditional post-OCR correction techniques rely on lexicon-based spell-checking [7], n-gram language models [3], and dictionary-based substitutions [9]. These methods detect and replace OCR-induced errors using pre-defined vocabularies but fail when encountering rare words, domain-specific terminology, or handwritten text. More recent methods leverage sequence-to-sequence models [5,6], which correct OCR text by modeling character-level dependencies. Transformer-based architectures, such as ByT5 [10], have demonstrated improvements in text restoration tasks. However, these models only refine textual inputs, making them ineffective when OCR errors lead to information loss.

To mitigate ambiguities in OCR text, researchers have explored hybrid correction approaches. OpenOCRCorrect [9] aligns multiple OCR outputs to construct consensus-based corrections, while DocRes [16] employs document image restoration to improve OCR quality. Despite these advancements, existing methods remain unimodal and struggle when text is heavily occluded, necessitating external sources of contextual knowledge.

2.2 Human-in-the-Loop Annotation Tools

Several human-assisted annotation platforms facilitate OCR correction through interactive interfaces. eScriptorium [12] and Pivan [13] provide tools for transcribing and segmenting historical texts, while Callico [19] enables collaborative

document annotation. DocVisor [8] integrates layout segmentation, supporting large-scale document correction workflows. However, these methods still rely heavily on manual annotation, making them inefficient for high-volume datasets.

Recent efforts have sought to integrate machine learning-assisted correction into annotation tools. LabelStudio [15] and Kili [1] employ deep learning models for preliminary OCR correction, but they lack adaptive, multimodal refinement mechanisms to handle ambiguous text segments. Unlike these approaches, MultiFOLD introduces speech-assisted corrections to complement manual edits, significantly reducing the annotation burden.

2.3 Multimodal Learning for Document Processing

Multimodal learning has gained prominence in OCR enhancement, where vision-based techniques and language models are jointly utilized to improve document readability. StainRestorer [18] employs image-based restoration to remove document artifacts, while DocDiff [17] uses self-attention mechanisms for document enhancement. However, these methods focus solely on preprocessing and do not refine OCR transcriptions post-recognition.

Integrating ASR and OCR for multimodal text recovery has been explored in limited contexts. Systems such as Speech2Text [4] attempt to align ASR outputs for improved recognition. However, these models lack context-aware refinement and rely on direct text substitution rather than learning contextual dependencies. Our work extends this idea by introducing a confidence-weighted OCR-ASR fusion strategy and a fine-tuned ByT5 model to ensure semantic consistency in OCR corrections.

Building upon these prior works, MultiFOLD bridges the gap between manual annotation tools and automated OCR correction by introducing a multimodal framework that fuses speech-assisted and text-based refinement. Unlike previous approaches, MultiFOLD (a) Dynamically integrates ASR transcriptions with OCR outputs using a confidence-weighted alignment strategy (b) Utilizes a fine-tuned ByT5 sequence-to-sequence model to refine transcriptions in contextually ambiguous settings (c) Reduces reliance on human annotation by allowing users to verbally correct OCR errors, achieving a 52.39% reduction in annotation time. By combining speech-based supervision with context-aware text modeling, MultiFOLD establishes a robust framework for post-OCR correction in cluttered document settings.

3 Design Principles

The design of MultiFOLD is guided by the need for adaptive, multimodal, and efficient OCR correction, ensuring that speech-assisted refinements seamlessly integrate with text-based corrections. To achieve this, MultiFOLD adheres to five key principles: openness and flexibility, multimodal user experience, adaptive document enhancement, context-aware text refinement, and scalability. These principles are motivated by the limitations of existing OCR correction tools and

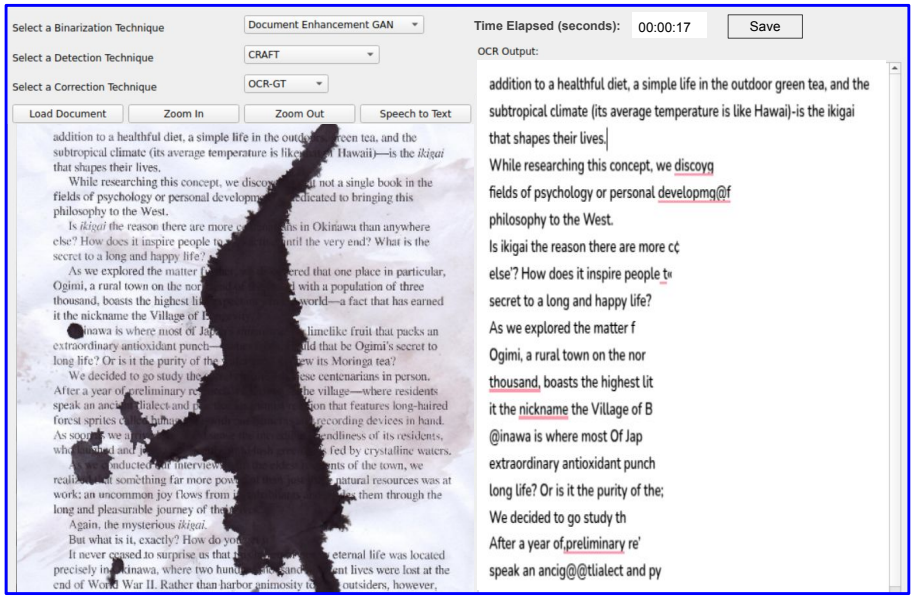

Fig. 2. Working of MultiFOLD tool on a cluttered document by setting its annotation mode. The OCR prediction (right) for the text image on the clutter, and on the right of the clutter (though readable), is missing. Such text is quickly added during the correction stage using Automatic Speech Recognition (ASR).

aim to bridge the gap between fully manual annotation and fully automated refinement.

3.1 Open-Source and Self-Hosting Flexibility

MultiFOLD is designed as an open-source framework[1] to facilitate research, development, and real-world adoption. Unlike proprietary OCR correction tools, which are often restricted to closed ecosystems, MultiFOLD allows users to self-host and customize the system for diverse annotation tasks. The modular architecture supports seamless integration with various OCR engines and ASR models, ensuring adaptability across different document types. Additionally, MultiFOLD is designed with privacy-aware deployment in mind, making it suitable for sensitive document digitization, such as legal, medical, and historical archives.

3.2 Multimodal User Experience with Speech and Text Corrections

Traditional OCR correction tools require annotators to manually edit recognized text, leading to cognitive overload and increased annotation time. Figure

[1] Code, Data and Model Weights will be made open source post-publication.

2 depicts MultiFOLD as a multimodal interface where users can switch between speech-based and text-based corrections, significantly reducing the need for manual typing. The annotation workflow is optimized for natural human interactions, allowing users to verbally correct misrecognized words while preserving the surrounding context.

To ensure smooth integration of speech-based corrections, MultiFOLD employs a confidence-weighted OCR-ASR fusion strategy. The system retains high-confidence OCR predictions while replacing uncertain segments with ASR transcriptions. This fusion mitigates error propagation from either modality, ensuring a more reliable correction pipeline. Unlike existing OCR correction methods relying solely on text-based refinements or standalone ASR transcription, MultiFOLD dynamically adapts to user preferences, enabling efficient human-in-the-loop correction.

3.3 Adaptive Document Enhancement for Cluttered Text

OCR performance degrades significantly in the presence of background clutter, low contrast, and document stains. MultiFOLD incorporates an adaptive preprocessing module that enhances document readability before text extraction to address this. Unlike static enhancement pipelines [17], MultiFOLD dynamically selects contrast adjustment, local thresholding, and noise reduction techniques based on the document's degradation level. This ensures optimal text recognition in varying cluttered conditions, improving OCR robustness while preserving document structure.

3.4 Context-Aware Text Refinement with ByT5

While ASR provides a valuable secondary modality for correcting OCR errors, it still struggles with contextually ambiguous phrases and homophones. MultiFOLD integrates a fine-tuned ByT5 sequence-to-sequence model that refines text predictions by learning context-aware dependencies to ensure accurate text reconstruction. ByT5 is trained on a corpus of OCR-ASR correction pairs, allowing it to: (a) Infer missing words caused by OCR misreads or ASR transcription gaps (b) Resolve homophones and ambiguous phrases using contextual cues (c) Preserve document-specific writing styles and formatting. This approach outperforms traditional rule-based spell-checking and standalone ASR corrections by ensuring that refinements remain consistent with surrounding document content. MultiFOLD processes documents in 8.75 seconds per sentence, making it suitable for real-time correction.

3.5 Scalability and Integration with Large-Scale Annotation Pipelines

MultiFOLD supports large-scale OCR annotation projects while maintaining high responsiveness and computational efficiency. The system is implemented

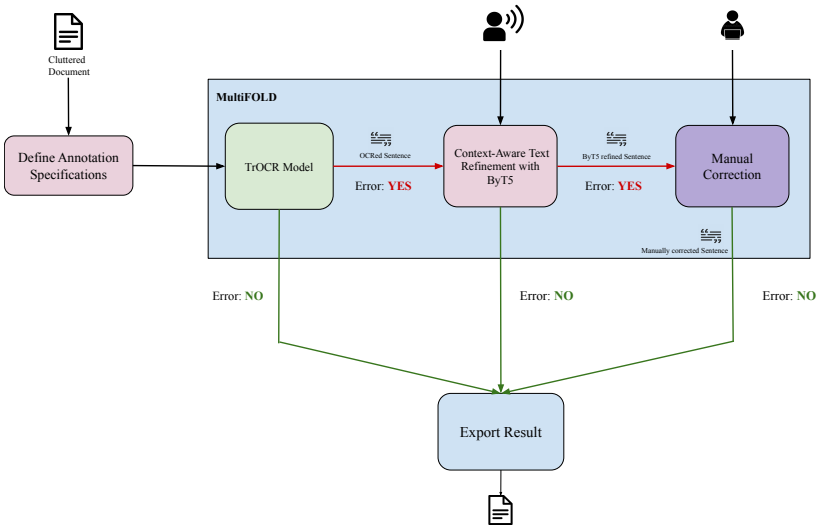

Fig. 3. Overview of MultiFOLD framework, integrating OCR, ASR, and context-aware text refinement.

using PyTorch for deep learning inference and PostgreSQL for structured data management, ensuring seamless handling of multimodal text refinements. Unlike conventional OCR correction tools that operate in a single-pass mode, MultiFOLD supports batch processing for efficient large-scale deployment.

MultiFOLD enables adaptive correction workflows as shown in Figure 3 to further enhance usability, where users can configure correction thresholds based on document complexity. This allows organizations to optimize annotation speed while maintaining high accuracy across various OCR use cases. Combining speech-based refinement, confidence-aware text fusion, and scalable deployment establishes MultiFOLD as a robust solution for real-world document digitization.

4 Experiments and Results

We conduct extensive experiments to evaluate the effectiveness of MultiFOLD in correcting OCR errors through multimodal refinement. Our evaluation focuses on four key research questions: (1) How does MultiFOLD compare to existing OCR correction methods in terms of accuracy? (2) Does integrating ASR with OCR refinement reduce human annotation time? (3) What types of errors does MultiFOLD effectively correct, and where does it struggle? (4) How significant is the contribution of the context-aware ByT5 model to overall performance?

To systematically address these questions, we benchmark MultiFOLD against multiple baselines, analyze its impact on annotation effort, and conduct an ablation study to quantify the role of different components in the system.

4.1 Dataset and Experimental Setup

We curate a dataset of 800 cluttered document pages spanning multiple domains, including legal contracts, historical manuscripts, and handwritten forms. Each document contains an average of **35 sentences** (total: 28k sentences). The dataset covers a diverse range of real-world noise conditions such as ink smudges, overlapping handwriting, stains, and artificial obfuscation. These variations ensure robustness in evaluation by mimicking challenging digitization scenarios.

For benchmarking, we compare MultiFOLD against widely used OCR and ASR models. Specifically, we use trocr-large-printed of all the state-of-the-art OCR systems inferred from the Table 1 and Whisper ASR for speech transcriptions. MultiFOLD's correction module is powered by a fine-tuned ByT5-small model, trained on a carefully curated dataset of OCR-ASR correction pairs. The fine-tuning corpus includes:

- OCR errors sampled from real-world document scans to ensure domain generalization.
- ASR transcriptions of human corrections to improve speech-based text alignment.
- Synthetic noisy inputs created by perturbing clean transcriptions, enhancing robustness.

Table 1. Comparison of OCR models based on Word Error Rate (WER) and Character Error Rate (CER).

OCR Models	**WER** (%)	**CER** (%)
trocr-large-printed	**62.15**	**43.26**
OCR-Donut-CORD	65.29	61.88
Llama-3.2-11B-Vision	146.88	138.22
Qwen2-VL-7B-Instruct	63.59	60.33
DeepSeek-VL2	66.58	68.01

To quantify performance, we compute the Character Error Rate (CER) and Word Error Rate (WER) two standard text correction metrics that measure deviation from ground truth text.

4.2 Comparison with Baseline Approaches

To understand the impact of MultiFOLD, we compare it with three baseline approaches:(1) Manual OCR correction, where annotators manually edit raw OCR outputs without ASR assistance. (2) ASR-only correction, where users read full sentences aloud, and an ASR system transcribes them. (3) OCR+ASR fusion without refinement, where raw OCR and ASR predictions are directly aligned without contextual modeling.

Table 2. Comparison of OCR-ASR correction performance.

Method	WER (%)	CER (%)
OCR-only (trocr-large)	58.17	38.32
ASR-only (Whisper)	34.25	10.28
OCR+ASR (no refinement)	19.25	6.32
MultiFOLD (Ours)	**14.04**	**2.28**

Table 3. Time reduction in annotation tasks on sentence level with MultiFOLD.

Annotator	Manual (s)	MultiFOLD (s)
Average	18.38	8.75

Table 2 presents WER and CER scores across methods. The results demonstrate that MultiFOLD outperforms all baselines, achieving a 75.86% reduction in WER and a 36.04% reduction in CER compared to traditional OCR correction. The ASR-only approach improves over raw OCR, but it still fails in cases where homophones or spoken ambiguities introduce new errors. By integrating ASR with OCR in a context-aware manner, MultiFOLD successfully bridges this gap.

4.3 Reduction in Annotation Time

In addition to improving accuracy, MultiFOLD significantly reduces human annotation effort. To quantify this reduction, we conduct a user study where 10 annotators correct OCR errors across 20 pages, both manually and using MultiFOLD.

Table 3 shows that MultiFOLD reduces annotation time by an average of 52.39% on the sentence level, allowing annotators to complete corrections significantly faster while maintaining accuracy. This improvement arises from the ability to vocalize corrections instead of manually typing them, along with the intelligent context-aware refinement enabled by ByT5.

Figure 4 shows a reduction in the annotation efforts in terms of time taken (s) to correct each page with and without MultiFOLD for the two annotators, one who had taken the overall maximum time and minimum time for 20 pages respectively.

4.4 Details for the Fine-Tuning of Byt5 Model

Learning Rate : 5e-5
Batch Size : 16
Dropout : 0.1
AdamW Optimizer $\beta_1 : 0.9, \beta_2 : 0.999, \epsilon$: 1e-08
Epochs (Early Stopping) : 20 (2)

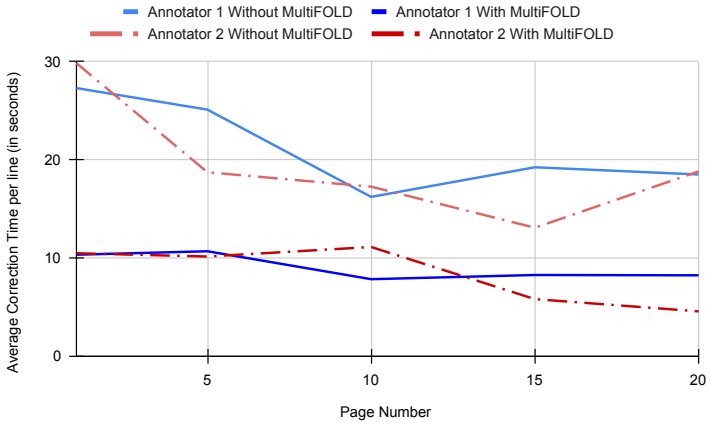

Fig. 4. Qualitative time analysis: MultiFOLD effectively corrects OCR errors by integrating ASR.

4.5 Error Analysis and Qualitative Insights

We analyze specific types of errors MultiFOLD corrects effectively. Key strengths include: (1) Homophone disambiguation, resolving issues like *"there" vs. "their"*. (2) Character restoration, recovering partially occluded or missing letters. (3) Phrase-level correction, ensuring grammatical coherence in OCR refinement.

However, MultiFOLD struggles in cases where ASR misinterpretations introduce incorrect substitutions, particularly in noisy environments. Future work will focus on confidence-based filtering mechanisms to address this issue. Figure 4 illustrates the corrections in the OCR of the sample cluttered document through the MultiFOLD tool.

4.6 Ablation Study: Impact of Context-Aware Refinement

To assess the contribution of the ByT5 correction module, we compare context-aware refinement of the MultiFOLD with different variants of the Byt5 model. Table 4 shows the difference between the results of the ByT5 variants, which had a 42.74% decrease in WER, validating the necessity to choose an optimal model for context-aware refinement.

Table 5 shows that the best Byt5-model inferred from Table 4 shows superior results when trained in OCR + ASR, rather than trained the OCR and ASR predictions on incorrect phrases separately.

5 Conclusions

In this work, we introduced MultiFOLD, a multimodal OCR correction framework that integrates speech-assisted refinement with context-aware text correction, enabling efficient and accurate post-OCR refinement in cluttered document

Table 4. Ablation study: Effect of ByT5-based refinement based on its different variants.

Method	WER (%)	CER (%)
MultiFOLD (ByT5-small)	48.73	9.44
MultiFOLD (ByT5-base)	**14.04**	**2.28**
MultiFOLD (ByT5-large)	49.05	11.37

Table 5. Ablation study: Effect of ByT5-Base refinement based on different modalities.

Method	WER (%)	CER (%)
OCR only	51.57	31.48
ASR only	24.79	9.09
OCR + ASR	**14.04**	**2.28**

settings. By leveraging a confidence-weighted OCR-ASR fusion strategy alongside a fine-tuned ByT5 sequence-to-sequence model, MultiFOLD significantly improves OCR accuracy while reducing human annotation effort. Our experimental results demonstrate a 36.04% reduction in Character Error Rate (CER) and a 52.39% decrease in annotation time, outperforming conventional OCR and ASR-based correction methods. These findings validate that multimodal learning can effectively mitigate OCR errors such as homophone ambiguities, missing characters, and phrase-level inconsistencies, establishing MultiFOLD as a robust and scalable solution for document digitization. While our approach substantially enhances OCR post-processing, several research directions remain open. Future work could explore adaptive confidence-based fusion models to dynamically adjust OCR-ASR alignment thresholds, self-supervised learning techniques to improve ASR robustness under noisy conditions, and extensions to multilingual and handwritten text recognition to support broader document digitization applications. Additionally, integrating active learning strategies could further optimize human-in-the-loop interactions by prioritizing intervention only in high-uncertainty cases, while real-time optimizations using efficient transformer variants could enhance deployment in large-scale archives and legal document processing. By bridging the gap between manual OCR correction and fully automated refinement, MultiFOLD lays the groundwork for future advancements in multimodal AI-driven document annotation systems, demonstrating that speech-assisted correction is a viable path forward for improving text recovery in complex, real-world scenarios.

References

1. Kili technology. `https://kili-technology.com/`, accessed: 2025-02-20
2. Fromthepage. `https://fromthepage.com/`, accessed: 2025-02-20
3. Tong, X., Evans, D.A.: A statistical approach to automatic OCR error correction in context. In: Fourth Workshop on Very Large Corpora, 1996.

4. Ahmed, I., et al.: Technique for automatic sentence level alignment of long speech and transcripts. In: Interspeech 2013.
5. Schmaltz, A., et al.: Sentence-Level Grammatical Error Identification as Sequence-to-Sequence Correction. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, 2016.
6. Schmaltz, A., et al.: Adapting Sequence Models for Sentence Correction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017.
7. Lin, J., Ledolter, J.: A Simple and Practical Approach to Improve Misspellings in OCR Text. arXiv preprint arXiv:2106.12030 (2021).
8. Belagavi, K., Tadimeti, P., & Sarvadevabhatla, R.K. (2021). DocVisor: A Multi-purpose Web-Based Interactive Visualizer for Document Image Analytics. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR). Cham: Springer International Publishing.
9. R. Saluja, D. Adiga, G. Ramakrishnan, P. Chaudhuri, and M. Carman, A Framework for Document Specific Error Detection and Corrections in Indic OCR". In: Proceedings of the 14th IAPR Int. Conf. Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017, pp. 25–30. https://doi.org/10.1109/ICDAR.2017.308
10. Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M.,... Raffel, C. (2022). Byt5: Towards a token-free future with pre-trained byte-to-byte models. Transactions of the Association for Computational Linguistics, 10, 291-306.
11. Souibgui, M. A., & Kessentini, Y. (2022). DE-GAN: A Conditional Generative Adversarial Network for Document Enhancement. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3), 1180–1191. https://doi.org/10.1109/TPAMI.2020.3022406
12. Kiessling, B., Tissot, R., Stokes, P., & Stökl Ben Ezra, D. (2019). eScriptorium: An Open Source Platform for Historical Document Analysis. In Proceedings of the 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), Sydney, NSW, Australia, 2019, pp. 19–19. IEEE. https://doi.org/10.1109/ICDARW.2019.10032
13. Constum, T., et al. (2023). Pivan: A Web-Platform for Document Annotation. In Proceedings of the Archiving Conference, Vol. 20, Society for Imaging Science and Technology, 2023.
14. Sulaiman, A., Omar, K., Nasrudin, M.F.: Degraded historical document binarization: A review on issues, challenges, techniques, and future directions. J. Imaging 5(4), 48 (2019).
15. Tkachenko, M., Malyuk, M., Holmanyuk, A., Liubimov, N.: Label Studio: Data labeling software (2020), open source software available from https://github.com/heartexlabs/label-studio
16. J. Zhang, et al., "DocRes: A generalist model toward unifying document image restoration tasks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 1–10.
17. Z. Yang, et al., "DocDiff: Document enhancement via residual diffusion models," in Proc. 31st ACM Int. Conf. Multimedia, 2023, pp. 1–10.
18. M. Li, et al., "High-fidelity document stain removal via a large-scale real-world dataset and a memory-augmented transformer," arXiv preprint arXiv:2410.229KDD' 024.
19. Kermorvant, C., et al.: Callico: a versatile open-source document image annotation platform. In: Proceedings of the International Conference on Document Analysis and Recognition, Cham, Springer Nature Switzerland (2024).

20. Baek, Y., Lee, J., & Shin, H. (2019). Character Region Awareness for Text Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).