ABSTRACT OF DISSERTATION

Jack Bandy

The Graduate School
University of Kentucky
2017

INTERACTIVE MACHINE LEARNING FOR WORD RECOGNITION ON
DAMAGED HANDWRITTEN DOCUMENTS

ABSTRACT OF DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By
Jack Bandy
Lexington, Kentucky

Director: Dr. your advisor, Professor of Mathematics
Lexington, Kentucky 2017

ABSTRACT OF DISSERTATION

INTERACTIVE MACHINE LEARNING FOR WORD RECOGNITION ON
DAMAGED HANDWRITTEN DOCUMENTS

an abstract

KEYWORDS: keywords go here

Author's signature: _____ Jack Bandy

Date: _____ October 31, 2017

INTERACTIVE MACHINE LEARNING FOR WORD RECOGNITION ON
DAMAGED HANDWRITTEN DOCUMENTS

By

Jack Bandy

Director of Dissertation: _____ your advisor

Director of Graduate Studies: _____ DGS name here

Date: _____ October 31, 2017

RULES FOR THE USE OF DISSERTATIONS

Name   Date

_____

_____

_____

_____

_____

_____

_____

DISSERTATION

Jack Bandy

The Graduate School
University of Kentucky
2017

INTERACTIVE MACHINE LEARNING FOR WORD RECOGNITION ON
DAMAGED HANDWRITTEN DOCUMENTS

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By
Jack Bandy
Lexington, Kentucky

Director: Dr. your advisor, Professor of Mathematics
Lexington, Kentucky 2017

ACKNOWLEDGMENTS

Acknowledge people/things here

Dedicated to things (optional)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Background

1.1 Project Components

There are two main components of the project. The first is a semi-supervised machine learning approach to document transcription, and the second is a word tracing tool for textual scholarship.

An Interactive Approach to Automated Transcription

Automated transcription is ideal and sometimes necessary for larger datasets of handwritten documents. Automated transcription for printed documents as well as handwritten documents is now on par with human performance. However, damaged historical documents present unique challenges.

I develop the following approach: given a small set of labeled samples, train a neural network in a semi-supervised manner using both labeled and non-labeled data. Once the initial model is trained, use it to create a transcription of the full document. During the transcription process, the model keeps track of difficult word images, prioritizing them for manual labeling afterwards.

Word Tracing

Once the transcription of a document is generated, many scholars wish to trace the outputted text back to the original manuscript image. Building on state-of-the-art word spotting techniques, I implement a tool that traces transcript text back to the original input image so that scholars can easily navigate and visualize transcriptions.

1.2 Literature Review

2009

- Finding words in alphabet soup: Inference on freeform character recognition for historical scripts [1].

2012

- A novel word spotting method based on recurrent neural networks [2].

- End-to-end text recognition with convolutional neural networks [3].

2013

- Handwritten word recognition using mlp based classifier: A holistic approach [4].

- Feature extraction with convolutional neural networks for handwritten word recognition [5].

2014

- A combined system for text line extraction and handwriting recognition in historical documents [6]

2015

- Efficient segmentation-free keyword spotting in historical document collections [7].

- Adapting off-the-shelf cnns for word spotting & recognition [8].

- Segmentation-free handwritten Chinese text recognition with LSTM-RNN [9].

2016

- On the Benefits of Convolutional Neural Network Combinations in Offline Handwriting Recognition [10].

- Reading text in the wild with convolutional neural networks [11].

- PHOCNet: A deep convolutional neural network for word spotting in handwritten documents [12].

- SpottingNet: Learning the Similarity of Word Images with Convolutional Neural Network for Word Spotting in Handwritten Historical Documents [13].

Surveys

- A survey of document image word spotting techniques [14].

- A survey on handwritten documents word spotting [15].

Chapter 2 The First Chapter

2.1   The First Section

Math goes here.

<div align="center">

Here's a figure

Figure 2.1: A Simple Figure

</div>

<div align="center">

Here	is
a	table

Table 2.1: A Simple Table

</div>

<div align="center">

Copyright© Jack Bandy, 2017.

</div>

Bibliography

[1] Nicholas R Howe, Shaolei Feng, and R Manmatha. Finding words in alphabet soup: Inference on freeform character recognition for historical scripts. *Pattern Recognition*, 42(12):3338–3347, 2009.

[2] Volkmar Frinken, Andreas Fischer, R Manmatha, and Horst Bunke. A novel word spotting method based on recurrent neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 34(2):211–224, 2012.

[3] Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng. End-to-end text recognition with convolutional neural networks. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3304–3308. IEEE, 2012.

[4] Ankush Acharyya, Sandip Rakshit, Ram Sarkar, Subhadip Basu, and Mita Nasipuri. Handwritten word recognition using mlp based classifier: A holistic approach. *International Journal of Computer Science Issues*, 10(2):422–427, 2013.

[5] Théodore Bluche, Hermann Ney, and Christopher Kermorvant. Feature extraction with convolutional neural networks for handwritten word recognition. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 285–289. IEEE, 2013.

[6] Andreas Fischer, Micheal Baechler, Angelika Garz, Marcus Liwicki, and Rolf Ingold. A combined system for text line extraction and handwriting recognition in historical documents. In *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, pages 71–75. IEEE, 2014.

[7] Marçal Rusiñol, David Aldavert, Ricardo Toledo, and Josep Lladós. Efficient segmentation-free keyword spotting in historical document collections. *Pattern Recognition*, 48(2):545–555, 2015.

[8] Arjun Sharma et al. Adapting off-the-shelf cnns for word spotting & recognition. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 986–990. IEEE, 2015.

[9] Ronaldo Messina and Jérôme Louradour. Segmentation-free handwritten chinese text recognition with lstm-rnn. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 171–175. IEEE, 2015.

[10] Dewi Suryani, Patrick Doetsch, and Hermann Ney. On the benefits of convolutional neural network combinations in offline handwriting recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*, pages 193–198. IEEE, 2016.

[11] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

[12] Sebastian Sudholt and Gernot A Fink. Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*, pages 277–282. IEEE, 2016.

[13] Zhuoyao Zhong, Weishen Pan, Lianwen Jin, Harold Mouchère, and Christian Viard-Gaudin. Spottingnet: Learning the similarity of word images with convolutional neural network for word spotting in handwritten historical documents. In *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*, pages 295–300. IEEE, 2016.

[14] Angelos P Giotis, Giorgos Sfikas, Basilis Gatos, and Christophoros Nikou. A survey of document image word spotting techniques. *Pattern Recognition*, 68:310–332, 2017.

[15] Rashad Ahmed, Wasfi G Al-Khatib, and Sabri Mahmoud. A survey on handwritten documents word spotting. *International Journal of Multimedia Information Retrieval*, 6(1):31–47, 2017.

Vita

A brief vita goes here.