Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Design of a Mechanically Integrated System for Pictorial Detection and Document Information Extraction using U-Net

[1]Dr. Lakshmi Kumari, [2]Dr. Raghavendra Joshi, [3]Prof. V Srinivasulu, [4]Prof. Shiva Kumar S Y, [5]Prof. E Manjunath, [6]Prof. Mohammed Fayaz K

[1]*Associate Professor, Department of Mechanical Engineering, Ballari Institute of Technology & Management, Ballari, Karnataka 583104, India*
thanulakshmi2003@gmail.com
[2]*Professor, Department of Mechanical Engineering, Ballari Institute of Technology & Management, Ballari, Karnataka 583104, India*
coe@bitm.edu.in
[3]*Assistant Professor, Department of Mechanical Engineering, Ballari Institute of Technology & Management, Ballari, Karnataka 583104, India*
srinivasulu@bitm.edu.in
[4]*Assistant Professor, Department of Mechanical Engineering, Ballari Institute of Technology & Management, Ballari, Karnataka 583104, India*
shivu070@gmail.com
[5]*Assistant Professor, Department of Mechanical Engineering, Ballari Institute of Technology & Management, Ballari, Karnataka 583104, India*
manjunatha.e@bitm.edu.in
[6]*Assistant Professor, Department of Mechanical Engineering, Ballari Institute of Technology & Management, Ballari, Karnataka 583104, India*
mohammedfayaz@bitm.edu.in

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

This paper presents a novel approach to pictorial detection and information extraction from documents employing the U-Net architecture. In the realm of computer vision and document analysis, our proposed methodology leverages the powerful capabilities of U-Net, a convolutional neural network renowned for semantic segmentation tasks. We delve into the intricacies of adapting U-Net for the specific challenges posed by document images, aiming to enhance the accuracy and efficiency of information extraction. The methodology involves a comprehensive training process, where the network learns to identify and isolate relevant pictorial elements within the document. Furthermore, our approach incorporates post-processing techniques to refine the extracted information, ensuring high precision in capturing critical details. We evaluate the performance of the proposed system through extensive experiments on diverse datasets, showcasing its robustness and versatility across various document types. The results demonstrate the efficacy of our approach in automating the extraction of valuable information from documents with pictorial content, paving the way for advancements in fields such as document analysis, information retrieval, and content understanding. This paper contributes to the growing body of research at the intersection of deep learning and document processing, offering a promising solution for efficient information extraction from visual documents.

Keywords: U-Net, Document Analysis, Image Segmentation, Pictorial Detection, Information Extraction, Object Localization

I. INTRODUCTION

Pictorial detection and information extraction from documents using U-Net is a fascinating topic in the realm of computer vision and deep learning [1-3].

U-Net is a convolutional neural network architecture that is particularly well-suited for image segmentation tasks. Image segmentation involves dividing an image into different regions based on certain characteristics. In the

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

context of document analysis, this can be incredibly useful for tasks such as text extraction, object detection, or layout analysis [4-6].

The U-Net architecture is named after its U-shaped structure, which consists of a contracting path, a bottleneck, and an expansive path. The contracting path captures the context and extracts features, while the expansive path helps in precise localization. This makes U-Net effective for tasks that require a detailed understanding of the spatial relationships within an image [7].

In the case of document analysis, the U-Net model can be trained to identify and extract specific elements from documents, such as text, tables, or images. The process typically involves labeling the training data to indicate the regions of interest and then training the model to learn the mapping from input images to these labeled regions [8-9].

Once trained, the U-Net model can be applied to new documents to automatically detect and extract relevant information. This has numerous applications, from document digitization and text recognition to data extraction for further analysis.

It's a powerful approach that has gained popularity in recent years due to its versatility and effectiveness in a wide range of image segmentation tasks [10].

II. LITERATURE REVIEW

In their paper [11], the authors introduce an innovative UAV benchmark that tackles complex scenarios with unprecedented levels of difficulty. Through careful selection and annotation of approximately 80,000 images from 10 hours of raw footage, the benchmark covers a wide range of attributes – including weather conditions, flight altitude, camera view, vehicle class, and occlusion – across three fundamental computer vision tasks: detection, single target tracking, and multiple target tracking. Each task is rigorously evaluated using the latest cutting-edge algorithms, revealing a significant gap between state-of-the-art methods and our dataset due to the unique challenges presented by real-world drone scenarios, such as high density and small objects.

Deep learning is a highly effective and affordable form of supervised learning that is capable of tackling complex problems [12]. Unlike other learning methods, it boasts a multitude of advantageous qualities and is not limited by a narrow range of solutions. Its exceptional performance and continued advancements have made it a popular tool in a wide range of applications. From image and face recognition to language processing and scientific analysis, deep learning's capabilities are virtually limitless. In fact, our research delves into the latest literature on deep learning to uncover its current trends and progress."

In the article [13], "A Novel Approach to HSI Modeling: The Hybrid Locally Dilated Convolutional Fusion Network", a groundbreaking HSI model called the LDFN is introduced. The LDFN revolutionizes perception by incorporating a combination of detailed local information and extensive spatial features. To achieve this, our approach utilizes a unique combination of local and hybrid dilated convolution fusion operations, carefully selected from an assortment of operations such as standard convolution, average pooling, elimination, and batch normalization. These operations effectively extract rich spatial and spectral information. The result is a powerful residual fusion network consisting of multiple convolutional layers, culminating in a softmax classification input.

In their study, [14], the author put forth a groundbreaking approach to lung CT segmentation - the fused U-Net with dilated convolution (DC-U-Net). This model was compared to the widely used Otsu and region growing methods, and was evaluated using key metrics such as Intersection Over Union (IOU), Data Coefficient, Precision, and Recall. The authors confidently suggested that their model can effectively and automatically segment original images, yielding superior results. Furthermore, their model has the capability to streamline segmentation processes and accurately capture important downstream structures such as pulmonary blood vessels and trachea.

In their paper [15], the authors present an ingenious and robust technique for identifying tables in both documents and websites. Instead of relying on complex deep learning methods that require extensive training, they have devised a rule-based approach that leverages the unique characteristics of tables, particularly their grid layout. The method consists of two stages: feature extraction and grid pattern recognition. Firstly, features are extracted from

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

the table content. Then, non-text elements and extraneous text are filtered out. In the second stage, a tree structure is constructed based on the extracted features, and a novel algorithm is employed to accurately identify the grid pattern of the table.

III. DOCUMENT DETECTION

The paragraph describes a comprehensive approach to document processing, including segmentation, text detection, image quantization, and extraction of valuable information.

Dataset Creation: Due to the unavailability of suitable datasets, custom datasets comprising pairs of images and masks were created, both artificially generated and manually photographed.

Text Detection: A specialized algorithm for text detection focuses on locating specific lines of text within small sections of the document, considering factors like position, orientation, and color.

Image Quantiles and Fullness: The concepts of image quantiles and fullness are introduced to quantify and evaluate features within images, aiding in the detection and assessment of text regions.

Local Maximum of Histogram: The algorithm involves identifying local maxima in the image histogram to locate potential text regions based on font color and height characteristics.

Finding Letters, Words, and Lines: Continuous components on the mask are identified and connected to form letters and lines, facilitating the extraction of text-based information.

Classification and Verification of Orientation: Document classification and orientation verification are performed using pre-defined anchor fields, ensuring accurate document processing.

Binarization and Reading: Binarization of text regions and subsequent reading using Tesseract library are employed to extract textual content from documents.

Extraction of Values by Regular Expressions: Structured fields with specified regular expressions are processed to extract multiple data items, such as street addresses, from text regions.

Image Binarization: Certain types of image data, like signatures and barcodes, are binarized to facilitate their processing and extraction.

Exit Process: The processed document and extracted data are structured into a JSON object, containing identification information, textual data, and image data for further analysis or use.

The found document and the read data are structured into a JSON object. The object contains identification information about the document taken from cards and models, a party object and information about whether the other side of the document is in the database. If the input is a front image, then front will be an object containing the fields read and back will be true or false. The client application can thus decide whether it is necessary to take a photo of the other party. The page object consists of two items – img and fields. In the first field, there is a string of the image of the transformed side of the document in jpg format encoded in base64. The fields field contains a list of all fields appearing on a given page of the document. For each field, its identifier, name in English, name on the document, type of field and order are given. Text and data fields have values stored in the value key. For data fields, the value is numeric and is a timestamp. If the text or data field was not read for any reason or the read text was invalid, its value is null. Image type fields have rgb and binary keys with images in jpg format, resp. png (Figs 1-3).

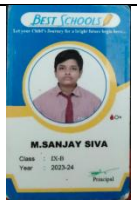
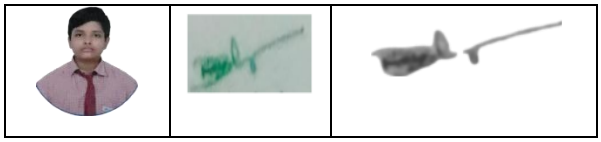


Fig. 1. Input photo

Fig. 2. Transformed document image

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

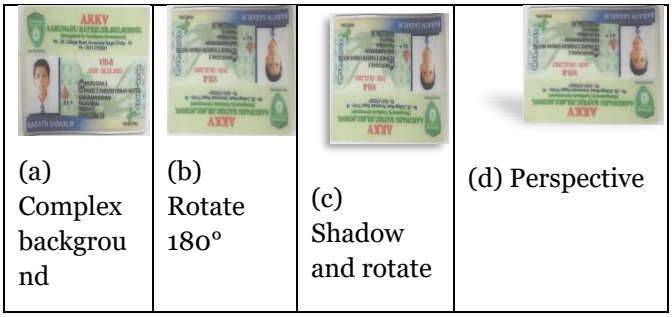


(a) Owner (b) Color signature (c) Binarized signature

Fig. 3. Image data from the document

IV. RESULTS

The custom solution recognized the document in 48 out of 51 photos. Smart IDReader recognized and correctly classified the document in 35 photos (Fig. 4). However, these numbers cannot be compared, as some photos were deliberately taken to show the weaknesses of one or the other solution.



(a) Complex background | (b) Rotate 180° | (c) Shadow and rotate | (d) Perspective

Fig. 4. Examples of photos where Smart IDReader does not find the document

Smart IDReader detects documents by combining edge and text detection. This means that scenes with many edges or large amounts of text may cause the document to go undetected. Photo 5.10a showing an ID card on a plaid skirt is proof of this. The image is sharp, the document on it occupies a large part of the surface and is not rotated too much, so the main reason why it is not found is the background.

Another problem can be caused by rotation or perspective distortion of the document (photos 9b, 9c and 9d). In photo 9b, the document is also turned 180° and the lighting conditions are not ideal. These may also be the reasons why it was not recognized. Half of the document in photo 9c is covered by a sharp shadow, which probably forms a higher quality edge than the right and top sides of the document.

Finally, Fig 4d shows the document under perspective distortion. The contrast between the brown background and the light document provides high-quality edges, but the document is not detected, most likely due to the limitation of the angles between the edges or poor text detection. Even the custom solution checks that the edges do not form angles too different from 90°, but the neural network detection is reliable enough that the condition can be more tolerant.

(a) Glare on anchor | (b) Covered anchor | (c) Missing edge

Fig. 5. Examples of photos where the custom solution does not find the document

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The custom solution (Fig 5) has two main disadvantages when detecting a document. The first is based directly on the method of document classification using anchors. Illegibility of all anchors of the document caused by reflection (10a) or obscuration by another object (10b) leads to the impossibility of verifying the document and rejection of the photograph. Another restriction is the position of the corners of the document, which must all be inside the photo. Parts of the document near the edge of the photo tend to be undetected, resulting in incorrect edge detection. In some cases, the given edge is detected so poorly that even the anchors are not verified, but other times the anchors can be verified and the result of reading the rest of the document has a low success rate. Example 10c demonstrates the rejection of a document whose upper left corner is outside the photo. Smart IDReader does not share this problem, because when detecting a document, it is not limited only to its edges, but also to the layout of the text. Based on the detected text, it reconstructs the missing edges of the document. It can even detect a document whose edges and corners are all outside the photo.

A. Reading results

Reading the photos, which she was able to detect both solutions, ended in a tie with her own solution slightly ahead. An important factor is the machine-readable zone on the back of the OP, which automatically means the loss of 92 characters for a given photo in the event of a failed detection. The custom solution failed to read the machine-readable zone on one photo, while Smart IDReader failed to read it on two. Therefore, results without MRZ are also included in the table. Omitting the MRZ reading results in a slight deterioration of precision and a more noticeable increase in recall. Smart IDReader has a recall better by 0.015 compared to its own solution.

Figs 6-7 show how many photos were read with what throughput and coverage. For example, the blue column on the far left shows how many documents were read to within 1.00. The blue column one place to the right of it shows how many documents were read with an accuracy of at least 0.98 (roughly 1-2 characters wrong). It can be seen from the graphs that the custom solution has a problem with some high-quality photos, where it makes a small mistake rather than Smart IDReader.

Fig. 6. Bar graph of the precision distribution function

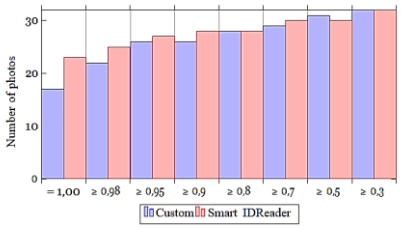


Fig. 7.Bar graph of the recall distribution function

V. DISCUSSION

The proposed algorithm finds the document in the segmentation mask created by the U-Net network, detects its edges and corners. Knowing the corners, country of origin and side of the document, it uniquely classifies it by reading all anchors on all documents of that country.

One of the disadvantages of reading anchors is poor scalability. Reading is an expensive operation, and as the number of documents in the database increases, so does the required number of processor cores or computing

time. This method is unsuitable for fully automatic classification, including the issuing country and the side of the document.

Even before reading, text is detected in the viewports. The text detector proposed in this work focuses on text written in a left-to-right typeface, where individual characters are separated. It does not work with Arabic in its current form.

After the document is classified, rotated and the text detected, the data on it is read. The reading implementation is blocking and does not allow the reading of multiple documents at once. In addition, some languages or photos are more complex and take longer to read. By the time the last thread detects or reads the text, the others wait unnecessarily, potentially slowing down the reading of multiple documents in a row.

Classification and structured reading of documents is only possible thanks to the database. Although JSON, the database format, is readable, some parts of the database may not be understandable to everyone. A good document editor that would allow changing field coordinates would speed up work when expanding the database. An actual database, not text files, would also be appropriate for project growth. Automatic tests are also absent at work.

The custom solution recognized the document in 48 out of 51 photos, indicating a high level of accuracy.

Smart IDReader, on the other hand, correctly classified the document in 35 photos, demonstrating slightly lower performance. Both solutions achieved comparable reading results, with the custom solution slightly ahead.

The presence of machine-readable zones (MRZ) significantly impacts reading accuracy, with failures resulting in the loss of important data characters when copmared to other methods [1-7]

VI. CONCLUSION

The result of the work was the documentation of existing approaches to text detection and its reading in a real scene and the design of an own algorithm for the detection and reading of documents. The algorithm detects the document using a neural network, removes perspective distortion, classifies it and preprocesses text fields to increase reading accuracy. As a demonstration, a Java program was designed and implemented, which achieves high-quality detection even on otherwise complex photographs. Using the Tesseract library, 98.4% correctness of read characters and 92.6% coverage of characters in the photo is achieved. The extracted information is converted into JSON format. The implementation is compared to the commercial mobile application Smart IDReader.

REFERENCES

[1] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, “Automatic text summarization: a comprehensive survey,” *Expert Systems with Applications*, vol. 165, pp. 113679, 2021.
[2] J. Bhatt, K. A. Hashmi, M. Z. Afzal, and D. Stricker, “A survey of graphical page object detection with deep neural networks,” *Applied Science*, vol. 11, pp. 5344, 2021.
[3] J. Younas, S. A. Siddiqui, M. Munir, M. I. Malik, F. Shafait, P. Lukowicz, and S. Ahmed, “Fi-Fo detector: figure and formula detection using deformable networks,” *Applied Science*, vol. 10, pp. 6460, 2020.
[4] J. C. W. Lin, and K. H. Yeh, “Security and privacy techniques in IoT environment,” *Sensors*, vol. 21, pp. 1, 2021.
[5] J. C. W. Lin, G. Srivastava, Y. Zhang, Y. Djenouri, and M. Aloqaily, “Privacy-preserving multiobjective sanitization model in 6G IoT environments,” *IEEE Internet of Things Journal*, vol. 8, pp. 5340–5349, 2020.
[6] J. M. T. Wu, G. Srivastava, A. Jolfaei, P. Fournier-Viger, and J. C. W. Lin, “Hiding sensitive information in eHealth datasets,” *Future Generation Computer Systems*, vol. 117, pp. 169–180, 2021.
[7] A. I. Kadhim, “Survey on supervised machine learning techniques for automatic text classification,” *Artificial Intelligence Review*, vol. 52, pp. 273–292, 2019.
[8] A. Guha, and D. Samanta, “Hybrid approach to document anomaly detection: an application to facilitate RPA in title insurance,” *International Journal of Automation and Computing*, vol. 18, pp. 55–72, 2021.
[9] F. U. Hassan, and T. Le, “Automated requirements identification from construction contract documents using natural language processing,” *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, vol. 12, pp. 04520009, 2020.

Research Article

[10] X. Zhai, K. Liu, W. Nash, and D. Castineira, "Smart autopilot drone system for surface surveillance and anomaly detection via customizable deep neural network". In Proceedings of the International Petroleum Technology Conference, Dhahran, Saudi Arabia, 13–15 January 2020.

[11] H. Yu, G. Li, W. Zhang, Q. Huang, D. Du, Q. Tian, and N. Sebe, "The unmanned aerial vehicle benchmark: object detection, tracking and baseline," *International Journal of Computer Vision*, vol. 128, pp. 1141–1159, 2020.

[12] S. Smys, J. I. Z. Chen, and S. Shakya, "Survey on neural network architectures with deep learning," *Journal of Soft Computing Paradigm*, vol. 2, pp. 186–194, 2020.

[13] C. Li, Z. Qiu, X. Cao, Z. Chen, H. Gao, and Z. Hua, "Hybrid dilated convolution with multi-scale residual fusion network for hyperspectral image classification," *Micromachines*, vol. 12, pp. 545, 2021.

[14] K. B. Chen, Y. Xuan, Y, A. J. Lin, and S. H. Guo, "Lung computed tomography image segmentation based on U-Net network fused with dilated convolution," *Computer Methods and Programs in Biomedicine*, vol. 207, pp. 106170, 2021.

[15] J. Kim, and H. Hwang, "A rule-based method for table detection in website images," *IEEE Access*, vol. 8, pp. 81022–81033, 2020.