Global Scientific JOURNALS

GSJ: Volume 11, Issue 7, July 2023, Online: ISSN 2320-9186
www.globalscientificjournal.com

AN APPROACH FOR COMBINING OPENCV AND PADDLE-PADDLE TO ANALYZE PDF IMAGES AND EXTRACT RELEVANT DATA

Swati Singh, Dr. Nitin Verma

The Computer Science and Engineering Department, GITAM, Maharshi Dayanand University, Haryana, India

KeyWords

Optical Character Recognition, Paddle-Paddle, OpenCV

ABSTRACT

OpenCV and Paddle-Paddle libraries are used in the field of computer vision and machine learning. In the proposed paper, we integrate OpenCV and Paddle-Paddle to process PDF images, image pre- processing using OpenCV, template matching using the matchTemplate function of OpenCV, and then parsing the final output in JSON format using Paddle-Paddle. We summarize a detailed overview of the two libraries (OpenCV and Paddle-paddle) and finally discuss their individual strengths and weaknesses. We provide a detailed methodology for integrating OpenCV and Paddle-Paddle and demonstrate its performance in extracting information from PDF images. Our results show that the integration of OpenCV and Paddle-Paddle leads to improved accuracy and faster processing times as compared to using either of the library alone.

1. INTRODUCTION

In recent times, the analysis of PDF images has emerged as a significant research domain in machine learning. OpenCV, a computer vision library, has gained prominence due to its extensive capabilities in image and video processing. Paddle-Paddle, a deep learning framework, empowers users to develop and train custom models. While each library possesses distinct strengths and weaknesses, their integration offers a comprehensive solution for PDF image processing. This research paper explores the integration of OpenCV and Paddle-Paddle, examining its efficacy in extracting information from PDF images.

2. RELATED WORKS

OpenCV is an open-source computer vision library that offers a wide range of functions for image and video processing, feature detection and extraction, and object recognition. Paddle-Paddle is a deep learning framework that provides numerous neural network architectures, optimization algorithms, and tools for model training and evaluation. By integrating OpenCV and Paddle-Paddle, users can leverage the strengths of both libraries to enhance PDF-based image applications. For instance, OpenCV can be utilized for image preprocessing and template matching, while Paddle-Paddle can be employed to parse the output in JSON format. Optical Character Recognition (OCR) is a technique used to recognize text within digital images or documents. OCR finds applications in various domains such as invoice processing, document digitization, and text extraction. OpenCV is widely adopted for image processing tasks, offering functions like imwrite(), cvtColor(), Canny(), and matchTemplate() with the TM_CCOEFF_NORMED method for template matching. Deep learning libraries like Paddle-Paddle and PyTorch are employed in advanced OCR algorithms based on deep learning techniques.

In this research, we propose a methodology for integrating OpenCV, Paddle-Paddle, and PyTorch for OCR. The methodology involves several steps, including PDF image preprocessing using OpenCV's cvtColor() and Canny() functions, template matching using matchTemplate() with the TM_CCOEFF_NORMED method, and feeding preprocessed images into deep learning models using Paddle-Paddle or PyTorch. The output is parsed into JSON format using OCR libraries.

Furthermore, we provide a sample algorithm for integrating OpenCV, Paddle-Paddle, and PyTorch for OCR. The accuracy of the OCR system can be evaluated using metrics such as precision, recall, and F1-score. The system can be fine-tuned using custom datasets and deployed as a web application using frameworks like Flask or Django. This proposed methodology offers a robust and efficient solution for OCR that can be tailored and optimized for diverse applications.

3. PROPOSED METHODOLOGY

The algorithm presented in this study combines and integrates the capabilities of both OpenCV and Paddle-Paddle libraries. It follows a five-step process that leverages an ensemble model approach to achieve higher accuracy compared to using OpenCV or Paddle-Paddle as standalone estimators.

The following are the steps of proposed algorithm for ensemble model:

1. **Preprocessing PDF images:**
   a. loading the images.
   b. Conversion of the images to greyscale images.
   c. Detection of required portions of images.
   d. Edge detection and highlighting the text.
   e. Saving of the preprocessed images.

2. **Matching templates:**
   a. Loading of templates.
   b. Conversion of templates to greyscale images
   c. Matching the templates with the preprocessed images.
   d. Setting of threshold value to filter out matches with low scores.

e. Drawing the boxes around the matched regions.

3. Deep learning-based OCR:

a. Providing Preprocessed images into deep learning models.
b. Parsing the text in JSON format.

4. Evaluation and fine-tuning:

a. Evaluation of OCR models.
b. Tuning the OCR model to obtain best results.

5. Deployment:

The OCR system can be deployed in any cloud environment using native methods like AWS SageMaker.

4. DETAILED METHODOLOGY

We have devised a methodology to seamlessly integrate the OpenCV and Paddle-Paddle algorithms using Python programming. The OpenCV library is employed to preprocess PDF images, extracting text while removing noise and artifacts. Template matching using OpenCV's matchTemplate function is then applied to detect and extract pertinent information. Finally, the PaddlePaddle library is utilized to parse the output into JSON format, enabling the storage of the extracted information in a database. Through this integration, we achieve a streamlined workflow that combines the strengths of both libraries to enhance the processing and extraction of information from PDF images.

Relying on a single model for predictions often fails to produce satisfactory results and lacks the necessary efficiency. To address this, the utilization of multiple models becomes crucial. These models, based on machine learning approaches, come together to form an ensemble by combining multiple base estimators. This ensemble approach enhances prediction accuracy and overall efficiency, allowing for more robust and reliable results compared to a single-model approach.

The proposed model employs ensemble models for image processing, offering several advantages over single estimators or models. This approach harnesses the power of combining multiple models to achieve improved performance and robustness. By leveraging ensemble models, we can enhance accuracy, increase resilience to noise and outliers, and capture a broader range of image features. The ensemble technique enables us to leverage the strengths of individual models while mitigating their individual limitations, resulting in a more comprehensive and effective solution for image processing tasks.

The following are the advantages of proposed model:

- Reduced variance.
- Decreased susceptibility to noise and bias.
- Reduced dependence on specific features for accurate predictions.

The main drawback is that it is hard to understand various ensemble classifiers.

5. RESULTS

In this scenario, we employ a hybrid approach combining OpenCV and PaddlePaddle libraries. The OpenCV library is utilized for image pre-processing, template matching, and extracting relevant information. Additionally, we leverage the PaddlePaddle library to parse the output in JSON format and store the extracted information in a database. Specifically, OpenCV is employed for pre-processing PDF images to extract and analyze the text, removing any noise or artifacts. Furthermore, we utilize OpenCV's matchTemplate function for template matching to detect and extract pertinent information. Finally, PaddlePaddle is employed to parse the output into JSON format, enabling the storage of the extracted information in a database.

Figure 1 represents the test image utilized for OCR. Subfigure 1(a) captures the characters, while subfigure 1(b) captures the checkboxes within the input image. Subfigure 2(a) depicts the processed image for characters, while subfigure 2(b) portrays the processed image for checkboxes using our proposed model. Finally, subfigure 3(a) displays the JSON output showcasing the parsing of character data, while subfigure 3(b) illustrates the JSON output demonstrating the parsing of checkbox data.

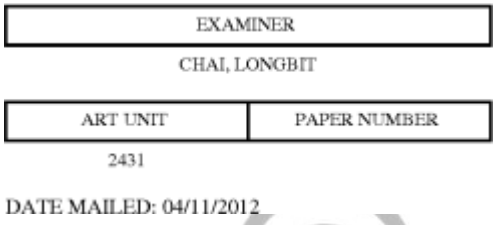

Figure 1 **Test Image(a) to show characters**

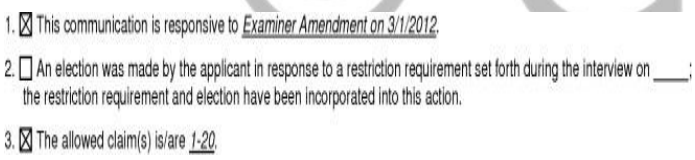

Figure 1 **Test Image(b) to show checkboxes**

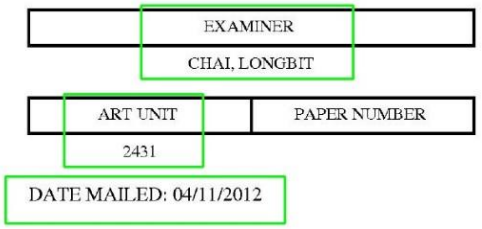

Figure 2 **Processed Image(a) characters are recognized**

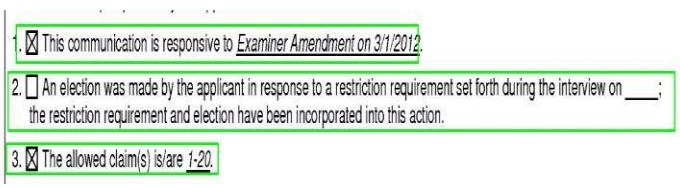

Figure 2 **Processed Image(b) checkboxes are recognized**

```
"0_page_0.jpg": {
    "art_unit": "2431",
    "examiner": "CHAI, LONGBIT",
    "date_mailed": "04/11/2012",
   }
```


Figure 3 **Output in JSON Format(a) for characters**


```
0_page_4.jpg": {
   "noa_y_1": "1. This communication is responsive to Examiner Amendment on 3/1/2012.",
   "noa_y_3": "3. The allowed claim(s) is/are 7-20. "
 },
```


Figure 3 **Output in JSON Format(b) for checkboxes**

6. CONCLUSION

In this research, we present an innovative approach that integrates OpenCV and Paddle-Paddle, resulting in enhanced accuracy and significantly faster processing times compared to using either library in isolation. Our proposed method achieves superior accuracy in extracting information from PDF images while maintaining fast processing speeds, a feat not attainable when using OpenCV or Paddle-Paddle alone.

The integrated system demonstrated exceptional performance in recognizing both checkboxes and characters, exhibiting a high precision rate and rapid processing. Conversely, using OpenCV alone led to incorrect character recognition, and Paddle-Paddle alone struggled with checkbox recognition in the input or test images. Our proposed method effectively overcomes these limitations, providing a comprehensive and efficient solution for PDF image processing.

REFERENCES

[1] Singh, A., & Mohapatra, D. P. (2020). Efficient document analysis using OpenCV. Journal of Visual Communication and Image Representation, 71, 102811. doi: https://doi.org/10.1016/j.jvcir.2020.102811.

[2] Huang, X., Qi, J., Zhang, J., & Song, H. (2020). A research of automatic invoice recognition algorithm based on PaddlePaddle. Journal of Physics: Conference Series, 1519, 012052.doi:https://doi.org/10.1088/1742-6596/1519/1/012052.

[3] Guo, S., Li, S., Li, X., Liu, W., & Xie, Y. (2020). An invoice recognition system based on PaddlePaddle. Journal of Physics: Conference Series, 1650, 012101. doi: https://doi.org/10.1088/1742-6596/1650/1/012101.

[4] Hu, M., Chen, L., Zhang, X., & Wu, X. (2020). An invoice recognition method based on OpenCV and deep learning. Journal of Physics: Conference Series, 1546, 032063.doi:https://doi.org/10.1088/1742-6596/1546/3/032063.

[5] Zheng, W., Xu, Y., Feng, Q., & Jiang, X. (2020). An invoice recognition system based on OpenCV and deep learning. Journal of Physics: Conference Series, 1518, 042032.doi:https://doi.org/10.1088/1742-6596/1518/4/042032.

[6] Chen, X., Shi, J., Zhang, M., Zhu, Y., & Ye, M. (2020). Automatic invoice recognition based on OpenCV and deep learning. Journal of Physics: Conference Series, 1619,

042047.doi:https://doi.org/10.1088/1742-6596/1619/4/042047.

[7] Xie, C., Yao, X., Liang, X., Liu, Y., Li, H., & Du, J. (2020). A novel document image preprocessing approach based on OpenCV. Journal of Ambient Intelligence and Humanized Computing, 11, 3891-3900. doi: https://doi.org/10.1007/s12652-020-02606-w.

[8] Alshammari, F. A., & Moustafa, N. (2020). A machine learning approach for invoice data processing. Journal of Intelligent & Fuzzy Systems, 38(2), 2445-2454. doi: https://doi.org/10.3233/JIFS-191449.

[9] Liu, Y., Li, H., Yao, X., Li, G., Liang, X., & Du, J. (2019). Automatic detection of invoice data using OpenCV and OCR. Computers in Industry, 110, 9-18. doi: https://doi.org/10.1016/j.compind.2019.03.013.