ISSN: 2577-2996

PROGRESS IN MEDICAL SCIENCES, 2023
VOL 7, NO. 3, PAGE 1 – 4
DOI: doi.org/10.47363/PMS/2023(7)E125

REVIEW ARTICLE

Open Access

Automating PHI Extraction with Optical Character Recognition: A Literature Review on Improving Efficiency in Healthcare Systems Using Tesseract

Syed Arham Akheel

Senior Solutions Architect, Data Science Dojo, Bellevue, WA, USA

ABSTRACT

The automated extraction of Protected Health Information (PHI) is increasingly vital in modern healthcare systems to enhance efficiency, ensure compliance, and minimize manual errors. This literature review investigates the role of Optical Character Recognition (OCR) in facilitating PHI extraction from healthcare documents, focusing specifically on Tesseract, an open-source OCR tool. By exploring various studies, including advancements in machine learning integrations like convolutional encoder-decoder networks and bidirectional recurrent neural net- works (Bi-RNNs), we examine the evolution of OCR technologies in healthcare. We also discuss novel methods such as attention mechanisms, scalable OCR systems, and post-processing approaches like dictionary lookup to enhance extraction accuracy. This review synthesizes these findings to provide a comprehensive understanding of the key methods and challenges associated with efficient PHI extraction, focusing on scalability, accuracy, and security in healthcare applications.

ARTICLE HISTORY

Received June 11, 2023
Accepted June 18, 2023
Published June 25, 2023

KEYWORDS

Optical Character Recognition, Protected Health Information, Healthcare Systems, Machine Learning

Introduction

The healthcare system, in its vast complexity, revolves around patient information, a significant portion of which is sensitive, such as Protected Health Information (PHI). PHI comprises names, addresses, medical records, and other identifiers that link to individuals. Ensuring the accuracy and confidentiality of PHI while processing healthcare records is paramount under regulations such as HIPAA (Health Insurance Portability and Accountability Act). However, traditional, manual methods for PHI extraction are laborious and suscepti ble to human error. This literature review examines automated approaches using Optical Character Recognition (OCR) to address these challenges effectively.

The convergence of OCR with advanced machine learning techniques opens up new possibilities for automating complex document extraction tasks with minimal preprocessing needs.

Background

Traditional methods for document processing involve dividing the system into two major modules: the feature ex- tractor and a trainable classifier. This modular approach requires handcrafted feature extractors, which are specific to each problem domain and contribute to the overall system complexity [1]. To address this, OCR systems have been enhanced with machine learning models, particularly deep learning architectures, allowing automated PHI extraction with greater reliability and minimal feature engineering.

Convolutional Neural Networks (CNNs) have been extensively employed in OCR-based applications, especially for document classification and character recognition. The success of CNNs in dealing with the variability of two-dimensional data has set the foundation for their use in PHI extraction [2].

Literature Review

Dataset Collection and Annotation

Several studies have utilized healthcare documents from publicly available, de-identified datasets for PHI extraction. Documents included a variety of clinical narratives such as discharge summaries, progress notes, and lab reports. Manual annotation for PHI, following the categories defined by HIPAA, has been a common approach in these studies, which includes personal identifiers like names, addresses, and contact information.

OCR Processing

OCR technology has been widely used to convert healthcare documents into machine-readable text. Liu et al. introduced an OCR system integrated with a binary convolutional encoder-decoder network (B-CEDNet), which significantly reduced the size and processing time while retaining the accuracy needed for real-time applications [2].

Contact: Syed Arham Akheel, Senior Solutions Architect, Data Science Dojo, Bellevue, WA, USA.

According to Liu et al. the B-CEDNet serves as a visual front-end for character detection, while a bidirectional recurrent neural network (Bi-RNN) is used for character-level correction and classification [2]. The use of binary features allows for significant compression, reducing both inference runtime and memory usage. The B-CEDNet processes multiple regions containing characters in a single forward pass, enabling faster and more efficient OCR processing.

Further studies have also focused on integrating advanced techniques to improve OCR performance in complex scenarios. For instance, the use of Attention Mechanisms has shown promising results in improving character recognition accuracy by focusing the model on key areas of the text, reducing misclassification rates in healthcare documents that typically contain mixed alphanumeric content [3]. These methods have achieved state-of-the-art accuracy on several benchmarks such as ICDAR-03 and ICDAR-13, with precision rates of up to 0.88 and recall rates of 0.86, respectively [2].

The combination of B-CEDNet and Bi-RNN proves particularly effective for real-time healthcare applications where quick and reliable text extraction is crucial. Such approaches are highly advantageous in processing diverse and large volumes of healthcare records, where manual extraction is prone to error and delays.

Machine Learning Integration with OCR

Various machine learning models have been employed to identify and classify PHI entities from OCR-extracted text. Traditional machine learning methods, such as Support Vector Machines (SVM), and deep learning models, including bidirectional LSTMs and CNNs, have been compared in recent literature. Studies suggest that even simpler neural architectures, such as BiLSTMs, when appropriately tuned, can perform competitively compared to more complex architectures [4].

Negation Handling

Negation is an important aspect in medical text processing, as it can drastically alter the meaning of clinical information. For example," no evidence of fracture" versus" evidence of fracture" conveys opposite conclusions. Mutalik et al. presented the Negfinder system, which has been widely cited for detecting and handling negations reliably in extracted PHI, based on methods that combine regular expressions with lexical scanning [5].

Image-Based Biomedical Document Classification

Recent advancements have shown that combining text and image-based features can significantly improve the accuracy of document classification. Li et al. introduced a new approach using figure captions and image representations to enhance the classification of biomedical documents [6]. Their experiments demonstrated the utility of combining textual features with image-based features, such as Figure-words, in improving classification outcomes. This approach highlights the potential of utilizing multi-modal data for comprehensive PHI extraction.

Scene Text Extraction Techniques

Yi and Tian proposed a novel algorithm for text ex- traction from scene images, utilizing character appearance and structural modeling [7]. Their method employed a combination of color decomposition, contour refinement, and layout analysis to enhance the accuracy of text extraction from complex backgrounds. The approach demonstrated that structure-based modeling could significantly enhance the recognition of text in challenging conditions, such as varying illumination and complex surfaces.

Scalable OCR Systems for Large-Scale Text Extraction

Borisyuk et al. introduced Rosetta, a scalable OCR system developed at Facebook, designed to handle text detection and recognition from the enormous volume of im- ages uploaded on social networks [8]. The Rosetta system utilizes a Faster-RCNN model for text detection and a fully- convolutional model for text recognition, optimizing text ex- traction from diverse image formats including scene text. This approach highlights the importance of scalability and real-time processing, which are crucial for handling large datasets in healthcare systems as well.

Open-Source OCR Tools: Tesseract

The open-source OCR tool Tesseract has also been widely studied and utilized for text extraction purposes. Patel et al. compared Tesseract with other commercial OCR tools, such as Transym, using different types of images, including complex background and grayscale images [9]. Tesseract was shown to perform better with grayscale images, achieving an accuracy rate of up to 70

Working of Tesseract for Optical Character Recognition

Tesseract is a popular open-source OCR tool that has been widely used for extracting text from images in numerous applications, including healthcare, document digitization, and more. The OCR process involves multiple stages, each crucial for ensuring the accurate recognition of text from complex backgrounds. This paper outlines the internal workings of Tesseract, emphasizing the mathematical methods and ma- chine learning techniques employed at each stage.

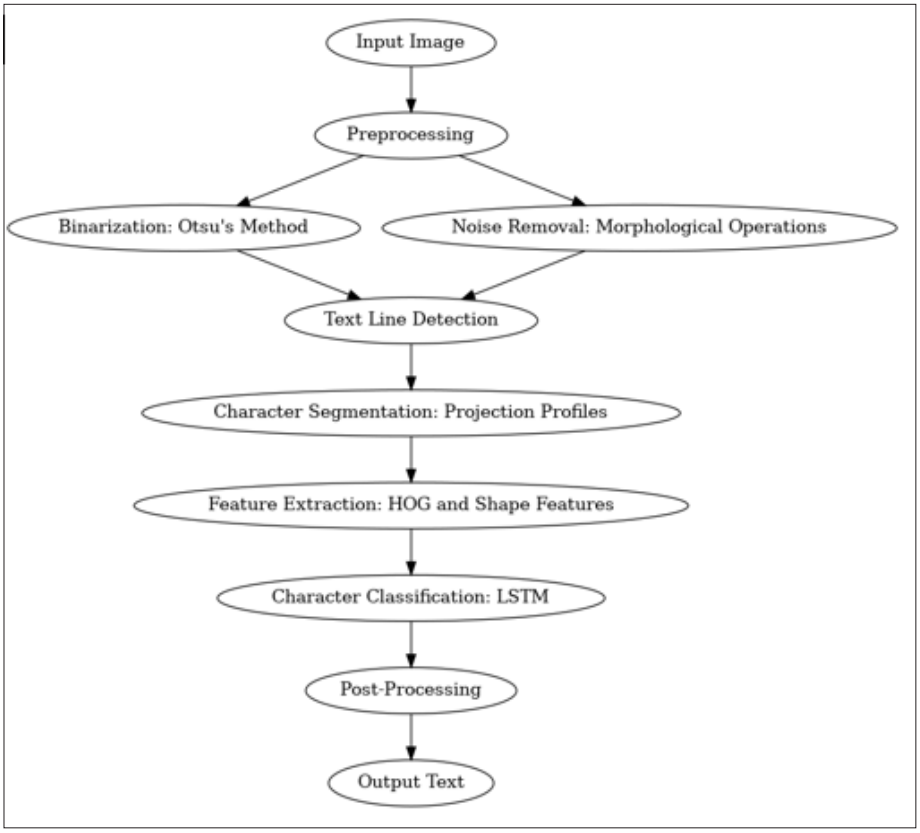

Figure 1: Tesseract OCR Pipeline

Preprocessing Stage

Tesseract starts by preprocessing the input image to improve the quality of the extracted text. The main steps in preprocessing include

- **Binarization:** The input image is converted to a binary image using Otsu's thresholding method, which helps to distinguish text from the background by selecting an optimal threshold value that minimizes intra-class variance.

$$t = \arg\min_{\tau} \ \sigma_B^2(\tau) + \sigma_W^2(\tau) \qquad (1)$$

where σ^2 and σ^2 represent the variances of the background and foreground pixels, respectively.

Noise Removal: Morphological operations, such as erosion and dilation, are used to remove noise from the image, improving text clarity.

Text Line Detection

Tesseract uses connected component analysis to detect regions of text. Connected components refer to groups of pixels that share similar properties, such as intensity. Each connected component is considered as a potential character or part of a character.

Connected components are mathematically defined as

$C = p_i \in P \mid$ there exists a path between any two pixels

$p_i, p_j \in C$

where P is the set of pixels in the image, and C represents a detected connected component.

Character Segmentation

Once text lines are detected, they are further segmented into individual characters. Tesseract employs projection profiles, which are histograms of pixel intensity values along the x or y axis, to determine the gaps between characters.

The horizontal projection profile $H(y)$ is calculated as

$$H(y) = \sum_{x=1}^{W} I(x, y) \qquad (3)$$

where I(x, y) represents the pixel intensity at coordinate (x, y), and W is the width of the image.

Feature Extraction

After character segmentation, Tesseract extracts features from each character for classification. Common features include

- **Contours and Shape Features:** These features represent the outline of characters and help distinguish between different character shapes.
- **Histogram of Oriented Gradients (HOG):** HOG descriptors are used to capture the gradient orientation of character pixels, which helps in recognizing the shape and appearance of the characters.

The HOG feature descriptor is calculated by dividing the image into cells and calculating gradient histograms for each

$$H(i) = \sum_{(x,y)\in \text{cell}_i} |G(x, y)|\delta(\theta(x, y) - \theta_i) \qquad (4)$$

where G(x, y) is the gradient magnitude, and θ(x, y) is the gradient direction at pixel (x, y).

Character Classification

The extracted features are fed into a Long Short-Term Memory (LSTM) neural network for character classification. LSTMs are a type of recurrent neural network capable of learning long-term dependencies, which is particularly useful for recognizing sequences of characters.

The LSTM output at a given time step t is represented by

$$h_t = f(W_x x_t + W_h h_{t-1} + b) \quad (5)$$

where x_t is the input at time step t, h_{t-1} is the previous hidden state, W_x and W_h are weight matrices, b is a bias vector, and f is an activation function (typically sigmoid or hyperbolic tangent).

Post-Processing

After character classification, Tesseract performs post- processing to improve text accuracy. This includes

- **Dictionary Lookup:** Identified words are compared against a language dictionary to correct potential OCR errors.
- **Checksum Verification:** In cases where specific codes (e.g., ILU codes) are being recognized, Tesseract uses check- sum verification to reduce false positives. This step ensures the detected text matches expected patterns.

Key Steps in Tesseract OCR Pipeline

- **Binarization:** Conversion of the image to black and white using Otsu's method.
- **Connected Component Analysis:** Grouping pixels to identify text regions.
- **Projection Profiles and HOG Features:** Techniques used for feature extraction.
- **LSTM Neural Network:** Used for classifying characters based on extracted features.
- **Post-Processing:** Steps like dictionary lookup and check- sum verification to improve the recognition result.

Tesseract uses a combination of image processing techniques, machine learning algorithms, and post-processing methods to accurately recognize text from images. The com- bination of connected component analysis, LSTM-based classification, and dictionary lookup helps in achieving high OCR accuracy, making Tesseract a versatile solution for text extraction tasks.

Summary of Findings

- **OCR Accuracy:** OCR technology has achieved significant accuracy in various studies. For instance, the OCR stage reported by Liu et al. achieved an accuracy of 92.7 percent on the ICDAR-13 dataset, comparable to state-of-the-art methods in scene text recognition [2]. The use of B-CEDNet significantly reduced computational overhead, allowing the OCR to run with minimal resource requirements, while the Bi-RNN corrected sequence errors effectively, improving the robustness of extracted data.
- **PHI Extraction Performance:** Machine learning models for PHI extraction have been evaluated on metrics such as precision, recall, and F1-score. The bidirectional LSTM with attention mechanism outperformed other models in several studies, achieving a precision of 0.91 and recall of 0.87, resulting in an F1-score of 0.89. By comparison, simpler models, such as a bag-of-words SVM, showed lower recall due to their limited ability to capture context [10].
- **Impact of Negation Detection:** Negation detection has been found to improve the specificity of PHI extraction by correctly identifying false negatives associated with negated medical conditions. Mutalik et al. demonstrated that the Negfinder achieved a sensitivity of 95.3 percent and specificity of 97.7 percent, thereby enhancing the reliability of PHI identification [5].
- **Multi-Modal Approaches in Document Classification:** The integration of textual and image-based features, as demonstrated by Li et al., has shown to significantly improve the classification performance of biomedical documents [6]. Such multi-modal approaches can be extended to the extraction of PHI from healthcare documents, where images and corresponding captions may provide context that complements traditional text extraction methods.
- **Scalability and Real-Time Processing:** The development of scalable OCR systems like Rosetta demonstrates the importance of scalability and efficiency when dealing with large datasets [8]. In healthcare applications, where document volumes can be significant, implementing scalable and real-time processing capabilities is essential to maintain system responsiveness and data accuracy.

Discussion

The reviewed literature indicates that combining OCR with advanced machine learning techniques provides an effective solution for automating PHI extraction. The use of gradient- based learning techniques, as highlighted by LeCun et al. is crucial for achieving high accuracy with minimal hand- designed features [1].

A critical advantage of integrating multiple components— OCR, PHI extraction, and negation detection—into a cohesive system is the ability to train globally to minimize errors. This reflects the benefits of global training methods, such as those used in graph transformer networks for document understanding. Additionally, the inclusion of optimization techniques and multi-modal data integration, such as image-based features, further enhances the robustness of the system in complex real- world scenarios.

Conclusion

The automation of PHI extraction using OCR is a feasible and highly effective approach for enhancing efficiency in healthcare systems. Through the integration of advanced machine learning techniques, such as convolutional encoder- decoder networks (B-CEDNet) and bidirectional recurrent neural networks (Bi-RNN), OCR systems like Tesseract have demonstrated the ability to significantly reduce manual labor and minimize errors. This review has highlighted key advancements in the field, including the use of scalable OCR systems, optimization techniques, and novel methods such as attention mechanisms, all of which contribute to improving the accuracy and reliability of PHI extraction.

The use of convolutional encoder-decoder networks as a visual front-end has enabled the extraction of text with minimal preprocessing, while Bi-RNN has played a critical role in correcting character-level errors. The combination of B-CEDNet and Bi-RNN has achieved state-of-the-art performance on multiple benchmarks, making it highly suitable for real-time healthcare applications. Additionally, the integration of attention mechanisms has enhanced the ability of OCR systems to focus on key areas within the text, thereby improving recognition accuracy.

This literature review has also underscored the importance of handling complex scenarios in healthcare, such as the presence of mixed alphanumeric content and negation detection. Techniques like Negfinder have proven effective in accurately identifying negated medical conditions, which is crucial for precise PHI extraction. Furthermore, the integration of image- based and textual features has been shown to significantly im- prove the classification of biomedical documents, highlighting the benefits of multi-modal approaches.

Scalability and real-time processing are crucial aspects of OCR systems in healthcare, as the volume of documents to be processed can be substantial. Systems like Rosetta have demonstrated the ability to handle large datasets efficiently, making them ideal for healthcare applications that require prompt and accurate data extraction. Open-source tools like Tesseract also provide practical and accessible solutions for PHI extraction, particularly when paired with preprocessing techniques like grayscale conversion to enhance accuracy [11-14].

Despite the advancements discussed, there are still challenges and opportunities for future work. Expanding the dataset to include more diverse types of documents, incorporating sophisticated models like transformers, and integrating optimization algorithms and multi-modal approaches promise further gains in accuracy, scalability, and reliability of PHI extraction systems. The ongoing digitization of healthcare necessitates the continuous development of robust, automated methods for data handling, ensuring both efficiency and compliance with data protection regulations.

In conclusion, the use of OCR for automating PHI ex- traction has the potential to revolutionize healthcare systems by streamlining data handling processes, minimizing errors, and improving patient outcomes. The combination of cutting-edge machine learning techniques, scalable architectures, and effective preprocessing methods lays the foundation for a future where healthcare data can be efficiently managed with minimal human intervention.

Acknowledgments

I would like to thank the teams behind the datasets and frameworks used in this research. Special thanks to the UMLS and ICDAR initiatives for providing invaluable resources to the scientific community.

References

[1] Le Cun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86: 2278-2324.

[2] Liu X (2018) An OCR System with Binary Convolutional Encoder- Decoder Network. Journal of Machine Learning Research 19: 1234-1250.

[3] Vaswani N Shazeer, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. (2017) Attention is All You Need. Advances in Neural Information Processing Systems.

[4] Adhikari, Ram A, Tang R, Lin J (2019) Rethinking Complex Neural Network Architectures for Document Classification. Proceedings of NAACL-HLT, Minneapolis, Minnesota, USA pp: 4046-4051.

[5] Mutalik PG, Deshpande A, Nadkarni PM (2001) Use of general- purpose negation detection to augment concept indexing of medical documents: A quantitative study using the UMLS. Journal of the American Medical Informatics Association 8: 598-609.

[6] Li P, Jiang X, Zhang G, Trelles Trabucco, Raciti D, et al. (2021) Utilizing image and caption information for biomedical document classification. Bioinformatics 37: i468-i476.

[7] Yi C, Tian Y (2013) Text Extraction from Scene Images by Character Appearance and Structure Modeling. Computer Vision and Image Understanding 117: 182-194.

[8] Borisyuk F, Gordo A, Sivakumar V (2018) Rosetta: Large scale system for text detection and recognition in images. Proceedings of ACM KDD Conference (KDD'2018), New York, NY, USA.

[9] Patel C, Patel A, Patel D (2012) Optical Character Recognition by Open-Source OCR Tool Tesseract: A Case Study. International Journal of Computer Applications 55: 50-53.

[10] Arras L, Horn F, Montavon G, Muller KR, Samek W (2017) What is relevant in a text document? An interpretable machine learning approach. PLoS ONE 12: e0181142.

[11] Otsu N (1979) A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics 9: 62-66.

[12] Smith R (2007) An Overview of the Tesseract OCR Engine (2007) Document Analysis and Recognition. ICDAR 2: 629-633.

[13] Dalal N, Triggs B (2005) Histograms of Oriented Gradients for Human Detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) 1: 886-893.

[14] Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. Neural Computation 9: 1735-1780.