JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Optimized OCR Data Extraction Using Custom-Trained NLP-NER Models for Enhanced Image Analysis

[1]Chandan S P, [2]K R Sumana, [3]Lahari S K

[1]PG Student, The National Institute of Engineering, Mysuru, Visvesvaraya Technological University, Belagavi, Karnataka, India
[2]Faculty, The National Institute of Engineering, Mysuru, Visvesvaraya Technological University, Belagavi, Karnataka, India
[3]Al/ML Developer, Qikk Automation Technologies PVT LTD, Mysuru, Karnataka, India

Abstract: It is becoming more apparent that scholarly work and many areas of business require proper and efficient data extraction through OCR. This paper has the following objectives of describing the strategies used in the development as well as optimization of the OCR data extraction and Barcode/QR Code through the construction of the NLP-NER models. Concerning the futures work of the proposed approach, the aim is to increase the efficiency of the image analysis in terms of quality and time by using the approaches based on the ML.

IndexTerms **- OCR, Barcode/QR Code, NLP, NER, Image Analysis, Machine Learning, Data Extraction, Custom Models.**

I. INTRODUCTION

OCR is a technology that changed the way text is extracted from images in industries such as document handling and data input. However, more conventional OCR's have significant difficulties with complicated or poor image quality to read. It can also be seen that the incorporation of NLP and NER with OCR can improve data extraction. NLP is important in analyzing human language while NER is important in finding features like name and dates. Custom trained NLP-NER models enhance the accuracy and feasibility of OCR for handling several issues of image-based texts.

II. LITERATURE SURVEY

The paper titled "Efficient Automated Processing of Unstructured Documents Using AI " by Dipali Baviskar, Swati Ahirrao, Vidyasagar Potdar and Ketan Kotecha is a systematic review of various techniques that have been made in employing AI in information extraction from unstructured documents. [1] The study employs SLR done systematically, following Kitchenham and Charters guidelines and utilising methods such as OCR, RPA, and NER.

An algorithm for QR code recognition using image processing techniques such as image binarization, tilt correction, orientation, geometric correction, and normalisation is proposed in the paper "QR Code Recognition Based on Image Processing" by J. Seetha, Dr. G. Manikandan, Ms. S. Hemalatha, and Ms. Vilma Veronica. [2]

Increasing the efficiency in real-time information obtained from ID cards has been the goal of the study by J. Wehr, et al. while paying more attention to the accuracy and speed of the process. [3] In the architecture research and development, it is suggested to apply deep learning methods like EAST detector and DNN in text segmentation and face detection, which exhibit higher efficiency as compared to the conventional methods.

The mane of this paper is "Text Extraction and Recognition from Image using Neural Network", authors are C. Misra, P. K Swain, and J. K Mantri et al. The aim of this work is to propose an unconstrained image indexing and retrieval on the basis of text extraction and recognition from images using neural networks. [4] The approaches used include HSV- colour reduction, geometrical/morphological feature-based ROI, singular value decomposition for feature selection, multilayer perceptron to decide blocks of text or non-text, connected component analysis, binarization and OCR for writing extraction and storage in a database.

In the article titled 'A custom-built deep learning approach for text extraction from identity card images' by Geerish Suddul, Rakshith Sastry, and Bala Bachina, the authors describe a brand-new DL approach to extract texts from the images of identity cards

to minimizing client onboarding time. The objective of the work is to identify where specific text fields including surname, first name, date of birth, gender and identity number embedded in Mauritian identity cards are. [5]

Tareq Al-Moslmi et al.'s literature review on named entity extraction for knowledge graphs provides in-depth analysis of the advancements made in named entity identification in text. Finding named entity mentions, classifying them, understanding the author's intention about the entities to be mentioned, and linking the mentions to the entities in a knowledge base constitute the study's task level. Various methods and strategies have been used to rank the entities, cluster the entities that belong to the NIL class, and generate candidate entities. These methods and strategies include [6]

The first part of the paper describes the functionality of the IE and its ability to process semi or unstructured data, which are critical aspects when dealing with the big data. In connection with the classic IE systems, it points to the complication of analyzing vast amounts of unstructured information and reiterates the need to update the computational platform to handle these problems. From the abstract, one may infer that there is no study that encompasses all sorts of data, although prior research has discussed IE questions related to certain forms of data. [7]

III. Proposed Work

In this respect, the proposed module brings together components into a coherent system for document analysis and information extraction. It starts with an input module for uploading images or selecting a directory. A pre-processing module prepares the images by resizing and enhancing them. Additionally, OCR is used in the main analysis module for getting texts from the documents, there is decoding of QR/barcodes and NER model trained to recognize entities within it. Furthermore, a data management module organizes and securely stores information while an evaluation module judges how well extractions were made as well NER predictions.

IV. Methodology

4.1 Data Collection and Preparation

Assembled a broad collection of text data from multiple sources, guaranteeing that numerous entity types, settings, and writing styles are represented. Properly annotated each named entity in the training, validation, and testing sets of data. The reel has several labels attached to it, each containing various pieces of important information. The reel has several labels attached to it, each containing various pieces of important information. These labels usually include part numbers, quantities, compliance information, manufacturing details, and barcodes/QR codes for efficient inventory tracking and management. The labels are strategically placed to be easily readable by both humans and automated scanning systems.

4.2 Text Pre-processing

Used methods such lemmatization or stemming, special character removal, tokenization, and lowercase conversion. generated word embeddings or employed pre-trained embeddings to numerically represent words. The purpose of the Text Detection Module is to locate and retrieve text from an already-processed image. First, a format appropriate for optical character recognition (OCR) is created from the pre-processed image, which has undergone grayscale conversion, noise reduction, edge enhancement, and contrast improvement.

4.3 Model Selection

researched several NER techniques and organised the investigation in term of scalability, practicable speed and precision. These included rule-based systems, Deep Learning Architectures (Like BERT-based models, and BiLSTM-CRF), and statistical models which were like Hidden Marko Model and Conditional Random Fields.

4.4 Model Training

The specific approach used in this work for training the selected NER model involved the application of supervised learning approaches. Usually, deep learning models embed the transfer learning from the trained language models. A few examples of hyperparameters that have to be optimised include the batch size, dropout rate and learning rate. used several optimization strategies such as early stopping, and learning rate schedule to enhance the model’s performance.

4.5 Model Evaluation

As an evaluation metric, I have also used F1-score, which gives an overview of the model’s efficiency in addition to individual kinds of the specified entities:.accuracy and loss gave general information about how well the model functioned in the classification process, while precision and recall rates demonstrated how well a specific type of the defined entities was classified. assessed the capability of the model to identify the articles in different settings and the model’s performance in a situation with uncertainty.

4.6 Model Validation

To guarantee generalization across various text kinds and topics not encountered during training, the learned model was validated on a different test dataset.

4.8 NER Architecture

Improved NLP-NER systems also improve the level of data extraction using OCR as they fine-tune a model for a client. These models are implemented by Trainer Named Entity Recognition and Natural Language and are processed by the Transformer-based architecture for entity recognition; These models undergo preprocessing for the enhancement of visibility of the text; These are

annotated datasets of different picture formats like barcodes, QR codes and so on. Batch normalisation aids in reducing processing costs and the issue with regard to the optimisation of point architectures as a result of bottleneck layers. Tuning of the model-tuning is the optimization of the model for localized improvements through the use of transfer learning. Correction of errors in OCR text and improvement of structure based on context The application takes errors manufactured by OCR and works at the flaws in interfaces. When employing these sophisticated procedures, the OCR data that is obtained is a lot more dogged in terms of accuracy and degree of implication.

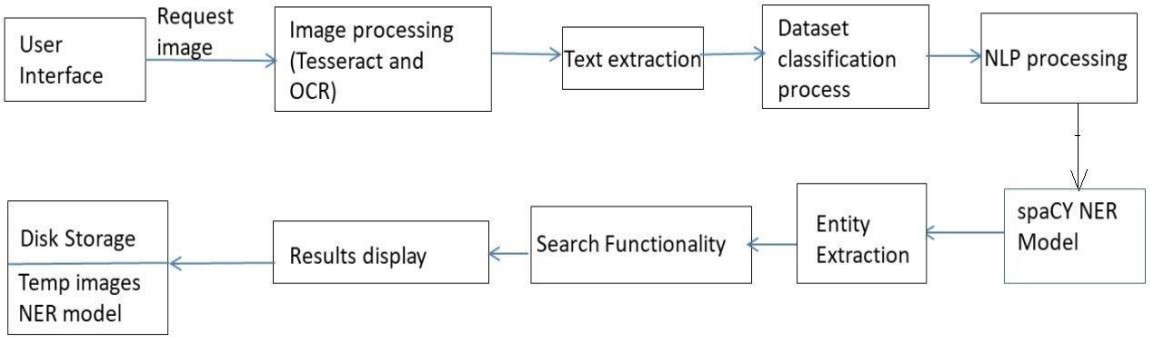

Fig 4.1 System Architecture

V. RESULT AND DISCUSSION

Text Extraction Module Analysis

To process photos and extract text, the text extraction module makes use of OCR. It extracts text accurately from high-resolution photos and performs well when tested with different image kinds (JPG, JPEG, and PNG) and quality levels. OCR performance declines with blurry or low-resolution photos, producing partially or inaccurate text along with an error notice. To avoid processing mistakes, the system additionally recognizes unsupported file types (such as PDF and DOCX) and shows the relevant error message.

Named Entity Recognition (NER) Module Analysis

The NER module takes the extracted text and the identification of the involved named entities depends with the type of the dataset. Based on the experiments, it is luminous in recognizing the datasets and using the corresponding NER model. It accurately searches for named items and delivers them with the structures smoothly and efficiently. Additional, through the parameters, users can enter certain keys and later use the search to separate the values of the relevant entities to the textual input, which guarantees its accuracy.

Results of System Testing

The following are the major findings made in the system testing stage: The technology is functional and capable of efficiently transfering and displaying several formats of photos while giving feedback to the users. From the observations made, it can be said that the OCR module is very effective in extracting text from high-quality photos; for other unsupported or low-quality images, an error message appears. It properly categorizes the datasets, provides named things in the right format, and processes the text using the NER model. In addition, the key-value pairs are retrieved through the search function accurately. The Streamlit application can be used to easily apply text display, entity recognition, and image upload. The target entities located have the option of being searched for within the application while other extracted entities are well aligned and can easily be retrieved.

4.1 Results of Descriptive Statics of Study Variables

Metric	Formula	Calculation	Result
Accuracy	(Number of Correctly Enhanced Images / Total Number of Usable Images) ×100	(7371 / 8190) ×100	90.00%
Precision	(True Positives / True Positives + False Positives)	(7371 / (7371 + 819))	90.00%

Recall	(True Positives / True Positives + False Negatives)	(7371 / (7371 + 819))	90.00%
F1-Score	(2 × Precision × Recall / Precision + Recall)	(2 × 90.00% × 90.00%) / (90.00% + 90.00%)	90.00%

Table 4.1: Descriptive Statics

In brief:

Accuracy= (7371 / 8190) ×100 Accuracy= (7371 / 8190) ×100 ≈ 90.00% accuracy
Precision= (7371 / 7371 + 819) Precision= (7371 / 8190) Precision ≈ 90.00%
Recall = (7371 / 7371 + 819) Recall= (7371 / 8190) Recall ≈ 90.00%. Recall percentage: 90.00%
F1-Score= (2 × 90.00% × 90.00% / 90.00% + 90.00%) F1-Score = (2 × 0.9 × 0.9 / 0.9 + 0.9) F1-Score= (1.62 / 1.8) F1-Score≈0.90. 90.00% is the F1 score.
The formulas utilized, the calculations made, and the outcomes are displayed in this table, which summarizes the computations for accuracy, precision, recall, and F1-score.

VI. Conclusion

The work combines OCR, NER, and QR/barcode text extraction from images with the help of Python libraries including spaCy, EasyOCR, pytesseract, and pyzbar. It first extracts the content of the images with high accuracy by using EasyOCR and pytesseract to convert the content to texts. It is then passed through a fine-tuning of the spaCy NER model tailored for the technological domain to identify entities such as organization name, TYPE, MARKING, PART_NUMBER, LOT_NUMBER, and QUANTITY. The workflow is implemented with Streamlit for easy utilization of the proposed tool for uploading images, extracted text, and searched entities; therefore it is suitable to be adopted in various industries with the need for accurate data extraction.

VII. Acknowledgment

I would especially like to thank the efficient workers of The National Institute of Engineering, Mysuru and my guide Smt. K R Sumana you have supported me till the end. In this I would like to convey my special thanks to Lahari S K from Qikk Automation in Mysuru for all their assistance and valuable advice. I honestly have a lot of gratitude to my parents, friends, and fellow students for the immense support I have got from them. Finally, it would be my honor to express gratitude to all people, who helped me directly or indirectly to finalize my project.

References

[1] DIPALI BAVISKAR, et. al. ‘’ Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions’’, doi.org/10.1109/ACCESS.2021.3072900.
[2] Shagun Davessar, “Data Extraction using Optical Character Recognition and Natural Language Processing”, doi.org/ JETIR2312486.
[3] Niloofar Tavakolian, et. al. “Real-time information retrieval from Identity cards’’, doi.org/ arXiv:2003.12103v1.
[4] C. Misra, et. al. “Text Extraction and Recognition from Image using Neural Network”, doi.org/ 10.5120/4927-7156.
[5] Geerish Suddul, et, al. “A custom-built deep learning approach for text extraction from identity card images”, doi.org/ 10.11591/ijict.v13i1.pp34-41
[6] TAREQ AL-MOSLMI, et. al. ‘’ Named Entity Extraction for Knowledge Graphs: A Literature Overview’’. Doi/org/10.1109/ACCESS.2020.2973928.
[7] Kiran Adnan, et. al. ‘’An analytical study of information extraction from unstructured and multidimensional big data’’. doi.org/ /10.1186/s40537-019-0254-8.
[8] Kaiming He, et. al. ‘’Deep Residual Learning for Image Recognition’’, doi.org/ arXiv:1512.03385v1
[9] Yiheng Xu, et.al. “LayoutLM: Pre-training of Text and Layout for Document Image Understanding”. Doi.org/ arXiv:1912.13318v5.
[10] Nishant Subramani, et. al. “A Survey of Deep Learning Approaches for OCR and Document Understanding”. Doi/org. arXiv:2011.13534v2.
[11] Yuntian Deng, et. al. ‘’Image-to-Markup Generation with Coarse-to-Fine Attention’’. **doi:** arXiv:1609.04938v2.
[12] Kalyani Pakhale, et. al. ‘’ Comprehensive Overview Of Named Entity Recognition:Models, Domain-Specific Applications And Challenges’. doi.org. arXiv:2309.14084v1.
[13] Xuezhe Ma, et. al."End-to-end Sequence Labelling via Bi-directional LSTM-CNNs-CRF". doi/org.arxiv.org/abs/1603.01354
[14] Ratapol Wudhlkarn, et. al. “Deep Learning in Barcode Recognition: A Systematic Literature Review”, doi.org/ 10.1109/ACCESS.2022.3143033