Recognition and Prominence Ranking of Alphanumeric Number Sequences in Images

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

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*
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
October 2017

Assoc. Prof. Andrew Cain
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.

Contents

List of Figures

List of Tables

List of Abbreviations

CC Connected Component. 2, 4, 5, 7–12

CNN Convolutional Neural Network. 5, 9

DSTIL Deakin Software and Technology Innovation Laboratory. 4

LPR License Plate Recognition. 2, 4

MSER Maximally Stable Extremal Region. 9, 10

NN Neural Network. 2, 4–7

OCR Optical Character Recognition. 1, 2, 5, 7

RBN Racing Bib Number. 2–6

SWT Stroke Width Transformation. 8, 9

TSR Traffic Sign Recognition. 2

Chapter 1

Introduction

Ever since the camera and phone were unified into smartphones, we have seen an increasing interest for image understanding (specifically to identify the content of an image) but text recognition still faces challenges within images of unstructured scenes. While successes in character recognition have a long history with Optical Character Recognition (OCR) engines [59], these are typically applied under strict conditions (e.g., flatbed scanners for documents without distracting backgrounds). Once applied within the context of a natural scene, real-world discrepancies pose serious shortcomings, such as illumination conditions, viewpoint and perspective differences, blur and glare variations, geometric and photometric distortion, and differences in font size and style [28, 70]. Overcoming these issues has motivated a variety of techniques to realise potential applications that make use of text recognition at scale.

With the ubiquity of smartphone cameras, practical applications of natural image processing have increased. In the last two decades, we have seen the development of point-and-shoot product recognition [17, 65], object detection in videos [56], building recognition [63], image feature extraction to improve visual-based search engines [3, 40], and translation services of American Sign Language gestures [24]. Nonetheless, embedded text within images contains indexable data on the image's semantics [57]; if text extraction is therefore not robust, information extraction suffers.

Text detection robustness is a factor which severely limits a text recognition pipeline. Research in overcoming such limitations have been competed numerous competitions [21, 41, 42, 53], where robustness is the key focus in the image processing pipelines proposed. This focus was reiterated by Chen et al. [10], 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, and therefore detection of these potential candidates must be robust. We can reduce errors, and thus robustness, in a pipeline where: (1) there are unwarranted stages (*excluding* unnecessary stages may also assist in reducing error cases) and (2) by piping through unmatched candidates to further pipelines, which can increase the detection.

Without the construct of robustness, we restrict these pipelines to very confined conditions, and its usefulness in products is not warranted. Therefore, the robustness of text extraction pipelines are imperative to gapping the semantic extraction of information from an image [57], and solving this issue can assist in applications of image processing and data indexing of content within images [**?**] of paramount proportions.

1.1 Background

This study focuses on character recognition in unstructured scenes (Figure 1.1): specifically, short, alphanumeric number sequences. Previous works present methods to extract these sequences in various areas, namely: License Plate Recognition (LPR) systems [1, 7]; Traffic Sign Recognition (TSR) [15, 30, 35, 51]; and, street number recognition, specifically a study by Netzer et al. [47], using Google Street View[1] to determine the numerical value of street numbers. Figure 1.2 highlights typical 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 Character Recognition (OCR) engines (such as Tesseract[2]) [4], machine learning algorithms [30, 32, 47] or deep Neural Network (NN) to classify the detected regions [25, 35, 52].

This study proposes the development of a learning-based detection and recognition pipeline using deep-learning neural networks within the context of unstructured photos, with a focus on marathon Racing Bib Numbers (RBNs)[3], 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.
[3] While referred to as numbers, some RBNs have alphabetic identifiers in them.

Figure 1.1: Four RBNs in a sample marathon photo.

(a) Successful LPR character segmentation [1]. *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 [15].

(c) Localisation of digits found from varying street view house numbers using the worker described in Netzer et al. [47].

Figure 1.2: Various sample alphanumeric sequences observed in literature.

1.2 Motivation

Detection is harder when the photo is unstructured. Early investigations in License Plate Recognition (LPR) systems were systematic in the subject material assessed; a detailed survey by [2] 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 limitations as the text components cannot be easily determined.

While further investigations in the area utilise enhanced Connected Component (CC)-based detection [10, 16, 55], performance is likely to degrade as image complexity increases [34]. This is especially relevant when text is geometrically obfuscated, such as malformed Racing Bib Numbers (RBNs) as worn on a marathon runner's torso. Malformed, in this sense, is caused by non-flat bib sheets that tend to follow the runner's body shape, in addition to images that are taken in dynamical contexts. Some studies have shown to overcome this by using facial recognition to find a more distinct candidate area [4], but nonetheless rely on a person's face to detect a number. Similarly, typical recognition techniques interpret text as segmented characters, rather than a single string, though there are exceptions such as in Zhu et al. [73].

We also identify subject prominence ranking within natural scenes as an area that has little exploration within literature. (For example, the prominence of a *specific* marathon runner within a scene of many runners.) Prominence ranking is an important field in the context of RBN recognition: runners typically choose not to purchase photos where they have been recognised in an image but are not in the foreground. There are also varying factors which influence purchase likelihood, such as face visibility, eye contact with the camera, and blurriness. An assessment into how the prominence of a runner can be ordered in hundreds of identified photos (based from their recognised RBN) can be used by use of a Neural Network (NN).

This study forms part of an industry project under the Deakin Software and Technology Innovation Laboratory (DSTIL). As a part of the research project, access has been made to a labelled dataset of hundreds of thousands of marathon photos.

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. The intention is to explore the viability of artificial deep-learning NNs—such as Convolutional Neural Networks (CNNs)—in the pipeline. Previous studies in RBN recognition [4] and similar areas [15, 30, 64] 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 [10, 15, 34]. We propose to apply these rule-based techniques to a large labelled dataset within the context of RBNs, and contrast them against a learning-based detection and recognition algorithms (using NNs). By benchmarking a against existing libraries and open source tools, we explore if heuristic-based detection algorithms (focusing namely on CC-based detection) outperforms learning-based detection methods. For this goal the research question is framed as:

RQ1) Do CNNs detect RBNs with equal or higher recall and precision rates than 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. In the context of marathon photos, we explore answers to the following:

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

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

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 a NN. We can therefore propose the followings research questions:

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

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

1.4 Thesis Organisation

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

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

Chapter 3 - Data Set Describes the data set to be used, data treatment steps, possible techniques in closer depth to develop a number recognition pipeline, and explores ways to develop prominence ranking techniques.

Chapter 4 - Benchmarking Collates results of a series of experiments using our dataset amongst existing open source tools and pipelines presented in previous work

Chapter 5 - Processing Pipeline Discusses the proposed processing pipeline developed that satisfies the aims of this study.

Chapter 6 - Findings Outlines the method used for validation and presentation of our results.

Chapter 7 - Discussion Presents implications that were found from the results of our findings and limitations.

Chapter 8 - Conclusions and Future Work Draws a number of conclusions and alleviates gaps in the findings of this work by presenting future studies.

Summary

In this chapter we identified some shortcomings in text recognition, developed the context of the study—namely RBN detection. We discussed the general stages that exist for text parsing within natural scenes, detection and recognition, and introduced typical techniques that are applied in this context. We outlined the research aims this study achieves, and how the thesis is organised. The following chapter will detail applications of image processing, using neural networks for image processing, and outline what techniques have been used in previous studies to achieve this.

Chapter 2

Background

We have introduced the context of image processing and neural networks in the related field, and discussed how text capturing within photos is typically achieved in two stages: detection and recognition. Detection techniques are generally classified as either CC- or learning-based. The recognition phases can be applied using traditional OCR engines or, more recently, using artificial NN.

This chapter surveys a number of broad applications where RBN recognition (and related works) are investigated using such phases. The various detection and recognition techniques discussed in the literature are also detailed. We also broadly define the applications of artificial deep-learning NN in these contexts.

2.1 Detection Strategies

Text extraction strategies have seen continuing interest in the literature, with many comprehensive surveys assessing the state of the art [8, 27, 28, 36, 70]. It is widely demonstrated that if text within an unstructured scene is *detected* reliably, then existing OCR engines can suitably extract these characters [58] once they exist in a structured context; thus not every extraction pipeline needs to self-contain a recognition strategy if commercial OCR packages suffice. A survey into the two prominent detection strategies is given in Sections 2.1.1 to 2.1.2.

These two prominent strategies have a varied nomenclature: (1) the CC-based (or *region*-based) approach, that utilise different region properties (e.g., colour, edges, CCs) [10, 16, 23, 29, 33, 34, 38, 39, 54, 55, 61, 62, 71, 72] for unsupervised extraction; and, (2) learning-based (or *texture*-based) approach, which uses unique texture properties to supervise extraction text from its background [11, 13, 19, 31, 66–68]. Additionally, some authors have proposed methods to combine these unsupervised and supervised techniques [5, 43, 46].

2.1.1 CC-based techniques

CC-based approaches generate separated CCs using properties such as stroke width, pixel colour and edges, typically applying geometric and texture filters to reduce false positives. Neighbouring pixels are then 'grouped' using an algorithm originally presented by Horn [20].

Previous work required the use of a scanning window [11, 26, 37] which is limited by a constant image scale and discrete orientations of the sliding (thereby preventing text strokes in non-linear directions). Subramanian et al. [61] overcame this limitation by implementing a algorithm to detect text strokes by scanning an image horizontally and looking for sudden changes in background intensity (Figure 2.1b). However, this algorithm assumes a darker text on a lighter background to find such intensity changes, and consequently there are numerous parameters that must be fine-tuned. Additionally, the algorithm is only able to detect horizontal text only, and detected strokes are not grouped into characters, words and sentences.

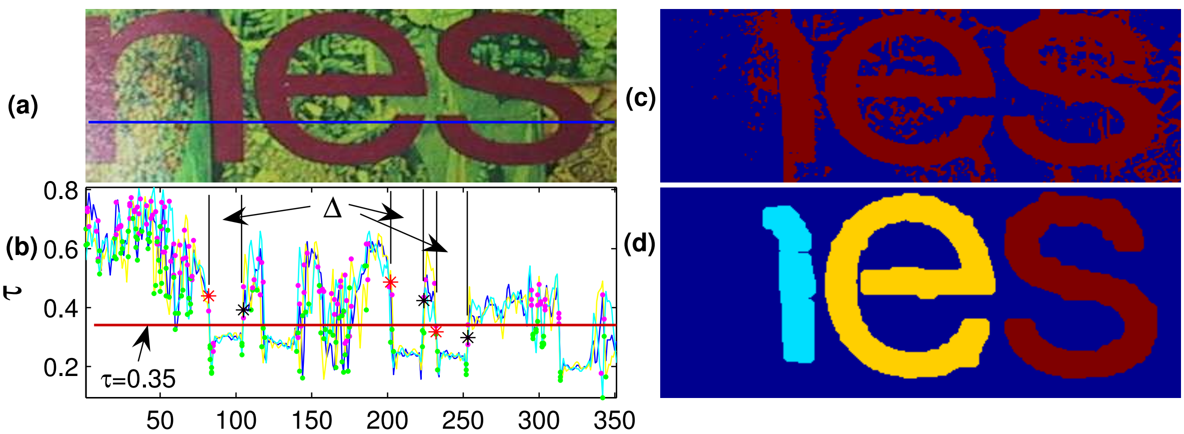

Figure 2.1: A study from Subramanian et al. [61] showed that stroke width could be determined from (a) the original image; (b) the intensity plots of the image to determine stroke regions; (c) the intensity at an optimal threshold; (d) the final thresholded image after morphological operations and CC analysis.

A study by Epshtein et al. [16] (and coincidentally Zhang and Kasturi [72]) built on the idea presented by Subramanian et al., and introduced the concept of Stroke Width Transformation (SWT), a local image operator that determines the most likely stroke of a given pixel by computing the per-pixel width. This was later expanded in Srivastav and Kumar [60]. The SWT approach overcame previous limitations by introducing a system that can detect text regardless of size, typeface, direction and language, making it one of the first widely cited multilingual text detection algorithms. Additionally, SWT overcame methods that required the use of an OCR filtering stage to reduce false positives [11, 12, 68]. A sample of SWT is shown in Figure 2.2.

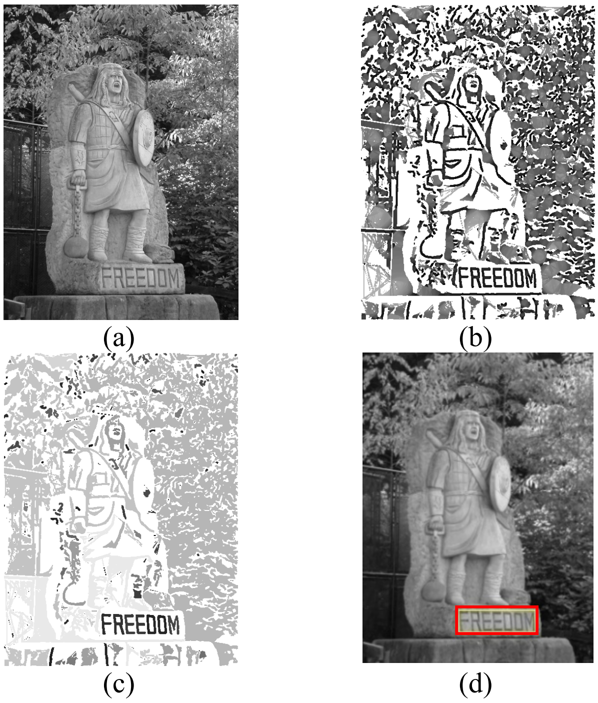

Figure 2.2: The Stroke Width Transformation (SWT) approach introduced in [16]. The original image (a) is converted to a binarised array with the most likely stroke width per-pixel (b), piping the information into geometric filtering (c) as text maintains fixed stroke width (excluding false positives such as foliage). The resulting detected text is shown in (d).

It is common to see edges computed from a raw image using the Canny-Edge Detection algorithm [6]. This was successfully applied in various CC-based studies [10, 16, 71]. While several papers have exploited SWT and adapted it further [55, 72], when opposite edges are not parallel, the SWT forms candidates with holes appearing in stroke curves or joints. This is due to candidates formed by shooting rays from detected the edges along the gradient found, removing the rays if terminated by another edge pixel of a perpendicular gradient. Further limitations include undetected stronger highlights, blurry text, and text with a wide curvature.

An alternate approach that overcomes this limitation was introduced by Chen et al. [10], where the complimentary properties of Canny-Edges [6] and Maximally Stable Extremal Regions (MSERs) [44] were combined. MSER is a detection mechanism suited for region-based detection, is robust against varying viewpoints, scales and illuminations [45], and can be extracted from images efficiently [48]. A limitation of MSER is its sensitivity to image blur [45], but Chen et al. demonstrated that MSER can be edge-enhanced using Canny-Edges on a contrast-enhanced image (Figure 2.3), achieving comparable results to SWT presented by Epshtein et al. [16]. Multiple works have utilised MSERs in a wide range of applications, such as their use in teaching CNNs and real-time text extraction [18, 22, 30, 34, 69].

Figure 2.3: Extracting text from a natural image shown in Chen et al. [10]. *From left to right:* Detected Maximally Stable Extremal Regions (MSERs) of black-on-white objects; text candidates grouped to formed text lines after geometric and stroke width filtering; false positives rejected using text verification showing detected text in the blue box.

A significant requirement of all CC-based techniques are the requirements to cluster extracted components back together again. This, in turn, also helps to remove any false positives by removing properties that don't meet set criteria. Various proposals have been made:

- Epshtein et al. [16] use basic geometric filtering based on the stroke width detected and height ratios of candidates. Additionally, colours of candidates are averaged as it is expected that words be written in the same colour. These are then clustered into candidates pairs (of at least three letters), chained together if they share a similar direction.

- Zhang and Kasturi [72] investigate the spacial relationship and property similarity of two neighbouring candidates, computing their link energies to compute *text energy* (the probability a candidate is a true positive). The distance of the text energies are computed and, where beyond a set threshold, will be eliminated if not met. This is presented in Figure 2.4.

- Zhang and Kasturi [71] expand the use of graph spectrum which has successfully been used in computer vision [50] to show image features in the form of a graph, use an adjacency matrix, then gather clusters of CCs based on the positive eigenvectors of the graph. This process is illustrated in Figure 2.5.

- Shivakumara et al. [55] propose the use of skeletal distance maps of a CC to remove small artificats and reduce false positives. They define *simple* and *complex* CCs respectively as: (a) a single text string or false positive; (b) multiple text strings which are connected to each other. The skeletonisation process is shown in Figure 2.6.

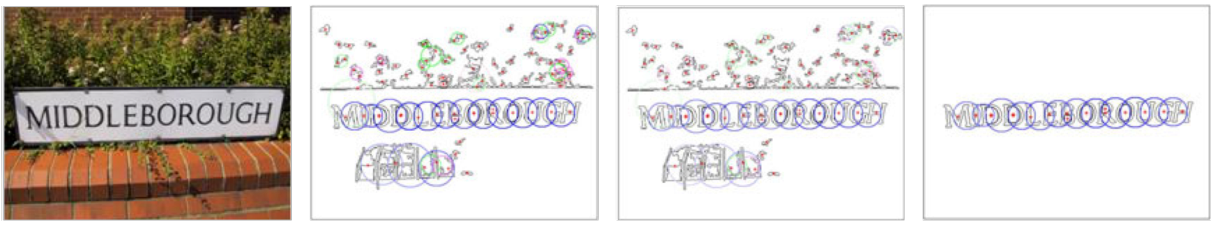

Figure 2.4: Using link and text energies for reconnecting character candidates shown in Zhang and Kasturi [72]. *From left to right:* original image; all link energies determined in a given image (note the false positives of background foliage); text energies calculated; all text energies greater than 0.5.

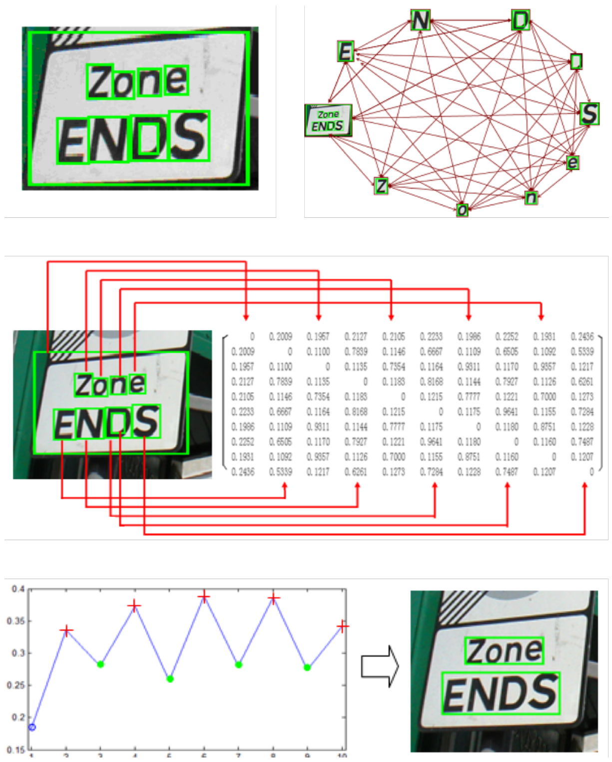

Figure 2.5: Process of grouping components via graph spectrum [71]. *Top-left:* 10 CCs detected. *Top-right:* generated graph from detected candidates. *Middle:* generated adjacency matrix. *Bottom-left:* Positive eigenvector resulting from the graph spectrum. *Bottom-right*: Resulting bounding boxes.

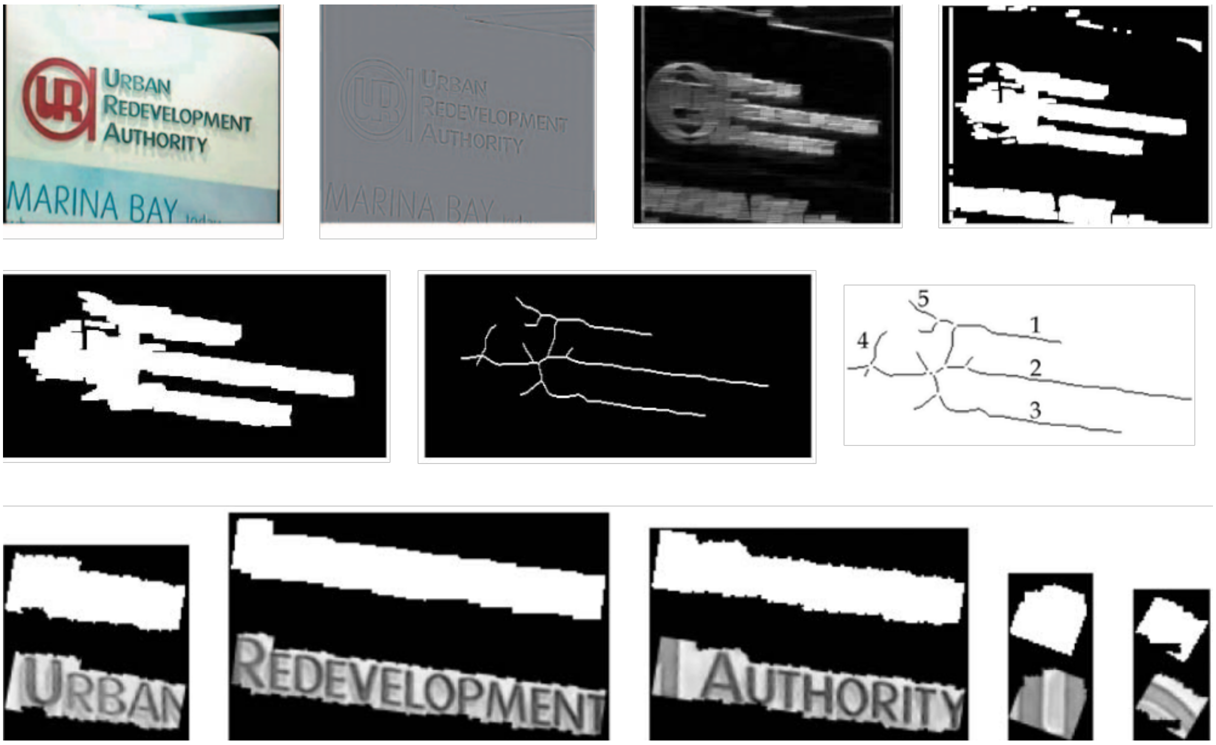

Figure 2.6: Developing a skeletal map using the process proposed by [55]. *Top row:* Original image is processed using Fourier-Laplacian filtering. A maximum distance map is developed and parsed through a morphological operation to remove smaller artefacts. *Middle row:* CC classification and further skeletonisation, showing five labelled subcomponents. *Bottom row:* The five sample subcomponents extracted from the skeletonisation process (in order). Note that subcomponents 4 and 5 are false positives.

2.1.2 Learning-based techniques

typically referred to as a *learning*-based approach, due to the common use of machine learning methods utilised

Typically, texture-based approaches utilise supervised learning methods, though it is typical for these classifiers to required thousands of training images [9]. Additionally, these methods

2.2 Recognition Strategies

2.3 Metrics

Throughout our survey, we have utilised the evaluation scheme first proposed for use in image processing in the ICDAR text extraction competitions [41, 42, 53]. This scheme was designed to be easy to understand and compute, reward text extraction useful for natural scenes, and heavily punish trivial solutions. The intention behind these metrics were to develop a measure of 'robustness' a text extraction pipeline can achieve.

2.3.1 Precision and Recall

Generally in information retrieval, the precision (p) and recall (r) metrics are used, first defined in the six evaluation criteria for information retrieval systems by Cleverdon et al. [14]. Precision refers to the proportion of relevant matches actually retrieved in the retrieved results, while recall refers to the proportion of relevant matches retrieved in total relevant instances. We use recall and precision metrics to assess the *effectiveness* of an information retrieval system [49].

In the context of image processing, systems that over-estimate are punished with a low precision score, while systems that under-estimate are punished with a low recall score [42]. Therefore, precision is the number of correct candidates (c) divided by the number of total estimates found (E):

$$p = \frac{c}{\mid E \mid}$$

And recall is defined as the number of correct estimates divided by the total number of ground-set truth targets (T):

$$r = \frac{c}{\mid T \mid}$$

However, it is not realistic for a given text extraction pipeline to *exactly* agree with the rectangle bounds manually tagged by a human. Lucas et al. [42] first proposed changes to these calculations to better suit their usage in the context of information extraction from within images. They adopt a more flexible notion of what a 'match' is. They define a new match measure (m_p) between two rectangles (i.e., the ground truth and the system's detected candidate) as "the area of intersection of both rectangles divided by the area of the minimum bounding box containing both rectangles" [42]. This allows for a match value of one when the candidate is identical to the ground truth, and zero where the candidate has no intersection at all to the ground truth.

Therefore, the best match, $m(r, R)$, of a rectangle r in a set of rectangles R is:

$$m(r, R) = \max m_p(r, r') \mid r' \in R$$

Lastly, we can redefine the recall and precision metrics to be more forgiving in the image extraction context:

$$p' = \frac{\sum_{r_e \in E} m(r_e, T)}{\mid E \mid}$$

$$r' = \frac{\sum_{r_t \in T} m(r_t, E)}{\mid T \mid}$$

2.3.2 The f-score

Common metrics used when developing text extraction pipelines utilise the use of the f-score, a single measure of quality that combines both precision and recall values computed above. We are able to compute this metric using the standard measure across many studies, as contrasted in ⟨ *reference the table* ⟩.

The f-score algorithm is given in the context of image processing in Lucas et al. [42]. Relative weights controlled by an α value of 0.5 give equal weight to both precision and recall metrics:

$$f = \frac{1}{\frac{\alpha}{p'} + \frac{1-\alpha}{r'}}$$

Summary

Chapter 3

Data Set

Chapter 4

Benchmarking

4.1 Open Source Tools

4.2 Existing Pipelines From Literature

4.3 Hermes Approach

Chapter 5

Processing Pipeline

Chapter 6

Findings

Chapter 7

Discussion

Chapter 8

Conclusions and Future Work

References

[1] 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*.

[2] 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*.

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

[4] 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.

[5] Bengio, Y., P. Lamblin, D. Popovici, and H. Larochelle (2006). Greedy Layer-Wise Training of Deep Networks. *NIPS*.

[6] Canny, J. F. (1986). A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*.

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

[8] Chen, D. and J. Luettin (2000). A survey of text detection and recognition in images and videos.

[9] Chen, D., J.-M. Odobez, and J.-P. Thiran (2004). A localization/verification scheme for finding text in images and video frames based on contrast independent features and machine learning methods. *Sig. Proc. - Image Comm.*.

[10] 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*.

[11] Chen, X. and A. L. Yuille (2004a). Detecting and reading text in natural scenes. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, pp. 366–373. IEEE.

[12] Chen, X. and A. L. Yuille (2004b). Detecting and Reading Text in Natural Scenes. *CVPR*.

[13] Chen, X. and A. L. Yuille (2005). A Time-Efficient Cascade for Real-Time Object Detection - With applications for the visually impaired. *CVPR Workshops*.

[14] Cleverdon, C., J. Mills, and M. Keen (1966). Factors Determining the Performance of Indexing Systems. Technical report, ASLIB Cranfield Research Project, Cranfield.

[15] 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.

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

[17] 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.

[18] Gonzalez, Á., L. M. Bergasa, J. J. Y. Torres, and S. Bronte (2012). Text location in complex images. *ICPR*.

[19] Hanif, S. M. and L. Prevost (2009). Text Detection and Localization in Complex Scene Images using Constrained AdaBoost Algorithm. *ICDAR*.

[20] Horn, B. (1986, January). *Robot Vision*. MIT Press.

[21] Hua, X.-S., L. Wenyin, and H. Zhang (2004). An automatic performance evaluation protocol for video text detection algorithms. *IEEE Trans. Circuits Syst. Video Techn.*.

[22] Huang, X., T. Shen, R. Wang, and C. Gao (2015). Text detection and recognition in natural scene images. In *2015 International Conference on Estimation, Detection and Information Fusion (ICEDIF)*, pp. 44–49. IEEE.

[23] Jain, A. K. and B. Y. 0002 (1998). Automatic text location in images and video frames. *ICPR*.

[24] 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.

[25] Jin, J., K. Fu, and C. Zhang (2014, September). Traffic Sign Recognition With Hinge Loss Trained Convolutional Neural Networks. *IEEE Transactions on Intelligent Transportation Systems 15*(5), 1991–2000.

[26] Jung, C., Q. Liu, and J. Kim (2009, January). A stroke filter and its application to text localization. *Pattern Recognition Letters 30*(2), 114–122.

[27] Jung, K., K. In Kim, and A. K Jain (2004, May). Text information extraction in images and video: a survey. *Pattern Recognition 37*(5), 977–997.

[28] Jung, K., K. I. Kim, and A. K. Jain (2004). Text information extraction in images and video - a survey. *Pattern Recognition*.

[29] Kim, H.-K. (1996). Efficient Automatic Text Location Method and Content-Based Indexing and Structuring of Video Database. *J. Visual Communication and Image Representation*.

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

[31] Lee, C. W., K. Jung, and H. J. Kim (2003, November). Automatic text detection and removal in video sequences. *Pattern Recognition Letters 24*(15), 2607–2623.

[32] Lee, E. R., P. K. Kim, and H. J. Kim (1994). Automatic Recognition of a Car License Plate using Color Image Processing. *ICIP 2*, 301–305.

[33] Lee, S., M. S. Cho, K. Jung, and J. H. Kim (2010). Scene Text Extraction with Edge Constraint and Text Collinearity. *ICPR*.

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

[35] 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.

[36] Liang, J., D. S. Doermann, and H. Li (2005). Camera-based analysis of text and documents - a survey. *IJDAR*.

[37] Lienhart, R. and A. Wernicke (2002). Localizing and segmenting text in images and videos. *IEEE Trans. Circuits Syst. Video Techn.*.

[38] Liu, Y., S. Goto, and T. Ikenaga (2006). A Contour-Based Robust Algorithm for Text Detection in Color Images. *IEICE Transactions*.

[39] Liu, Z. and S. Sarkar (2008). Robust outdoor text detection using text intensity and shape features. *ICPR*.

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

[41] Lucas, S. M. (2005). ICDAR 2005 text locating competition results. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pp. 80–84 Vol. 1. IEEE.

[42] 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.

[43] Mairal, J., F. R. Bach, J. Ponce, G. Sapiro, and A. Zisserman (2008). Discriminative learned dictionaries for local image analysis. *CVPR*, 10.

[44] Matas, J., O. Chum, M. Urban, and T. Pajdla (2002). Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. *BMVC*.

[45] Mikolajczyk, K., T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. J. Van Gool (2005). A Comparison of Affine Region Detectors. *International Journal of Computer Vision*.

[46] Mutch, J. and D. G. Lowe (2006). Multiclass Object Recognition with Sparse, Localized Features. *CVPR*.

[47] Netzer, Y., T. Wang, and A. Coates (2011). Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.

[48] Nistér, D. and H. Stewénius (2008). Linear Time Maximally Stable Extremal Regions. *ECCV*.

[49] Rijsbergen, C. J. V. (1979, January). *Information Retrieval*. Butterworth-Heinemann.

[50] Sarkar, S. and K. L. Boyer (1996). Quantitative Measures of Change based on Feature Organization - Eigenvalues and Eigenvectors. *CVPR*, 478–483.

[51] 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.

[52] Sermanet, P. and Y. LeCun (2011). Traffic sign recognition with multi-scale Convolutional Networks. *IJCNN*.

[53] Shahab, A., F. Shafait, and A. Dengel (2011). ICDAR 2011 Robust Reading Competition Challenge 2: Reading Text in Scene Images. In *2011 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1491–1496. IEEE.

[54] Shivakumara, P., W. Huang, T. Q. Phan, and C. L. Tan (2010). Accurate video text detection through classification of low and high contrast images. *Pattern Recognition*.

[55] 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.

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

[57] Smeulders, A. W. M., M. Worring, S. Santini, A. Gupta, and R. C. Jain (2000). Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell.*.

[58] Smith, R. (2007). An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, pp. 629–633. IEEE.

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

[60] Srivastav, A. and J. Kumar (2008). Text detection in scene images using stroke width and nearest-neighbor constraints. In *TENCON 2008 - 2008 IEEE Region 10 Conference (TENCON)*, pp. 1–5. IEEE.

[61] Subramanian, K., P. Natarajan, M. Decerbo, and D. A. Castañón (2007). Character-Stroke Detection for Text-Localization and Extraction. *ICDAR*.

[62] Sun, Q., Y. Lu, and S. Sun (2010). A Visual Attention Based Approach to Text Extraction. *ICPR*.

[63] 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*.

[64] Torresen, J., J. W. Bakke, and L. Sekanina (2004). Efficient recognition of speed limit signs. In *The 7th International IEEE Conference on Intelligent Transportation Systems*, pp. 652–656. IEEE.

[65] 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*.

[66] Tu, Z., X. Chen, A. L. Yuille, and S. C. Zhu (2003). Image Parsing - Unifying Segmentation, Detection, and Recognition. *ICCV*.

[67] Wang, X., L. Huang, and C. Liu (2009). A New Block Partitioned Text Feature for Text Verification. *ICDAR*.

[68] Ye, Q., Q. Huang, W. G. 0001, and D. Zhao (2005). Fast and robust text detection in images and video frames. *Image Vision Comput.*.

[69] Zagoruyko, S. and N. Komodakis (2015). Learning to compare image patches via convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4353–4361. IEEE.

[70] Zhang, J. and R. Kasturi (2008). Extraction of Text Objects in Video Documents: Recent Progress. In *2008 The Eighth IAPR International Workshop on Document Analysis Systems (DAS)*, pp. 5–17. IEEE.

[71] Zhang, J. and R. Kasturi (2010). Text Detection Using Edge Gradient and Graph Spectrum. *ICPR*.

[72] Zhang, J. and R. Kasturi (2011). Character Energy and Link Energy-Based Text Extraction in Scene Images. In *Computer Vision – ACCV 2010*, pp. 308–320. Berlin, Heidelberg: Springer Berlin Heidelberg.

[73] 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