Chapter 45
Efficient Image Denoising for Effective Digitization Using Image Processing Techniques and Neural Networks

K.G. Srinivasa
CBP Government Engineering College, India

B.J. Sowmya
M.S. Ramaiah Institute of Technology, India

D. Pradeep Kumar
M.S. Ramaiah Institute of Technology, India

Chetan Shetty
M.S. Ramaiah Institute of Technology, India

ABSTRACT

Vast reserves of information are found in ancient texts, scripts, stone tablets etc. However due to difficulty in creating new physical copies of such texts, knowledge to be obtained from them is limited to those few who have access to such resources. With the advent of Optical Character Recognition (OCR) efforts have been made to digitize such information. This increases their availability by making it easier to share, search and edit. Many documents are held back due to being damaged. This gives rise to an interesting problem of removing the noise from such documents so it becomes easier to apply OCR on them. Here the authors aim to develop a model that helps denoise images of such documents retaining on the text. The primary goal of their project is to help ease document digitization. They intend to study the effects of combining image processing techniques and neural networks. Image processing techniques like thresholding, filtering, edge detection, morphological operations, etc. will be applied to pre-process images to yield higher accuracy of neural network models.

DOI: 10.4018/978-1-5225-5204-8.ch045

1. INTRODUCTION

In our country, most of the documents are in the form of paper records or documented in registers. Due to time factor, most of the documents have become old, dirty and unreadable. The information contained in these documents mainly consists of patient data, population statistics and few other important information, if lost might have an adverse impact on important decisions like budget. Developing efficient methods for digitizing text documents would result in availability of information in ancient and medieval text, besides serving as a safe storage mechanism. This digitization would make the editable, searchable contents and easier to share. Optical Character Recognition (OCR) techniques are currently used to extract text from handwritten and typed documents. Input to the OCR is an image of the document. It uses image processing and machine learning techniques to detect text. Image processing is used to analyse various colours in the text, and using machine learning techniques we plan to remove the noise using a dataset of images containing such scanned text. Machine learning aids in filtering those colours to yield the text. Despite much advancement, the accuracy is low due to various difficulties in processing the image.

Digitizing of numerous documents are put on hold. Some books have coffee stains, water, paint marks, faded sun spots, dog-eared pages, faded sunspots, lots of wrinkles etc. Lot of wrinkles is the reason that has kept some printed documents from being digitized. These drastically affect accuracy of OCR making it impossible to use in some cases. Hence, our project focuses on eliminating this noise from scanned text images. Feature engineering and use of neural networks seem to be of greater promise. Our work would involve examining these methods, using them in whole or part, to find an effective solution. There exist many approaches to convert these dirty documents to clean document. Naive solutions include Least Square Regression, and thresholding techniques. Background removal from damaged documents to increase accuracy of OCR technique. The main idea is to convert these dirty documents to ones that have scanned text only content.

The primary goal of our project is to help improve the ease of document enhancement. By doing so, the time taken to convert the ancient manuscripts and texts to a digital format will be reduced with a high accuracy. The dataset consists of two sets of images, train dataset and test dataset. These images hold different styles of text, to which synthetic noise has been introduced to simulate real-world, messy artifacts. The objective of this project is to design and implement a model with a high degree of accuracy and draw a comparison regarding the performance of models developed using different neural networks.

Goal can be achieved by developing several different models using CNN, DNN, and boosted trees for background removal using information leakage. The architecture to train and run the model would be developed either using Theano (depending on computing resources available) or lua's torch library.

Hence we plan to make a website for GUI (Graphical User Interface)using python flask so that it would be easy for the end user to upload the dirty images or unclean images from their local machine or from cloud storage. Post processing and computation of the dirty image, clean image will be displayed on the website so that the user will be able to download the clean image onto his local machine or choose to share it on cloud.

Hence the current scope of the project is to help OCR by removing dirty stains or noise from the old documents and digitalize it so that retrieving and processing the information is much faster and efficient thereby preventing any loss in the information or knowledgebase of any organization.

The future scope of the project would be to go a step further, to try and retrieve the text from the image. That is, when an end user feeds a dirty image, the clean image will be produced and the text in the image will be retrieved. Hence saving a lot of time and manual work of trying to digitalize a docu-