A Survey Paper On Template Extraction and Template Matching for Medical Reports and Prescriptions

Kushal. B. R,*Student*
Dept of Computer Science and engineering
Atria Institute of Technology
Bangalore,India
kushal.br12@gmail.com

Lakshmi Govind Joshi,*Student*
Dept of Computer Science and engineering
Atria Institute of Technology
Bangalore,India
joshilakshmi2002@gmail.com

Akarsh, *Student*
Dept of Computer Science an engineering
Atria Institute of Technology
Bangalore,India
akarsh.ravikumar6@gmail.co

Dr. Devi Kannan, *HoD*
Dept of Computer Science and engineering
Atria Institute of Technology
Bangalore,India
devi.kannan@atria.edu

***Abstract*— Similar document templates are widely used in various domains, such as medical, legal, and financial. However, the use of similar templates can also pose a risk of fraud, especially when the content of the documents is modified to deceive the recipients or claimants. In this work, we suggest a novel algorithm for matching and detecting similar document templates, with applications to the medical domain. Our algorithm comprises three main steps: template extraction, template comparison, and fraud detection. First, we develop a mechanism to extract and standardize templates from medical invoices, prescriptions, and reports on lab tests obtained from providers and customers. Second, we implement an algorithm to compare and identify similarities among different document templates, accounting for minor modifications and variations in content. Third, we enhance the algorithm to flag instances of potentially fraudulent claims where similar templates are used, with modifications made to customer details while maintaining the overall structure and design. We evaluate our algorithm on a large dataset of real-world medical documents and demonstrate its effectiveness and efficiency in matching and detecting similar document templates. We also discuss the implications and limitations of our algorithm for fraud prevention and detection.**

I. INTRODUCTION

In various professional fields, the widespread use of standardized document templates has significantly improved operational efficiency. However, the adoption of similar templates has also given rise to the potential for fraudulent activities, particularly in sensitive domains such as healthcare. This paper addresses the specific challenges posed by the misuse of similar document templates within the medical domain and proposes an innovative algorithm tailored for matching and detecting these templates with a focus on preventing fraud.

The algorithm unfolds in three distinct phases, meticulously designed to address the unique considerations of medical documents. First, the authors introduce a template extraction mechanism tailored to medical invoices, prescriptions, and lab test reports. This step is crucial for standardizing templates from diverse sources, setting the stage for subsequent comparisons.

The second phase involves the implementation of an algorithm adept at comparing medical document templates, accounting for minor modifications and variations in content. This nuanced approach recognizes the legitimate need for flexibility in medical templates while ensuring the recognition of similarities among documents, even in cases where certain details are changed.

The third and final phase enhances the algorithm's functionality to serve as a fraud detection tool within the medical domain. By flagging instances where similar templates are used with modifications to patient details, the algorithm becomes a valuable asset in identifying potential fraudulent claims. The authors validate the effectiveness and efficiency of their algorithm through rigorous testing on a substantial dataset of real-world medical documents.

This work does more than just a technological advancement tailored for the medical field but also discusses the broader implications of the proposed algorithm for fraud prevention and detection in healthcare. The authors acknowledge the limitations of their approach These offer directions for additional study, emphasizing the importance of document security and integrity in the sensitive realm of medical documentation.

II. METHODOLOGIES – TEMPLATE EXTRACTION

A. Optical Character Recognition

Optical Character Recognition (OCR) is a technology that enables the conversion of printed or handwritten text in images or scanned documents into machine-readable and editable text. Using sophisticated algorithms, OCR software recognizes individual characters and interprets them, allowing for the extraction of textual information from sources like documents, images, or PDFs. OCR has a vital part in automating data entry, digitizing printed content, and facilitating text-based analysis by transforming non-editable text into a searchable and manipulable format, and processed by computers. It finds widespread applications in various fields, including document management, data extraction, and accessibility services.

Implementing Optical Character Recognition (OCR) for template extraction from medical reports involves leveraging OCR technology to convert scanned or digital medical documents into machine-readable text. The first step is selecting a robust OCR tool such as Tesseract OCR, Google Cloud Vision OCR, or Microsoft Azure OCR, capable of accurately transcribing characters from images. Following document preparation, the OCR software extracts text, necessitating preprocessing to refine the accuracy and consistency of the extracted information.

The next phase centers on defining a structured template for the medical reports. This entails identifying key details like patient names, dates, medical histories, and test results. Text analysis and parsing methods, such as regular expressions and natural language processing (NLP), are then

employed to categorize and extract relevant information according to the predetermined template structure. The third step involves mapping the parsed data to specific fields within the template, creating a cohesive representation. Subsequently, validation checks and quality control measures ensure the accuracy and completeness of the extracted information, leading to the generation of a final template. Integration with databases or Electronic Health Record (EHR) systems facilitates seamless storage, retrieval, and analysis of the structured medical data. Regular testing, iteration, and potential updates contribute to the system's adaptability to varying report formats and languages. Ultimately, deploying the OCR-based template extraction system enhances the efficiency of processing medical reports while maintaining data integrity.

B. Document Image Matching

Document Image Matching is a computational process that involves utilizing computer vision techniques to compare and analyze the visual content of documents. This technique is especially helpful for jobs like extracting templates from documents, including medical records. In relation to medical reports, Document Image Matching can involve identifying and aligning specific sections or templates within the documents, enabling the extraction of key information like patient details, diagnoses, and treatment plans. Techniques may include image feature extraction, pattern recognition, and deep learning algorithms, allowing for the automated recognition and understanding of document structures. The application of Document Image Matching streamlines information retrieval from diverse document formats, contributing to improved data organization and analysis in various fields, including healthcare.

To implement Document Image Matching for extracting templates from medical reports, begin by collecting a diverse dataset of medical reports in various formats. Employ Optical Character Recognition (OCR) to convert image-based reports into machine-readable text. Following data preprocessing, design templates representing key elements such as patient information, diagnosis, and medications. Utilize computer vision techniques, including feature extraction and pattern matching, to match the visual layout of these templates within the reports. Simultaneously, leverage natural language processing (NLP) methods, such as Named Entity Recognition, to extract textual content from identified regions. Iterate on your algorithms using feedback and validation mechanisms, ensuring scalability, robustness, and adherence to privacy regulations. Finally, integrate the system with healthcare workflows while prioritizing data security.

III. Methodologies – Template Matching

A. Clustering

Template matching in medical reports involves identifying patterns or structures within the reports that follow a certain template or format. ***Clustering algorithms play a vital part in this process by automatically grouping similar medical reports together in light of their content and structure.*** One common application is the clustering of radiology reports, where different reports may share similarities in terms of the types of observations, diagnoses, or recommended treatments. Clustering helps organize these reports, simplifying it for healthcare professionals to review and analyze similar cases collectively.

One notable clustering algorithm employed in medical template matching is k-means clustering. K-means groups data points into k clusters based on their similarities, and in the context of medical reports, it can identify common patterns in terms of terminology, structure, and content. By applying k-means clustering to a dataset of medical reports, healthcare providers can quickly categorize reports that adhere to specific templates, such as those for different imaging modalities or clinical specialties. This not only aids in organizing large volumes of medical data but also facilitates the development of more efficient and accurate reporting systems.

Additionally, hierarchical clustering can be valuable for template matching in medical reports. This algorithm builds a tree-like structure of clusters, allowing for the identification of both broad categories and more specific subgroups within the medical reports. When working with complex medical data that may show nested patterns or hierarchical linkages, hierarchical clustering is especially helpful. Healthcare practitioners can acquire insights into the minute variances in reporting styles or specifics within particular medical specialties thanks to its more comprehensive explanation of template variants. Overall, clustering algorithms enhance the efficiency of template matching in medical reports, enabling better organization, retrieval, and analysis of valuable healthcare information.

B. Component Block Listing

Component block listing algorithms are an effective tool for matching templates in medical reports, allowing for the extraction and identification of specific sections or components within the reports. In the context of medical documentation, reports often follow a standardized structure with distinct sections for patient information, clinical findings, diagnoses, and treatment plans. Component block listing algorithms can automatically parse these reports, dividing them up into useful parts or sections. This approach facilitates template matching by isolating relevant information, making it easier to compare and categorize reports based on their constituent elements.

One key advantage of using component block listing algorithms in medical template matching is their ability to

handle variations in report formats. Medical reports may come in diverse styles and structures, and component block listing algorithms can adapt to these variations by identifying and extracting specific content blocks regardless of their placement or formatting. This flexibility is particularly valuable in healthcare settings where report templates may evolve or differ across different departments or institutions. By focusing on the components that contain critical information, such as diagnostic codes or patient demographics, these algorithms enhance the accuracy and reliability of template matching processes.

Furthermore, component block listing algorithms can aid in the creation of intelligent healthcare information retrieval systems. By automatically categorizing and indexing report components, these algorithms enable efficient retrieval of relevant information from vast medical datasets. This capability is especially important for research purposes, quality assurance, and decision support systems, where quick access to specific information within medical reports is crucial. In summary, component block listing algorithms play a pivotal role in template matching for medical reports by efficiently identifying and extracting meaningful components, promoting consistency in data analysis, and supporting the development of advanced information retrieval systems in the healthcare domain.

C. Quality Aware Template Matching

Quality Aware Template Matching (QATM) is a sophisticated approach to template matching in medical reports that focuses on not only identifying templates but also assessing the quality or accuracy of the matched templates. In medical documentation, the accuracy and completeness of information are paramount, and QATM addresses this Taking into account the caliber of the match additionally to the structural alignment with a predefined template. This approach helps in filtering out false positives and ensuring that the matched templates are present, but they also trustworthy, which is particularly crucial in healthcare settings.

One key aspect of QATM in medical template matching involves incorporating quality metrics for individual components within the template. Rather than treating the entire document uniformly, the algorithm assesses the standard of each section or block within the medical report. This fine-grained evaluation allows for a more nuanced understanding of The accuracy of the information contained in different parts of the report, ensuring that critical elements such as diagnoses or treatment plans are accurately identified and prioritized. QATM, therefore, contributes to the overall data integrity and reliability of template matching in medical reports.

Additionally, QATM can be instrumental in handling variations in writing styles or document formats commonly found in medical reports. By taking the match's quality into account, the algorithm can adapt to different linguistic expressions or variations in report structures without sacrificing accuracy. This adaptability is crucial in the healthcare domain, where diverse medical specialties may have unique reporting conventions. Overall, Quality Aware Template Matching brings a nuanced and quality-focused perspective to template matching in medical reports, enhancing the reliability and applicability of the results in clinical decision-making, research, and healthcare administration.

D. Deformable Diversity Similarity

Deformable Diversity Similarity (DDS) represents an innovative approach to template matching in medical reports, emphasizing the adaptability of the matching process to accommodate deformations or variations in the arrangement of documents. In relation to medical reporting, documents may deviate from a standard template due to differences in reporting practices, diverse medical specialties, or evolving documentation standards. DDS addresses these challenges by incorporating a deformable matching a plan that permits flexibility in aligning and comparing templates with varying structures. This adaptability is crucial in the dynamic landscape of healthcare documentation, where the content and formatting of medical reports can exhibit significant variability.

One notable feature of DDS in medical template matching is its consideration of diversity in the matching process. Rather than relying solely on rigid structural alignment, DDS introduces a diversity component that accounts for variations in the representation of information within medical reports. This enables the algorithm to recognize and adapt to different ways of expressing the same information, enhancing its ability to accurately match templates across a diverse range of reports. In the medical field, where standardized reporting practices may not be universally adopted, DDS provides a robust solution for template matching that can accommodate the inherent diversity in document structures and contents.

Moreover, DDS contributes to the robustness and generalizability of template matching in medical reports. The deformable nature of the algorithm allows it to handle not only variations in the organization of information but also transformations such as scaling or skewing. This is especially helpful when handling medical images or other types of reports where the spatial arrangement of elements may differ. By incorporating deformable matching and diversity considerations, DDS stands out as a versatile and effective tool for template matching in the complex and evolving landscape of medical documentation.

IV. Literature Survey

In recent years, template extraction and matching have emerged as critical techniques in the domain of document analysis and image recognition. The foundational works, such as those presented in [1], [2], [14], and [24], have laid the groundwork for template matching techniques, addressing challenges in document image recognition and spatial template extraction. These studies form the

cornerstone for exploring innovative applications of template-based approaches in diverse domains.

Researchers have explored template extraction approaches for image recognition, as evidenced by papers like [4], [6], and [12]. These works offer valuable insights into techniques that can be adapted for extracting information from medical images. The adaptability of these methods to heterogeneous data sources is a recurring theme, as seen in [10], [11], [13], [20], and [28]. This adaptability is particularly relevant for medical reports, which often vary in structure and content.

In the realm of natural language generation, [9] and [13] delve into unsupervised template extraction for generating natural language. These studies provide potential avenues for summarizing medical reports and making them more accessible. Additionally, incorporation of template matching with deep learning, as explored in [16] and [17], indicates the possibility of advanced machine learning methods to enhance template-based approaches in the medical document analysis domain.

Quality-aware template matching and deformable diversity similarity are introduced in [16], [17], and [22]. These methodologies present promising advancements in achieving more accurate template matching, a crucial aspect for extracting relevant information from medical documents. The importance of fast and high-performance template matching is underscored by [25] and [30], offering efficiency improvements that could prove essential for processing large-scale medical datasets.

Beyond document analysis, the literature includes studies that touch upon biomedical literature and chemical data extraction ([7], [30]), suggesting applications for processing scientific information within medical reports. Logo detection and recognition ([21], [22], [29]) showcase the adaptability of template-based approaches in identifying specific patterns within medical documents, providing a glimpse into potential use cases for structured data extraction.

In conclusion, the literature survey reveals a rich tapestry of research exploring template extraction and matching techniques across diverse domains. The adaptability of these approaches to heterogeneous data, integration with deep learning, and advancements in accuracy and efficiency collectively contribute to the ongoing exploration of template-based methods in medical data analysis and information extraction. As the field continues to evolve, these insights serve as a valuable foundation for future research endeavors in relation to medical reports and documents.

V. Conclusion

To sum up, this research study has explored a robust and multifaceted approach to template extraction and matching for medical documents, employing a suite of methodologies including Optical Character Recognition (OCR), Document Image Matching, Clustering, Component Block Listing, and Quality Aware Template Matching. The integration of OCR proved crucial for converting scanned medical documents into machine-readable text, forming the foundational step for subsequent template matching processes. Document Image Matching techniques facilitated the alignment and comparison of documents, allowing for the identification of common templates across diverse formats and layouts. The incorporation of clustering methodologies organized extracted information into meaningful categories, contributing to a more systematic and insightful analysis of medical data.

In summary, the amalgamation of OCR, Document Image Matching, Clustering, Component Block Listing, and Quality Aware Template Matching has demonstrated promising results, showcasing the potential for practical implementation in real-world healthcare scenarios. This research contributes to the evolving landscape of medical document analysis, providing a comprehensive framework that leverages the strengths of each methodology. The findings offer insights into improved medical information extraction and template matching, with implications for streamlined healthcare workflows and more informed decision-making processes. Future research may explore additional refinements and potential integration of machine learning algorithms to further enhance pattern recognition and adaptability to evolving medical document structures.

References

[1] Dr S. Vijayarani and Ms. A. Shakila, “Template Matching Technique For Searching Words In Document Images”, International Journal on Cybernetics & Informatics (IJCI), 2015.

[2] Hanchuan Peng; Fuhui Long; Zheru Chi, “Document Image Recognition based on template matching of component block projections” IEEE, 2003.

[3] Jun-Wei Hsieh, W. Eric L. Grimson, “Spatial template extraction for image retrieval by region matching”, ScienceDirect.com, 2003

[4] Gang Wang, Hui-chuan Duan, “A Template Extraction Approach For Image Recognition”, IEEE, 2012.

[5] Fanman Meng, Bing Luo, Chao Huang, Liangzhi Tang, Bing Zeng, Nini Rao, “Favorite object extraction using web images”, IEEE, 2014

[6] Shankar Shivappa, Patrick Nguyen and Geoffrey Zweig, “Discriminative Template Extraction For Direct Modeling”, Microsoft Research, 2014.

[7] Jan P. Unsleber, “Accelerating Reaction Network Explorations with Automated Reaction Template Extraction and Application”, American Chemical Society(acs.org), 2023.

[8] Aicha Ghoulam, Fatiha Barigou, Ghalem Belalem, “Information Extraction in the Medical Domain”, Journal of Information Technology Research, 2015.

[9] Daniel Duma, Ewan Klein “Generating Natural Language from Linked Data: Unsupervised template extraction”, School of Informatics, University of Edinburgh (psu.edu), 2015.

[10] Chulyun Kim, Kyuseok Shim, “TEXT: Automatic Template Extraction from Heterogeneous Web Pages” IEEE, 2010.

[11] Vinay Aggarwal, Praneetha Vaddamanu, Bhanu Prakash Reddy Guda, Balaji Vasan Srinivasan, Niyati Chhaya, Vishwa Vinay, “Template-Based Information Extraction without the Templates”.

[12] Satoshi Kamegai, Kenji Satou, Akihiko Konagaya, “Automated Template Discovery for Information Extraction from Biomedical Literature”, International Conference on Cybernetics and Information Technologies (fui.edu), 2004.

[13] Daniel Duma, “Natural Language Generation For The Semantic Web: Unsupervised Template Extraction”, 2012.

[14] Hanchuan Peng, Fuhui Long, Zheru Chi, Wan-Chi Siu, “Document image template matching based on component block list”, Sciencedirect.com, 2001.

[15] F. Cesarini, M. Gori, S. Marinai, G. Soda, ” INFORMys: a flexible invoice-like form-reader system”, IEEE,1998.
[16] Jiaxin Cheng, Yue Wu, Wael Abd-Almageed, Premkumar Natarajan, “QATM:Quality-Aware Template Matching For Deep Learning”, USC Information Sciences Institute, 2019.
[17] Itamar Talmi, Roey Mechrez, Lihi Zelnik-Manor, “Template Matching with Deformable Diversity Similarity”, USC Information Sciences Institute, 2019.
[18] Amit Adam, Ehud Rivlin, and Ilan Shimshoni, “Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition”, IEEE, 2006.
[19] Alireza Alaei and Mathieu Delalandre. A complete logo detection/recognition system for document images. In Proceedings of International Workshop on Document Analysis Systems, IEEE,2014.
[20] Min Bai, Wenjie Luo, Kaustav Kundu, and Raquel Urtasun. Exploiting semantic information and deep matching for optical flow. In European Conference on Computer Vision. Springer,2016.
[21] Raluca Boia, Corneliu Florea, Laura Florea, and Radu Dogaru. Logo localization and recognition in natural images using homographic class graphs. Machine Vision and Applications,2016.
[22] Tali Dekel, Shaul Oron, Michael Rubinstein, Shai Avidan, and William T Freeman. Best-buddies similarity for robust template matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015.
[23] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016.
[24] J.-H. Chen, C.-S. Chen, and Y.-S. Chen. Fast algorithm for robust template matching with m-estimators. IEEE Transactions on Signal Processin,2003.
[25] P.F. Felzenszwalb, R. B. Girshick, D. Mc Allester , and D. Ramanan. Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010.
[26] Y. Hel-Or, H. Hel-Or, and E. David. Matching by tone mapping : Photometric in variant template matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.
[27] Michael Ryan and Novita Hanafiah. An examination of character recognition on id card using template matching approach. ProcediaComputerScience,2015
[28] Denis Simakov, Yaron Caspi, Eli Shechtman, and Michal Irani. Summarizing visual data using bidirectional similarity. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,IEEE,2008.
[29] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Buster net: Detecting copy-move image forgery with source/target localization. In European Conference on Computer Vision, 2018.
[30] A. Sibiryakov. Fast and high-performance template matching method. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2011.