Detection and Recognition Algorithms for Extracting Text from Stained Documents using Deep Learning

1st Marie Emmanuelle Tacsay
Department of Computer Engineering
Technological Institute of the Philippines
Quezon City, Philippines
qmettacsay@tip.edu.ph

2nd Daniel Kimbell Yu
Department of Computer Engineering
Technological Institute of the Philippines
Quezon City, Philippines
qdkmyu@tip.edu.ph

3rd Roman Richard
Department of Computer Engineering
Technological Institute of the Philippines
Quezon City, Philippines
rrichard.cpe@tip.edu.ph

Abstract*—Documents are one of the most important physical records which can contain vital information which are vulnerable to degradation such as stains, fading, and spilled ink. Retrieving information from these would be crucial for the maintenance of the records of such events or knowledge. This study was conducted for the purpose of retrieving the important information from stained documents, and to help retrieve this information from disappearing through the application of deep learning and OCR models. The stained documents would be denoised through autoencoders and other deep learning models. Among the deep learning models used, Autoencoder showed great performance when it came to denoising images. Once complete, other preprocessing methods would be applied to the image such as: binarization, edge detection, and so on, before feeding to the OCR model. The OCR models would be used for both detection and extraction of the information printed onto the document and output into readable text. This study tested three models: Tesseract OCR, Easy OCR, and Keras OCR, with Tesseract OCR being the best in terms of performance among the three. When comparing all three models, Tesseract OCR was observed to have the lowest Character Error Rate (CER), as well as, Word Error Rate (WER). With this being the case, the researcher concluded that from among the three, Tesseract OCR performed best in terms of detection and recognition.

Keywords - Denoising, Deep Learning, OCR, Scanned Documents, Autoencoder, Tesseract, CNN, Text Extraction, Text Recognition

I. Introduction

Dating back to the stone age period, manuscripts since the beginning have been referred for various vital information to this day. These manuscripts take form in different ways, examples of such are palm leaf, paper, stone, etc. These manuscripts contained different information ranging from history to medicine to astronomy, the paper manuscript in particular is still used up to this day, but though these manuscripts or documents rather contain information regarding a wide variety of information, is prone to the threat of degradation, due to physical and chemical effects [1].

As previously mentioned, documents contain various useful information in the form of text, graphics and pictures that depict the records of various contents that impose significant information [3]. These printed documents are commonly stored as images which then can be retrieved using tools or softwares [4]. With this being said, there are several factors which affect the condition of recognition of these texts, such conditions include: low quality paper, color deformation, low contrast of the text, different shapes of alphabets due to the use of fonts that are no longer standard, and so on. These factors, along with the possible fading of the ink prove to challenge the recovery of the contents of the information within these documents and causes failures in even state-of-the-art Optical Character Recognition (OCR) methodologies [2].

For the OCR model to return high accuracy results, the quality of the scanned documents must first be considered and be rendered in high quality. With this, denoising images would be implemented and this technique has attracted the interest of researchers which still remains challenging and open for improvement [5]. This study aims to use deep learning models for denoising or cleaning of images, as well as OCR algorithms for detection and recognition of text to solve and improve the retrieval of information from stained documents. Through the mentioned methods, text extraction would be possible which can be of use to retrieve information from aforementioned documents.

II. Methodology

The methods used to conduct this study are composed of description of the dataset, image pre-processing, visualization of the dataset, optical character recognition (OCR), denoising models, and the performance metrics used. These make up the whole methodology or methods which the study would follow to generate and analyze results.

A. Dataset

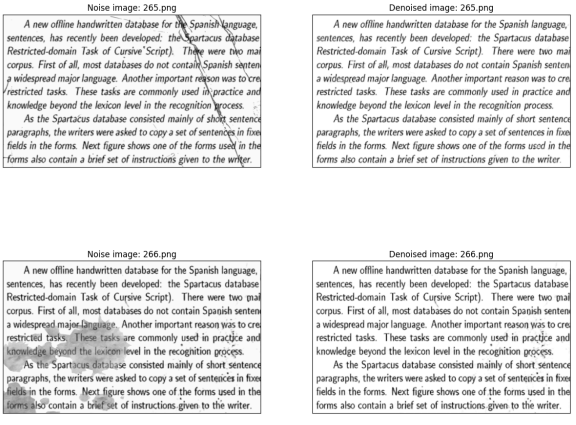

Fig. 1. Sample Image of Denoised Documents

The dataset was composed of 504 images which came from the dataset used in the study, Denoising Text Image Documents using Autoencoders conducted by Gregory, F [6] and mixed with the data from the authors of this study. The data were used for training and testing of the model

used in this study. The images included in the dataset are purely scanned printed documents in which other instances such as handwritings were not included. In addition, the document images have its noise such as bleeding of ink, marks of folded paper, fading ink, etc., which this study would be needed for training denoising autoencoder model to enhance its quality for text detection and extraction.

B. Image Pre-processing

The images in the dataset were subject to various preprocessing methods. These methods consisted of resizing the images to (360x612). These images would later be grayscaled and fed into the denoising models for cleaning, other preprocessing methods were also done, such methods include: Image Sharpening, Binarization, and Edge Detection. These methods would be used to optimize the results of the Optical Character Recognition (OCR) Model.

C. Optical Character Recognition (OCR)

Optical Character Recognition or otherwise known as OCR, for short, refers to the process in which the conversion of image into a machine-readable text format, examples of such documents are not limited to documents but can also refer to signs, billboards, and logos. OCR can be seen in different applications, ranging in various languages, and not being limited to only English. OCR can be broken down into four different steps to achieve it: 1) Image Acquisition, 2)Text Detection, 3) Text Recognition, and 4) Post Processing. The three (3) models that were used and compared in the study were as follows: Tesseract OCR, Easy OCR and Keras OCR.

Tesseract Technology supports short-short memory neural networks, which can be utilized for the recognition of english text [7] Tesseract is an open-source project which is owned by google. The reason why Tesseract is more accurate than other OCR engines, is due to it checking for any missing words or mispelled words and if there are any present and then correcting it [8]. Easy OCR utilizes Python and Pytorch deep learning library this OCR uses the Character Region Awareness for Text (CRAFT) algorithm for detection, and Convolutional Recurrent Neural Network (CRNN) for recognition, furthermore, the OCR consists of three parts: decoding which uses a Connectionist Temporal Classification (CTC) algorithm, feature extraction that utilizes ResNet VGG and Long Short-Term Memory (LSTM) network for interpretation of these features and sequence labeling [9][10]. In addition, Keras OCR supported by Keras and Tensorflow, is based on the architecture of the CRNN model which provides the OCR its ability to detect and recognize the text in an image [11]. Also, it provides a framework for training and deploying other OCR models using deep learning techniques and an end-to-end training pipeline for building new OCR models [12].

D. Deep Learning Models for Denoising Images

Autoencoder is considered as one of the deep learning models that is commonly used in various applications and one of these applications is to clean images. This is often used as an image denoiser of various images which generates new cleaned images for a better visualization of the data [14]. They have layers where the images will be processed to generate new and close-to-original images. In addition, this study used CNN or Convolutional Neural Network models to compare and observe the performance of these models in terms of cleaning images. CNN models are known and proven to be a good choice in classification of data [13] in which a good example is classifying images of handwritten digits. According to Zhang, J., Niu, Y., Shangguan, Z., et al, (2023) CNN image denoising models have many layers, require many parameters and have high computational cost during training of the data.

With this, using CNN models would be efficient in denoising images but it would require a large size of resources needed in order to properly execute the processes. The two (2) CNN models used in this study are U-Net and Noise2Void models. First, the U-Net model has been introduced for the purpose of Biomedical Image Segmentation [15]. There are various studies wherein U-Net was modified and expanded its applications other than for the purpose of image segmentation [16]. Because of its flexibility, the model was able to learn to denoise images and having variants based on its architecture in such example of the models conducted in the study of Komatsu, R. and Gonsalves, T. Furthermore, the U-Net model used upsampling operations instead of pooling operations as its layers [13] which made the model a candidate for denoising images.

The last second CNN model used in this study is the Noise2Void model which has convolutional layers that are proposed to use in training images. The model presents novel training methods that do not require clean target images for the model to use as a basis in denoising dirty images [17]. This model also showed promising performance in terms of denoising as it is also used by various studies to denoise images.

E. Performance Evaluation Metrics

To measure the performance of the denoising models the used metrics are the following: Mean Squared Error (MSE) Loss and Mean Absolute Error (MAE). As for measuring the performance of the Optical Character Recognition (OCR) models, Character Error Rate (CER) and Word Error Rate (WER) were used, where lower CER and WER is preferred. Furthermore, when calculating for the accuracy of the model, it would simply be the difference between the error rate attained to 100%. This would in turn be the indicator of the study on the performance of the model.

Mean Squared Error (MSE) Loss is a loss function that quantifies the magnitude of the error between the predicted and actual value through taking the squared difference between the prediction and the target values. This metric is calculated through the following formula below:

$$MSE = \left(\frac{1}{n}\right) \bullet \Sigma(actual - forecast)^2 \qquad (1)$$

Whereas, the Mean Absolute Error (MAE) also refers to the magnitude of the error between the predicted and actual values; however, it takes the average of the absolute errors of the predictions and observations as the measurements of the magnitude of errors of a group of data or values. It can be calculated through the following formula below:

$$MAE = \left(\frac{1}{n}\right) \bullet \Sigma|actual - forecast| \qquad (2)$$

Character Error Rate (CER) is a metric which measures the accuracy of predicted text through – substitutions, deletions. This differs from word-level errors, this metric is

useful in surfacing mispronunciations and erroneous phonemes. This metric is calculated using the following formula:

$$CER = \frac{Substitutions + Deletions + Insertions}{\#\ of\ words\ in\ Reference} \quad (3)$$

Similarly to CER, Word Error Rate (WER), is a metric that also measures the text based on three different types of errors - substitutions, deletions, and insertions. Though CER and WER both use these three types of errors as their basis, Word-level error surface mispredicted words, This metric is useful to visualize the common word-level failures that exist in the model, as so we are able to flesh out the weaknesses present. WER is be calculated using the following formula:

$$WER = \frac{Substitutions + Deletions + Insertions}{\#\ of\ words\ in\ Reference} \quad (4)$$

III. Results and Discussion

Before being subjected to Optical Character Recognition, three different models were tested and compared when denoising the images. These models being: Autoencoders, U-Net, and Noise2Void. The following are the results obtained from the different denoising models. Firstly the three models were used with the visualization of images.

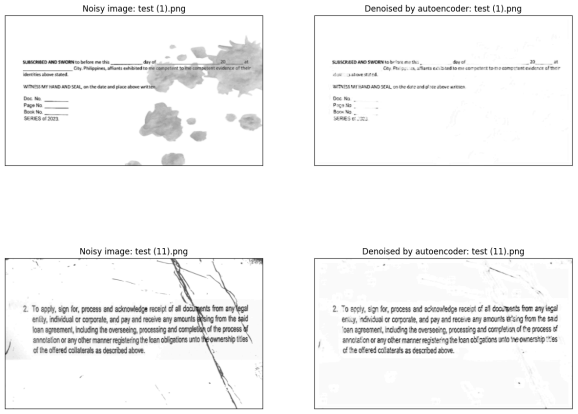

Fig 2. Autoencoder Denoising Visualization

As seen in Figure 2, the autoencoder model was able to effectively clean the image. The autoencoder was able to fully clear out the dark stain on the image making it easier to read.

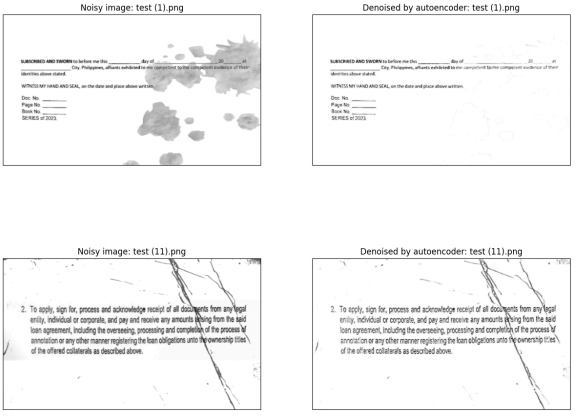

Fig 3. U-Net Denoising Visualization

Based on Figure 3, U-Net proved its quality when it comes to denoising images which can be seen from the images shown above. Most of the document images were cleaned because of the model and its ability to denoise or clean images.

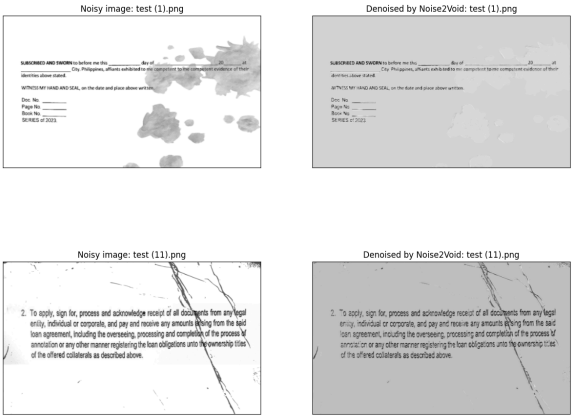

Fig 4. Noise2Void Denoising Visualization

According to Figure 4 above, Noise2Void did not properly denoised the images provided because there are still staines left and the generated denoised images have gray background.

From all the results generated by the three (3) models tested in this study, autoencoders gave the best performance compared to others. The denoised images of the autoencoder model being cleaner as compared to the other models. In addition, U-Net and Noise2Void models still left stains on the images after processing and denoising them which made the autoencoder be the best to perform when it comes to denoising images. Thus, the autoencoder model would be used to denoise images for text detection and extraction using OCR and will be used for the model deployment in web applications.

To determine the accuracy of the result, the CER and WER of the predicted test was taken and compared with the ground truth. Below are the results of the following tests.

TABLE I. Accuracy Results of OCR Model in CER

Model	Image 1		Image 2	
	WER	***CER***	***WER***	***CER***
Tesseract OCR	98.63%	99.77%	91.30%	98.97%
Easy OCR	91.78%	98.62%	95.65%	99.31%
Keras OCR	16.44%	37.33%	39.13%	65.64%

Table I above shows the accuracy results of the OCR model between two images for both CER and WER, with Tesseract OCR being the best performing of the three models.

To summarize, from the given results of the above, the top performing models, when tested by the researchers, for denoising were Autoenders and Tesseract OCR for the text recognition portion; these two models showed promising results from among the different models. As seen in the visualization of results for autoencoders, it was able to effectively clean most of the noise, and TesseractOCR having

the lowest CER and WER from among the three models, entails that it is able to closely predict the words on the image.

To test the best performing denoising model and OCR algorithm used, this study utilized the web application deployed through Streamlit, to verify the consistency of the performance of the mentioned methods.

Fig 5. Web Application

After the model was deployed, the researchers had gone through multiple images to verify the performance of the model when faced with a use-case. The number of images tested on were a total of 10 images.

TABLE II. Accuracy Results of Deployed Model

Image	WER	CER
1	93.48%	97.59%
2	98.63%	98.16%
3	98.95%	97.11%
4	97.5%	96.8%
5	97.30%	96.58%
6	95%	96.85%
7	96.85%	97.63%
8	98.59%	96.51%
9	94.44%	96.76%
10	97.30%	97.38%
Total	**96.80%**	**97.14%**

Based on Table II above, the model deployed was tested on ten different images, and was able to achieve an overall 96.80 % accuracy in detection and recognition in terms of word-level wise or the inverse of WER. Whereas, the model attained a 97.14% accuracy in detection and recognition in terms of character-level wise or the inverse of CER. The results indicate to us that the model is able to accurately predict the text within the image, at a text-wise level.

IV. Conclusion

From the different tests that were performed on different models, it can be concluded that from these tests, Autoencoders and Tesseract OCR respectively performed the best for denoising and text recognition. Autoencoders were able to generally clean the document, whilst maintaining the text printed onto the document with minimal erasures. Whereas Tesseract OCR yielded the lowest CER and WER Rate, which would mean that from among the three, it had the least amount of substitutions and erasure when compared to the actual text. This was further supported by the results of the deployed model which also yielded similar results.

References:

[1] T. J. Alexander, S. S. Kumar and B. Sowmya, "Performance Analysis of Fuzzy based Restoration Technique for Ink Bleed-through Degraded Documents," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2020, pp. 1429-1434, doi: 10.1109/ICECA49313.2020.9297394.

[2] C. Biswas, P. S. Mukherjee, K. Ghosh, U. Bhattacharya and S. K. Parui, "A Hybrid Deep Architecture for Robust Recognition of Text Lines of Degraded Printed Documents," 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 2018, pp. 3174-3179, doi: 10.1109/ICPR.2018.8545409.

[3] D. Ghai and N. Jain, "Text Extraction from Document Images- A Review," *International Journal of Computer Applications*, vol. 84, no. 3, p. 40, 2013, Accessed: Apr. 28, 2024. [Online]. Available: https://www.academia.edu/50741354/Text_Extraction_from_Document_Images_A_Review?sm=b

[4] Gangeh, M. J., Tiyyagura, S. R., Dasaratha, S. V., Motahari, H., & Duffy, N. P. (n.d.). *Document enhancement system using auto-encoders*. OpenReview. https://openreview.net/forum?id=S1Mnzp9qLB

[5] M. Ibrahim *et al.*, "Denoising Letter Images from Scanned Invoices Using Stacked Autoencoders," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 1371–1386, 2022, doi: https://doi.org/10.32604/cmc.2022.022458.

[6] F. Gregory, "Denoising Text Image Documents using Autoencoders," *ResearchGate*, 2021. https://www.researchgate.net/publication/356423394_Denoising_Text_Image_Documents_using_Autoencoders (accessed May 25, 2024).

[7] J. Li, "Research on English Automatic Recognition System Based on OCR Technology," 2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT), Jiaxing, China, 2023, pp. 1061-1066, doi: 10.1109/ACAIT60137.2023.10528556.

[8] S. Sundara Pandiyan and C. Kelvin, "A Translator for Indian Sign Boards to English using Tesseract and SEQ2SEQ Model," 2021 International Conference on Simulation, Automation & Smart Manufacturing (SASM), Mathura, India, 2021, pp. 1-4, doi: 10.1109/SASM51857.2021.9841215.

[9] C. Jeeva, T. Porselvi, B. Krithika, R. Shreya, G. S. Priyaa and K. Sivasankari, "Intelligent Image Text Reader using Easy OCR, NRCLex & NLTK," 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), Chennai, India, 2022, pp. 1-6, doi: 10.1109/ICPECTS56089.2022.10047136.

[10] A. Mahajan, "EasyOCR: A Comprehensive Guide," *Medium*, Oct. 29, 2023. https://medium.com/@adityamahajan.work/easyocr-a-comprehensive-guide-5ff1cb850168

[11] faustomorales, "faustomorales/keras-ocr," *GitHub*, Dec. 08, 2019. https://github.com/faustomorales/keras-ocr

[12] "OCR COMPARISONS - TESSERACT, EAST, AND KERAS OCR," *www.linkedin.com*. https://www.linkedin.com/pulse/ocr-comparisions-tesseract-east-keras-serkan-erdonmez/ (accessed May 27, 2024).

[13] C. Wiraatmaja, K. Gunadi and I. N. Sandjaja, "The Application of Deep Convolutional Denoising Autoencoder for Optical Character Recognition Preprocessing," 2017 International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT), Denpasar, Indonesia, 2017, pp. 72-77, doi: 10.1109/ICSIIT.2017.32.

[14] L. Gondara, "Medical Image Denoising Using Convolutional Denoising Autoencoders," 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 2016, pp. 241-246, doi: 10.1109/ICDMW.2016.0041.

[15] M. Tripathi, "Facial image denoising using AutoEncoder and UNET," *Heritage and Sustainable Development*, vol. 3, no. 2, pp. 89–96, Jul. 2021, doi: https://doi.org/10.37868/hsd.v3i2.71.

[16] R. Komatsu and T. Gonsalves, "Comparing U-Net Based Models for Denoising Color Images," *AI*, vol. 1, no. 4, pp. 465–487, Oct. 2020, doi: https://doi.org/10.3390/ai1040029.

[17] T. -A. Song and J. Dutta, "Noise2Void Denoising of PET Images," 2020 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), Boston, MA, USA, 2020, pp. 1-2, doi: e10.1109/NSS/MIC42677.2020.9507875.