IMPLEMENTATION OF OCR TECHNOLOGY IN SIGNATURE VERIFICATION AND CONVERTION OF HANDWRITTEN NOTES INTO DIGITAL TEXT

Syahrul Hidayat[1]

[1]Departement of Informatics, National Development University "Veteran" East Java,
Surabaya, Indonesia
20081010076@student.upnjatim.ac.id

ABSTRACT

This paper delves into the challenges and potential of optical character recognition (OCR) technology within an industrial context, focusing on text recognition, archiving, and converting handwritten notes to digital text. It aims to develop reliable OCR capable of recognizing writing style variations and efficiently processing large-scale data while addressing ethical privacy concerns. Using quantitative and qualitative methods, the research addresses three key problem formulations. Experiments compare CNN algorithms with and without OCR in signature recognition, demonstrating CNN's superior accuracy, precision, and F1-Score without OCR. The study highlights ethical concerns related to privacy, accuracy, and potential discrimination in OCR use. Recommendations include enhancing data quality, further evaluating OCR performance, and safeguarding processed data privacy. To advance OCR technology, the research suggests assessing classification quality, improving training data, and enhancing data security. This underscores the imperative of enhancing OCR security to ensure information privacy and accuracy.

KEYWORDS

Optical Character Recognition (OCR), Handwritten Notes, Text Recognition, Ethical Issues

1. INTRODUCTION

1.1. Background

In the industrial sector, the application of OCR (Optical Character Recognition) technology plays an important role in various digitalization processes such as text form recognition, verification, and the process of converting hand notes into digital text. However, as with many technological innovations, the use of OCR is not without various challenges that need to be overcome to ensure optimal performance.

One of the main challenges facing OCR technology is its technological complexity in recognizing characters and words in handwritten documents. Wide variability in writing styles causes difficulties in accurately recognizing text, resulting in a decrease in the quality of the data produced. In addition, on a larger scale, there is a decrease in the efficiency of the OCR algorithm when archiving data. In this context, the importance of algorithm optimization becomes important to ensure the ability to process large amounts of data with sufficient speed and accuracy.

Not only technically but also in terms of policy, ethical issues have a key role in the application of OCR technology. Protection of personal data and efforts to maintain ethics in the use of this technology are important aspects related to existing regulations and privacy policies.

In facing these complex challenges, research in this area has great relevance. The main focus of this research is to develop more sophisticated and reliable OCR technology, especially for the recognition and processing of large documents. By continuing to explore and overcome existing challenges, it is hoped that the development of OCR technology can expand the scope of its application, increase visibility, and have a broad positive impact in various industries, especially those that depend heavily on document processing.

1.2. Problem Statement

1. How can we improve the reliability of OCR technology in recognizing characters and words in handwritten documents that have a wide variety of writing styles?

2. How can we optimize the OCR algorithm so that it can process data on a large scale with optimal speed and accuracy in the archiving process?

3. How can we overcome ethical issues related to the use of OCR technology in the context of protecting personal data and complying with existing privacy regulations?

1.3. Objectives

1. Develop OCR technology which is able to increase reliability in character and word recognition in handwritten documents with a variety of writing styles.

2. Develop an optimized OCR algorithm to process large amounts of data quickly and accurately in an archiving context.

3. Look for solutions and implementation that consider ethical issues related to personal data privacy in the use of OCR technology, in line with applicable privacy regulations.

2. Literature Review

2.1. Theory and Concepts

2.1.1. Character and Word Recognition Accuracy in OCR Documents

Character and word recognition accuracy in OCR documentsSeveral studies have been carried out to improve the accuracy of character and word recognition in OCR documents. In research by [1], the use of deep learning models, especially convolutional neural networks (CNN), succeeded in significantly increasing the accuracy of character and word recognition in OCR documents. They highlighted that by applying transfer learning, OCR was able to overcome complex representation constraints with 1000 object classes and was able to identify written characters well in offline and online modes.

Another study using a deep learning model was conducted by [2]. They show that approaches via artificial neural networks can strengthen character and word recognition in OCR documents. Apart from that, the use of character and word embedding is also able to increase the ability of artificial neural networks to recognize characters and words in OCR documents.

2.1.2. OCR Accuracy in Converting Characters and Words to Digital Text

Several studies have also tried to improve the accuracy of the process of converting characters and words to digital text in OCR. Research conducted by [3] highlights that the accuracy of the OCR process can vary depending on the quality of the image received by the OCR tool. They asserted that the accuracy of OCR can be improved through the use of image pre-processing methods such as binarization, noise reduction, and skew correction before performing OCR.

[4] also conducted similar research, which showed that the Otsu binarization algorithm was able to facilitate the thresholding process optimally. This allows images to be detected and identified more accurately without significant errors. This improvement certainly has an impact on the level of accuracy in the results of converting characters and words to digital form.

2.1.3. Use of OCR technology on a large scale

A number of studies have also been conducted to increase the effectiveness of using OCR technology on a wide scale. [5] conducted research that illustrates how machine learning and deep learning technology can improve the ability of OCR technology to handle large volumes of data. They highlight that the use of machine learning techniques such as decision trees, nearest neighbors, random forests, artificial neural networks, and support vector machines can make a significant contribution to the development of OCR technology on a wide scale.

On the other hand, research conducted by [6] emphasizes the importance of considering compliance with data privacy regulations such as GDPR in the European Union or applicable data privacy rules in relevant jurisdictions in the use of OCR technology regarding consumer data. They emphasize that ethical aspects and privacy protection in the use of consumer data must be the main focus of any research involving consumer data analysis.

2.2. Research Methodology

This research will be conducted using quantitative and qualitative research methods. Quantitative methods will be used to answer the first and second problem formulations. Meanwhile, qualitative methods will be used to answer the third problem formulation. The following are the research steps that will be carried out:

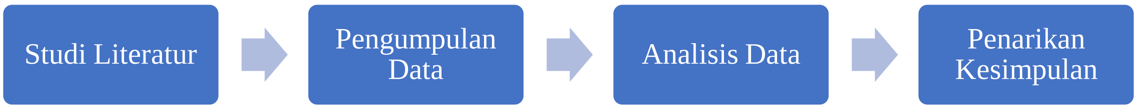


Figure 1. Research Stages

To answer the first problem formulation, an experiment will be carried out to compare the accuracy of conventional technology that uses deep learning methods with OCR technology that uses deep learning methods. This experiment will be carried out using an OCR document dataset consisting of various types of handwritten documents, such as text documents, image documents, and scanned documents. The data will be divided into two groups, namely the group using the conventional method and the group using the OCR method. Character and word recognition accuracy will be calculated for each group. The difference in accuracy between the two groups will be analyzed to find out whether the deep learning method with OCR can increase the reliability of the system in recognizing characters and words in handwritten documents.

To answer the second problem formulation, a regression analysis will be carried out to determine the factors that influence the performance of the OCR algorithm in processing data on a large scale. Regression analysis will be carried out using an OCR document dataset, which consists of various factors that can influence the performance of the OCR algorithm, such as document type, document quality, and document size. The data will be analyzed to determine the factors that have a significant influence on the performance of the OCR algorithm.

To answer the third problem formulation, literature studies and interviews will be conducted with experts and users of OCR technology to determine the impact and solutions for dealing with ethical issues related to the use of OCR technology in the context of protecting personal data and complying with existing privacy regulations. A literature study will be conducted to gather information relating to ethical issues related to the use of OCR technology. Interviews will be conducted with experts and users of OCR technology to obtain more in-depth information and opinions. Data obtained from literature studies and interviews will be analyzed to determine the impact and solutions to overcome ethical issues related to the use of OCR technology.

3. Result and Discussion

3.1. Comparision Experiment and Performance Analysis

A comparison was carried out on two program codes, which both carry out the process of recognizing and verifying signatures. The first program code uses the CNN algorithm to verify the signature, while the second program code uses the CNN algorithm combined with OCR (Optical Character Recognition) technology via the `pytesseract` library.

To carry out this experiment, the Handwritten signatures dataset was used, which was taken from Kaggle.com. This dataset contains real and fake signature samples from 30 people. Each person has five original signatures made by himself and five fake signatures made by other people. Below is a sample signature.

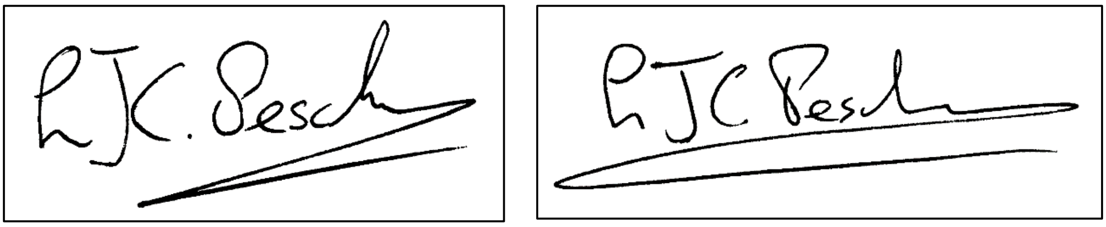


Figure 2. Fake and real signature samples

Before accuracy testing, the dataset is divided into training data and validation data with a ratio of 8:2. Next, a model evaluation was carried out using validation data to calculate accuracy, precision, recall, f1-score, and ROC AUC values. The following is a comparison table of evaluation results between the two program codes.

Table 1. Comparison of Test Matrix Results

Matriks	CNN tanpa OCR	CNN dengan OCR
Accuracy	0.72	0.53
Precision	0.72	0.52
Recall	0.70	0.73
F1-Score	0.71	0.61
ROC AUC	0.75	0.57
Confusion Matrix	$\begin{bmatrix} 22 & 8 \\ 9 & 21 \end{bmatrix}$	$\begin{bmatrix} 10 & 20 \\ 8 & 22 \end{bmatrix}$

Based on the results of the test matrix above, it can be explained as follows:

1. **Accuracy**

Accuracy is a measure of how often an OCR algorithm makes correct predictions. Accuracy can be calculated using the following formula:

Accuracy = (Number of Correct Predictions) / (Total Number of Predictions)

Based on the results of the measurement matrix, the accuracy of the CNN algorithm without OCR is 0.72, while the accuracy of the CNN algorithm with OCR is 0.53. This shows that the CNN algorithm without OCR has higher accuracy than the OCR algorithm with OCR.

2. Precision

Precision is a measure of how often an OCR algorithm's predictions are correct. Precision can be calculated using the following formula:

Precision = (Number of Positive Correct Predictions) / (Number of Positive Predictions)

Based on the test matrix results, the precision of the CNN algorithm without OCR is 0.72, while the precision of the CNN algorithm with OCR is 0.52. This shows that the CNN algorithm without OCR has higher precision than the CNN algorithm with OCR.

3. Recall

Recall is a measure of how many positive samples are detected by the OCR algorithm. Recall can be calculated using the following formula:

Recall = (Number of Positive Correct Predictions) / (Number of Positive Samples)

Based on the test matrix results, the recall of the CNN algorithm without OCR is 0.70, while the recall of the CNN algorithm with OCR is 0.73. This shows that the CNN algorithm with OCR has a higher recall than the CNN algorithm without OCR.

4. F1 Score

F1-Score is a combined measure of precision and recall. F1-Score can be calculated using the following formula:

F1 Score = (2 * Precision * Recall) / (Precision + Recall)

Based on the test matrix results, the F1-Score of the CNN algorithm without OCR is 0.71, while the F1-Score of the CNN algorithm with OCR is 0.61. This shows that the CNN algorithm without OCR has a higher F1-Score than the OCR algorithm with OCR.

5. ROC AUC

ROC AUC is a measure of the performance of an OCR algorithm in differentiating between positive classes and negative classes. ROC AUC can be calculated using the following formula:

ROC AUC = Area under the ROC curve

Based on the test matrix results, the ROC AUC of the CNN algorithm without OCR is 0.75, while the ROC AUC of the CNN algorithm with OCR is 0.57. This shows that the CNN algorithm without OCR has a higher ROC AUC than the CNN algorithm with OCR.

6. Confusion Matrix

Confusion Matrix is a table that shows how the OCR algorithm's predictions compare to the actual labels of the test data. Confusion Matrix can be used to understand the weaknesses and strengths of OCR algorithms.

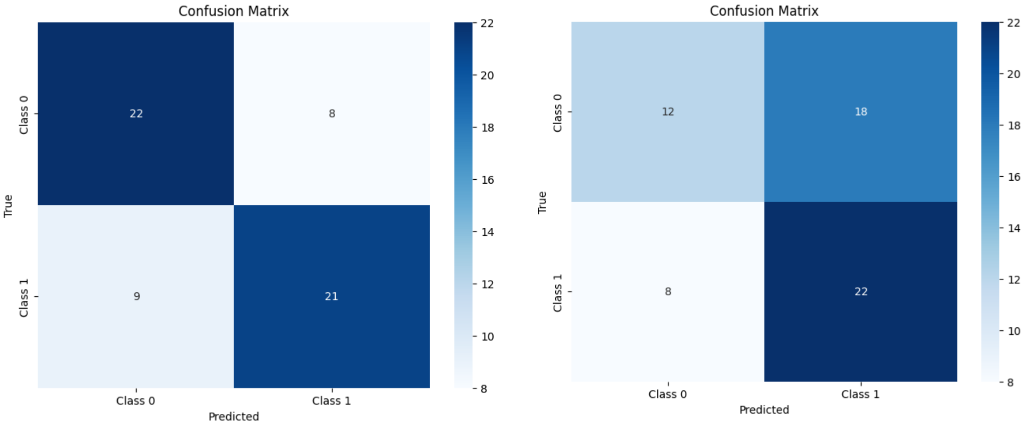
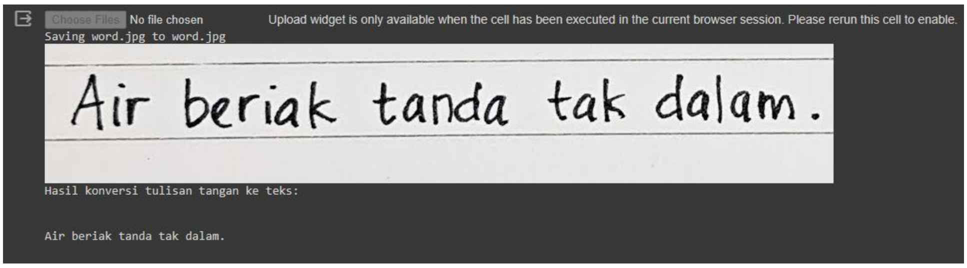


Figure 3. Matrix convolution results

Based on the confusion matrix above, it can be seen that the CNN algorithm without OCR more often makes negative class errors (false negative), while the CNN algorithm with OCR more often makes positive class errors (false positive).

As for the handwriting conversion program, no comparison was made because the code only applies OCR technology without any combination with any classification algorithm. In addition, creating a handwriting conversion program that does not implement OCR can be more complicated. A common method is to use advanced image processing techniques and associated modeling to extract text from images. Some approaches include text segmentation, character detection, pattern recognition, processing, and text normalization.

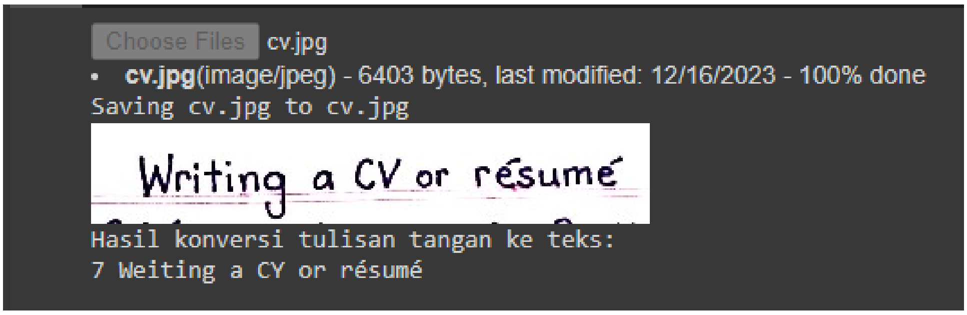
Even so, the output resulting from using OCR technology in the handwriting conversion process is almost perfect. Below are some samples of test results for writing conversion using OCR technology.


Choose Files
No file chosen
Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Air beriak tanda tak dalam.

```
Saving word.jpg to word.jpg
```

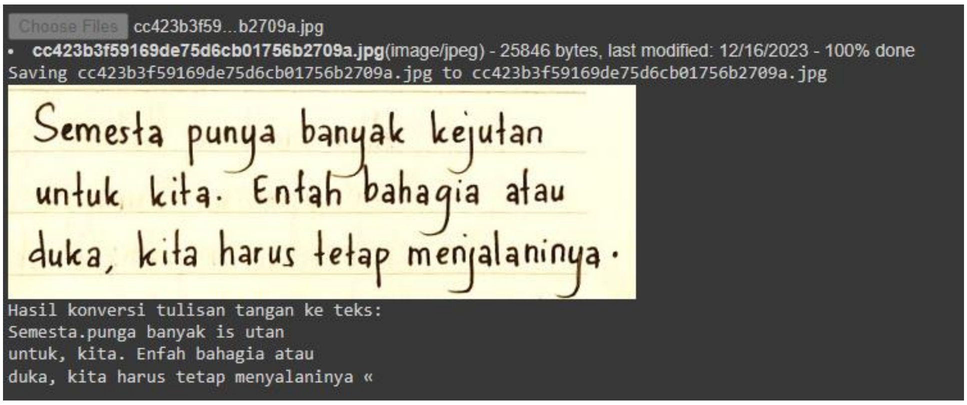
```
Hasil konversi tulisan tangan ke teks:


Air beriak tanda tak dalam.
```

Figure 4. First attempt at signature verification



```
cv.jpg
• cv.jpg(image/jpeg) - 6403 bytes, last modified: 12/16/2023 - 100% done
Saving cv.jpg to cv.jpg
Hasil konversi tulisan tangan ke teks:
7 Weiting a CY or résumé
```

Figure 5. Second attempt at signature verification


Choose Files
Semesta punya banyak kejutan
untuk kita. Entah bahagia atau
duka, kita harus tetap menjalaninya.

```
cc423b3f59...b2709a.jpg
• cc423b3f59169de75d6cb01756b2709a.jpg(image/jpeg) - 25846 bytes, last modified: 12/16/2023 - 100% done
Saving cc423b3f59169de75d6cb01756b2709a.jpg to cc423b3f59169de75d6cb01756b2709a.jpg
Hasil konversi tulisan tangan ke teks:
Semesta.punga banyak is utan
untuk, kita. Enfah bahagia atau
duka, kita harus tetap menyalaninya «
```

Figure 6. Third attempt at signature verification

3.2. Ethical Issues Related to the Use of OCR Technology

Based on a literature study regarding ethical issues related to the use of OCR technology, several relevant issues were found. One of the most important ethical issues regarding the use of OCR technology is privacy. OCR technology can be used to extract text from documents that contain personal information, such as names, addresses, telephone numbers, and bank account numbers. This information can be used for undesirable purposes, such as to commit fraud or misuse personal data.

Another ethical issue related to the use of OCR technology is accuracy. OCR technology is not always accurate, especially for documents that are of poor quality or contain handwriting. Errors in OCR can lead to an incorrect interpretation of the information contained in the document.

OCR can also be used to discriminate against people based on their characteristics, such as race, gender, or age. For example, OCR can be used to reject job or loan applications based on inaccurate information about an applicant's race or gender. OCR can also be used to access confidential or sensitive information. If the OCR system is not properly secured, this information can fall into the hands of unauthorized persons.

OCR technology is a useful technology, but it is important to be aware of the ethical issues associated with its use. Users of OCR technology must take steps to protect the privacy, accuracy, fair treatment, and security of information processed by OCR technology. Here are some recommendations for addressing ethical issues related to the use of OCR:

1. Users of OCR technology must be transparent about how they use the technology and what information is collected.
2. Users of OCR technology must give individuals control over their information. Individuals must be able to access, update, and delete their information.

3. Users of OCR technology must secure the information processed by the technology.

By addressing these ethical issues, we can ensure that OCR technology is used responsibly and does not pose risks to individuals' privacy, accuracy, fair treatment, or security.

4. Conclusion

From the comparison of the two programs, the CNN algorithm without OCR exhibited superior performance with higher accuracy, precision, and F1-Score compared to the CNN algorithm incorporating OCR. Despite a slightly higher recall in CNN with OCR, the algorithm without OCR excelled in identifying negative classes. Data quality, document size, and hardware specifications significantly impacted OCR performance. Ethical issues, particularly privacy, accuracy, and potential discrimination, are crucial concerns. Recommendations encompass enhancing data quality, further evaluating OCR performance, and implementing measures to safeguard data security and privacy.

1. Further evaluation of the reasons for low classification quality when using OCR.
2. Consider updates to OCR technology or improvements in training data to enhance classification performance.
3. Ensure the security and privacy of data processed by OCR technology to mitigate the risks of misuse or leakage of sensitive information.

Given this comparison, continuous development and security of OCR technology remain imperative for individual information privacy, accuracy, and security.

Acknowledgements

The author would like to thank God Almighty and his parents who always provide support in the form of moral and material.

References

[1] K. O. Mohammed Aarif and S. Poruran, “OCR-Nets: Variants of Pre-trained CNN for Urdu Handwritten Character Recognition via Transfer Learning,” in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 2294–2301. doi: 10.1016/j.procs.2020.04.248.

[2] S. Gajendran, M. D, and V. Sugumaran, “Character level and word level embedding with bidirectional LSTM – Dynamic recurrent neural network for biomedical named entity recognition from literature,” *J Biomed Inform*, vol. 112, Dec. 2020, doi: 10.1016/j.jbi.2020.103609.

[3] A. Ignasius, J. C. Chandra, R. Oscadinata, and D. Suhartono, “Image Pre-Processing Effect on OCR’s Performance for Image Conversion to Braille Unicode,” *Procedia Comput Sci*, vol. 227, pp. 922–931, 2023, doi: 10.1016/j.procs.2023.10.599.

[4] A. Y. Sugiyono, K. Adrio, K. Tanuwijaya, and K. M. Suryaningrum, “Extracting Information from Vehicle Registration Plate using OCR Tesseract,” *Procedia Comput Sci*, vol. 227, pp. 932–938, 2023, doi: 10.1016/j.procs.2023.10.600.

[5] A. Singh and A. S. Bist, “A Wide Scale Survey on Handwritten Character Recognition using Machine Learning,” *International Journal of Computer Sciences and Engineering*, vol. 7, no. 6, pp. 124–134, Jun. 2019, doi: 10.26438/ijcse/v7i6.124134.

[6] M. M. Mariani, M. Borghi, and B. Laker, “Do submission devices influence online review ratings differently across different types of platforms? A big data analysis,” *Technol Forecast Soc Change*, vol. 189, Apr. 2023, doi: 10.1016/j.techfore.2022.122296.