Bachelor Thesis:

Digitizing Receipt using Machine Learning

Department of Computer Science

Author:

Nguyen Tien Hung - 9972

Supervisor:

Dr. Huynh Trung Hieu

Dr. Nguyen Viet Linh

September 1, 2019

Abstract

Computer Vision and Natural Languages Processing has archived accomplishments in digitizing document and document classification. This paper present a machine learning approach for recognizing scanned receipts. The recognition system has two main modules: text detection based on Connectionist Text Proposal Network (CTPN) and text recognition based on Long short-term memory (LSTM) network.

Acknowledgement

This is the Acknowledgement line

Contents

1 Introduction **7**

1.1 Motivation . . . 7

1.2 Purpose . . . 8

1.3 Related work . . . 8

1.4 Approach and methodology . . . 8

1.5 Scope and Limitation . . . 9

1.6 Outline . . . 9

2 Research Background **10**

2.1 Data Processing . . . 10

2.2 Computer Vision . . . 10

2.2.1 Overview . . . 10

2.2.2 OpenCV . . . 10

2.2.3 Gaussian Blur . . . 11

2.2.4 Canny edge detection . . . 11

2.2.5 Hough transform . . . 12

2.3 Machine Learning . . . 12

2.3.1 Overview . . . 12

2.3.2 Deep learning . . . 12

2.3.3 EAST . . . 13

2.3.4 Connectionist Text Proposal Network . . . 13

2.3.5 Recurrent neural network . . . 14

2.3.6 Long short-term memory . . . 14

2.3.7 Generative adversarial network . . . 14

2.4 Optical Character Recognition . . . 15

2.4.1 Tesseract OCR . . . 15

2.5 Server . . . 15

3 Methodology **16**

3.1 Image processing . . . 16

3.1.1 . . . 17

3.2 Text-line detection . . . 17

3.2.1 EAST . . . 17
3.2.2 CTPN . . . 17
3.2.3 Comparison . . . 18
3.2.4 Merge Boxes . . . 18
3.3 Text Recognition . . . 18
3.3.1 Tesseract OCR . . . 18
3.4 Intent Classification . . . 18

4 Implementation **19**
4.1 Image processing . . . 19

5 Result and Discussion **20**

6 Conclusion **21**
6.1 Conclusions . . . 21
6.2 Further research . . . 21

List of Figures

2.1 Image applied Gaussian Blur 11
2.2 Image applied Canny Edge Detector 11
2.3 FCN structure of EAST text detection FCN 13
2.4 Architecture of CTPN 14
2.5 CTPN Text detection on a receipt image 15

3.1 Background filtered out image 16
3.2 CTPN detection result on receipt image 17
3.3 Merge Boxes 18

List of Tables

Abbreviations

AI Artificial Intelligence. 7

BiLSTM bi-directional Long short-term memory. 13, 14

CTPN Connectionist Text Proposal Network. 1, 4, 5, 13–18

EAST An Efficient and Accurate Scene Text Detector. 4, 5, 13, 17

FCN Full Convolutional Network. 5, 13

ICDAR International Conference on Document Analysis Recognition. 13, 14

LSTM Long short-term memory. 1, 14

NLP Natural Language Processing. 7

OCR Optical Charater Recognization. 10

R-CNN Region Convolutional Neural Networks. 13

RNN Recurrent neural network. 14

VGG16 16-layer vggNet. 13

Chapter 1

Introduction

Computer vision is a branch of Artificial Intelligence (AI) technology that has already entered our lives and businesses in ways many of us may not be aware of. For example, computer vision allows banking customers deposit remotely by captures the image of the check destined for deposit in the bank, then verifies if the signature on the check is genuine. In 2017, Amazon released Echo Look for offering then compares the outfit with options it suggests and delivers the user an overall style rating based on user's full-body picture.

Natural Language Processing (NLP) is also another branch of AI that allows computers manipulate, interpret and understand human language. NLP has participated in our daily life from small add-ons such as auto-complete or predictive-typing to spam email recognition or document classification. Online translator like Google Translate has included NLP for understanding text situation to produce fluent text in output language.

Digitizing Receipt is a Machine Learning approach for recognizing receipt and saving as structured data inspired by the success of Computer Vision and NLP. The purpose of the system is for converting receipt image to structured data for enables a variety of applications.

1.1 Motivation

A collection of facts called data is an unlimited resource represent in many forms. Previously, people used paper to keep data such as contract, report in human-readable also known as unstructured data such a text or image, it requires a person to interpret it. Nowadays, the size of data is grown significantly, we changed the data storage approach to structured data refers to information that computer can process, manipulating data from a set of instructions. The problems in this transformation is previous data in the unstructured way still need human to converting the data structure, this is a time-consuming work and requires much human resource.

Recognizing the problem has inspired me to participate in the data revolution, save human resource for other issues and save the data structure converting time.

One of my behavior is tracking my personal finance in order to reach further financial target. The tracking process is not require much effort but I have to remember save in every time I have spent money. In

this research, I have found a more convenient approach for financial tracking that automatically converting receipt image into digital data and computer understandable.

1.2 Purpose

Currently people are changing the information storage method, data is saved under structured data, computer-understandable as the very first step to enable variety of application. As more and more information becomes digitized, it means that analysts can start to use it as data for different purposes. For example, in the business, with the data, we can observe the changes over years and might understand customer's demands. In this thesis, machine learning technique will be used in text line detection, text recognition and data field detection from extracted text. Further more, an approach on converting data into structured form for different applications.

1.3 Related work

An idea that automatically digitizing data which converting unstructured data into computer-readable type is not new. The Electrical Telegraph is an point-to-point system for telegraphy, using the Morse Code In 1816, In 1976, Ray Kurzweil has developed a machine that read characters and converted them into standard telegraph code [14].

1.4 Approach and methodology

This thesis focus on using machine learning method digitizing receipt for extract fields data from receipt image. The process is divided into two main steps, conversion of receipt image into machine-encoded text and field extraction.

First, the receipt image will be processed for eliminating noise and meaningless area before using Tesseract OCR[16] for converting cropped image into machine-encoded text. For enhancing accuracy of area elimination from the first step, this work focus on detect paper and text line inside the paper for increasing data quality.

In the second part of the digitizing process is field extraction from machine-encoded data. Before extracting field, intent classification is applied for allocating part into suitable extractor from five intents:

- Brand Name
- Information
- Index
- Content
- Total
- Thank you

The reason and purpose of each intention will be discussed in chapter 3.

- Field extraction

1.5 Scope and Limitation

Due to the thesis time fixed, limitations is set for finished this thesis on time:

- This work is focus on digitizing Vietnamese receipt only.
- The output accuracy usually depend on the input image quality including image resolution, clearness, receipt orientation and contrast of background area. These dependencies can be solved by enhancing image processing methodology in further work.
- This work is focus on machine learning method, other methods such as regular expression will not be applied.
- Receipt with table in the content will affect the accuracy of the process because of the program usually detect vertical border column of the table as an character.
- The domain diversity of receipt training data is not enough for covering all of receipt types. This research is focus on sales invoice only.

1.6 Outline

The following chapter 2 will provide research backgrounds used in this thesis focusing on data processing, computer vision, machine learning approaches. In chapter 3, applied methodologies will be taken for solving the topic and these implementations is addressed in chapter 4. The output and evaluation of this work will be discussed in chapter 5. Finally, chapter 6 contains a summary of the project and further expansion plan for this topic.

Chapter 2

Research Background

This chapter presents theoretical backgrounds related with this work. It is divided in five main parts in order

2.1 Data Processing

- Vector
- Image vector
-

2.2 Computer Vision

2.2.1 Overview

In human vision system, we perceived the three-dimension structure from many properties such as shading or pattern of light of the world around and effortlessly recognize shape or object from our vision. Computer vision is trying to reconstruct human vision properties from one or more digital images is a challenge and a topic for many researchers[17]. Applications of computer vision is being used today such as Optical Charater Recognization (OCR) for reading handwritten or printed text, surveillance for monitor and analyze traffic.

2.2.2 OpenCV

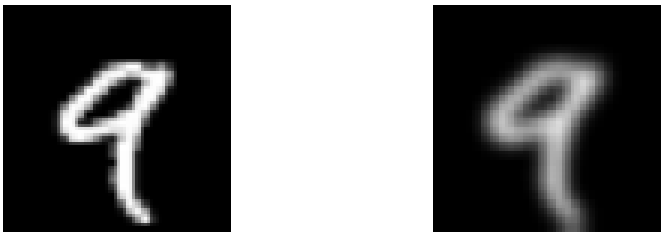
In this project, an open source computer vision library, OpenCV is used for implement image processing and preparation. OpenCV has began since 1998 as a research project at Intel [3] and available since 2000. Digital image in this library is represented as a matrix $m\ x\ n$ with m and n are number of row and column pixel of image, value of cell is for color attribute. OpenCV has built-in many functions image processing including features in following sections in this section 2.2.

2.2.3 Gaussian Blur

In digital image processing, Gaussian Blur is a type of image-blurring filter using Gaussian function. This is an common technique for reduce image noise, detail or enhance image structure in another scale in image processing[4]. This technique calculates pixel's transformation by using Gaussian function:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

with $x$and y for distance from horizontal axis, vertical axis to origin and σ for standard deviation[15][12].

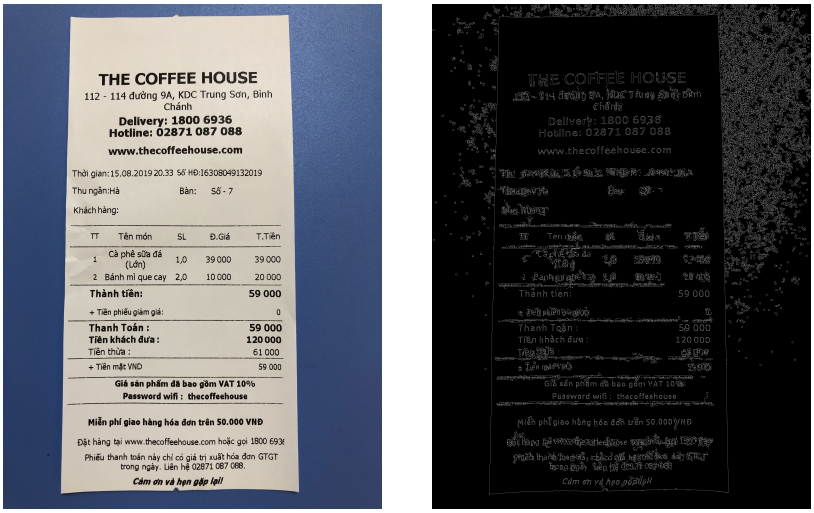


(a) Origin image **(b)** Applied Gaussian Blur, $\sigma = 5$

Figure 2.1: Apply Gaussian Blur on an image from MNIST data set[2]

2.2.4 Canny edge detection

Canny is an edge detection technique for recognize the edge of receipt paper in this thesis. By including collection of mathematical method, edge detection identify discontinuities in digital image. The Canny detector uses adaptive thresholding according to the amount of image's noise for eliminating streaking edge contours. To face with varying image signal-to-noise ratio, various of operator widths is applied and combining operator outputs using a method called feature synthesis [5].



(a) Origin image **(b)** Applied Canny Edge Detector

Figure 2.2: Apply Canny Edge Detector on an capture receipt image

Canny edge detection applied in this work in the image processing part, section 3.1, highlight the edge of image and apply Hough transform and Morphing for detection the receipt paper in image.

2.2.5 Hough transform

Another technique for feature extraction for finding imperfect object in computer vison, image processing is Hough Transform, developed by Paul Hough in 1962 and patented by IBM[15]. Parameterizing a description of a feature at any given location in the space of original image consists in the transformation. Then these parameter generates a mesh the space is defined, a value is accumulated at each mesh point and indicating how, indicating the result of the parameters defined object fits the given image. Mesh points that acquire relatively larger values and describe features that may be projected back onto the image, fitting to some degree the features actually present in the image[1].

Hough tranfrom is applied after receipt edge detection, for solving noise and connect discontinue edges before finding the contours.

2.3 Machine Learning

2.3.1 Overview

By study, instruction or experience, humans and animals improve their knowledge or skill for enhacing result quality variety of tasks. This biological learning behavior is applied into machine from changes of input data, structure or program for future improvement by using machine learning approach[10]. These are two major types of learning in this method are *supervised* and *unsupervised* learning. *Supervised learning* is suppose to given a set of data including input and output, for example, in sentence intent classification, a set of having both sentence and their intention output. In the other setting, *unsupervised learning* only requires a set of input only and apply different approach for classify data into meaning categories.

- Supervised and unsupervised
- Different between human and machine in predicting an problem
- Machine learning tasks - Classification and regression - Grouping or clustering
- Importance of data in machine learning
- Optimization
- Statistic
- Application

2.3.2 Deep learning

- A broader family of machine learning -

2.3.3 EAST

Text detection is an essential step in extracting and understanding text, for distinguish text from background. One of approaches is An Efficient and Accurate Scene Text Detector (EAST), directly produces word or text-line level predictions by using a Full Convolutional Network (FCN) model, improving speed and responsive[19]. EAST support different orientations and directions text. This uses an trained neural network model for finding presence and geometric of text in and image or an frame of video.

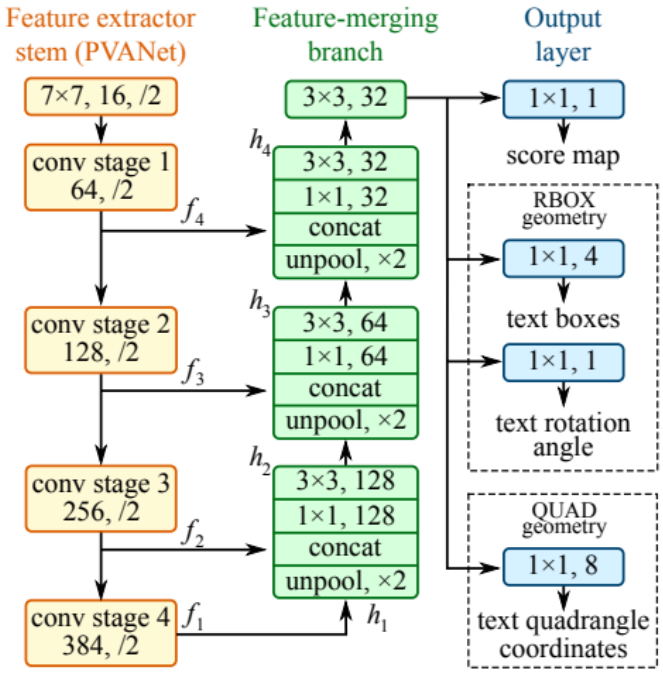


Figure 2.3: FCN structure of EAST text detection [19]

On International Conference on Document Analysis Recognition (ICDAR) 2015 data set, EAST method achieves an Fscore of 0.7820 on original image scale and 0.8072 in F-score in multiple scale. The advantage of this text detection approach is effective combining between speed and accuracy for suitable result, achieve state-of-the-art performance on benchmarks.

2.3.4 Connectionist Text Proposal Network

For detect a text line in a sequence, a vertical anchor mechanism has been developed that jointly predicts location and text/non-text score of each fixed-width proposal, considerably improving localization and connected by a recurrent neural network called Connectionist Text Proposal Network (CTPN) [18]. This approach able to handle multi-scale and multi-lingual text in a single process, avoiding further post filtering or refinement. Architecture of CTPN mostly base on Faster Region Convolutional Neural Networks (R-CNN) with a bi-directional Long short-term memory (BiLSTM). From the beginning, CTPN detects a text line by densely sliding a small window in the convolutional feature maps into a sequence of fine-scale text-proposal using 16-layer vggNet (VGG16). The sliding method is for adapting detecting text line in

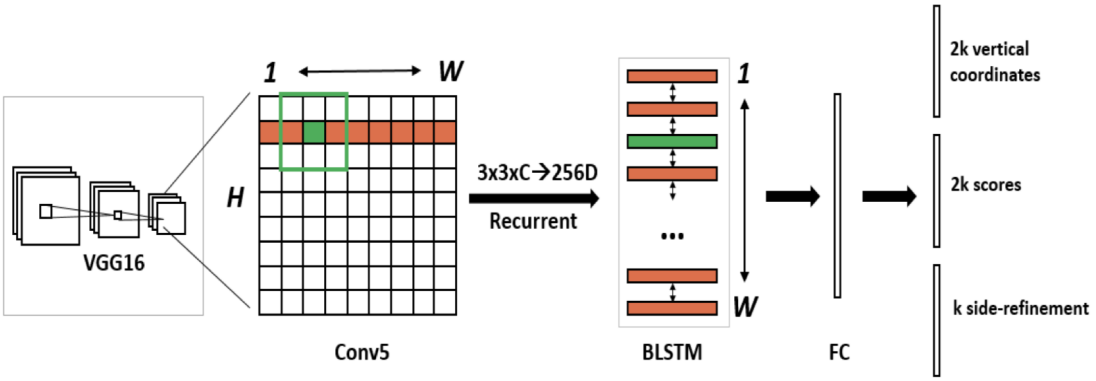


Figure 2.4: Architecture of CTPN[18]

difference sizes. In each 16-pixel width image might contains a single or part of a character. The vertical anchor mechanism is scoring the probability of text/non-text y-axis location every text-proposal. A sequence anchors of text/non-text having score larger than 0.7 generates the detected text proposals. For avoiding false detections on non-text objects such as leaves or some text-like pattern, CTPN uses bi-directional Long short-term memory (BiLSTM) for its Recurrent neural network (RNN) layer, which enables regards the text-proposal independently. The LSTM was proposed specially to address vanishing gradient problem, this RNN architecture is described detailed in subsection 2.3.6. After applied text-proposal detector and bi-directional Long short-term memory (BiLSTM), this detection model constructs the text-line by connecting continuous text/non-text fields having score larger than 0.7, but in the predited localization of 16-pixel width divided image is may inaccuracy with the a ground truth text line area. The last step of CTPN is solve this problem by a side-refinement approach that estimates the offset for each anchor/proposal in both left and right horizontal sides.

In the result CTPN has archived 0.88 and 0.61 using F-measure on ICDAR 2013 and ICDAR 2015.

2.3.5 Recurrent neural network

[7]

2.3.6 Long short-term memory

[9] [11]

2.3.7 Generative adversarial network

[13] [6] [8]

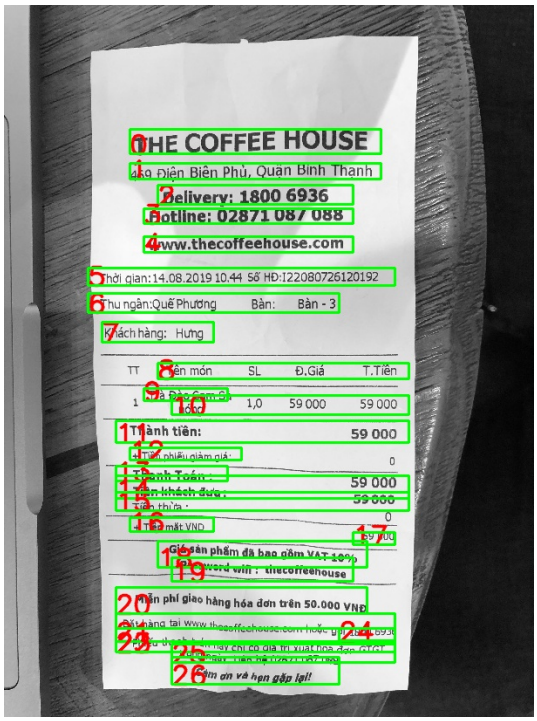


Figure 2.5: CTPN Text detection on a receipt image

2.4 Optical Character Recognition

2.4.1 Tesseract OCR

2.5 Server

Chapter 3

Methodology

In this chapter will present three main steps in this work to archive the receipt feature extraction from receipt image, including crop receipt image in image processing stage and detect text-line using CTPN approach before apply text recognition.

3.1 Image processing

The capture receipt image input is contain the background which is meaningless and cause false detection in further step, pre-processing input data is required for selecting usable image area. To accurately recognize the receipt area, input image is transformed for highlighting receipt features from the background

(a) Origin input image **(b)** Background filtered image

Figure 3.1: Image after apply image processing for background filter

The first applied filter is Gaussian Blur

3.1.1

3.2 Text-line detection

After filtered out the background scene, the input data now is clarify for avoiding false detection. At the beginning, input image is sent to Text Recognition which will be discussed in section 3.3, for converting image to text but the result accuracy is low on a torn or not spared receipt paper which usually happens on screen text. Text-line detection is a step for solving this problem, EAST and CTPN are two applied methods will be discussed in following section.

image without text line detection

3.2.1 EAST

EAST is an screen text detection approach was mentioned in subsection 2.3.3 for extract text from receipt image.

3.2.2 CTPN

Figure 3.2: CTPN detection result on receipt image

3.2.3 Comparison

3.2.4 Merge Boxes

After using CTPN for detect texts, most of texts is detected as Figure 3.2, but shorten text such as text *1* in Figure 3.2 in *SL* column is not detected. I apply the Merge Boxes for detect missing texts in line.

The idea of Merge Boxes is calculating overlap rate of detected boxes and merge all of them into the bigger box, including not detected short text between.

Box geometric is defined by 2-dimension of four points. This work compute the overlap rate on *y-coordinate* as:

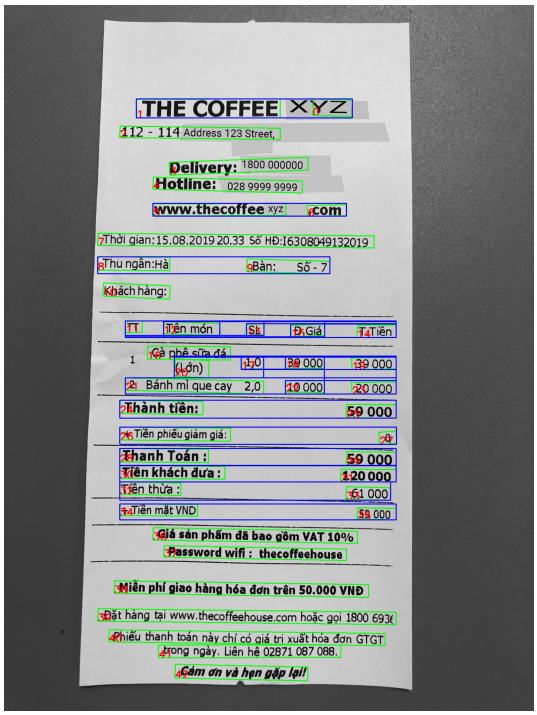


Figure 3.3: For CTPN detection with merge boxes (blue boxes) and without merge boxes (green boxes)

3.3 Text Recognition

3.3.1 Tesseract OCR

3.4 Intent Classification

Chapter 4

Implementation

4.1 Image processing

Chapter 5

Result and Discussion

Chapter 6

Conclusion

6.1 Conclusions

6.2 Further research

References

[1] Hough transform. https://planetmath.org/houghtransform. Accessed on 2019-08-25.

[2] MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges.

[3] OpenCV. https://opencv.org. Accessed on 2019-08-20.

[4] Ferhat Bozkurt, Mete Yaanolu, and Faruk Baturalp Gnay. Effective Gaussian Blurring Process on Graphics Processing Unit with CUDA. *International Journal of Machine Learning and Computing*, 5(1):57–61, February 2015.

[5] John Canny. A computational approach to edge detection. *Ieee Transactions on Pattern Analysis and Machine Inligence*, 1986.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[7] A. Graves, M. Liwicki, S. Fernndez, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, May 2009.

[8] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *CoRR*, abs/1606.03476, 2016.

[9] Sepp Hochreiter and Jrgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[10] Nils J. Nilsson. *Introduction to Machine Learning.* Stanford University, 1998.

[11] Xiangang Li and Xihong Wu. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. *CoRR*, abs/1410.4281, 2014.

[12] Mark S. Nixon and Alberto S. Aguado. *Feature Extraction and Image Processing.*

[13] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016.

[14] Herbert F. Schantz. Recognition technologies users association. *The history of OCR, optical character recognition.*, 1982.

[15] Linda G Shapiro and George Stockman. *Computer Vision.*

[16] Ray Smith. An overview of the tesseract ocr engine. In *Proc. Ninth Int. Conference on Document Analysis and Recognition (ICDAR)*, pages 629–633, 2007.

[17] Richard Szeliski. Computer Vision: Algorithms and Applications. page 979.

[18] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. Detecting text in natural image with connectionist text proposal network. *CoRR*, abs/1609.03605, 2016.

[19] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. EAST: An Efficient and Accurate Scene Text Detector. *arXiv:1704.03155 [cs]*, April 2017.