Recognition and Prominence Ranking of Alphanumeric Number Sequences in Images

By Alex Cummaudo
BSc *Swinburne*

Supervised by Prof. Rajesh Vasa, Assoc. Prof. Andrew Cain

A thesis submitted in partial fulfilment of the requirements for the
Bachelor of Information Technology (Honours)

Deakin Software and Technology Innovation Laboratory
School of Information Technology
Deakin University, Australia

27 October 2017

Abstract

Text detection in natural images is a growing area with increasing applications, including traffic sign and license plate recognition, and text-based image search. Robustly detecting and recognising text is especially challenging when text is deformed, such as the photometric and geometric distortions of text worn by a moving subject in unstructured scenes. Existing methods of text detection in such cases are classified as learning-based or connected component (CC)-based, applying a mix of enhanced detection techniques—such as stroke width transformation (SWT), canny-edge detection and maximally stable extremal regions (MSERs)—and feeding candidates into optical character recognition (OCR) engines or neural networks to recognise the text. This study proposes applying a learning-based approach using deep-learning strategies to automate the recognition of racing bib numbers (RBNs) in a natural image dataset of various marathons, and then ranking detected subject's photos in order of prominence. Experimental results showed that these deep-learning strategies performed favourably against other methods using a consistent dataset, prompting further investigation in the generality of the technique developed to other similar subject material.

Declarations

I certify that the the thesis entitled “Recognition and Prominence Ranking of Alphanumeric Number Sequences in Images” submitted for the degree of Bachelor of Information Technology (Honours) is the result of my own work and that where reference is made to the work of others, due acknowledgement is given. I also certify that any material in the thesis which has been accepted for a degree or diploma by any university of institution is identified in the text.

Alex Cummaudo, BSc *Swinburne*
27 October 2017

We certify that the thesis prepared by Alex Cummaudo entitled “Recognition and Prominence Ranking of Alphanumeric Number Sequences in Images” is prepared according to our expectations and that the honours coordinator can proceed to accept this submission for examination.

Prof. Rajesh Vasa
27 October 2017

Assoc. Prof. Andrew Cain
27 October 2017

Acknowledgements

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

I would also like to thank Andrew Cain for his extraordinary efforts over many years to teach hundreds of students (myself included) and who has developed a valued mentorship with me in guiding me throughout my academic life.

Contents

List of Figures

List of Tables

List of Abbreviations

ADAS Advanced Driver Assistance Systems.

CC Connected Component.

CNN Convolutional Neural Network.

LPR Licence Plate Recognition.

NN Neural Network.

OCR optical character recognition.

RBN Racing Bib Number.

TSR Traffic Sign Recognition.

Chapter 1

Introduction

Ever since the camera and phone were unified into smartphones, we have seen an increasing interest for image processing analysis, but text recognition still faces challenges within images of unstructured scenes. While successes in character recognition have long been developed and improved upon with Optical Character Recognition (OCR) engines (Smith, 1987), these are typically applied under strict conditions (e.g., flatbed document scanners). Once applied within the context of a natural scene, real-world discrepancies pose serious shortcomings, such as illumination and viewpoint conditions, blur and glare variations, geometric and photometric distortion, and differences in font size and style. Overcoming these issues has motivated a variety of different proposed techniques in order to realise potential applications that make use of text recognition at scale.

Dissect the practicality of image processing within natural scenes, along with the ubiquity of smartphone cameras, and potential applications become clearer. In the last two decades, we have seen the development of point-and-shoot product recognition (Tsai et al., 2010; Girod et al., 2011), object detection in videos (Sivic and Zisserman, 2003), building recognition (Takacs et al., 2008), image feature extraction to improve visual-based search engines (Lowe, 2004; Bay et al., 2008), and translation services of American Sign Language gestures (Jin et al., 2016). Nonetheless, embedded text within images reveals the largest form of informative features about the image; if text extraction is therefore not robust, information extraction suffers.

Research in overcoming such limitations were competed in Lucas et al. (2003), where robustness was a key focus in the image processing pipelines proposed. This focus was reiterated by Chen et al. (2011), who state the primary prerequisite for text-based recognition (especially within natural scenes) is the text location must be robustly located. As with any data processing pipeline, false negatives increase where early stages of the pipeline fail. Detection of potential candidates must therefore be robust, and we can reduce cases of errors where we potentially include stages within a pipeline that could be warranted unnecessary, increasing the robustness by parsing unmatched candidates through further pipelines.

Robustness is therefore a key consideration made in our assessment of how useful that pipeline may be. Without the construct of robustness, we restrict these pipelines to very confined conditions, and its usefulness in products is not warranted.

1.1 Background

This study presents character recognition within the context of short, alphanumeric number sequences. We define alphanumeric number sequences as short fragments of digits within an unstructured scene.

Certain literature have focused on these kinds of sequences, namely: License Plate Recognition (LPR) systems (Cano-Perez and Pérez-Cortes, 2003; Anagnostopoulos et al., 2006); Traffic Sign Recognition (TSR), namely speed limit recognition, to better realise Advanced Driver Assistance Systems (ADAS) (Eichner and Breckon, 2008; Kundu and Mackens, 2015; Seo et al., 2015; Lian et al., 2016); and, street number recognition, specifically a study by Netzer et al. (2011), using Google Street View[1] to determine the numerical value of street numbers. Figure 1.2 summarises the usage of these sequences.

Different applications apply varying methods to parse short alphanumeric characters. There are typically two stages of any parsing method: *detection* and *recognition*. Detection refers to locating possible candidates and recognition refers to the representation of the text itself. Detection techniques usually are categorised as either connected component (CC)-based or learning or texture-based. CC-based detection will typically use a set of distinct properties on the image to detect relevant areas (such as width, stroke and colour) while learning-based feed images into a classifier that can distinguish candidates from false positives. The recognition phase can typically be achieved using optical image recognition OCR engines (such as Tesseract[2]), machine learning algorithms or deep neural networks (NNs) to classify the detected regions.

This study proposes the development of a learning-based detection and recognition pipeline using deep-learning neural networks within the context of unstructured photos, namely focusing within the context of marathon Racing Bib Numbers (RBNs), as shown in Figure 1.1.

[1] `https://www.google.com/streetview/` last accessed 13 May 2017.

[2] `https://github.com/tesseract-ocr/tesseract` last accessed 14 May 2017.

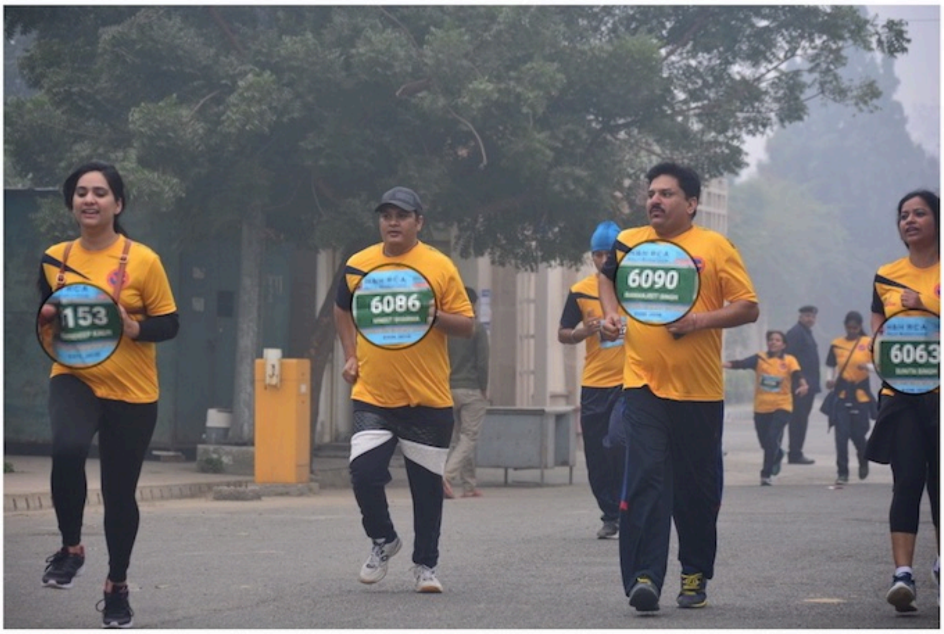


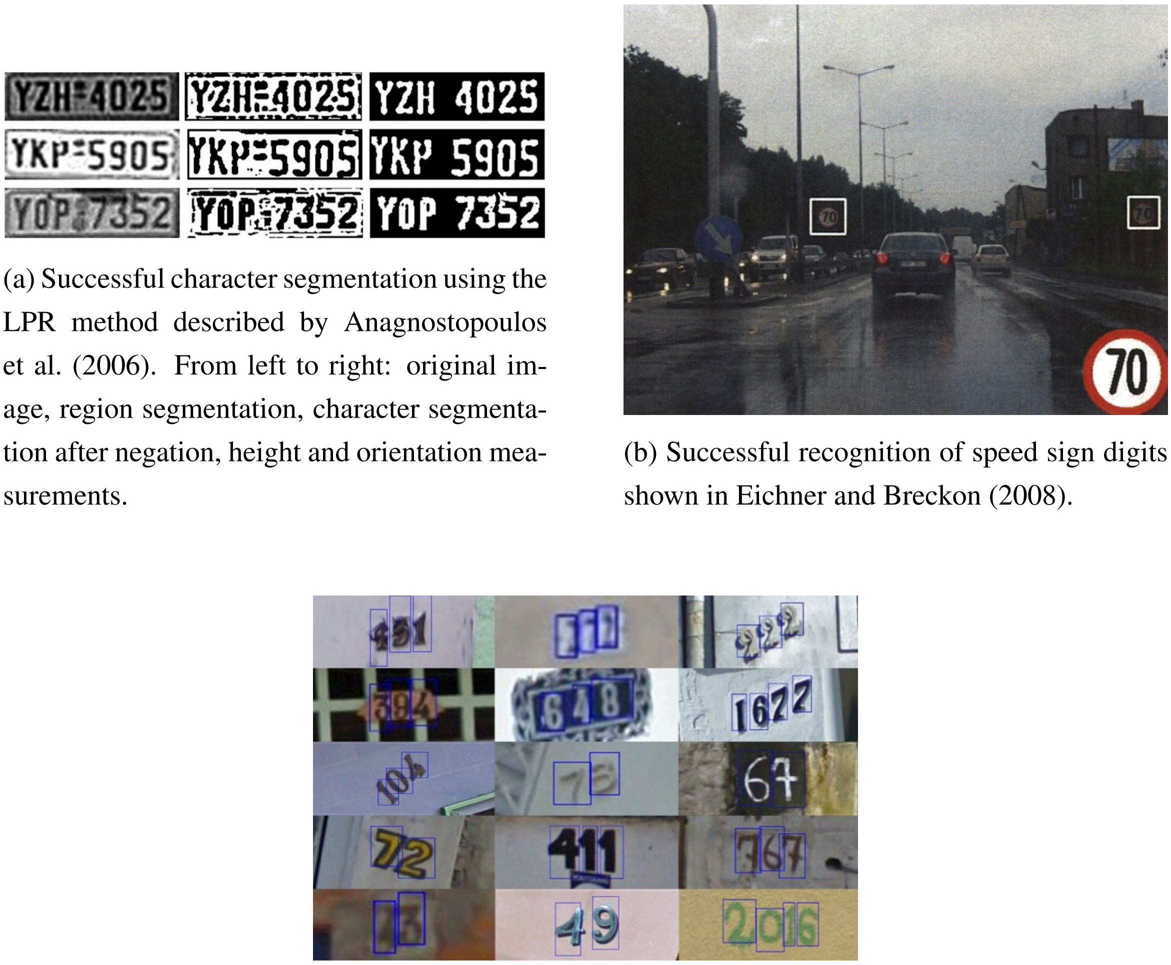
Figure 1.1: Four RBNs in a sample marathon photo.

1.2 Motivation

Detection becomes difficult when the photo is unstructured. Early investigations in LPR systems were systematic in the subject material assessed; a detailed survey by Anagnostopoulos et al. (2008) showed that they work best with consistent lighting, specific colour and typeface detection, fixed detection regions, and non-noisy backgrounds. When applied in the context of images with unstructured backgrounds, these systematic approaches begin to have sever limitations as the text components cannot be easily determined.

While further investigations in the area utilise enhanced CC-based detection (Chen et al., 2011; Shivakumara et al., 2011; Epshtein et al., 2010), performance is likely to degrade as image complexity increases (Li and Lu, 2012). This is especially relevant when text is geometrically obfuscated, such as malformed RBNs as worn on a marathon runner's torso. Some studies have shown to overcome this by using facial recognition to find a more distinct candidate area (Ben-ami et al., 2012), but nonetheless relies on such properties like a person's face to detect a number. Similarly, most recognition techniques interpret text as segmented characters, rather than a single string, though there are exceptions such as in Zhu et al. (2016).

We therefore identify gaps using a learning-based approach in *both* detection and recognition using deep-learning artificial NNs. We also identify gaps in prominence ranking that may also potentially use NNs to order prominence of runners based from their detected RBN.



(a) Successful character segmentation using the LPR method described by Anagnostopoulos et al. (2006). From left to right: original image, region segmentation, character segmentation after negation, height and orientation measurements.

(b) Successful recognition of speed sign digits shown in Eichner and Breckon (2008).

(c) Localisation of digits found from varying street view house numbers using the worker described in Eichner and Breckon (2008).

Figure 1.2: Various sample alphanumeric sequences observed in literature.

1.3 Research Goals

This study aims to develop a processing pipeline that both detects and recognises RBNs on a marathon runner, and then ranks the prominence of each runner detected in the photo. Using artificial deep-learning NNs (such as Convolutional Neural Networks or CNNs) in this pipeline is the main objective, unlike previous studies in RBN recognition that were heavily heuristic and rule driven. This primary aim is developed into three key objectives:

Goal 1: ***Detect RBNs using a CNN***

Literature has shown that heuristic-based detection algorithms (that are CC-based) are able to detect text within photos. We propose to apply these techniques to a large labelled dataset within the context of RBNs, and contrast them against a learning-based detection and recognition algorithm. By benchmarking a against existing libraries and open source tools, we suggest that heuristic-based detection algorithms (focusing namely on CC-based detection) outperforms learning-based detection methods. Therefore, we suggest the following research question:

RQ1) Do CNNs detect RBNs with equal or higher recall and precision rates than CC-based methods?

The findings from these experiments offer metrics (namely, recall and precision rates) to compare the merit of learning-based versus CC-based detection methods within our dataset of marathon photos. We observed that the precision and recall rates of learning-based detection methods compared ⟨ *favourably* | *worse than* ⟩ those of CC-based methods.

Goal 2: ***Design a CNN that can recognise RBNs***

Typically, traditional alphanumeric sequence parsing can be performed by character segmentation, and then piping those characters into OCR engines. This begs the following research questions:

RQ2) Does CNN-based OCR outperform or is at parity with traditional OCR with higher or equal recall and precision rates?

RQ3) Does CNN-based OCR perform *without* the use of character segmentation?

The findings from these experiments observed that the development of our learning-based OCR pipeline outperformed that of a traditional OCR engine by a factor of ⟨ *value* ⟩.

Goal 3: ***Rank prominence of alphanumeric sequences***

Our research objective is aimed to compare if humans are always better at ranking the prominence of an RBN than that of a NN. We can therefore propose the following research questions:

RQ4) Can a deep-learning NN be trained to rank marathon runners by prominence?, and if so

RQ5) Does a deep-learning NN rank prominence of a runner better or equal to that of a human?

The findings from these experiments observed that the development of a ranking system for RBNs was able to match human characteristics with a similarity factor of ⟨ *value* ⟩.

1.4 Thesis Organisation

This thesis is organised into the chapters as outlined below. An appendix follows with additional supplementary material.

Chapter 2 - Background Work Provides an overview of prior studies broadly around the areas of number recognition in image processing and artificial NNs.

Chapter 3 - Related Work Documents a number case studies within the literature directly or closely related to the aims of this research.

Chapter 4 - Research Methodology Describes possible techniques in closer depth to develop a number recognition pipeline, and explores ways to develop prominence ranking techniques.

Chapter 5 - Benchmarking Collates results of a series of experiments using our dataset amongst other tools and pipelines currently developed.

Chapter 6 - Processing Pipeline Discusses the proposed processing pipeline developed that satisfies the aims of this study.

Chapter 7 - Deep Learning Comparison Compares our deep-learning approach with those benchmarked.

Chapter 8 - Validation of Results Highlights a number of validation techniques used to ensure results found in the comparison are correct.

Chapter 9 - Discussion and Limitations Presents implications that were found from the results of our findings and possible limitations.

Chapter 10 - Conclusions and Future Work Draws a number of conclusions and alleviates gaps in the findings of this work by presenting future studies.

Chapter 2

Background Work

In this chapter, we survey a range of literature to explore

2.1 Applications of Image Recognition

2.2 Detection Strategies

2.3 Recognition Strategies

2.4 Prominence Strategies

Chapter 3

Related Work

3.1 RBN Recognition

Fu et al. (2015)

3.2 Speed Limit Sign Recognition

3.3 License Plate Recognition

Chapter 4

Research Methodology

4.1 Overview

4.2 Prominence Ranking Survey

This section encapsulates an experiment to capture prominence rankings of a given sample of the dataset. In this context, prominence is defined as the prominence of a particular marathon runner is within a photo, as identified by the runner's RBN. Results gathered from this experiment will assist in developing a quantitative measure of humans identify prominence within our context. We present participants with a number of subjects and ask to rank them by a prominence Likert scale. The aggregated results of the findings are used as a prominence training dataset fed into a deep-learning neural network.

4.2.1 Survey Design

The survey published for the experiment was collected online via Google Forms[1]. The collection period was for ⟨ *number of months* ⟩ months between ⟨ *survey start date* ⟩ and ⟨ *survey end date* ⟩.

Previous chapters indicated that

Images

4.2.2 Ethics Approval

4.2.3 Demographics

[1] `http://forms.google.com` last accessed 8 May 2017.

Chapter 5

Benchmarking

5.1 Open Source Tools

5.2 Existing Pipelines From Literature

5.3 Hermes Approach

Chapter 6

Processing Pipeline

Chapter 7

Deep Learning Comparison

Chapter 8

Validation of Results

Chapter 9

Discussion and Limitations

Chapter 10

Conclusions and Future Work

References

Anagnostopoulos, C.-N., I. Anagnostopoulos, V. Loumos, and E. Kayafas (2006). A License Plate-Recognition Algorithm for Intelligent Transportation System Applications. *IEEE Trans. Intelligent Transportation Systems*.

Anagnostopoulos, C.-N., I. Anagnostopoulos, I. D. Psoroulas, V. Loumos, and E. Kayafas (2008). License Plate Recognition From Still Images and Video Sequences - A Survey. *IEEE Trans. Intelligent Transportation Systems*.

Bay, H., A. Ess, T. Tuytelaars, and L. J. Van Gool (2008). Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*.

Ben-ami, I., T. Basha, and S. Avidan (2012). Racing Bib Numbers Recognition. In *British Machine Vision Conference 2012*, pp. 19.1–19.10. British Machine Vision Association.

Cano-Perez, J. and J. C. Pérez-Cortes (2003). Vehicle License Plate Segmentation in Natural Images. *IbPRIA 2652*(Chapter 17), 142–149.

Chen, H., S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod (2011). Robust text detection in natural images with edge-enhanced Maximally Stable Extremal Regions. *ICIP*.

Eichner, M. L. and T. P. Breckon (2008). Integrated speed limit detection and recognition from real-time video. In *2008 IEEE Intelligent Vehicles Symposium (IV)*, pp. 626–631. IEEE.

Epshtein, B., E. Ofek, and Y. Wexler (2010). Detecting text in natural scenes with stroke width transform. *CVPR*.

Fu, C., C.-W. Cheng, W.-H. Shen, Y.-L. Wei, and H.-M. Tsai (2015). LightBib: Marathoner Recognition System with Visible Light Communications. In *2015 IEEE International Conference on Data Science and Data Intensive Systems (DSDIS)*, pp. 572–578. IEEE.

Girod, B., V. Chandrasekhar, D. Chen, N.-M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. Tsai, and R. Vedantham (2011). Mobile Visual Search. *IEEE Signal Processing Magazine 28*(4), 61–76.

Jin, C. M., Z. Omar, and M. H. Jaward (2016). A mobile application of American sign language translation via image processing algorithms. In *2016 IEEE Region 10 Symposium (TENSYMP)*, pp. 104–109. IEEE.

Kundu, S. K. and P. Mackens (2015). Speed Limit Sign Recognition Using MSER and Artificial Neural Networks. *ITSC*.

Li, Y. and H. Lu (2012). Scene text detection via stroke width. *ICPR*.

Lian, Z., X. Jing, S. Sun, and H. Huang (2016). Frequency Selective Convolutional Neural Networks for Traffic Sign Recognition. In *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring*, pp. 1–5. IEEE.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision 60*(2), 91–110.

Lucas, S. M., A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young (2003). ICDAR 2003 robust reading competitions. In *Seventh International Conference on Document Analysis and Recognition*, pp. 682–687. IEEE Comput. Soc.

Netzer, Y., T. Wang, and A. Coates (2011). Reading digits in natural images with unsupervised feature learning. *NIPS workshop on ...* .

Seo, Y.-W., J. Lee, W. Zhang, and D. Wettergreen (2015). Recognition of Highway Workzones for Reliable Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems 16*(2), 1–11.

Shivakumara, P., T. Q. Phan, and C. L. Tan (2011). A Laplacian Approach to Multi-Oriented Text Detection in Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*(2), 412–419.

Sivic, J. and A. Zisserman (2003). Video Google - A Text Retrieval Approach to Object Matching in Videos. *ICCV*.

Smith, R. W. (1987). *The Extraction and Recognition of Text from Multimedia Document Images*. Ph. D. thesis, University of Bristol.

Takacs, G., V. Chandrasekhar, N. Gelfand, Y. Xiong, W.-C. Chen, T. Bismpigiannis, R. Grzeszczuk, K. Pulli, and B. Girod (2008). Outdoors augmented reality on mobile phone using loxel-based visual feature organization. *Multimedia Information Retrieval*.

Tsai, S. S., D. M. Chen, V. Chandrasekhar, G. Takacs, N.-M. Cheung, R. Vedantham, R. Grzeszczuk, and B. Girod (2010). Mobile product recognition. *ACM Multimedia*.

Zhu, L., C.-S. Yang, and J.-S. Pan (2016). Detection and Recognition of Speed Limit Sign from Video. *ACIIDS*.

Appendix A

Ethics Clearance

Appendix B

Prominence Ranking Survey Results