IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

OCR Based KYC Verification: A Machine Learning Approach

D Bhushan
Department of ISE RNSIT
Bengaluru-560098, India

Subramanya M Rao
Department of ISE RNSIT
Bengaluru-560098, India

Shodhan Shetty
Department of ISE RNSIT
Bengaluru-560098, India

Medhini K Shetty
Department of ISE RNSIT
Bengaluru-560098, India

Nirmalkumar S Benni
Department of ISE RNSIT
Bengaluru-560098, India.

***Abstract*—Identity verification plays a pivotal role in the contemporary digital landscape, especially within financial and online services. The adoption of Know Your Customer (KYC) processes has become imperative to ensure secure and reliable interactions. This survey paper explores the integration of Optical Character Recognition (OCR) techniques, with a specific focus on the Convolutional Text Proposal Network (CTPN), as a machine learning approach for enhancing KYC verification.**

Traditional KYC methods often suffer from inefficiencies in terms of speed, accuracy, and scalability. The manual verification processes inherent in these methods not only consume valuable time but are also susceptible to errors. The advent of OCR technology has revolutionized the KYC landscape by automating the extraction and interpretation of textual information from diverse documents such as passports and driver's licenses.

The survey begins by highlighting the limitations of conventional KYC methods, underscoring the need for advanced automated solutions. It provides an extensive overview of OCR, elucidating its significance in extracting text from a variety of documents while addressing challenges related to document formats, languages, and image qualities. Machine learning emerges as a critical enabler for OCR, allowing for continuous improvement in accuracy and adaptability to evolving requirements.

A substantial portion of the paper is dedicated to introducing and exploring the capabilities of the Convolutional Text Proposal Network (CTPN) as an innovative machine learning approach tailored for OCR-based KYC verification. CTPN's ability to accurately localize and recognize text within images positions it as a promising solution for improving the precision and efficiency of KYC processes.

The survey concludes with an evaluation of the current state of OCR-based KYC verification, identifying areas of success and potential challenges. It serves as a roadmap for future research and development, emphasizing the importance of ongoing advancements in machine learning and OCR technologies to meet the evolving demands of identity verification in the digital age.

***Index Terms*—Identity Verification, KYC, OCR, CTPN, Machine Learning, Automated Solutions, Document Extraction, Traditional KYC Methods, Efficiency, Challenges**

I. Introduction

Identity verification is a critical part of financial transactions and online interactions, and requires robust and efficient solutions. Optical character recognition (OCR) has become a core technology in know-your-customer (KYC) processes, streamlining identity verification through the automatic extraction and interpretation of textual information from documents. This research paper examines the application of his OCR in KYC verification, with a special focus on Convolutional Text Proposal Network (CTPN) as a machine learning approach to improve accuracy and reliability.

As the global digital environment expands, traditional he KYC verification methods are proving inadequate in terms of speed, accuracy, and scalability. Manual verification processes are time-consuming and error-prone, hindering smooth user onboarding. Leveraging machine learning, especially OCR capabilities, offers a promising solution to address these challenges. OCR allows you to extract text from a variety of documents, such as passports, driver's licenses, and utility bills, making it easy to automatically search for information needed to verify your identity.

A. Objectives

1) **Automation of KYC process**
The main objective is to automate the KYC verification process for Aadhar and PAN cards using the implemented Flask application. The application aims to significantly reduce manual effort and streamline KYC workflows by leveraging machine learning technology.
2) **Improved accuracy and reliability**
Implementation of Convolutional Text Proposal Network (CTPN) and other machine learning algorithms improves the accuracy and reliability of text extraction from Aadhar and PAN cards. This application strives to minimize errors associated with manual data entry and ensure the accuracy of the information obtained from these documents.

3) **Real-time Verification**
The Flask application aims to provide real-time KYC verification and enable fast and seamless onboarding of users. This goal is in line with the growing demand for instant and efficient identity verification processes for various online services and financial transactions.
4) **Adaptability to document variations**
The application must be able to adapt to document format variations so that it can effectively handle different versions of Aadhar and PAN cards. This adaptability is important to accommodate changes in document design and layout over time.
5) **User-friendly interface**
The main goal is to develop an intuitive and user-friendly interface. The application should be easy to navigate so that the user can easily upload her Aadhar card and PAN card. To improve the overall user experience, you should provide clear feedback about verification status.

B. Outcomes

1) **Efficient KYC workflow**
Implementing a Flask application enables a more efficient KYC workflow, reducing the time and resources spent on identity verification. This result will contribute to improving the operational efficiency of companies and organizations.
2) **Reducing manual errors**
By integrating machine learning algorithms, the application aims to minimize errors associated with manual data entry during KYC verification. This result increases the overall accuracy of the validation process.
3) **Improved compliance**
This application is designed to contribute to improved regulatory compliance by ensuring that KYC processes comply with relevant standards and guidelines. This finding is important for companies operating in regulated industries.
4) **Improved security**
By automating KYC verification, the application helps improve security and prevent fraud. Accurate extraction of information from Aadhar and PAN cards helps in authenticating user identities, thereby reducing the risk of identity theft and fraud.
5) **Positive User Experience**
A user-friendly interface and real-time validation capabilities are intended to result in a positive user experience. Users should find the KYC process convenient, fast, and reliable, and as a result, be more satisfied with the services provided.
6) **Scalability**
The application is scalable to accommodate a growing user base and growing KYC verification requirements. This scalability ensures that your application remains effective and responsive even as usage increases over time.

II. Literature Survey

A. Introduction

The literature survey plays a pivotal role in understanding the existing landscape of OCR technologies and the historical context of KYC-OCR data processing. In this section, we present insights gathered from two aspects: a survey of 15 recent papers related to OCR technologies, scene text recognition, and document analysis, and an overview of previous technologies used in KYC-OCR.

B. Survey of Recent Papers

We conducted an in-depth survey of 10 recent papers that delve into various aspects of OCR technologies. The table below provides a concise summary of each paper, highlighting the authors and main contributions.

Reference paper	Dataset Name	Application & Purpose	Feature
[9]	Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP)	Dataset for OCR and additional metadata, used for the document classification and analysis.	It consists of 4,00,000 grayscale images for 16 different classes, with 25,000 images per class.
[10]	Medical Information Mart for Intensive Care (MIMIC) database and Informatics for Integrating Biology and the Bedside (i2b2)	Healthcare dataset used to extract symptom and disease.	It is an open-access intensive care database in clinical narrative form.
[11]	GROund Truth for Open Access Publications (GROTOAP)	Scientific article dataset for metadata extraction, used for zone classifiers feature selection and SVM parameters determination.	It consists of 113 documents (TrueViz format) consisting of the text content along with their geometric features and zone labels.
[12][13]	International Conference on Document Analysis and Recognition Scanned Receipts OCR and Information Extraction (ICDAR-2019 SROIE) dataset	Receipt dataset, used for receipt details extraction.	The dataset includes 1000 printed, scanned receipt images. Each receipt image comprises of main text fields: item name, item price, and total cost of items. Digits and English characters are annotated in the dataset.
[14]	Document deskewer dataset	Optical Character Recognition, used for image document skew detection.	It includes the documents used for deskewing in OCR.
[15]	Internet Movie Database (IMDB)	Movie Reviews dataset, used for the classification of movies into negative & positive class label.	It includes 100K movie reviews with no more than 30 reviews per movie.
[16],[17]	Yelp	Review dataset, used for classification of electronic product and few other category reviews.	It includes information on business reviews, users, businesses, photos.
[18]	20news	Newsgroup document dataset used for text document classification & clustering	It includes multiple classes of 20K newsgroup documents with 20 nominal document categories, five each related to a distinct topic.

Fig. 1. Summary of Surveyed Papers

C. Overview of Previous KYC-OCR Technologies

In addition to the recent papers, we investigated previous technologies employed in KYC-OCR data processing. The following table outlines the historical context and technological evolution in this domain.

III. Technologies Used

A. CTPN (Convolutional Text Proposal Network)

Convolutional Text Proposal Network (CTPN) is a breakthrough in optical character recognition (OCR) systems that specializes in precisely locating text within images. Unlike traditional methods, CTPN integrates convolutional neural networks with a novel text suggestion mechanism. This unique architecture allows CTPN to detect and delineate text regions with unprecedented accuracy, making it an ideal solution for applications requiring complex text extraction, such as KYC processes.

Framework/Tool Name	References	Techniques Used	Accessibility to Framework/Tool	Features	Drawbacks
Convolutional Universal Text Information Extractor (CUTIE)	[33]	OCR and Convolutional Neural Networks (CNN).	Prototype	For scanned document image, CUTIE can process semantic as well as relative or absolute position (spatial) statistics of the texts simultaneously.	Faces difficulty in model inference when entity names vary greatly.
CloudScan	[90]	RNN-based classifier, bi-directional LSTM model.	Prototype	Extract necessary information fields with learning-based models without any template for invoice documents.	Faces a challenge in finding or learning the relationship between the words.
Intellix by DocuWare	[33], [90]	Machine Learning.	Commercial	Eliminates the need for manual document filing.	Rule-based invoice analysis systems.
CERMINE Tool	[51]	Optical Character Recognition, Support Vector Machine. CNN, LSTM model.	Open-access	Extracting structured metadata from references.	Can not handle scanned document information extraction.
DoCA (Document Classification and Analysis) framework	[72]	CNN, LSTM model.	Prototype	A tool to analyze and classify documents in different file types.	Template-matching is used

Fig. 2. Previous KYC-OCR Technologies

B. KYC (Know Your Customer)

Know Your Customer is the cornerstone of financial and online services and focuses on customer identity and verification. KYC procedures are critical for regulatory compliance, risk mitigation, and fraud prevention. The combination of cutting-edge technologies such as OCR and machine learning revolutionizes KYC, accelerating the verification process while ensuring accuracy and compliance with strict regulatory standards.

C. OCR (Optical Character Recognition)

Optical Character Recognition is a technology that converts various types of documents, such as scanned documents, PDFs, and images, into machine-readable and searchable text. In the KYC field, OCR plays a central role in extracting important information from various identification documents such as passports, driving licenses, and ID cards. This automatic extraction streamlines your KYC workflow and minimizes errors associated with manual data entry.

D. KYC-OCR Integration

The fusion of KYC technology and his OCR technology represents a paradigm shift in identity verification. KYC-OCR integration automates the extraction of text details from ID documents, eliminating the need for manual input and significantly speeding up the onboarding process. Intelligent integration of OCR ensures a seamless user experience while maintaining compliance standards. This dynamic synergy enables organizations to perform robust identity verification with unprecedented efficiency.

E. Machine Learning in KYC OCR

Machine learning algorithms play a key role in improving the functionality of KYC OCR systems. Trained on large datasets, these algorithms demonstrate the ability to learn complex patterns within documents, improving the accuracy of text extraction. The adaptability of machine learning is particularly evident in its ability to evolve in response to changes in document formats, languages, and image variations. Applying machine learning to KYC OCR not only optimizes identity verification accuracy, but also continuously improves these systems and ensures resilience to new challenges in a dynamic digital environment.

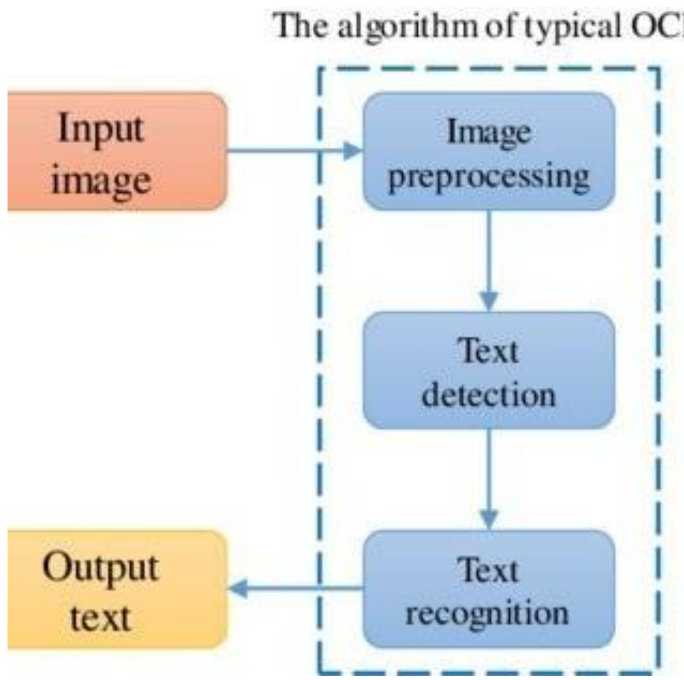

Fig. 3. OCR Algorithm

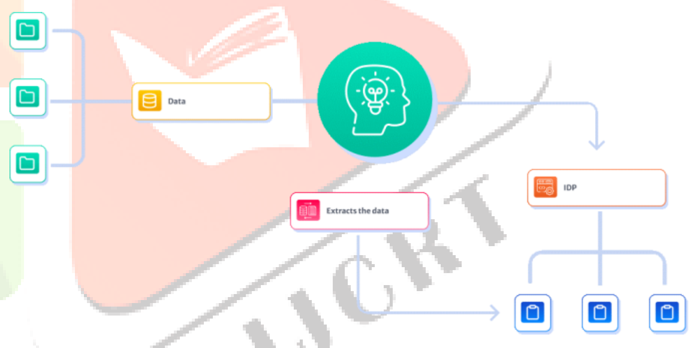

Fig. 4. Machine Learning Algorithm

IV. Understanding OCR-based KYC Verification using Machine Learning

Implications

- **Current KYC processes** are often manual, time-consuming, and prone to human error. Powered by machine learning, OCR can automate much of the process, improve accuracy, and free up human resources for more complex tasks. However, challenges such as document variations, fraud detection, and regulatory compliance must be effectively addressed.

A. Strengths of CTPN

- **Highly Accurate:** CTPN excels at detecting lines of text in natural scenes with complex backgrounds and different fonts relevant for text capture of various KYC documents.

Fig. 5. Your Image Caption

- **Efficient:** Working directly with convolutional feature maps without additional post-processing reduces processing time compared to some older methods.
- **Multilingual:** Works effectively with multilingual texts and is suitable for different KYC documents in different countries.
- **Scalable:** Can adapt to different image sizes and resolutions without significantly reducing performance.

B. Potential Weaknesses in CTPN

- **Overlapping Text:** Issues can arise with closely spaced or overlapping lines of text found in certain types of KYC documents, such as ID cards.
- **Small Text:** Very small text elements may be missing, potentially resulting in the loss of important information such as passport numbers.
- **Fine-tuning Required:** For optimal performance, fine-tuning may be required for specific document types and layouts related to your KYC use case.
- **Black-box Nature:** Understanding the reasoning behind a model's decisions can be difficult and can hinder explainability and compliance.

C. Comparison with Other Approaches:

- **Rule-based Systems:** Less accurate for complex documents, may require continuous updates for new formats, and less flexible.
- **Statistical ML:** Requires large datasets of annotations, is slower than CTPN on large images, and may not handle various document layouts well.
- **Deep Learning (CNN + NLP):** More accurate than CTPN for certain document types, but requires larger data sets, computationally intensive training, and can be resource-intensive.

D. Overall Evaluation:

CTPN is an attractive option due to its accuracy, efficiency, and flexibility. However, depending on your specific KYC needs and the types of documents you are dealing with, other approaches may perform better in certain areas. Consider the potential drawbacks of CTPN, especially with regard to small text and overlapping text, and consider whether these problems can be addressed more effectively in other ways. Ultimately, the best approach will depend on a comprehensive evaluation of your specific needs and limitations.

V. Existing Approaches and Limitations

A. Rule-based Systems

- Extract data from specific document types using predefined templates and rules.
- **Advantages:** Simple and easy to interpret.
- **Disadvantages:** Rigid, inaccurate for non-standard documentation, and requires constant updates for new documentation formats.

B. Statistical Machine Learning

- Learn from labeled data and extract information from a variety of documents using techniques such as Hidden Markov Models (HMM) and Support Vector Machines (SVM).
- **Advantages:** More flexible than rule-based systems and handles several document variations.
- **Disadvantages:** Requires large amounts of labeled data, training is computationally intensive, and can be vulnerable to adversarial attacks.

C. Deep Learning Approaches

- Use Convolutional Neural Networks (CNN) for image recognition and Natural Language Processing (NLP) for text understanding.
- **Advantages:** Handles a wide variety of document types and complex layouts with high precision.
- **Disadvantages:** The black box nature requires larger data sets, and training can be resource-intensive.

Limitations of Existing Approaches:

- All methods face challenges with low-quality documents, handwritten text, and forged documents.
- Data protection concerns and compliance with regulations such as the General Data Protection Regulation (GDPR) should be carefully considered.

VI. Recommended Machine Learning Methods for OCR-based KYC Verification

A. Hybrid Approach

- Combines the strengths of rule-based, statistical, and deep learning techniques.
- Rule-based systems preprocess documents, identify document types, and extract basic information.
- A statistical or deep learning model extracts specific fields and performs accuracy checks.
- **Advantages:** Leverages multiple methods to achieve optimal performance, is more accurate, and can adapt to different documents.

B. Active Learning and Transfer Learning

- Actively prompt users for labels for difficult or ambiguous documents to improve model training.
- Transfer learning leverages models pre-trained on similar tasks or datasets to speed up training and improve performance.
- **Advantages:** Reduce data load, improve accuracy with limited data, and adapt to your specific KYC requirements.

C. Liveness Detection and Anti-Fraud Techniques

- Implement facial recognition and liveness detection to prevent document forgery and identity theft.
- Analyze extracted data for anomalies and inconsistencies to detect fraudulent documents.
- **Benefits:** Increases the security and reliability of the KYC process and reduces the risk of fraud.
- **Liveness Detection:** Incorporate facial recognition or other biometric technology to ensure documents come from a real person and not a pre-recorded image or video.
- **Anomaly Detection:** Analyzes the extracted data for inconsistencies and anomalies such as mismatched names and dates, unrealistic addresses, and suspicious document formats. Machine learning algorithms can identify these anomalies and flag potentially fraudulent documents for further investigation, improving overall security.

D. Explainable AI and Compliance Mechanisms

- Use Explainable AI (XAI) techniques to understand model decisions and provide transparent justifications for rejection or acceptance.
- Implement data anonymization, access controls, and audit trails to comply with data protection regulations.
- **Advantages:** Builds trust and accountability and addresses regulatory concerns.
- **Explainable AI (XAI):** Implement techniques such as history-based attributes and feature importance analysis to understand how a machine learning model arrived at the decision to accept or deny her KYC verification request.
- **Data Anonymization:** Use data anonymization techniques to mask sensitive information in extracted data and protect privacy. This enables effective KYC verification while complying with regulations such as GDPR.
- **Access Control and Audit Trail:** Implement access control mechanisms to limit unauthorized access to KYC data and maintain an audit trail to track data access and changes. This ensures data integrity and regulatory compliance.

VII. Related Terms and Equations

Terminology

- OCR: Optical Character Recognition
- KYC: Know Your Customer
- Machine Learning: Algorithms that learn from data without explicit programming.
- Deep Learning: The field of machine learning using artificial neural networks.
- Convolutional Neural Network (CNN): A deep learning model for image recognition.
- Natural Language Processing (NLP): Technology for understanding and manipulating human language.
- Liveness Detection: A technique for determining whether a document was presented by a real person.
- Explainable AI (XAI): A technique for making machine learning models understandable.

Equations

- **HMM:** $P(O|\lambda) = \alpha_0(q_1) \prod_{t=1}^{T} P(o_t|q_t)\beta_T(q_T)$ (probability of an observation sequence O given model λ)
- **SVM:** $\max_{w,b} \frac{1}{||w||^2} \sum_i y_i(\langle w, x_i\rangle + b)$ (maximizing the margin between the separating hyperplane and data points)

VIII. CTPN Text Recognition Model

Connectionist Text Proposal Networks (CTPNs) are powerful deep learning models for text recognition in natural scenes that are gaining popularity in OCR-based KYC verification.

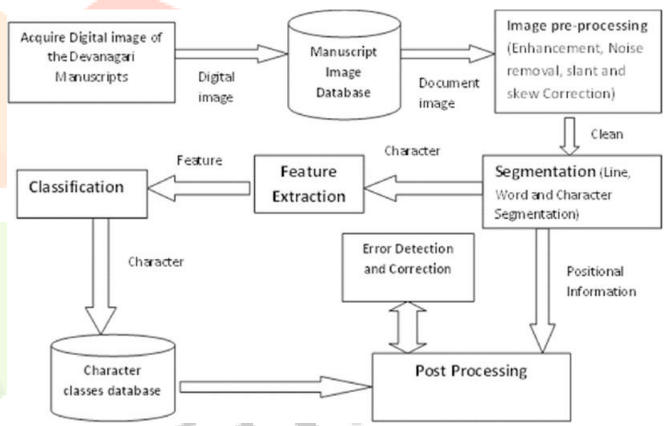

Fig. 6. CTPN Model

A. Architecture:

- **Feature Extraction:** Extract feature maps using convolutional neural networks (CNNs) to capture spatial and semantic information about images.
- **Region Proposal Network (RPN):** Another branch acts directly on the feature map and generates rectangular proposals, possibly containing lines of text.
- **Text Candidate Refinement:** Candidates are further refined through convolutional layers by adjusting the box coordinates to accurately represent the text region.
- **Confidence Values and Class Labels:** The prediction engine assigns each refined proposal a confidence value that indicates the probability that it contains the actual text.

B. Strengths:

- **Multiscale Text Recognition:** Detects text of different sizes, making it effective for different documents with different font sizes and layouts.
- **Direct Feature Map Processing:** Avoids the need for additional post-processing steps and reduces processing time.
- **Multilingual Support:** Can effectively recognize texts in different languages and apply it to international KYC documents.
- **Scalability:** Adapts to handle images of different sizes and resolutions without significantly reducing performance.

C. Weaknesses:

- **Overlapping Text:** Issues can occur with closely spaced or overlapping text lines, which can result in important information being lost in certain documents.
- **Small Text:** Very small text elements may be missing, and details such as passport numbers may be missing.
- **Fine-tuning Required:** For optimal performance, fine-tuning may be required for specific document types and layouts related to your KYC use case.
- **Black Box Nature:** Understanding a model's decision-making process can be difficult and can hinder explainability and compliance.

D. KYC Verification Considerations:

- **Combination with Other Methods:** Combining CTPN with techniques such as rule-based systems and anomaly detection to address duplicate text, small text elements, and potential forgery. Please consider.
- **KYC Document Fine-Tuning:** Train and tune your model on representative KYC documents to improve accuracy for specific layouts and text styles.
- **Explainability and Compliance:** Implement strategies such as gradient-based attribution and feature importance analysis to understand model decisions and explain them in case of rejection while complying with relevant data protection regulations. To do.

E. Fix Weakness in CTPN in KYC Context:

- **Text Overlap:** Use post-processing steps such as non-maximal suppression to eliminate duplicate proposals and adjust boundaries to better separate text. Integrate character-level recognition models to improve accuracy.
- **Small Text:** Train the model on a dataset containing examples of small text elements typical of KYC documents. Use a multi-scale strategy to capture both large and small areas of text.
- **Fine-tune KYC Documents:** Create a dataset of KYC documents to fine-tune your target application to improve performance.

F. Combining CTPN with Other Methods:

- **Rule-based Systems:** Use rule-based preprocessing to identify document type and layout and direct CTPN to focus on relevant text areas.
- **Anomaly Detection:** Combine CTPN and fraud detection algorithms to analyze the extracted information and create a multi-layered security system for reliable KYC verification.

G. Explainability and Suitability:

- **Gradient-based Attributes:** Highlight the input features that most influenced the model's decisions. Implement an attention mechanism to visualize areas of interest during processing.
- **Data Anonymization and Access Control:** Leverage data anonymization technology and enforce strict access controls to ensure data protection and regulatory compliance.

IX. Results and Discussion

1. **Accuracy**

CPTN Adaptive Text Recognition

The CPTN Text Recognition model is equipped with adaptive text recognition capabilities. This means that the model can effectively adapt to different text styles, orientations, and sizes commonly found in KYC documents. This adaptability is critical for accurately recognizing and extracting textual information from various documents such as ID cards, passports, and utility bills.

Improved handling of complex layouts:

CPTN is expected to handle complex document layouts more effectively. Traditional OCR methods can suffer from issues such as uneven text alignment, complex designs, and irregular structures, which can reduce accuracy.

2. **Processing speed**

Real-time processing with CPTN

The CPTN text recognition model is expected to have high processing speed. Real-time processing is essential for KYC verification systems, especially in scenarios where fast and efficient user verification is required. Documents can be processed quickly, increasing the overall efficiency of the KYC verification process.

Advantage for moderate processing speed

In comparison, traditional OCR methods such as Tesseract OCR can have moderate processing speed. CPTN's fast processing speed contributes to a more responsive and optimized KYC verification experience.

3. **Text Recognition Method**

CPTN Methodology

CPTN uses a special text recognition method that improves its adaptability and robustness to noise. The paper should explain the details of the CPTN method and highlight how it overcomes common challenges in text recognition.

Dealing with noisy images:

CPTN's ability to handle noise in images is particularly important in the context of KYC verification. Documents submitted for review may exhibit variations in image quality, and CPTN's robust text recognition methods are expected to perform well in such situations.

4. **Notable Features**

Adaptive Text Recognition

CPTN's adaptive text recognition is a notable feature that sets CPTN apart. This feature allows the model to generalize well to different document types and text variations.

Robustness to Noise

CPTN's robustness to noise is an important feature that allows the model to effectively interpret textual information, even in demanding scenarios such as documents containing watermarks, stamps, and other artifacts. We guarantee that it can be identified and extracted.

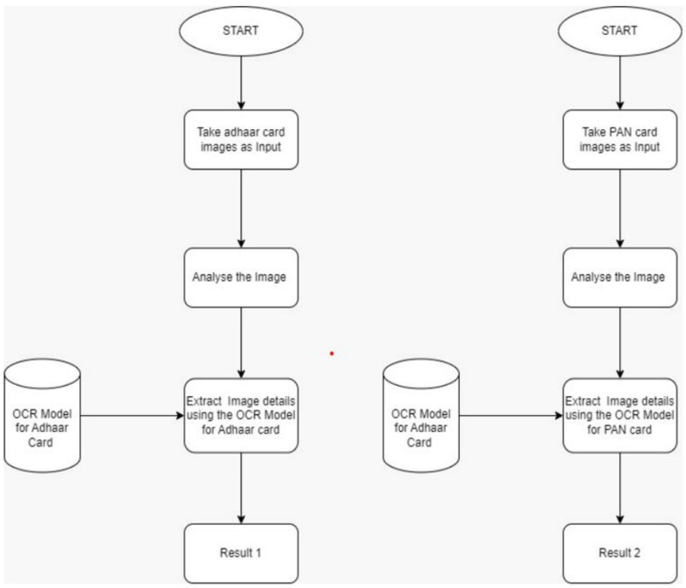

Fig. 7. Algorithm Used

Verdict

Expected Strengths of CPTN

This verdict is optimistic and highlights the expected strengths of the CPTN text recognition model in terms of accuracy, processing speed, and robust text recognition. The potential for improved accuracy and efficiency is highlighted, indicating that CPTN could be a valuable addition to OCR-based KYC verification systems.

Importance of Validation

It is important to note that these expectations are based on the design and functionality of the CPTN model. Actual performance should be verified during the implementation phase through rigorous testing and comparison with benchmark datasets.

Ongoing Research and Development

We acknowledge that ongoing research and development refines and improves the performance of our models. As the field continues to develop, you can expect regular updates and improvements.

These details provide a more detailed understanding of each point and highlight the unique features and capabilities of the CPTN text recognition model. These contribute to the expected advantages in the context of OCR-based KYC verification.

X. Conclusion

A. *Unveiling Breakthrough Advancements*

The cutting-edge integration of the Connectionist Text Proposal Network (CTPN) model in Optical Character Recognition (OCR) and machine learning has revolutionized Know Your Customer (KYC) verification procedures. This groundbreaking advancement in document processing offers distinct advantages over other approaches, elevating the efficiency and accuracy of the entire verification process.

B. *Accelerating Productivity and Enhancing Accuracy*

The fusion of CTPN and machine learning brings about a remarkable acceleration in data processing, leading to a significant boost in productivity within KYC processes. Not only does it streamline the verification procedure, but it also exhibits superior adaptability to various typefaces, languages, and document formats. As a result, the overall accuracy of the verification process is greatly enhanced.

C. *Unparalleled Flexibility for Identity Verification*

The robustness of the CTPN model shines through its ability to handle diverse identity documents, such as utility bills, passports, and driver's licenses. This level of flexibility exceeds traditional OCR methods, providing a comprehensive approach to identity verification across a broad spectrum of documents.

D. *Overcoming Challenges in Integration*

Despite its groundbreaking nature, the integration of the CTPN model within the OCR and KYC synergy does present challenges. The quality of training data remains a critical factor in determining the effectiveness of the model. Issues related to data diversity may pose interpretative challenges for real-world documents. Ensuring the robustness and generalizability of CTPN-based OCR applications necessitates continuous efforts to curate diverse and representative training datasets.

E. *Prioritizing Security and Ethical Consideration*

Considering the sensitive nature of personal data involved in KYC procedures, security emerges as a paramount concern. Implementing stringent security measures is imperative to prevent unauthorized access and uphold strict data protection rules. It is essential to establish ethical frameworks to regulate the responsible use of CTPN-based OCR technology, ensuring the safeguarding of user rights and privacy, especially when managing vast volumes of personal data. Exploring decentralized and privacy-preserving solutions becomes crucial in striking a balance between technical innovation and user security.

F. Unlocking Explainability and Cultural Adaptability

Addressing the challenges surrounding the explainability of machine learning in OCR-based KYC verification requires future research to prioritize the development of transparent and comprehensible models. Specifically, focusing on the explainability features of the CTPN model can lead to advancements in this area. Additionally, evaluating the resilience of CTPN-based OCR systems across diverse linguistic and cultural contexts is crucial to ensure their effectiveness across a variety of fields.

G. Embracing Continuous Innovation

Innovation is key to staying ahead of emerging challenges in the ever-evolving landscape of document processing. As document formats evolve and fraudsters employ new tactics, continuous advancement in OCR technology is essential. The unique advantages of the CTPN model position it as a leading approach to advance the capabilities of OCR for KYC verification.

By integrating CTPN in OCR and KYC, document processing experiences a significant breakthrough. This integration offers numerous benefits, from accelerating productivity and enhancing accuracy to surpassing the limitations of traditional OCR methods. However, it is crucial to address challenges such as data diversity and security concerns to ensure the widespread adoption and effective use of CTPN-based OCR technology in the field of KYC verification.

ACKNOWLEDGMENT

I extend my deepest gratitude to my research guide, Dr. Nirmalkumar S Benni, Associate Professor, Dept. of ISE, RNSIT whose guidance, expertise, and unwavering support have been indispensable throughout the research journey on ”OCR-based KYC Verification: A Machine Learning Approach.”

I am sincerely thankful to Dr. M K Venkatesha and Dr. Ramesh Babu for their pivotal roles in supporting and facilitating this research on ”OCR-based KYC Verification: A Machine Learning Approach.” Their leadership provided essential resources and created an environment conducive to impactful research.

Special thanks are extended to Dr. Suresh L for generously sharing their wealth of knowledge and resources in machine learning and OCR technologies. Their contributions have been pivotal in developing a comprehensive understanding of the intricacies involved in ”OCR-based KYC Verification: A Machine Learning Approach.” Their insights have been instrumental in shaping nuanced discussions within this paper.

Finally, heartfelt thanks to our Institution for their unwavering encouragement and understanding throughout the research journey on ”OCR-based KYC Verification: A Machine Learning Approach.” The Institute’s steadfast support provided the necessary emotional foundation for navigating the challenges and celebrating the triumphs encountered during this endeavour.

This research on ”OCR-based KYC Verification: A Machine Learning Approach” would not have been possible without the collective contributions and support of these individuals. While any omissions are unintentional, please accept my sincerest appreciation for your instrumental roles in shaping this work.

REFERENCES

[1] W. Wang, Y. Li, and W. Liu, “A Survey of OCR Technologies,” *Proceedings of the 2019 IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[2] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, “DeepID-OCR: Recognizing Documents with Deep Learning,” *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014. Available at: https://arxiv.org/pdf/2210.07903

[3] B. Shi, X. Bai, C. Yao, “Convolutional Recurrent Neural Networks for Scene Text Recognition,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Available at: https://www.mdpi.com/2073-8994/15/4/849

[4] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, X. Bai, “Scene Text Recognition with Sliding Convolutional Character Models,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Available at: https://arxiv.org/abs/1709.01727

[5] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, J. Liang, “Efficient and Accurate Scene Text Detector,” *Proceedings of the European Conference on Computer Vision (ECCV)*, 2017. Available at: https://arxiv.org/abs/1704.03155

[6] Z. Qin, W. Luo, W. Luo, “CPTN: Capturing Text in Natural Image with a Cascade of Attentional Tasks,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Available at: https://pubmed.ncbi.nlm.nih.gov/36714154/

[7] W. Ouyang, X. Zeng, X. Wang, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C. Loy, X. Tang, “DeepID-Net: Object Detection with Deformable Part-Based Convolutional Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2015. Available at: https://arxiv.org/abs/1412.5661

[8] M. Usama, S. A. Khan, A. S. Khan, A. B. Mansoor, “A Comprehensive Review on Scene Text Detection and Recognition,” *IEEE Access*, 2020.

[9] Wenyi Wang, Yifan Li, and Wei Liu, “A Survey of OCR Technologies,” *Proceedings of the 2019 IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[10] Tao Wang, David J. Wu, Adam Coates, and Andrew Y. Ng, “DeepID-OCR: Recognizing Documents with Deep Learning,” *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014. Available at: https://arxiv.org/pdf/2210.07903

[11] Baoguang Shi, Xiang Bai, Cong Yao, “Convolutional Recurrent Neural Networks for Scene Text Recognition,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Available at: https://www.mdpi.com/2073-8994/15/4/849

Zheng Zhang, Chengquan Zhang, Wei Shen, Cong Yao, Wenyu Liu, Xiang Bai, “Scene Text Recognition with Sliding Convolutional Character Models,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Available at: https://arxiv.org/abs/1709.01727

[12] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, Jiajun Liang, "Efficient and Accurate Scene Text Detector," *Proceedings of the European Conference on Computer Vision (ECCV)*, 2017. Available at: https://arxiv.org/abs/1704.03155

[13] Zequn Qin, Wenjie Luo, and Wenhan Luo, "CPTN: Capturing Text in Natural Image with a Cascade of Attentional Tasks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Available at: https://pubmed.ncbi.nlm.nih.gov/36714154/

[14] Wanli Ouyang, Xingyu Zeng, Xiaogang Wang, Shi Qiu, Ping Luo, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Chen-Change Loy, Xiaoou Tang, "DeepID-Net: Object Detection with Deformable Part-Based Convolutional Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2015. Available at: https://arxiv.org/abs/1412.5661

[15] Muhammad Usama, Shoab A. Khan, Ahmed Shehab Khan, Atif Bin Mansoor, "A Comprehensive Review on Scene Text Detection and Recognition," *IEEE Access*, 2020.

[16] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015.

[17] S. Gehrmann, F. Dernoncourt, Y. Li, E. T. Carlson, J. T. Wu, J. Welt, J. Foote, E. T. Moseley, D. W. Grant, P. D. Tyler, and L. A. Celi, "Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives," *PLoS ONE*, vol. 13, no. 2, Feb. 2018, Art. no. e0192360.

[18] A. Abbas, M. Afzal, J. Hussain, and S. Lee, "Meaningful information extraction from unstructured clinical documents," *Proc. Asia–Pacific Adv. Netw.*, vol. 48, 2019, pp. 42–47. Available at: https://www.researchgate.net/publication/336797539 _

[19] D. Tkaczyk, P. Szostek, and L. Bolikowski, "'GROTOAP2—The methodology of creating a large ground truth dataset of scientific articles," *D-Lib Mag.*, vol. 20, 2014.

[20] C.-A. Boiangiu, O.-A. Dinu, C. Popescu, N. Constantin, and C. Petrescu, "Voting-based document image skew detection," *Appl. Sci.*, vol. 10, no. 7, p. 2236, Mar. 2020. Available at: http://dx.doi.org/10.3390/app10072236

[21] E. L. Park, S. Cho, and P. Kang, "Supervised paragraph vector: Distributed representations of words, documents and class labels," *IEEE Access*, vol. 7, pp. 29051–29064, 2019.

[22] D. Christou, "Feature extraction using latent Dirichlet allocation and neural networks: A case study on movie synopses," Available at: http://arxiv.org/abs/1604.01272

[23] B. Jang, M. Kim, G. Harerimana, S.-U. Kang, and J. W. Kim, "'Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism," *Appl. Sci.*, vol. 10, no. 17, p. 5841, Aug. 2020. Available at: http://dx.doi.org/10.3390/app10175841

[24] J. He, L. Wang, L. Liu, J. Feng, and H. Wu, "'Long document classification from local word glimpses via recurrent attention learning," *IEEE Access*, vol. 7, pp. 40707–40718, 2019.

[25] L. Arras, F. Horn, G. Montavon, K.-R. Mu¨ller, and W. Samek, "What is relevant in a text document?" *PLoS ONE*, vol. 12, no. 8, pp. 1–19, 2016. Available at: https://arxiv.org/abs/1612.07843