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 and transfer learning to automate the recognition of racing bib numbers (RBNs) in a natural image dataset of various marathons, with the intention to then rank the subject's photos in order of prominence. Experimental results showed that these deep-learning strategies performed favourably, with RBN detection accuracy beyond 95%. This prompts 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 that 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

Dedicated to Tom Fellowes.

Acknowledgements

I thank everyone at DSTIL for fostering a highly supportive and motivating environment to work in: you make every day enjoyable and are all my second family. Because of you, I have developed a love of research that no one can take from me. I especially express my gratitude toward Simon Vajda and Rodney Pilgrim for their direction throughout my honours year; the exceptional research guidance from Raj Vasa and Nicola Pastorello, including our fruitful discussions and their well-valued feedback on this thesis; the extraordinary assistance provided by the entire DS/R team at DSTIL; and Jake Renzella and Reuben Wilson for their much-needed personal support and friendship since our undergraduate years at Swinburne, as well as the increasing number of coffees shared at Fusion as the submission date drew nearer! I am also infinitely indebted to Andrew Cain for his remarkable teaching efforts over many years, inspiring thousands of students to cherish the art of programming (myself included) and with whom I have developed a highly-valued mentorship to guide me throughout life. And lastly, I thank my loving family and partner Tom Fellowes, without whom this thesis—indeed this year—would not have been possible.

Contents

Abstract iii

Declaration v

Acknowledgements ix

Contents ix

List of Abbreviations xiv

List of Figures xvii

List of Tables xxi

List of Listings xxiii

1 Introduction 1

1.1 Background 2

1.2 Motivation 4

1.3 Research Goals 4

1.4 Thesis Organisation 6

1.5 Summary 6

2 Background 7

2.1 Detection Strategies 7

2.1.1 Connected Component-based techniques 8

2.1.2 Learning-based techniques 13

2.2 Recognition Strategies 16

2.3 Metrics 21

2.3.1 Precision and Recall 21
2.3.2 The f-score 22
2.4 Summary 24

3 Dataset **25**
3.1 A Data-Capturing Metamodel 25
3.2 Methodology 26
3.3 Motivating Case Study: What to Capture? 28
3.3.1 Feature 1: Image-Level Features 30
3.3.2 Feature 2: Bibs 31
3.3.3 Feature 3: Faces 31
3.3.4 Feature 4: Prominence 33
3.3.5 Feature 5: Colours 33
3.3.6 Overview of Features 35
3.4 Describing the Metamodel 37
3.4.1 Image Frame 37
3.4.2 Feature 38
3.4.3 Annotation 38
3.4.4 Construction Rules 38
3.4.5 Attributes 39
3.5 The Data-Capturing Process 39
3.5.1 Informing Argus 39
3.5.2 Extracting Features 42
3.5.3 Transformation 45
3.5.4 Load 45
3.6 Data Postprocessing 47
3.6.1 Flagging 47
3.6.2 Data Augmentation 50
3.7 Metamodel Evaluation 52
3.7.1 MS COCO and COCO-Text JSON Format 53
3.7.2 PASCAL VOC XML Format 53
3.7.3 ICDAR XML Formats 54
3.7.4 MATLAB-Encoded Binaries 54

3.8 Summary 55

4 Argus **57**

4.1 Design 57

4.2 System Evaluation 57

4.2.1 Annotation Throughput 58

4.2.2 Annotation Likelihood of Purchase Bias 61

4.2.3 Annotation Quality Evaluation 61

4.3 Comparative Annotation Tools 65

4.3.1 LabelMe 65

4.3.2 Annotation and Performance Evaluation Platform 66

4.3.3 Amazon Mechanical Turk 66

4.3.4 VATIC 67

4.3.5 ScaleAPI 67

4.4 Summary 69

5 Processing Pipeline **71**

5.1 Bib Detection 71

5.1.1 Existing Approaches 72

5.1.2 Deep-Learning Approaches 72

5.1.3 Person Filtering 73

5.2 Text Detection 77

5.3 Text Recognition 77

5.4 Runtime Performance 78

5.5 Summary 79

6 Evaluation **81**

6.1 Evaluation Strategies 81

6.2 Metrics 82

6.2.1 Bib Detection Performance 82

6.2.2 Text Detection Performance 83

6.2.3 Character Recognition 83

6.2.4 Overall Performance 84

6.3 Results 86

6.3.1 Summarised Performance over all Evaluations 86
6.3.2 Performance in Manual Evaluation 88
6.3.3 Overall Performance 91
6.4 Summary 92

7 Conclusions **93**
7.1 Primary Contributions 93
7.2 Future Work 94
7.3 Wider Applicability 95
7.4 Closing Remarks 96

References **111**

A Ethics Clearance **113**

B Usability Notes **115**
B.1 Notes 115
B.2 System Usability Scale Scores 116

C Metamodel Class Diagrams **117**

D Dataset Schemas **121**

E Dataset Mappings **133**

F Supplementary Data **139**

List of Abbreviations

ACL Argus Constraint Language. 39, 41, 42, 44, 45

ADF Argus Data Format. 39, 43, 45–48, 51–54, 134–137

AI Artificial Intelligence. 25, 28, 30, 39, 45, 47, 48, 55, 61, 62, 71, 96

AMT Amazon Mechanical Turk. 52, 66, 67

APEP Annotation and Performance Evaluation Platform. 52, 66

API Application Programming Interface. 67

AWS Amazon Web Services. 79

CC Connected Component. 2, 4, 5, 7–12, 16, 20, 23, 24, 93

CNN Convolutional Neural Network. 5, 9, 14, 15, 24, 72, 92, 94

COCO Common Objects in COntext. 52, 53, 66, 77, 122, 124, 134

CSV Comma-Separated Values. 45

CVC Computer Vision Center. 52, 66

DSTIL Deakin Software and Technology Innovation Laboratory. 4, 27, 71

HIT Human Intelligent Task. 66

HOG Histogram of Oriented Gradient. 13, 18, 23, 72

HTML HyperText Markup Language. 41

ICDAR International Conference on Document Analysis and Recognition. 16, 21, 24, 52–54, 66, 125, 127, 135

IoU Intersection over Union. 22

JSON JavaScript Object Notation. 45, 52, 53, 68, 122, 124

LBP Local Binary Patterns. 13, 23

LoP Likelihood of Purchase. 33, 34, 36, 46, 48, 57, 61, 94

LPR License Plate Recognition. 2, 4, 18, 55, 93, 95

LSTM Long Short-Term Memory. 78, 83

MDE Model-Driven Engineering. 26

MLP Multilayer Perceptron. 13, 17

MSER Maximally Stable Extremal Region. 9, 10, 17, 23

NLP Natural Language Processing. 55

NN Neural Network. 2, 4–7, 13, 17–19, 23, 24, 28, 30, 39, 45, 61, 62, 71, 72, 78, 79, 92–94

OCR Optical Character Recognition. 1, 2, 5, 7, 16–19, 23, 24, 71, 77–79, 82, 85, 91–95

PASCAL Pattern Analysis, Statistical Modelling and Computational Learning. 52, 53, 128, 130, 136

PNN Probabilistic Neural Network. 18

R-CNN Region-based Convolutional Neural Network. 14, 15, 72–74, 76, 77, 79, 81, 86, 88, 92–95

RBN Racing Bib Number. 2–7, 16, 18–20, 24, 28, 30, 31, 36, 38, 59, 62, 69, 71, 73, 77–79, 81, 83, 84, 93–95

RLE Run Length Encoding. 53, 66

RPN Region Proposal Network. 72

SaaS Software as a Service. 66, 67

SETL Supervised Extract-Transform-Load. 39

SIFT Scale Invariant Feature Transform. 13, 72

SUN Scene UNderstanding. 52, 65

SURF Speeded-Up Robust Features. 13

SUS System Usability Scale. 28, 116

SVHN Street View House Numbers. 18, 52, 54, 137

SVM Support Vector Machine. 13, 14, 23

SWT Stroke Width Transformation. 8, 9, 19, 20, 23

TSR Traffic Sign Recognition. 2, 17, 18, 24, 55, 95

UI User Interface. 41, 42, 57, 58, 63, 69

UML Unified Modelling Language. 27, 37, 42, 45, 47

VATIC Video Annotation Tool from Irvine, California. 67

VOC Visual Object Classes. 52, 53, 128, 130, 136

XML eXtensible Markup Language. 41, 45, 52, 53

XSD XML Schema Definition. 125

YAML YAML Ain't Markup Language. 45

YOLO You Only Look Once. 72–74, 76, 79

List of Figures

1.1 Sample racing bib numbers . . . 3
1.2 Alphanumeric sequences observed in literature . . . 3

2.1 Stroke analysis from Subramanian et al. [144] . . . 8
2.2 Stroke Width Transformation from Epshtein et al. [38] . . . 9
2.3 Using contrast-enhanced MSERs to detect text . . . 10
2.4 Text energy for connecting candidates back together . . . 11
2.5 Using graph spectrum to cluster CCs . . . 11
2.6 Skeletonisation process of CCs . . . 12
2.7 A pipeline for text extraction using CNNs . . . 15
2.8 The Mask R-CNN framework for instance segmentation . . . 15
2.9 Use of CNNs for object detection . . . 15
2.10 A NN designed to recognised speed limit signs . . . 17
2.11 A PNN used to recognise license plate characters . . . 18
2.12 Text recognition pipelines dependent on heuristics . . . 19
2.13 Character processing for feeding an OCR engine . . . 20
2.14 Drawbacks of heuristic-driven approaches for RBN detection . . . 20
2.15 Overlapping areas of ground truth and estimated targets . . . 23

3.1 An overview of systems, models and technical spaces . . . 26
3.2 A layered approach to develop our metamodel . . . 27
3.3 Implementation methodology . . . 29
3.4 Various image-level features . . . 30
3.5 Bib Sheet segment-level features . . . 32
3.6 Face Bounds segment-level features . . . 32
3.7 Face visibility and its effect on prominence . . . 34

3.8 Variant Purchase Likelihoods 34
3.9 Class diagram of our proposed metamodel 37
3.10 Product testing issues identified 40
3.11 Concrete workflow example using Argus 44
3.12 Sample representation of an ADF 46
3.13 State diagram to represent flagging 47
3.14 Overview of processing made on our dataset 49
3.15 Various augmented images from our dataset 51

4.1 An overview of the Argus user interface 58
4.2 Bib and Face feature annotation with Argus 59
4.3 Prominence and Colour feature annotation with Argus 60
4.4 Throughput of images using Argus 61
4.5 Quality assurance using Argus 63
4.6 Quality of tagging related the number of runners per photo 64
4.7 Quality of tagging and seconds evaluating the photo 64
4.8 The LabelMe user interface 65
4.9 The APEP user interface 66
4.10 The VATIC web-based user interface 67

5.1 Overview of our processing pipeline 71
5.2 Bib detection results using FRCNN 73
5.3 Invalid intersection to union both regions 74
5.4 Person filter using YOLO to improve accuracies of bib detection 76
5.5 Text region detection pipeline 77
5.6 Image post-processing applied to optimise input into an OCR engine 78

6.1 Euler diagram to illustrate OCR performance 85
6.2 Bib, text and character performance 87
6.3 Bib, text and character performance of manual inspection 88
6.4 Distribution of manual inspection evaluation 90
6.5 OCR performance results 91

7.1 Potential use of Mask-RCNN on to recognise an RBN 95

C.1 Type class hierarchy of annotations 118
C.2 Type class hierarchy of features 119
C.3 Type class hierarchy of attributes 119
C.4 Type class hierarchy of construction rules 119

E.1 Mapping ADF to COCO-based annotation formats 134
E.2 Mapping the ICDAR annotation formats to ADF 135
E.3 Mapping the PASCAL VOC annotation formats to ADF 136
E.4 Mapping the binarised MATLAB format to ADF 137

List of Tables

2.1 Top-scoring word recognition results from ICDAR 2011–2015. 16
2.2 Survey of text extraction literature 23

3.1 Summary of annotations captured in the dataset 36
3.2 Sample UI elements to extract features in Argus 42
3.3 Breakdown of training and validation data 50
3.4 Various datasets and respective annotation formats 52
3.5 Components present in the ADF compared to other annotation formats 52

4.1 Mistakes using Argus 63

6.1 Summary of Evaluations. 82
6.2 Comparison of various f-score values for cropping versus evaluation sets. . . 86

B.1 SUS Scores 116

F.1 Summary of Bib Detection 140
F.2 Summary of Text Detection 141
F.3 Summary of OCR 142
F.4 Summary of Runtime 143
F.5 Summary of Manual Inspections 144

List of Listings

3.1 Sample Argus Constraint Language File Format 41
4.1 Sample ScaleAPI request 68
4.2 Sample ScaleAPI response 68
D.1 MS COCO Schema 122
D.2 COCO-Text Schema 124
D.3 ICDAR 2011-2015 XSD 125
D.4 ICDAR 2003-2011 Sample Annotation File 127
D.5 PASCAL VOC 2007 SampleAnnotation File 128
D.6 PASCAL VOC 2012 SampleAnnotation File 130

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 [139], 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 [75, 170]. 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 [49, 151], object detection in videos [136], building recognition [147], image feature extraction to improve visual-based search engines [7, 103], and translation services of American Sign Language gestures [70]. Embedded text within images contains indexable data on the image's semantics [137]; a further potential application.

Text detection robustness is a factor that severely limits a text recognition pipeline. Research in overcoming such limitations have been competed numerous competitions [63, 104, 105, 133], where robustness is the key focus in the image processing pipelines proposed. This focus was reiterated by Chen et al. [24], who state the primary prerequisite for text-based recognition (especially within natural scenes) is that 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 [137]. Solving this issue can assist in applications of image processing and data indexing of content within images [43] 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 [2, 17]; Traffic Sign Recognition (TSR) [37, 85, 94, 131]; and, street number recognition [117]. (Specifically a study by Netzer et al., 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]) [9], machine learning algorithms [85, 89, 117] or deep Neural Networks (NNs) to classify the detected regions [71, 94, 132].

This study proposes the development of a learning-based detection and recognition pipeline using deep-learning neural networks (within the context of unstructured photos) focusing 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 [2]. *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 [37].

(c) Localisation of digits found from varying street view house numbers using the worker described in Netzer et al. [117].

Figure 1.2: Various sample alphanumeric sequences observed in literature.

1.2 Motivation

Detection is harder when a photo is unstructured. Early investigations in License Plate Recognition (LPR) systems were systematic in the subject material assessed; a detailed survey by Anagnostopoulos et al. [3] 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 [24, 38, 135], performance is likely to degrade as image complexity increases [92]. 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 dynamic contexts. Some studies have shown to overcome this by using facial recognition to find a more distinct candidate area [9], 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 [173].

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 that 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 determining ways to rank 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 [9] and similar areas [37, 85, 149] 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 [24, 37, 92]. We propose to apply these rule-based techniques to a large labelled dataset within the context of RBN detection, and contrast them against a learning-based detection and recognition algorithms (using NNs). By developing an end-to-end recognition pipeline, we explore if learning-based detection methods can detect these sequences just as heuristic-based detection algorithms can (focusing namely on CC-based detection). For this goal the research questions are framed as:

RQ1) Can CNNs be taught to recognise RBNs?

RQ2) What is a systematic methodology for developing annotated data required for training the CNN?

Goal 2: ***Use a NN-based OCR engine to recognise the detected RBN***

Typically, traditional alphanumeric sequence parsing can be performed by character segmentation, and then piping those characters into Optical Character Recognition (OCR) engines. In the context of marathon photos, we explore answers to the following:

RQ3) Can a NN-based OCR approach recognise extracted RBNs detected from an image?

RQ4) Can a CNN-based OCR algorithm perform this *without* the use of character segmentation?

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

Our research objective focuses on whether prominence of RBNs can be captured and then fed into a NN. We can therefore propose the followings research questions:

RQ4) How can prominence be quantified and collected for the purposes of data collection?

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 - Dataset Describes the dataset to be used, data treatment steps, and possible techniques to develop a number recognition pipeline in closer depth. Furthermore, we explore ways to develop a metamodel to gather large, annotated datasets.

Chapter 4 - Argus Describes an implementation of our metamodel for the purposes of curating a training dataset in the context of marathon running photography.

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

Chapter 6 - Evaluation Discusses several evaluation strategies used to assess the accuracy of our processing pipeline, and presents the implications and limitations found from the results of our findings.

Chapter 7 - Conclusions Draws a number of conclusions and alleviates gaps in the findings of this work by proposing future studies.

1.5 Summary

In this chapter we identify some shortcomings in text recognition and develop the context of this study—namely RBN detection. We discuss the general stages that exist for text parsing within natural scenes, detection and recognition, and introduce typical techniques that are applied in this context. We outline the research aims of this study and how this thesis is organised.

Chapter 2

Background

Text capturing within unstructured photos from the wild are typically achieved in two stages: detection and recognition. Detection techniques are classified as either Connected Component (CC)- or learning-based. The recognition phase uses traditional OCR engines or, more recently, artificial Neural Networks (NNs).

In this chapter, we survey different applications where Racing Bib Number (RBN) recognition (and related works) are investigated. Various detection and recognition techniques discussed in literature are detailed.

2.1 Detection Strategies

Text extraction strategies have seen continuing interest in literature, with many comprehensive surveys assessing the state of the art [20, 74, 75, 95, 170]. It is widely demonstrated that if text within an unstructured scene is *detected* reliably, then existing OCR engines can suitably extract these characters [138] 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 are given in Sections 2.1.1 and 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) [24, 38, 69, 80, 90, 92, 101, 102, 134, 135, 144, 145, 171, 172] for unsupervised extraction; and, (2) learning-based (or *texture*-based) approach, which uses unique texture properties to supervise text extraction from its background [27, 29, 34, 52, 56, 87, 91, 97, 112, 122, 123, 152, 159, 168]. Some authors have proposed methods that mix both supervised and unsupervised techniques [10, 107, 116].

2.1.1 Connected Component-based techniques

Connected Component (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 such as the one originally presented by Horn [61].

Previous work required the use of a scanning window [27, 73, 97] that is limited by a constant image scale and discrete orientations of the sliding (thereby preventing text strokes in non-linear directions). Subramanian et al. [144] overcame this limitation by implementing an 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.

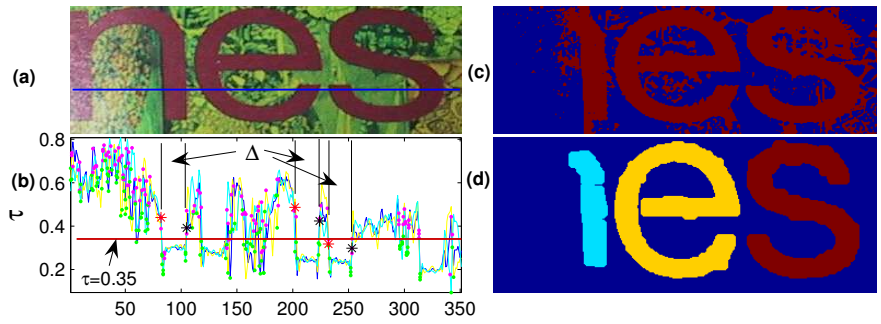


Figure 2.1: A study from Subramanian et al. [144] showed that stroke width could be determined from (a) the original image; (b) intensity plots of the image to determine stroke regions (τ is the intensity threshold and Δ is the stroke width); (c) the intensity at an optimal threshold; (d) the final thresholded image after morphological operations and CC analysis.

A study by Epshtein et al. [38] (and coincidentally Zhang and Kasturi [172]) 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 [143] by using inherent properties of scene text. 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 [27, 28, 168]. A sample of SWT is shown in Figure 2.2.

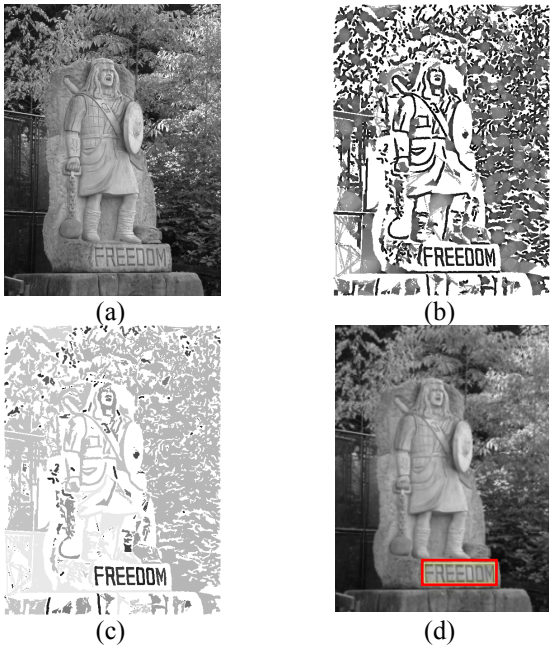


Figure 2.2: The Stroke Width Transformation (SWT) approach introduced in [38]. The original image (a) is converted to a binarised array with the most likely stroke width per-pixel (b), piping the information into geometric filtering (c) as text maintains fixed stroke width (excluding false positives 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 [16]. This was successfully applied in various CC-based studies [24, 38, 171]. While different researchers have exploited SWT and adapted it further [135, 143, 172], 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 edges along the gradient found and 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. [24], where the complimentary properties of Canny-Edges [16] and Maximally Stable Extremal Regions (MSERs) [108] were combined. MSER is a detection mechanism suited for region-based detection, is robust against varying viewpoints, scales and illuminations [111], and can be extracted from images efficiently [119]. A limitation of MSER is its sensitivity to image blur [111], 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. [38]. Multiple works have utilised MSERs in a wide range of applications, such as their use in teaching CNNs and real-time text extraction [53, 64, 85, 92, 169].

Figure 2.3: Extracting text from a natural image shown in Chen et al. [24]. *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. [38] 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 [172] 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 [171] expand the use of graph spectrum, which has successfully been used in computer vision [130] to show image features in the form of a graph. This uses 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. [135] propose the use of skeletal distance maps of a CC to remove small artefacts and reduce false positives. They define *simple* and *complex* CCs respectively as: (a) a single text string or false positive or; (b) multiple text strings that are connected to each other. The skeletonisation process is shown in Figure 2.6.

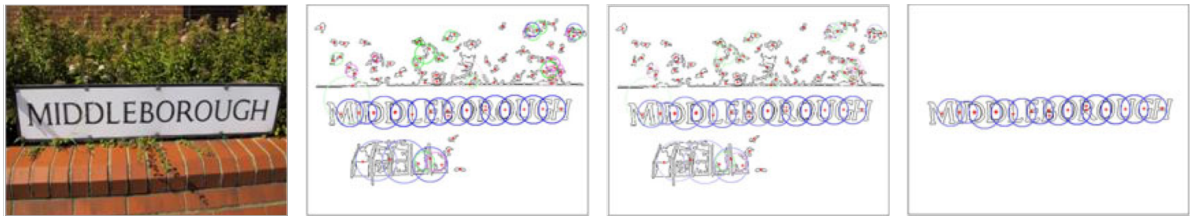


Figure 2.4: Using link and text energies for reconnecting character candidates shown in Zhang and Kasturi [172]. *From left to right:* original image; all link energies determined in a given image (note the false positives of background foliage); text energies calculated; all text energies greater than 0.5.

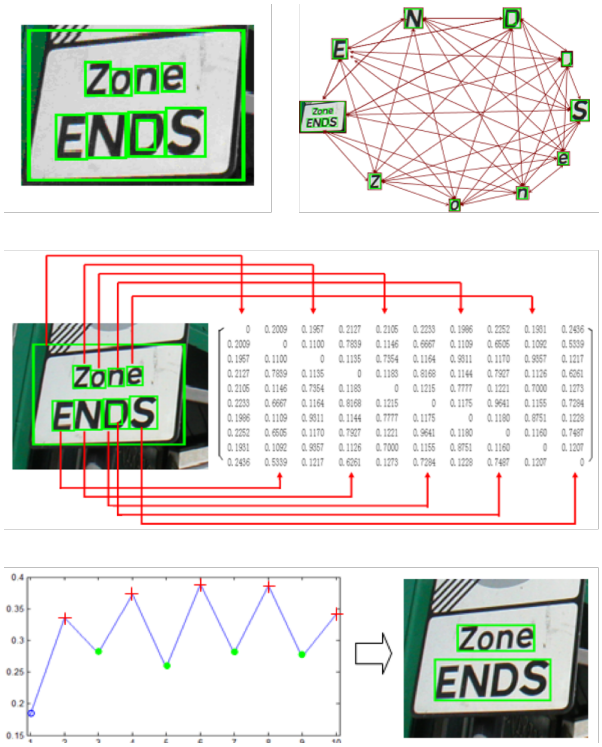


Figure 2.5: Process of grouping components via graph spectrum [171]. *Top-left:* 10 CCs detected. *Top-right:* generated graph from detected candidates. *Middle:* generated adjacency matrix. *Bottom-left:* Positive eigenvector resulting from the graph spectrum. *Bottom-right*: Resulting bounding boxes.

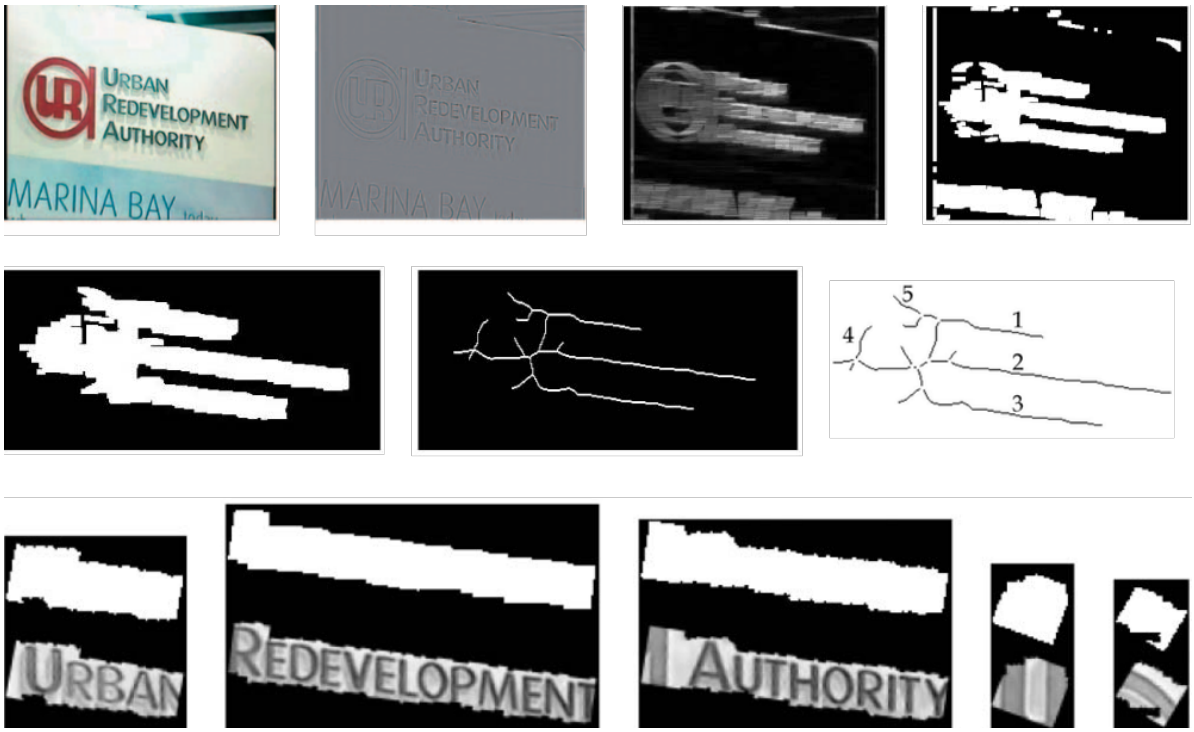


Figure 2.6: Developing a skeletal map using the process proposed by Shivakumara et al. [135]. *Top row:* Original image is processed using Fourier-Laplacian filtering. A maximum distance map is developed and parsed through a morphological operation to remove smaller artefacts. *Middle row:* CC classification 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

The learning-based approach has had varied popularity over time. While also referred to as a *texture*-based approach, it typically utilises learning-based methods to train a classifier using these textures, considering the text as a special texture within the image. This is done by extracting certain features over a portion of the image (i.e., the texture) either via heuristics or—recently—machine learning. The texture is typically extracted by scanning the image at varying scales, then classifying areas of pixels based on certain features. A classifier uses these features to identify text from non-text to extract text from its background.

Early works in the area focused primarily on how to classify a region as either text or non-text [81, 91, 97, 140]. However, while it may be easy to procure training samples of text, it is often more difficult to procure non-text training samples [59, 146]. This is due to the wide variance of what 'non-text' samples are, and these need to be well-represented in the training set of the classifier.

A varied range of features have been utilised to mark textures. A recent feature made popular is the Histogram of Oriented Gradient (HOG), a object detection technique that utilises gradient orientation, first conceptualised in [110] and later coined by Freeman and Roth [44]. However, its popularity did not become widespread until [34], where Dalal and Triggs showed its applicability on pedestrian detection. Later, HOG was shown to be useful in text detection by Hanif et al. [57]. Further features include: Local Binary Patterns (LBP) [120], shown to improve detection when combined with HOG [158]; wavelet energy [113, 155], applied to extract subtitles [52] and worked optimally with HOG in text extraction [122] (when compared to other features); Gabor filters [99]; Scale Invariant Feature Transform (SIFT) descriptors [103]; grey scale features [81]; Speeded-Up Robust Features (SURF) [7]; edge map features [21]; and shape contexts [8].

Many of these features can be combined and fed into a single or multiple classifiers [52, 56, 57, 122, 152, 159, 168]. Example classifiers include Support Vector Machines (SVMs) [15, 32, 153], the Adaptive Boosting (AdaBoost) classifier [45]—and variants thereof [46, 56, 141]—or NNs, such as Multilayer Perceptrons (MLPs), though Chen et al. [22] report that the use of SVMs show better text texture verification than MLPs.

As shown in the wide range of features and classifiers, a main limitation of learning-based methods is the difficulty in selecting which combinations of features to use, the inability to detect sufficiently slanted text, as well as its reported high computational complexity due to

the need to scan the image multiple times at different scales [38, 92]. Furthermore, it is typical for these classifiers to required thousands of training images [23].

More recently though, deep-learning CNNs have been utilised for instance classification and per-pixel segmentation, as emphasised in Figure 2.9. While CNNs are a relatively old concept (see LeCun et al. [86]), they lost interest within image processing to Support Vector Machines (SVMs) and AdaBoost throughout the late 1990s and 2000s. In 2012, however, Krizhevsky et al. [83] re-sparked interest by demonstrating far increased classification accuracy within images in the ImageNet Large Scale Visualisation Recognition Challenge [35], with only a few modifications made to the CNNs proposed by LeCun et al. more than a decade prior.

This has since led to the rise of enhanced CNNs such as FICS+++ [93], Region-based Convolutional Neural Network (R-CNN) [50] and Fast/Faster R-CNN [51, 127]. Furthermore, a preprint for a proposed Mask R-CNN [58] appeared in early 2017. A comparison of two of these networks using images in the Microsoft Common Objects in Context [98] dataset are shown in Figure 2.9, with the Mask R-CNN framework illustrated in Figure 2.8. The rise of these recent developments have furthered the interest in multi-language machine learning libraries, such as MXNet[1], which are able to combine these academic works for use in large-scale industry-focused production code [25].

These methods are increasingly improving per-pixel classification of objects within natural scenes. However, applicability of these deep-learning networks in the sole context of *text extraction* is yet to be widely explored, though a 2016 paper by Jaderberg et al. [68] did show the viability of a real-time text reading pipeline based on CNNs (Figure 2.7) that successfully extracts any text query from 2.3 million frames of BBC News footage[2].

[1] `http://mxnet.io` last accessed 30 June 2017.

[2] `http://zeus.robots.ox.ac.uk/textsearch` last accessed 14 August 2017.

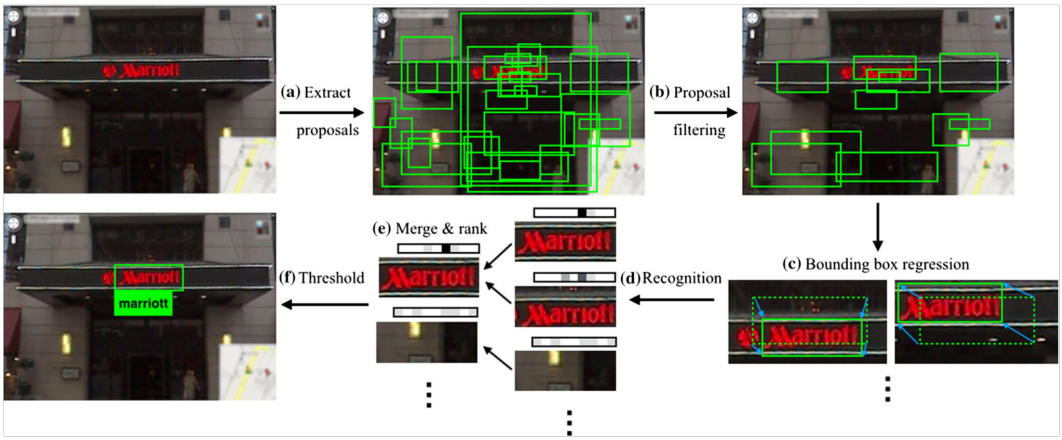


Figure 2.7: The pipeline for text extraction proposed in [68].

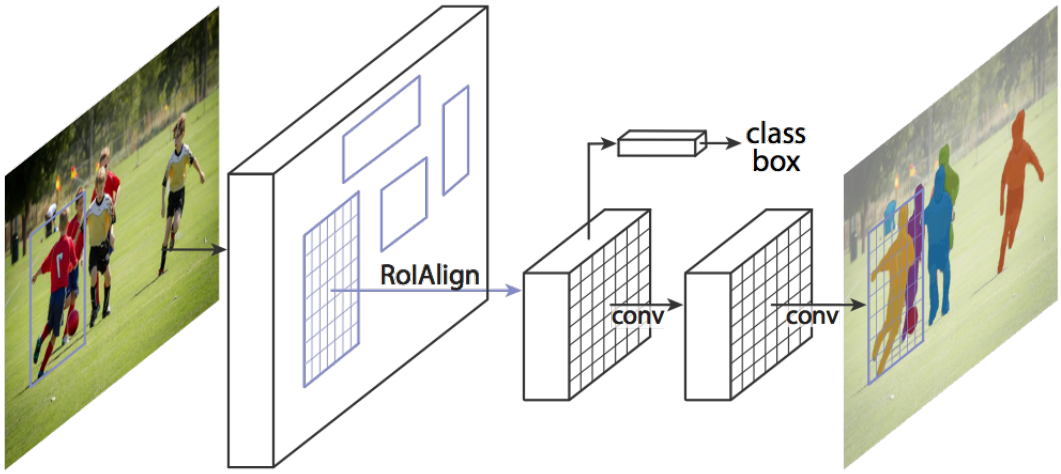


Figure 2.8: The Mask R-CNN framework for instance segmentation [58].

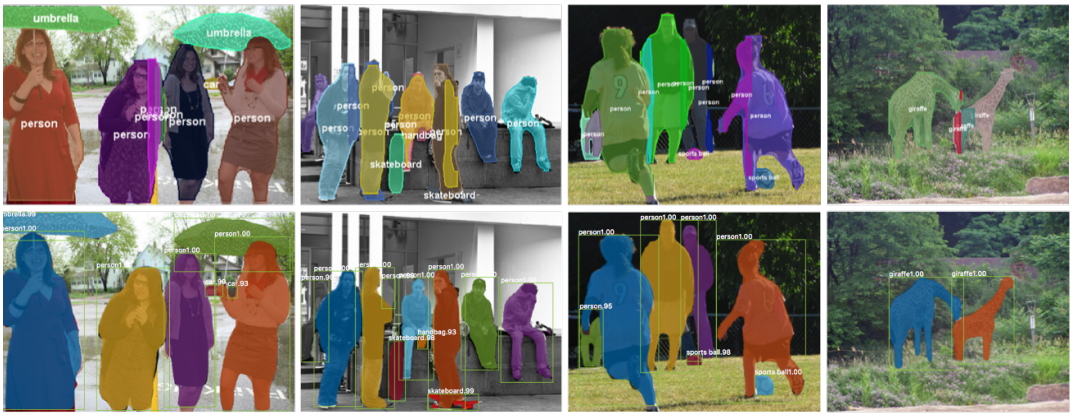


Figure 2.9: Use of CNNs have recently been shown to have accurate per-pixel object detection and classification. *Top row:* FCIS+++ [93]. *Bottom row:* Mask R-CNN [58]. Mask R-CNN outperforms FICIS+++ for overlapping instance segmentation.

2.2 Recognition Strategies

The International Conference on Document Analysis and Recognition (ICDAR) Robust Reading Competitions [77, 79, 104, 105, 133] broke down the issue of text extraction into two sub-problems: text locating and character recognition. Most of the literature discussed in Section 2.1 focused within the text locating sub-problem. For character recogntion, the *Focused Scene Text* word recognition task[3] received three entries in 2011 and four in 2013. In 2015, the recognition challenge was redesigned[4] using a new (and more challenging) *Incidental Scene Text* dataset of photos captured in-the-wild using Google Glass. The evaluation scheme uses the *Total Edit Distance* metric (described in [79]) and additionally the number of correctly recognised words for qualitative analysis. Table 2.1 summarises the top-scoring recognition rates from these competitions.

Table 2.1: Top-scoring word recognition results from ICDAR 2011–2015.

Year	Method	Dataset	Total Edit Distance	Correctly Recognised Words (%)
2011	TH-OCR System [100]	Focused	176.23	41.2
2013	PhotoOCR [13]	Focused	122.7	82.83
2015	MAPS [84]	Incidental	1128.0	32.93

It has been widely demonstrated that off-the-shelf commercial and open source OCR packages are able to correctly recognise text once the characters are extracted. In their conclusions of the ICDAR 2015 competition, Karatzas et al. [77] note that the top performing methods will make use of commercial OCRs and conclude that conventional shape-based OCR engines can produce competitive results with pre-processed images.

This conclusion is emphasised in further works. The Open Source Tesseract OCR engine was used in Ben-ami et al. [9] to extract RBNs after preprocessed CC-based extraction. Leading commercial OCR engines used by Chen and Yuille [27] were able to achieve 93% recognition from binarised text regions after AdaBoost non-text classification, using ABBYY FineReader[5], TOCR[6] and Readiris Pro[7]. Similarly, Gatos et al. [48] used ABBYY FineReader and showed their extraction method gave a 50% improvement for indoor and outdoor scene images, an approach also applied by Chen and Yuille [27]. This is not to say that OCR engines

[3] See Challenge 2, Task 3 in [79, 133].
[4] See Challenge 4, Task 3 in [77].
[5] `https://www.abbyy.com` last accessed 3 July 2017.
[6] `http://www.transym.com` last accessed 3 July 2017.
[7] `http://www.irislink.com/readiris` last accessed 3 July 2017.

are always needed.

Recent interest in developing novel general character recognition strategies have also been investigated. Wang et al. [157] compared ABBYY FineReader with their novel PLEX approach and demonstrated that their text recognition system can outperform traditional OCR engines without the use of a text detector. This inspired more recent works: in 2016, Lee and Osindero [88] developed a lexicon-free photo OCR framework. Furthermore, application-specific recognition has been investigated using variant techniques.

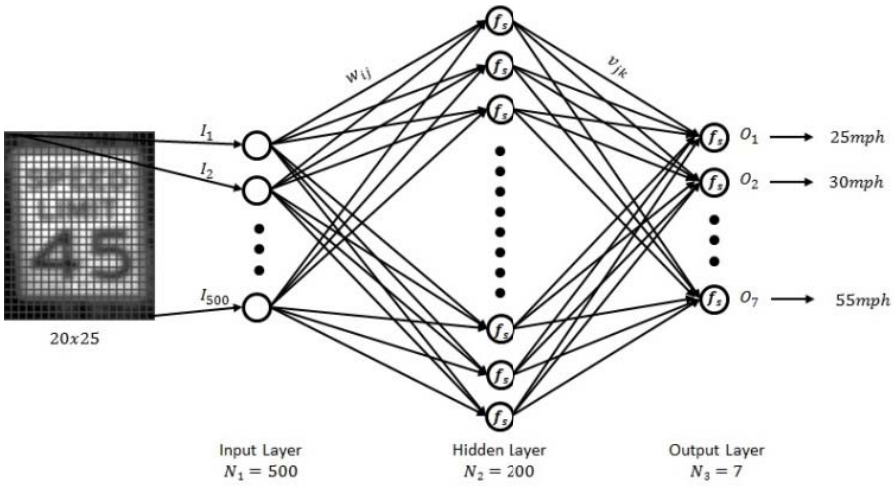


Figure 2.10: The artificial Multilayer Perceptron (MLP) NN designed in Kundu and Mackens [85] to recognise US-style speed limit signs.

A 2015 study into TSRs systems to detect US-style speed limit signs achieved recognition without the use of any OCR packages. In [85], Kundu and Mackens were able to extract a speed limit sign via the use of MSERs and template matching. The resulting detected signs were scaled to a grayscaled size of 20×25 pixels and fed into a Multilayer Perceptron (MLP) NN of 200 neurons in the hidden layer. The output layer of the network consisted of seven nodes, each representing the seven kinds of speed limit signs in US cities (25, 30, 35, 40, 45, 50, and 55 miles per hour). This architecture is shown in Figure 2.10. When trained with 13,289 images of text cases and 4,319 non-text cases, the results showed that their recognition classier was able to correctly recognise speed limit signs with an accuracy of 98.04%. Similar results were achieved using a feed-forward MLP in [37], using UK/Poland style speed limits scaled to 20×20 pixels (grayscale) and 12 output layer neurons ($10 \ldots 100$ kilometres per hour, the national speed limit sign, and non-sign neurons).

However, works in TSR systems that utilise NNs are generally non-generalisable, and only work in a limited context (i.e., by classifying speed limit signs of known outputs). In our

context of RBN recognition, we have a known character output range of 36 possibilities: 0–9 and A–Z = 10 + 26. Additionally, we consider the figure dash as these also consist within our dataset: thus 37 possibilities. We do not consider lowercase letters.

Beyond TSR systems, however, we see the use of more generalisable networks: Netzer et al. [117] trained neural network to recognise street number characters from Google Street View (using the Street View House Numbers (SVHN) dataset). They produced higher precision and recall than that of HOG and the Tesseract OCR engine, showing that the applicability of NNs for recognition can outperform traditional means. Anagnostopoulos et al. [2] used a Probabilistic Neural Network (PNN) to recognise single characters of the same 36 possibility range (i.e., uppercase alphanumeric characters) in the context of LPR, corresponding to the input grayscale vector of 9×12 pixels ($9 \times 12 = 108$ input neurons) for a single character. Figure 2.11 illustrates the PNN architecture used in this study. Furthermore, investigations in comparing different architectures of NNs for this context is given in Lee and Osindero [88].

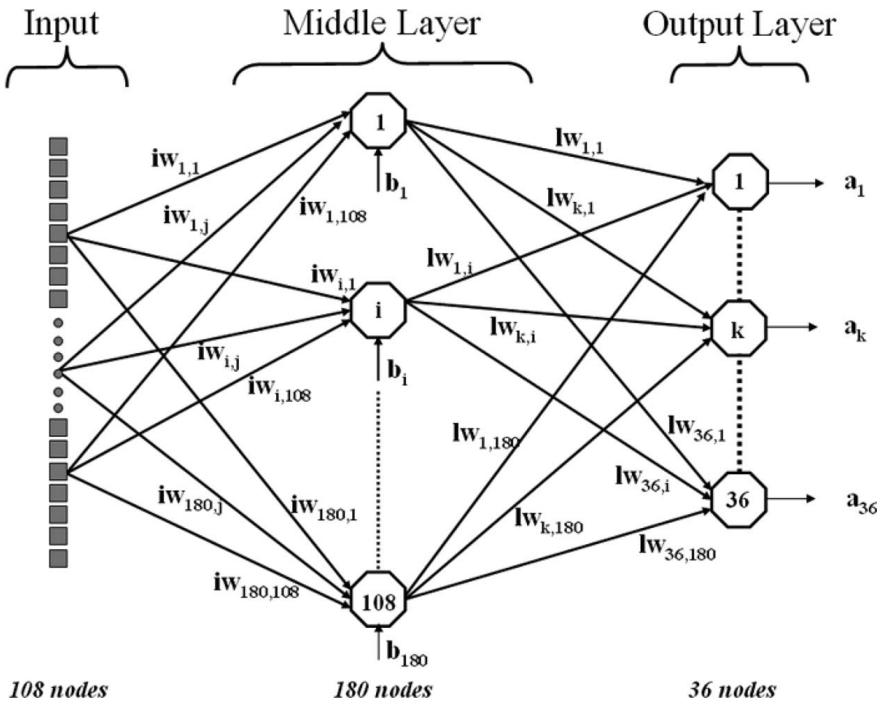


Figure 2.11: Anagnostopoulos et al. [2] developed the architecture of a PNN to determine a single character from within a license plate.

However, these pipelines are still entirely dependent on good detection strategies, and in the case where classifiers are used, quality training data must be supplied. The pipeline developed by Ben-ami et al. [9] for RBN detection (Figure 2.12) is dependent on quality facial detection: the OpenCV implementation [96] was used in their study. In order to detect

the torso region—and thus detect the bib sheet itself—a heuristically-driven calculation is used to hypothesise where the torso bounds ($T_b = T_h \times T_w$) are, given by the face height (F_h), face width (F_w):

$$T_b = (3 \times F_h) \times (\frac{7}{3} F_w)$$

The location of T_b is horizontally located at the centre of the face midpoint and vertically halfway below the face height ($0.5 \times F_w$). Hence, the dependency on these heuristics are heavily driven by a tolerant bounding box, which in itself depends on the face detection. Additionally, these heuristics are not always accurate—Figure 2.14 highlights a case where the face detection approach has missed the RBN entirely. As Fu et al. [47] have also noted, even if the face is detected, the bib sheet can be obfuscated (e.g., by another runner) or if the camera is capturing on the runner's side, the OCR engine or face-detection algorithm may produce poor results. In our study, we ignore these limitations for the intention of capturing prominent runners.

Furthermore, the use of Tesseract in the study meant significant filtering, separation and character alignment is needed for the OCR engine to read characters correctly (Figure 2.13). This is yet another step that can possibly be avoided via the use of well-trained NNs, as demonstrated in previous works. Investigation into applying such networks in this context (i.e., alphanumeric sequences on *human* subjects) is largely lacking.

Figure 2.12: The RBN recognition pipeline by Ben-ami et al. [9]. *From left to right:* Input image; face detection results in red and the estimated RBN region hypothesis in green; SWT of the hypothesis region; tag region detection in blue; tag region after digit processing.

Figure 2.13: Character processing in [9]. *From left to right:* A detected tag; separation via SWT CC analysis; binarised characters; CC analysis and filtering; separation and alignment.

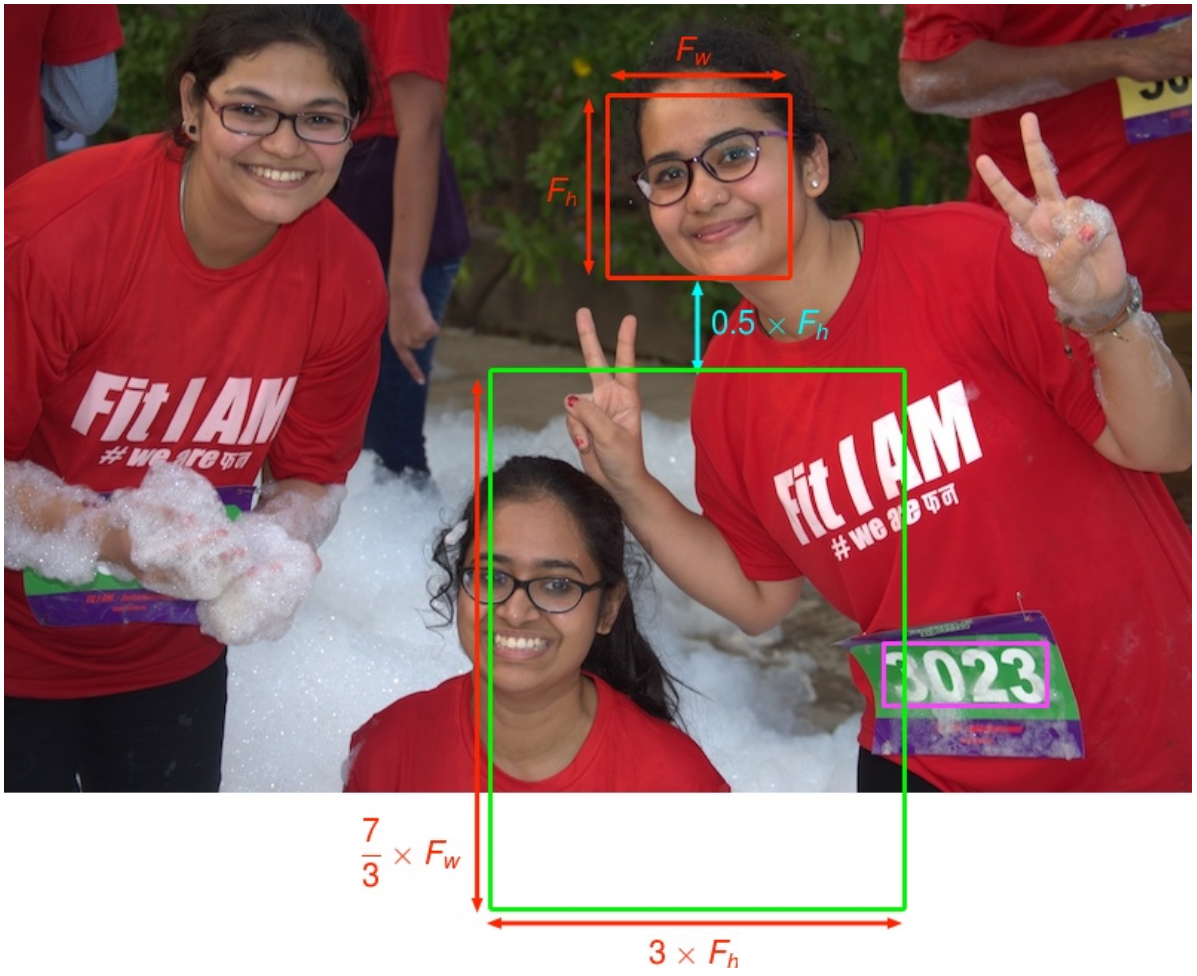


Figure 2.14: Cases where false negatives occur with the heuristic-based approach proposed in [9]. The expected RBN region (shown in magenta) is missed as the subject is leaning in the photo. Face detection in this photo uses the same OpenCV implementation [96] that was used in the study.

2.3 Metrics

Within the field of text detection, various evaluation protocols exist. For an extensive survey of these protocols, see [163, 167]. Throughout our survey, we note the standard the evaluation scheme for text extraction first proposed for use in image processing in the International Conference on Document Analysis and Recognition (ICDAR) competitions [77, 79, 104, 105, 133]. 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' that a text extraction pipeline can achieve.

In the ICDAR competitions, two metrics are used: the initial ICDAR 2003 evaluation protocol proposed by Lucas et al. [105], and, more recently, the *DetEval* evaluation protocol [135] (based on [163]), as used in the ICDAR 2011 and 2013 robust reading competitions. The simplicity and continued wide use of the ICDAR 2003 for text localisation evaluation is chosen over the *DetEval* approach.

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. [31], precision refers to the proportion of relevant matches actually retrieved in the 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 [128].

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 [105]. Therefore, precision is the number of correct candidates (c) divided by the number of total estimates found (E):

$$p = \frac{c}{|\,E\,|}$$

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

$$r = \frac{c}{|\,T\,|}$$

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. [105] 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, the match area m_a, 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 minimum bounding box containing both rectangles (i.e., the union of both) [104–106]. This metric is otherwise commonly referred to as the Intersection over Union (IoU) [68, 77, 98]. Using $a(r)$ to denote the area of the rectangle, we can represent this as:

$$m_a(r_1, r_2) = \frac{a(r_1 \cap r_2)}{a(r_1 \cup r_2)}$$

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:

$$m(r, R) = \max \{ m_a(r, r') \mid r' \in R \}$$

Lastly, we can redefine the recall and precision metrics to be more forgiving:

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

One issue with the calculation is that it assumes a rectangular bounding box. Words of non-rectangular nature, such as curved or rhomboidal text, are not well-represented.

2.3.2 The f-score

A common metric used when developing text extraction pipelines is the f-score, a single measure of quality that combines both precision and recall values. We are able to compute this metric using the standard measure across many studies, as contrasted in Table 2.2.

The f-score algorithm is given in the context of image processing in Lucas et al. [105]. Relative weights controlled by an α value of 0.5 give equal weight to both precision and recall metrics:

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

As Chen et al. [24] report, even when all text is correctly localised, it is likely that the f-score will vary between 0.8–1.0. This is because E boundaries are unlikely to match *exactly* with the manually labelled T boundaries. An illustration of this is phenomena is shown in Figure 2.15.

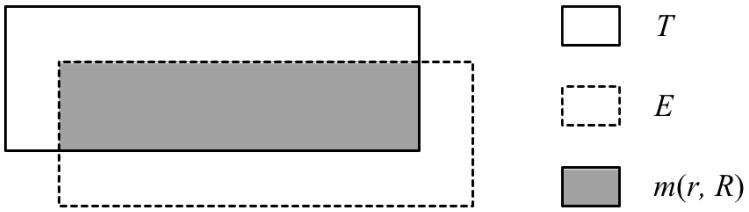


Figure 2.15: Overlapping areas of the ground truth targets, T, the estimated target boundaries E and the best match $m(r, R)$. (Adapted from [168].)

Table 2.2: A survey of text extraction literature, separated into CC- and learning-based detection methods.

Ref.	Year	Scientific Background		Precision	Recall	f-score	Dataset(s)	Performance	
		Detection	Recognition					Platform	Time (s)
[9]	2012	*CC-based*: SWT; Canny-Edges; Binary Conversion; Geometric Filtering; CC-Alignment	OCR (Tesseract)	0.65	0.62	0.63	N/A	N/S	N/S
[24]	2011	*CC-based*: Canny-Edges; MSER; SWT, Geometric, Template Matching	OCR (N/S†)	0.73	0.60	0.66	[104, 105]	2.5 GHz	0.20
[92]	2012	*CC-based*: MSER; geometric and SWT filters; skeletal distance mapping	N/S	0.59 [105] 0.59 [133]	0.59 [105] 0.62 [133]	0.59 [105] 0.61 [133]	[104, 133]	N/S	N/S
[172]	2011	*CC-based*: character detection with HOG and SWT; link energies via spacial relationships between characters	N/S	0.73	0.62	0.67	[105]	N/S	N/S
[135]	2011	*CC-based*: Fourier-Laplacian Filtering; clustering based on maximum distance via K-Means; skeletal distance mapping	N/S	0.76 [105] 0.81 [63]	0.86 [105] 0.93 [63]	0.81 [105] 0.87 [63]	[63, 105]	2.0 GHz	7.80
[38]	2010	*CC-based*: Canny-Edges; SWT; Modified CC algorithm; Geometric Filtering	OCR (N/S)	0.73	0.60	0.66	[104, 105]	N/S	0.94
[171]	2010	*CC-based*: Canny-Edges; HOG; geometric filtering; graph spectrum	N/S	0.67	0.46	0.55	[105]	N/S	N/S
[168]	2005	*Learning-based*: SVM to reject non-text; wavelet movement; HOG; OCR filtering	N/S	N/S	0.97	N/S	[63]	1.6 GHz	8.30
[122]	2010	*Learning-based*: Waldboost Classifier; HOG; LBP; Gabor Wavelets	N/S	0.56	0.70	0.68	[105]	3.4 GHz	0.37
[52]	2004	*Learning-based*: Wavelet transformation; feature estimation; pixel-block classification	N/S	0.87	0.90	0.88	[62]	N/S	N/S
[56]	2009	*Learning-based*: Mean Difference, Standard Deviation and HOG features; AdaBoost and CAdaBoost classifiers; NN-baed localiser	N/S	0.25	0.35	0.35	[105]	2.2 GHz	N/S

†Not Specified

2.4 Summary

In this chapter, we present a survey of literature in various application contexts: RBN and TSR recognition, recognition of alphanumeric sequences 'in the wild', and additionally object instance segmentation. We also present the varied range of techniques used to both detect and recognise the text, using both heuristic-based and NN-based approaches. This said, we acknowledge that datasets within the survey are not always recent (ranging as far back as 2003) due to tendencies for researchers to use the more popular (though dated) ICDAR datasets.

The state of the art of learning-based detection approaches, such as CNNs, have gained wide popularity, albeit for object segmentation. These approaches are yet to be applied within the context of alphanumeric sequences. Recent years have had a heavier focus on heuristic-based detection strategies using CC-based methods, while a majority of learning-based detection methods have had far fewer recent investigations. Furthermore, recognition of characters using NNs are not yet widely used, and off-the-shelf OCR packages are still standard.

Chapter 3

Dataset

How do we efficiently create, organise and use labelled datasets to permit useful machine learning? What is a label, and how can we constrain them? In this chapter, we present a novel metamodel that offers the conceptual vocabulary and an organising principle to improve efficiency of labelling.

3.1 A Data-Capturing Metamodel

Why should we care about the architecture of our dataset capturing, so long as the Artificial Intelligence (AI) is well-trained? This question fundamentally leads to the *provenance* or *lineage* of our data: not all training data is initially perfect. At times, we will need to modify our training data, or in the cases of some AI systems, training data is re-fed back into the system from the AI itself.

Hence, it is imperative to understand how the transition and flow of data occurs [14, 33, 65] in these AI systems. Where does it comes from? How is it derived and updated, and how might this affect our trained AI models over time? Without a thorough understanding of data provenance, tracing potential defects within training models becomes difficult. It may be a difference in the data training format, or a mismatch of values on the same training data. Systematic record-keeping is critical to overcome this challenge.

We propose a grammar to capture the vocabulary of our metamodel and the approach of developing a data-capturing architecture. As such a grammar is textual, it can be version controlled: therefore data provenance is recorded, which allows us to source how the mutations in training data can occur and and what effect this has our AI systems. From an exploratory analysis of developing a concrete implementation of this system (Section 3.2), we propose a conceptual metamodel that can be used to annotate and formally describe image data for train-

ing purposes (e.g., frames of a video, images of natural scenes). This metamodel is largely inspired by the works of Wickham [160, 161] and Moody [114].

3.2 Methodology

To understand the methodology on how we captured our dataset, we must first introduce the three key notions behind Model-Driven Engineering (MDE): technical spaces, models and systems. A system is a concrete "group or set of related or associated elements perceived or thought of as a unity or complex whole" [121]. Technical spaces were introduced by Ivanov et al. [66] as a model management framework based on algebraic structures (e.g., trees, (hyper)graphs, and categories). Technical spaces are usually based on a three-tier conjecture: metametamodels, metamodels and models. Whereas a model is an abstract representation of a concrete system of specific purpose, a *meta*model, in contrast, describes the way to describe those models. A *metameta*model can be used to describe the representation structure of our metamodels and defines a type system [18] that supports all underlying layers [11]. Figure 3.1 captures these concepts in further detail.

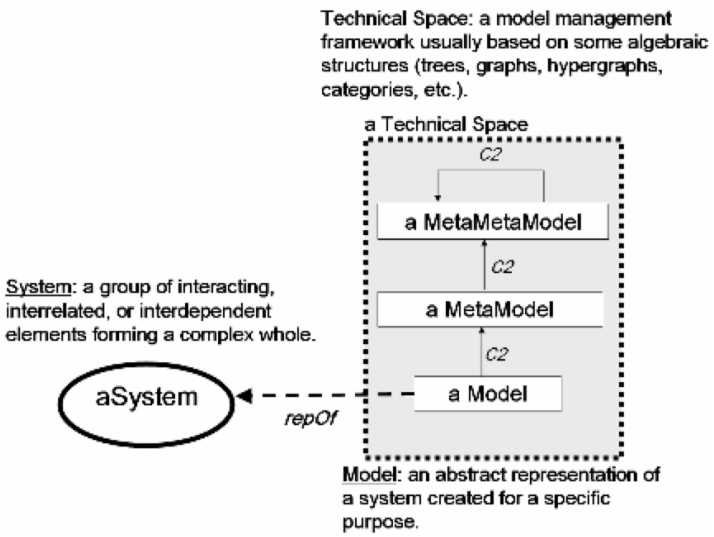


Figure 3.1: Systems, models and technical spaces. (From [11].)

We conceptualise our technical space as a series of layers (Figure 3.2), whereby the innermost layer is the metamodel and the outermost layer is the system. To develop our metamodel, we adopt a pragmatic approach by working *backwards* from the system to the metamodel: using a prototypical exploration process, we developed a prototype system (*Argus*, described in Chapter 4) iteratively with a data tagging team who manually curated the data by observing

what features were missing. We use the term *iteration* to refer to the process of deploying Argus to the data tagging team, annotating a set of photos from various races, and sending the data back for quality assurance. Once this prototype was stable, we conceptualised a model (Section 3.3) from the prototype to describe what it is that Argus captures. We then generalised that model, expanding from a model that was marathon-specific, into a more generalisable metamodel (Section 3.4). We represent this overall process as as a Unified Modelling Language (UML) activity sequence diagram, shown in Figure 3.3.

Formally, we adopt a mixed methods approach that combines a quasi-experimental design with an observational study [54, 150]: each of these iterations are considered as a quasi-experiment, where the previous iteration acts as a control to the following iteration without random assignment. Between iterations, we apply a treatment (software improvement) that we deduce from an observational assessment on the behaviour on how people use the software, and therefore to determine how to improve our tagging tool.

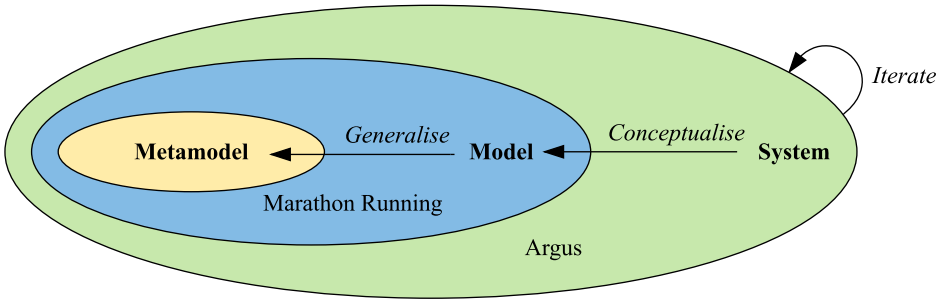


Figure 3.2: A layered discovery process from iteratively designing a system, conceptualising the system to a model, and generalising the model into a metamodel.

The initial phase in our dataset development was to create a prototype system with the specific purpose of tagging our marathon runner context (using the dataset provided by the industry client).

To develop a model (that informed our metamodel), we began by discussing what requirements were needed—that is, the key features that we thought were necessary for extracting from our image. Five features were decided: (1) the crowdedness of the photo, (2) the visible bib sheets within the photo, (3) the faces corresponding to those bib sheets, (4) the prominence of runners of the photo and (5) the colours of runners' clothing in the photo.

Once an initial prototype was developed, we conducted informal usability tests with members from DSTIL[1], which captured minor flaws in the workflow (namely required conditions

[1]These members were research colleagues. Refer to Appendix A.

that were missing in annotations) as well as general usability enhancements. We achieved an average System Usability Scale (SUS) score [72] of 70 of 100. These are noted in Appendix B. Once the internal testing had concluded, we developed instructional video guides on how to use Argus, and then deployed the tool to a remote data tagging team.

We ran four iterations of tagging with our data taggers. These iterations consisted of photos from different marathon races in our dataset. (To ensure variance in tagging, there were some races at night and some races with alphanumeric components in the RBN.) After each iteration, the research team assessed the tagging for quality. We explain this quality evaluation further in Section 4.2.3. Feedback identified further restrictions that needed to be placed on Argus as poor approximations or incorrect tagging would cause our AI to learn poorly.

The following issues were identified:

- the face region tagged was too far away from the bib region,
- the face region tagged was overlapping the bib region,
- some RBNs had spaces, and
- some RBNs were misidentified, where the alphabetic 'I' was tagged as the numeric '1'.

To prevent further training photos from being labelled with incorrect data, and therefore improperly train the NN, we added geometric restrictions and conditions into the tool to prevent these errors from occurring at all (i.e., ensuring faces could only be marked above and horizontally near bibs). We also ensured that only alphanumeric characters can be entered as RBNs. Some of these issues are highlighted in Figure 3.10.

3.3 Motivating Case Study: What to Capture?

Let us consider the motivating case study of marathon races and determine the model of our dataset. What features do we want to capture from images within our dataset? To answer this question, we need to consider what information we see as relevant in a standard marathon photo (such as Figure 1.1). We will consider five distinct features in the following sections.

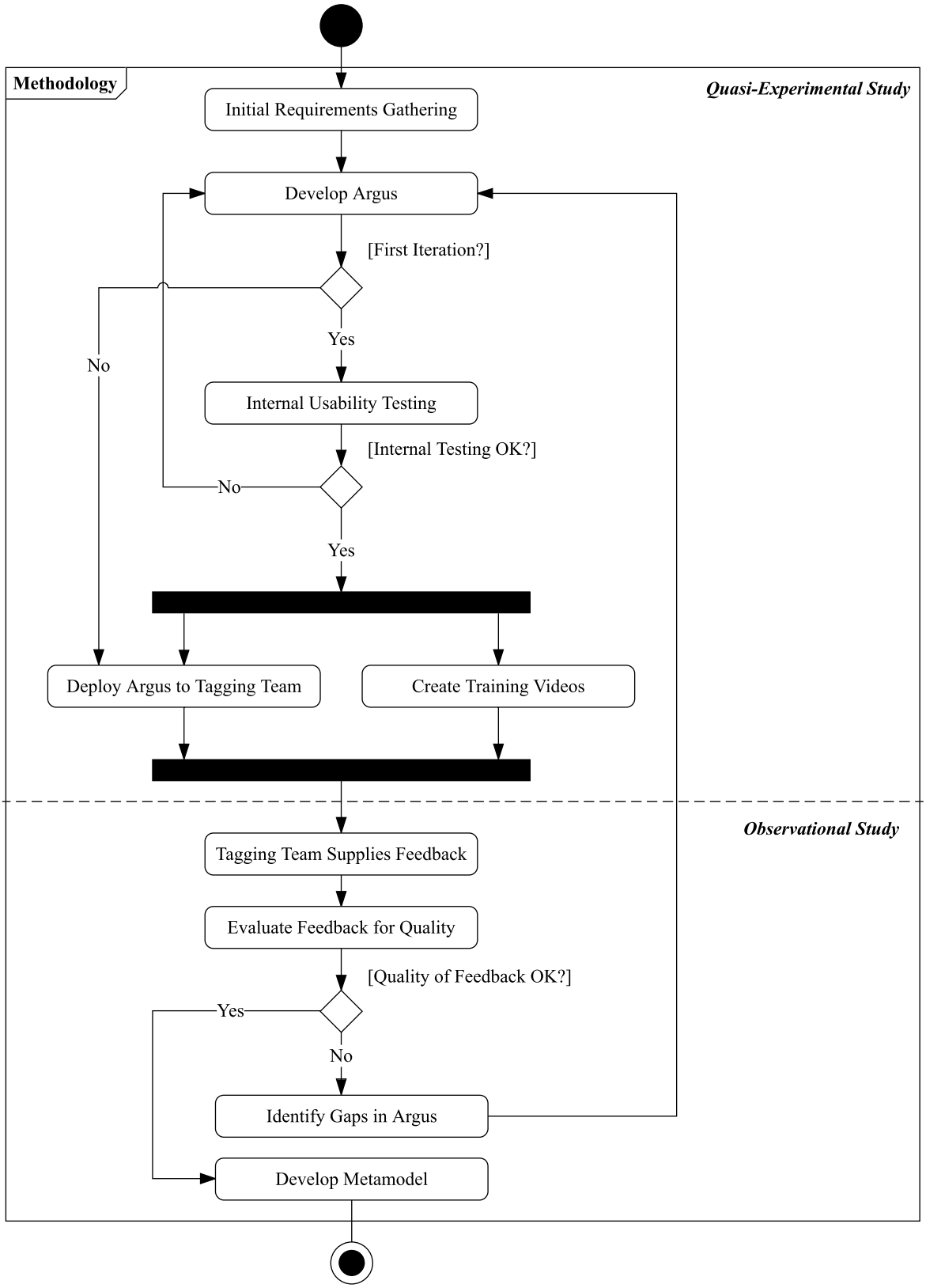


Figure 3.3: An overview of the methodology used to discover our metamodel. Formally, we consider our approach to be a mixed methods approach between a quasi-experimental and observational study for the relevant states.

3.3.1 Feature 1: Image-Level Features

There are two features to describe components of the entire photo that we call *Image-level* features:

1. whether or not the photo is considered as **crowded** (TRUE or FALSE), and
2. an *optional* collection of **runners** in the photo, given that the photo is not crowded.

(a) A crowded photo.

(b) A photo where all RBNs are cropped.

(c) A photo where there are no RBNs present.

Figure 3.4: Image-level features. In (a), the *PhotoCrowded* annotation would be marked as TRUE. Note the obstruction of all RBNs in this photo. In (b) and (c), *PhotoCrowded* is FALSE, but there are no *Runners* to tag.

We train a NN based on photos that are not crowded; photos that are considered crowded are typically not desirable as runners prefer photos where they are the key subject. We therefore discard these photos to remove any potential bias that the AI could learn. We can only tag runners on the condition that the photo is not marked as crowded (Figure 3.4a). Additionally, in some photos, all RBNs are missing within the photo (e.g., hidden behind other runners, cropped out of view) and some photos contain no runners at all (Figures 3.4b and 3.4c, respectively). We therefore identify this as an *optional* collection as we must associate an RBN

to a runner—the photo is not crowded, but there are no runners to tag, thereby making the annotation optional. Thus, a *derived attribute* exists with such a collection: the *count* of the runners in a photo is something we can automatically count.

3.3.2 Feature 2: Bibs

We identify the following two annotations of what we would label for a given runner's bib sheet. This is summarised in Figure 3.5. Within this feature we identify the following annotations for RBNs:

1. a polygon around the runner's **bib sheet** that contains the x and y coordinates, given that there are exactly four vertices of this polygon, and
2. a string with the runner's **RBN**, once the bib sheet is tagged (exists).

3.3.3 Feature 3: Faces

A further feature that is important is the runner's face region (Figure 3.6), which compromises of five annotations:

1. a rectangle around the runner's **face** that contains the two x and y coordinates of the two opposite vertices, given that the $bottom$ of this rectangle is above the top of the bib sheet,

and once this face bounds has been tagged, more annotations can be extracted:

2. whether or not the runner is **wearing** a **hat** (TRUE or FALSE),
3. whether or not the runner is **wearing (sun)glasses** (TRUE or FALSE), and
4. the **gender** of the runner (MALE, FEMALE, or UNSURE).

The bib sheet polygon would contain four *explicit attributes* required as input by the tagger: the coordinates of its four vertices, $\{(x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4)\}$. Similarly, there are two explicit attributes required for the runner's face rectangle: the coordinates of the top left and the bottom right of the rectangle (i.e., the two opposite vertices) of the runner's face bounds, $\{(x_1,y_1),(x_2,y_2)\}$. As both of these are shapes, we can automatically calculate derived attributes from these coordinates: $\{top, left, bottom, right, width, height\}$.

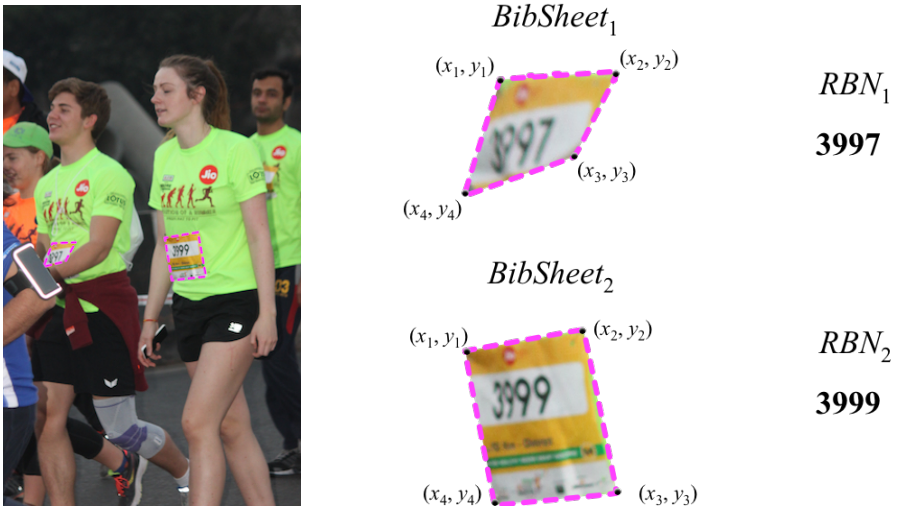


Figure 3.5: The bib segment-level feature. *From left to right*: The manually annotated bib sheets in the original photo (outlined in magenta); the relative *BibSheet* annotations with the four respective vertices; the detected *RBN* input strings for the corresponding *BibSheet*.

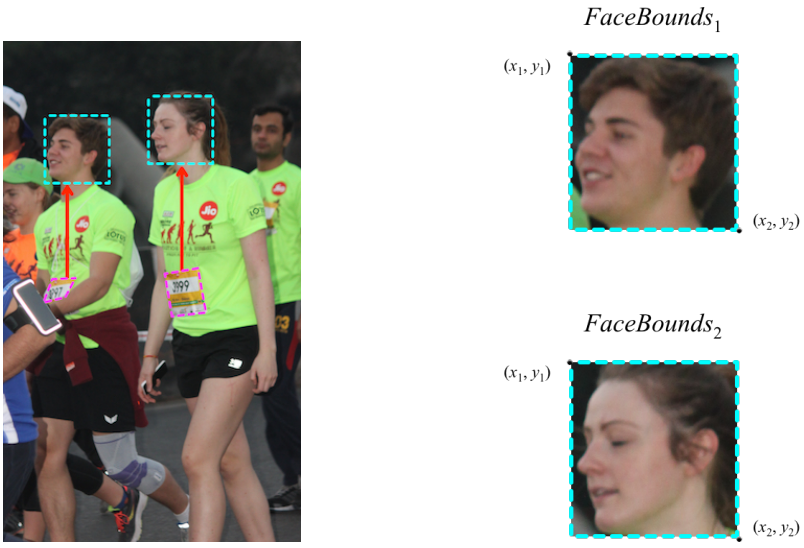


Figure 3.6: The face segment-level feature. *Left*: The manually tagged face regions (cyan) that comply with the conditions that the bottom must be above the top of the bib (red arrows). *Right*: the relative *FaceBounds* annotations with the two opposite vertices. Other annotations in this feature would be: $WearingHat1,2 = WearingGlasses1,2 = \text{FALSE}$, $Gender_1 = \text{MALE}$, $Gender_2 = \text{FEMALE}$.

3.3.4 Feature 4: Prominence

We identify another feature, the prominence of the runner within the photo. These consist of three annotations:

1. whether or not the runner's **face** is **visible** (TRUE or FALSE),
2. whether or not the runner is **blurry** (TRUE or FALSE), and
3. the **Likelihood of Purchase (LoP)** that the runner buys the photo (NO, MAYBE, YES).

A runner's face is considered 'invisible' if it is obstructed by another object, cropped out of the image, or if the runner is looking down. See Figure 3.7. Runners are more likely purchase a photo if they are looking at the camera, and likewise if they are not blurry in the image; these therefore have a significant impact on their prominence. We also define the Likelihood of Purchase (LoP) value as a qualitative metric. We ask the data tagger to 'picture' themselves as runner, and we ask if they would purchase it. We then use this metric to train positive samples of good photos (LoP = YES) and samples of bad photos (LoP = NO). Where MAYBE is provided, the runner is ignored to prevent training the network with indeterminate samples. Refer to Figure 3.8 for examples of the varying LoP values.

3.3.5 Feature 5: Colours

Another feature we can capture is a way to identify runners by the colour of their clothing. (Within our dataset, various clothing items of distinct colours are given to runners for particular races.) This feature comprises of four annotations:

1. an *optional* **colour** of the runner's **hat**, given that they were annotated as wearing a hat,
2. the **colour** of the runner's **shirt**,
3. an *optional* **colour** of the runner's **shorts**, and
4. an *optional* **colour** of the runner's **shoes**.

The colour of a runner's shirt is required, as we expect that a bib would be detected on the shirt of a runner (which would be tagged). The other colour annotations are all optional, as it is likely that some of these clothing items may be cropped out of the photo or are not visible. Furthermore, the hat colour can only be specified on the condition that the tagger has marked the runner has wearing a hat.

(a) Face obstructed

(b) Face cropped

(c) Face looking down

Figure 3.7: The *FaceVisible* annotation plays a significant impact on the prominence feature. Above are cases where the face would be marked as not visible.

(a) LoP = YES

(b) LoP = MAYBE

(c) LoP = NO

Figure 3.8: Various examples of the *LoP* values. In (a), the woman is clearly posing and is the primary subject of the photo. In (b), the runner is in focus in a pose, but is behind another runner and there are other runners behind him. In (c), the woman is out of focus, is blurry, and this photo has very poor lighting conditions.

3.3.6 Overview of Features

We segment groups of the annotations we extract into features of two categories: annotations that feature at the *image*-level and those that are not, which we call *segment*-level features. The image-level features are those which apply to the entire image (i.e., *PhotoCrowded*, *Runners*). Those that apply at the segment level can be grouped into what types of features we are extracting: the runner's bib, face, their prominence and their colour identification.

In summary, we have identified a total of 5 features of 15 annotations, which are summarised within Table 3.1. Each of these annotations can be classified with:

- a name and type,
- conditions for the annotation to be valid,
- dependencies on other annotations to exist before the annotation can be made,
- explicit attributes that the tagger must specify,
- derived attributes that we can automatically compute,
- possible values that limit the range of data for that annotation, and
- whether or not the annotation is optional.

Table 3.1: A summary of the annotations we wish to capture from our dataset. Image and segment-level features are separated using the double line.

Feature	Annotation	Type	Conditions	Dependencies	Explicit Attributes	Derived Attributes	Possible Values	Default	Optional
Image-Level	*PhotoCrowded*	Boolean	N/A	N/A	N/A	N/A	{TRUE, FALSE}	FALSE	No
	Runners	Collection	$\{PhotoCrowded = \text{FALSE}\}$	N/A	N/A	$\{count\}$	N/A	NULL	Yes
Bib	*BibSheet*	Polygon	$\{vertices = 4\}$	N/A	$\{x_1 \dots x_4, y_1 \dots y_4\}$	$\{top, left, bottom, right, width, height\}$	N/A	NULL	No
	RBN	Label	N/A	*BibSheet*	N/A	N/A	N/A	NULL	No
Face	*FaceBounds*	Rectangle	$\{bottom > BibSheet_{top}\}$	*BibSheet*	$\{x_1, y_1, x_2, y_2\}$	$\{top, left, bottom, right, width, height\}$	N/A	NULL	No
	WearingHat	Boolean	N/A	*FaceBounds*	N/A	N/A	{TRUE, FALSE}	FALSE	No
	WearingGlasses	Boolean	N/A	*FaceBounds*	N/A	N/A	{TRUE, FALSE}	FALSE	No
	Gender	Category	N/A	*FaceBounds*	N/A	N/A	{MALE, FEMALE, UNSURE}	MALE	No
Prominence	*LoP*	Category	N/A	*FaceBounds*	N/A	N/A	{NO, MAYBE, YES}	MAYBE	No
	FaceVisible	Boolean	N/A	*FaceBounds*	N/A	N/A	{TRUE, FALSE}	TRUE	No
	Blurry	Boolean	N/A	*FaceBounds*	N/A	N/A	{TRUE, FALSE}	FALSE	No
Colours	*ShirtColor*	Color	N/A	*FaceBounds*	$\{red, green, blue\}$	N/A	N/A	NULL	No
	ShoeColor	Color	N/A	*FaceBounds*	$\{red, green, blue\}$	N/A	N/A	NULL	Yes
	ShortsColor	Color	N/A	*FaceBounds*	$\{red, green, blue\}$	N/A	N/A	NULL	Yes
	HatColor	Color	$\{WearingHat = \text{TRUE}\}$	*FaceBounds*	$\{red, green, blue\}$	N/A	N/A	NULL	Yes

3.4 Describing the Metamodel

In the example given in Table 3.1, we describe a model that was derived using information discovered from Argus' incremental development. In this section, we generalise this information and develop a metamodel . This metamodel is represented as a UML class diagram, shown in Figure 3.9.

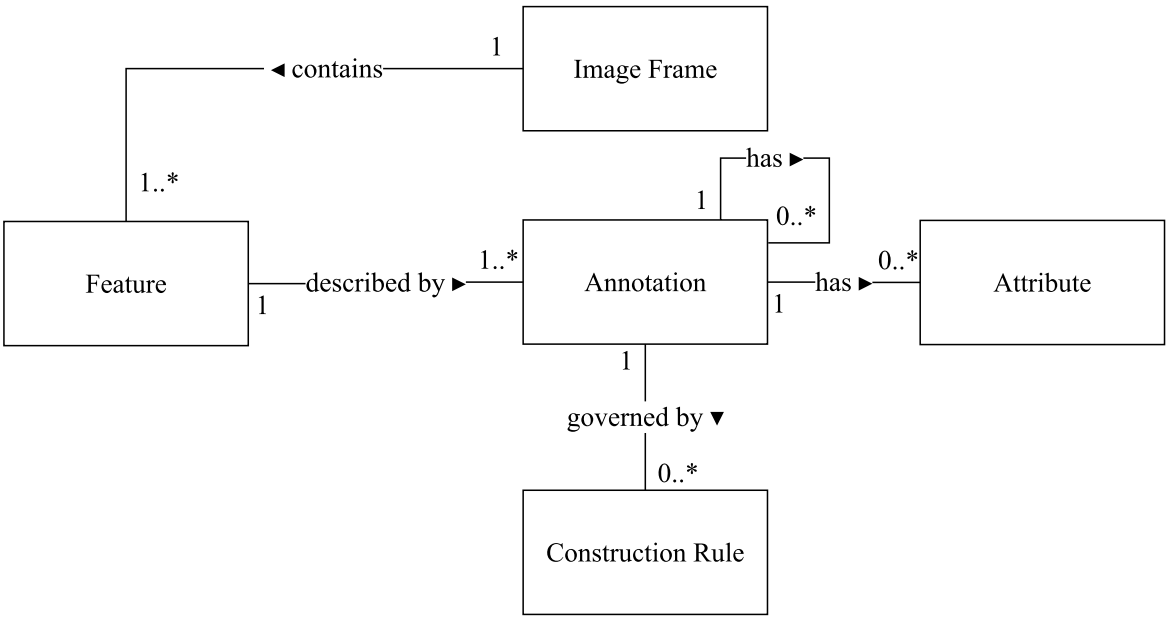


Figure 3.9: A class diagram of our proposed metamodel.

The core concept of this metamodel is an *Annotation* that (collectively) describe a feature within an image. Annotations may contain additional attributes to further extend what information they contain, and are governed by a set constrictions that we call *Construction Rules*.

These form layers of data within images that we annotate that can be represented in a hierarchical data format (e.g., a tree), thereby describing the layering of a fully annotated photo's features. These classes have further depth that are captured via inheritance models, presented in Appendix C.

We describe each of these classes in the following sections.

3.4.1 Image Frame

An image frame is any visual representation of a graphic. We extend the possibilities of our metamodel beyond just static images—these frames may be those found in videos also.

3.4.2 Feature

Image frames contain a certain number of features. These features are the key 'concepts' within the image that we want to capture. There are two main features: (1) image-level features, and (2) segment-level features. Image-level features are those that apply to the entire frame and do not exist in a broken down context—we can't break down these features into their own subjects. This contrasts to segment-level features: features that only apply to a segmented region of the image. We identified four segment-level features in our example (all of which relate to one runner): Bib, Face, Prominence and Colours.

3.4.3 Annotation

Annotations are a set of feature descriptors that describes what the feature is comprised of. From Table 3.1, we describe a simple type system consisting of the following: *Boolean*s, a restricted form of the *Category* type (where possible values are $\{\text{TRUE}, \text{FALSE}\}$); *Polygon*s and *Rectangle*s, generalised into a high-level *Boundary* type; and *Colour*s, represented as an RGB represented hexadecimal *Label* (a generic labelling type). This only leaves our *Collection* of runners: a data type that encompasses multiple annotations. Attributes can therefore be described by a type system that is generalisable from *Labels*, *Boundaries*, *Collections* and *Categories*. We model this as a type tree (Figure C.1).

3.4.4 Construction Rules

Construction Rules govern how attributes can be made. We break these down into four types: Dependencies, Conditions, Optionality and Restrictions (Figure C.4).

Some annotations are dependent on others existing—these are rules that we name *Dependencies*. For instance, we cannot annotate an RBN if there is no *BibSheet* annotated. Likewise, a *FaceBounds* is dependent on where the *BibSheet* is, and so the *BibSheet* must be annotated first. These dependencies add *order* to what features are being extracted and is important to the tagger marking up the photo by structuring a workflow.

A *Condition* is a construction rule that disallows a data tagger to create an annotation until a condition is met by the annotator. For instance, we can never have a *BibSheet* above a *FaceBounds*, so when we can specify this as a condition. These are not bound to just geometric constraints; we also see that we cannot tag *Runners* if the *PhotoCrowded* annotation is marked as TRUE.

Not all annotations are needed—in these cases the annotation is restricted by an *Optional* construction rule; by default, all annotations are required, but we can specify these to be optional using such a rule.

Lastly, *Restrictions* exist to prevent invalid data from being tagged. In our example, we implemented two restrictions in Argus: (1) a face region must be tagged within an aspect ratio of $3 \times BibSheet_{width} : 5 \times BibSheet_{height}$, and (2) the opposite edges of a face region must be at least 15 pixels apart. This helps prevent issues shown in Figure 3.10.

3.4.5 Attributes

Attributes exist as a means to capture further information about an annotation. We represent attributes in our metamodel as either derived or explicit. As stated in Section 3.3, those attributes that are required for the tagger to manually markup are explicit (i.e., manually must be added) whereas those that can be computed by the system are derived (i.e., inferred from explicit attributes). A useful example of derived attributes may be the bounding box and area that inferred from a polygon, both of which are key values required in training NNs.

3.5 The Data-Capturing Process

Now that we have defined *what* to mark up, we now describe the process to define *how* we capture it. To do this, we need to encode both the *description* and *constraints* of the data we wish to capture within our dataset. The following sections refer to *Argus* as a generic labelling tool, rather than its concrete implementation referred to in Chapter 4.

This can be done in a four-step process: (1) informing Argus about what it is that we want to capture; (2) extracting information from our dataset; (3) transforming the data into an encodable format; and, (4) loading the data into an AI model. Thus, the following sections describe the data capturing process as a Supervised Extract-Transform-Load (SETL) process.

3.5.1 Informing Argus

The data captured from Section 3.1 can be represented as a hierarchical data structure, reinforced by our metamodel. For convenience, we refer to the data structure persistence format as the Argus Data Format (ADF). We restrict the kind of data captured using a set of constraint rules. These are encoded into the Argus Constraint Language (ACL) file, which can also be represented by our metamodel.

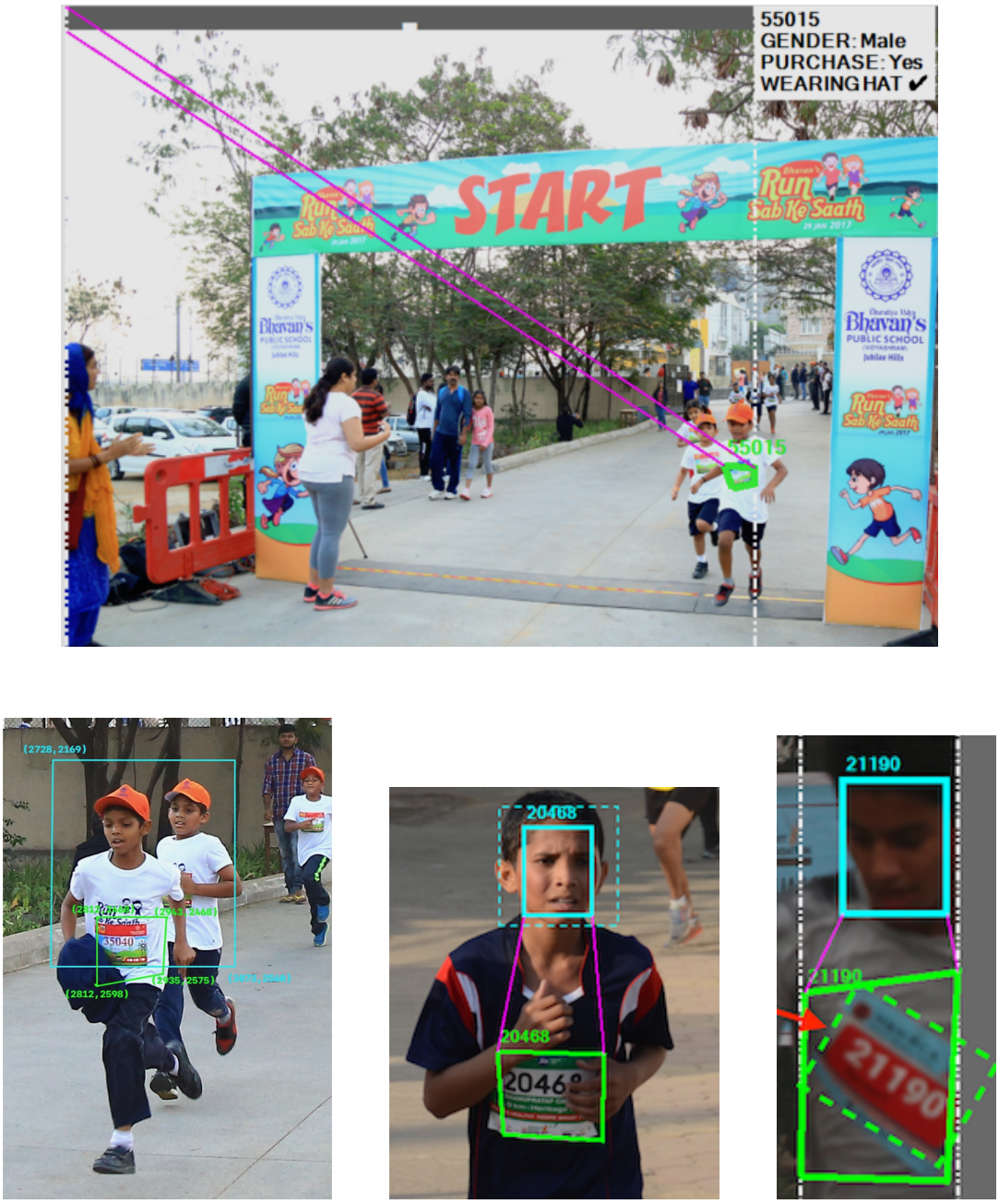


Figure 3.10: Issues identified from rounds of external testing to the tagging team. Cyan lines indicate face regions, lime lines indicate bib regions, magenta lines indicate bib-to-face guidelines. *Clockwise:* Face region is too far away from bib region; bib region is poorly marked (dotted lines expected); face region is too small (dotted lines expected); face region overlaps bib region.

The first step is to feed in the constrains into a data capturer that restricts what kind of data we wish to read from our image dataset. This declarative format informs the capturer of the potential features, annotations and construction rules that is marked up by the annotator. A concrete implementation of an ACL file in XML schema is given in Listing 3.1, where we define the features, annotations, conditions and dependencies of the image-level and *Bib* segment-level features, using ID referencing inspired by HTML.

Listing 3.1: An Argus Constraint Language (ACL) file describing the image-level and *BibSheet* segment-level features as represented in an XML schema.

```
<?xml version="1.0" encoding="utf-8"?>
<capture>
  <!-- Define the Image-Level Features -->
  <feature level="image" id="image">
    <annotation id="photo-crowded" type="boolean">
    </annotation>
    <annotation id="runners" type="collection">
      <!-- Condition that photo crowded is false -->
      <condition context="photo-crowded" value="self == false" />
    </annotation>
  </feature>

  <!-- Define the Bib Segment-Level Feature -->
  <feature level="segment" id="bib" owner="runners">
    <annotation id="bib-sheet" type="polygon">
      <!-- Condition that vertices must be 4 -->
      <condition context="bib-sheet" value="self.vertices == 4" />
    </annotation>
    <annotation id="rbn" type="label">
      <!-- Dependent on bib sheet to exist -->
      <dependency on="bib-sheet" />
    </annotation>
  </feature>
</capture>
```

Argus is therefore told what it is that the annotator needs to capture, and—by using this declarative language—a User Interface (UI) can automatically be generated to indicate to the annotator the flow by which we must mark up information.

Table 3.2: Sample UI elements that can be made to extract features.

Annotation Type	UI Element(s)	Usage
Boolean	Checkbox	Check to indicate TRUE value, FALSE otherwise.
Category	Dropdown-Box Radio Button Group	All variant Possible Values are shown as distinct options within a dropdown box's list items or as separate radio buttons.
Label	Text Box	Text box shown to enter in label value.
Colour	Eye-Dropper Colour Wheel Colour Palette	Selection of colour can be made from eye dropping a distinct colour in the image, from a selection of all colours in the colour wheel, or from a set of colours from a set colour palette.
Polygon	Clicking $n_{vertices}$ times	Boundary is selected by clicking the required number of vertices around the selection area.
Rectangle	Drag-And-Drop	A drag-and-drop around the required boundary is made.

3.5.2 Extracting Features

Raw images are loaded into an 'Annotate' mode of Argus, whereby the user interface is generated using the constraints discussed in Section 3.5.1. Additionally, errors are now detected when conditions or restrictions are not met. Each of the specific features can now be extracted using a number of the elements within the generated UI as per those shown in Table 3.2, whereby the annotator follows the required steps (i.e., in dependency order) to carry out annotations. This, therefore, becomes the workflow of extraction which the annotator runs through—an example of such a workflow for our case is shown in Figure 3.11, represented as a UML state diagram.

Additionally, the UI elements can provide guidance to the tagger in the form of an instruction. For instance, annotations of a boolean type can be asked as a question: "Is this photo crowded?". Polygons can be specifically guided: "Left click four times around the bib sheet region". Similarly, the same can be done with a colour selection using an eye-dropper: "Click the shorts of the runner to select their shorts colour". These labels and UI elements can be automatically generated based on the information supplied in the ACL file.

Keyboard shortcuts can be specified to improve keyboard-driven navigation, reducing the need to use a mouse and automating responses (i.e., enter 'Y' for 'Yes' and 'N' for 'No').

This is ultimately to speed up the progression by which taggers extract these features.

Lastly, annotators can mark the image as ‘complete’, indicating they are finished with all possible annotations in the dataset. This then produces an Argus Data Format (ADF) file based on the image (Section 3.5.3). Once all images in the dataset are fully ‘complete’, Argus informs that the labelling process on the entire dataset is now complete.

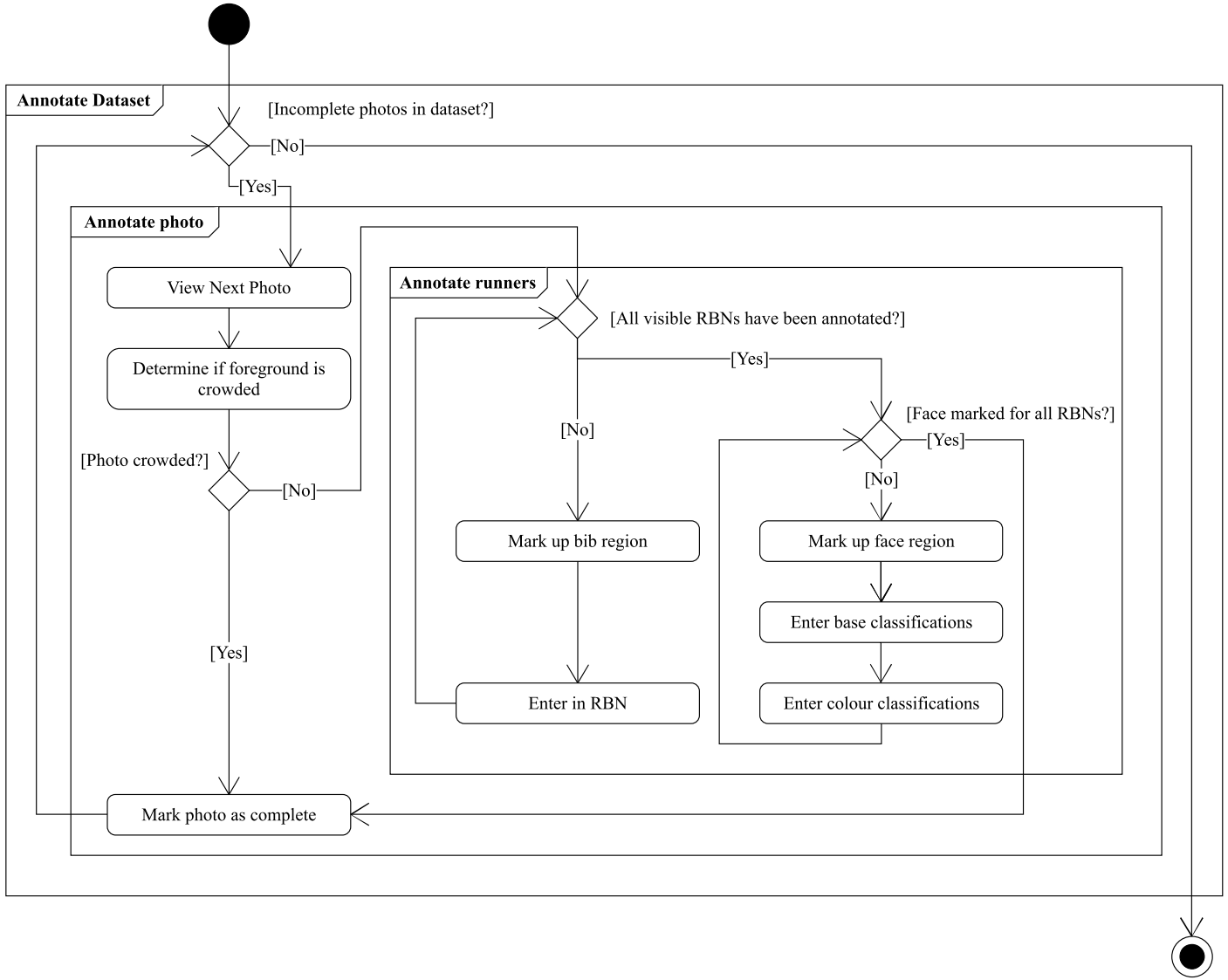


Figure 3.11: Concrete workflow of Argus in our bib annotation context. The ACL structures this process by defining dependency ordering.

3.5.3 Transformation

The features, once extracted, are encoded into an Argus Data Format (ADF) upon marking the image as complete. To prevent data loss, files are encoded as the annotator progresses onto the next feature within the workflow of extraction. All derived attributes can be calculated given the explicit attributes manually entered by the runner at this stage.

A sample of an ADF in a language-agnostic tree format is shown in Figure 3.12 for a photo with one runner. This could be serialised into any human-readable hierarchical format (such as a YAML, JSON, eXtensible Markup Language (XML) file or machine-readable compressed formats). Additionally, workflows from the ACL input file may be generated into a UML state diagram (as we have done in Figure 3.11) to assist annotators in understanding how to mark up the photo. Thus, the transformation step *universalises* the annotated data into a format readable by any supported parser of the ADF or ACL files. We propose examples of how this data is loaded in the following section.

3.5.4 Load

Once the data is transformed into a consistent format (i.e., an ADF), we are able to supply it to a number of different tools. In our research, we process the ADF in four tools:

1. Argus is executed in 'Review' mode, whereby a second annotator reviews the parsed annotations by loading a directory containing multiple ADFs. There, they can run through a number checks to ensure the annotated data is accurate for training the AI models.

2. Custom augmenters load the ADFs as well as the source images to produce a number of variations on the image, essentially turning one image and its respective ADF into many variant images and ADFs. This is imperative when training our Neural Network (NN), as explained in Section 3.6.2.

3. An adapter created for our NN is used to load in an ADF and convert it into a CSV of file paths to images, bounding boxes and labels. The NN reads this file and uses it to learn where bib locations are in an image. This is later discussed in Chapter 5.

4. We used Git[2] to track changes within our ADF. The benefit of serialising the data into a text file is that any version control management software can handle provenance for

[2] `https://git-scm.com` last accessed 11 Aug 2017.

- **Image** `001`
  - PhotoCrowded : *Boolean*
    - FALSE
  - Runners : *Collection*
    - Runner `1`
      - **Bib**
        - Bounds : *Polygon*
          - $x_1 = 1164, y_1 = 2990$
          - $x_2 = 2147, y_2 = 2990$
          - $x_3 = 2229, y_3 = 3745$
          - $x_4 = 1215, y_4 = 3807$
        - RBN : *Label*
          - `21090`
      - **Face**
        - Bounds : *Rectangle*
          - $x_1 = 512, y_1 = 1034$
          - $x_2 = 1671, y_2 = 2162$
        - Gender : *Category*
          - MALE
        - WearingHat : *Boolean*
          - TRUE
        - WearingGlasses : *Boolean*
          - FALSE
      - **Prominence**
        - LoP : *Category*
          - NO
        - FaceVisible : *Boolean*
          - FALSE
        - Blurry : *Boolean*
          - FALSE
      - **Colours**
        - ShirtColor : *Color*
          - $r = 255, g = 248, b = 239$
        - ShoeColor : *Color*
          - N/A
        - ShortsColor : *Color*
          - $r = 66, g = 63, b = 58$
        - HatColor : *Color*
          - $r = 248, g = 255, b = 134$

Figure 3.12: Sample hierarchical data representation of an ADF, encoding the features (bolded) of the runner on the left. Derived attributes are removed.

us (as suggested in Section 3.1), thereby sourcing changes in our AI models easier by sourcing how the training data (i.e., ADF files) has derived over time.

3.6 Data Postprocessing

Once our dataset is annotated, a number of actions can be applied—our images and annotations undergo a series of post-processing and eventually are used by the AI model to make predictions on new datasets. We can broadly categorise these into two sections: flagging the annotations/predictions and augmenting the data. The overall process is illustrated within Figure 3.14.

3.6.1 Flagging

Flagging helps indicate 'check'-states that the annotated or predicted ADF data has undergone. These checks can be considered as another way to maintain data provenance. We have considered four states of data: (1) ground truth annotations; (2) *reviewed* ground truth annotations; (3) predicted regions from an AI model; (4) *validated* predictions. These help track data provenance of training and inference samples, as presented in Figure 3.13 as two UML state diagrams.

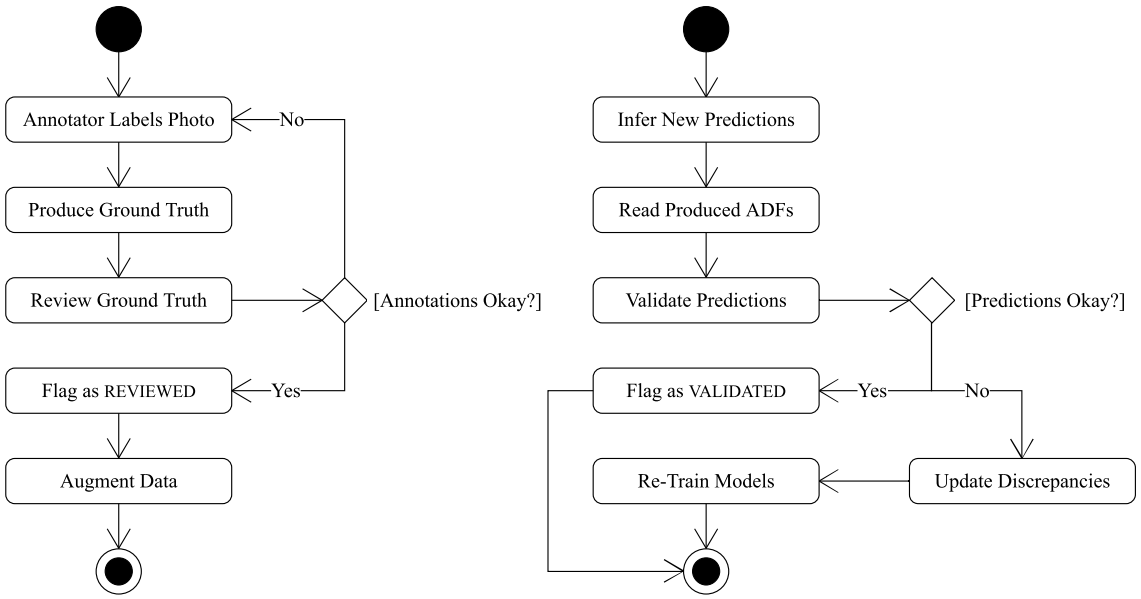


Figure 3.13: State diagram of for training data (left) and inference data (right).

Ground truth annotations encompasses the general process that data taggers annotate data.

This process is largely the initial data capture that occurs within our dataset, described by the process outlined in Section 3.5. As mentioned in Section 3.5.4, once data is annotated, a second parse is done on the annotations by another reviewer to check for any mistakes. Some of our data is also annotated multiple times, and thus multiple ADFs are produced. A reviewer will check these files and use the best matches to merge the best data together—for instance, in a single photo, if three annotators mark a given runner's LoP as YES whereas one annotator marks it as MAYBE, the reviewer may choose to update the MAYBE to the more likely value.

Once ground truth annotations are inspected for a second time, the annotations in the ADF file are *flagged* as 'REVIEWED'. We only allow reviewed data to be augmented (Section 3.6.2) and thus trained in our AI model. This method allows us to maintain a high level of quality and integrity of input data.

The second flag is set on *predicted* ADF data generated by the AI. As the AI model produces ADF files (see Chapter 5), we are able to perform a second parse on the predicted data by loading the predicted ADFs into Argus' validation mode. Here, a human validates that the predictions made by the model are indeed correct, and if not, we are able to update any discrepancies where needed (via Argus modifications) and feed this corrected invalid data into the model again to correct any mistakes it is making. Here we can flag the predicted data as 'VALIDATED'. In turn, predications that are validated help ensure that mistakes made by the AI model are amended.

Additional flags that may be considered for future work include 'CREATOR', a first class concept that captures the source of who initially created the data, reviewed the data and validated the data. In doing so, quality analysis of annotators can be tracked. Furthermore, a 'MERGE' flag could be considered, whereby the merge of multiple ADFs on the same source image are made. This would help track when such merges occur and how frequently they occur. We leave such suggestions open for future improvements of Argus.

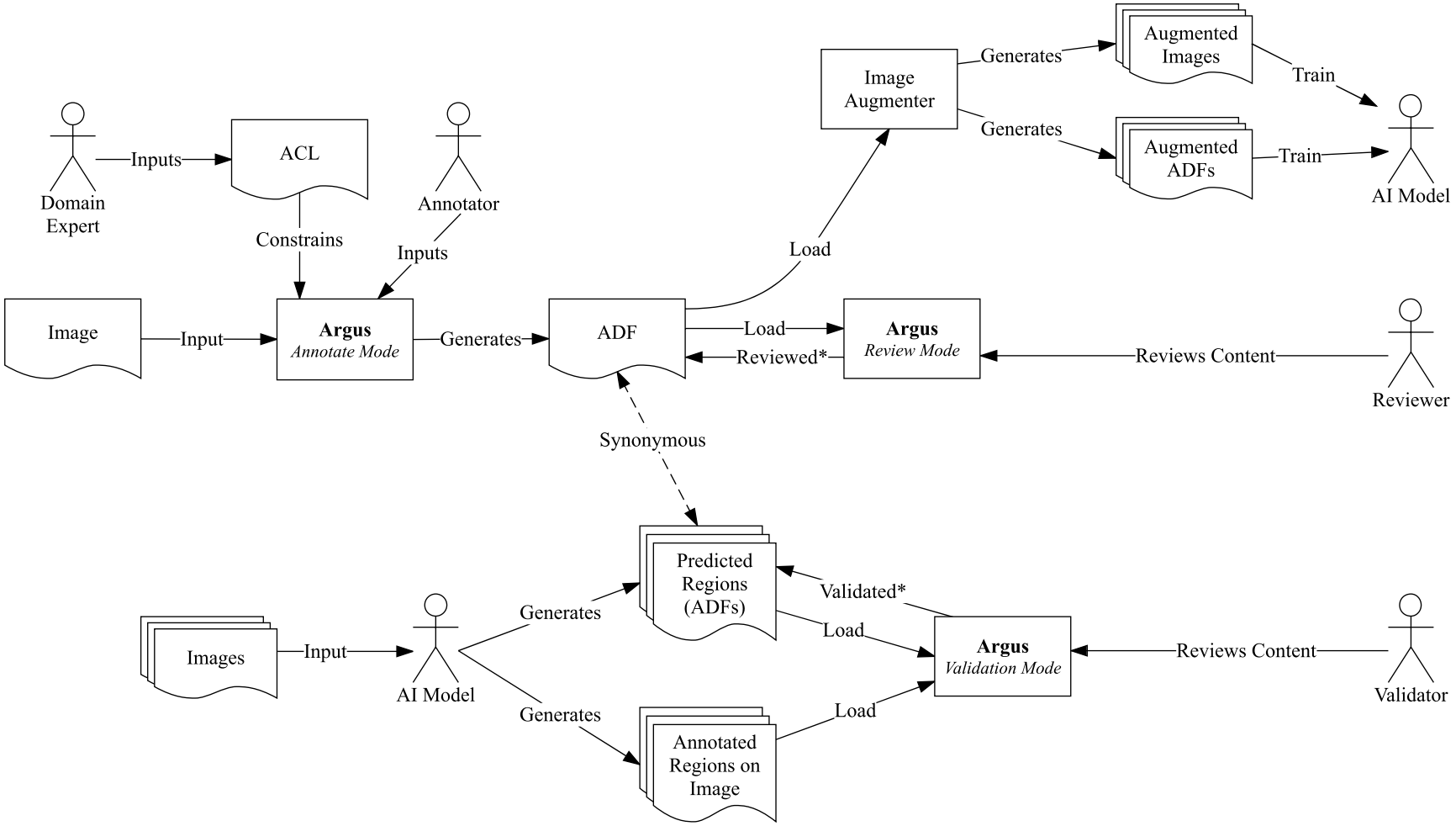


Figure 3.14: A general outline of the processing done on the dataset and all relevant actors and associated files. Asterisks indicate where flags have marked annotations as REVIEWED and predictions as VALIDATED.

Table 3.3: Breakdown of images and annotations (runners) used for training and validation.

Image Type	Training			Validation		
	Images	**Annotations**	**Annotations / Image**	**Images**	**Annotations**	**Annotations / Image**
Original	722	850	1.177	81	109	1.346
Augmented	36100	33098	0.917	4050	4266	1.053
Total	36822	33948	0.922	4131	4375	1.059

3.6.2 Data Augmentation

Previous works show that augmenting data makes a classifier more robust in its detection [4, 164, 166]. A solid augmentation strategy is applied to make the training data for our models more extensive and variable by generating deformations and defects and thereby producing synthetic data. We therefore motivate the need to augment our datasets by producing varied images of the same source image, rapidly fortifying our training data with more images in our dataset.

For our network, we manually tagged 803 images over a total of 14 different marathon events using our concrete Argus implementation (Chapter 3) within our dataset. We decided on an augmentation strategy using a 1:50 ratio (thereby to produce a further 40,150 augmented images from these original 803 images). We split the dataset into 90% for training and 10% for validation. The specific quantities of images and annotations chosen for training and validation are described in Table 3.3. An annotation to image ratio less than 1 indicates there were more crowded photos in the sample set.

Our augmentation strategy consisted of the following:

1. perform affine transformations on every image,
2. distort colour channels by adding intensities of $\pm 45\%$ applied 50% of the time,
3. distort colour channels by multiplying intensities between a range of $[0.5, 1.5]$ applied 50% of the time,
4. apply one of a gaussian, average or median blur 30% of the time.

Our affine transformations consisted of translations in both the x and y axes shifted between $\pm 35\%$, rotated between $\pm 45^{\circ}$, and sheared between $\pm 5\%$. These transformations were implemented using the *imgaug* library[3] for Python. See Figure 3.15 for examples.

[3] `https://github.com/aleju/imgaug` last accessed 11 August 2017.

As these transformations occur randomly, there are cases where translation, shearing and rotation occur extremely (e.g., Figures 3.15e and 3.15f). This causes some bibs to be cropped off the image. Therefore, we remove these from the transformed ADF file as we cannot have a bounding box referencing outside the image.

(a) Original annotated image.

(b) Slight rotation with blur.

(c) Darkened channels

(d) Slight translation.

(e) Darkened channels with translation.

(f) Large translation with strong rotation.

Figure 3.15: Various instances of our augmented dataset. Here the ADF annotation files are drawn directly onto the source image, with the *BibSheet* polygon shown in green and its respective derived bounding box attribute in red, dotted.

3.7 Metamodel Evaluation

We analysed seven popular datasets recently used in literature seen as benchmarks for text extraction or object recognition. Our criteria for selection was generally open, but we ensured that usage of these datasets were from papers within at least the last 5 years. We assessed their annotation format to determine the heuristic overlap between the dataset's annotation model and our metamodel. An overview of the datasets, annotation formats and capturing software is shown in Table 3.4. We summarise our analysis of directly mapping components from these formats within the ADF in Table 3.5.

Table 3.4: Various datasets and their respective annotation formats which we have assessed for comparison with our metamodel. See Appendix D for serialisable annotation formats.

Dataset	Reference	Challenge	Format	Listing	Annotation Tooling
MS COCO	[26, 98]	Object Recognition	MS COCO JSON	D.1	Custom Tool via AMT
COCO-Text	[154]	Text Reading	MS COCO JSON	D.2	Tool from Matera et al. [109] via AMT
ICDAR 03–11	[24, 104, 105]	Text Reading	XML	D.4	Custom Tool via Java Applet
ICDAR 11–15	[77, 79, 133]	Text Reading	XML	D.3	CVC APEP [78]
ImageNet	[35]	Object Recognition	PASCAL VOC	D.5	Custom Tool via AMT
SUN	[165]	Object Recognition	PASCAL VOC	D.5	LabelMe [129]
PASCAL-Context	[115]	Object Recognition	PASCAL VOC (2012)	D.6	Custom Tool similar to LabelMe [129]
Synth90K	[55, 67, 68]	Text Reading	MATLAB Binary	N/A	SynthText[4]
SVHN	[117]	Text Reading	MATLAB Binary	N/A	Custom Tool via AMT

Table 3.5: Presence of components of popular annotation formats within the ADF metamodel. See Appendix E for detailed mapping.

ADF		COCO JSON	PASCAL VOC XML	ICDAR XML	MATLAB Binaries
Feature	Image-Level	•			
	Segment-Level	•	•	•	•
Annotation	Label	•	•	•	•
	Category	•	•	•	
	Collection		•	•	•
	Boundary	•	•	•	•
Attribute	Derived	•	•	•	•
	Explicit				•

[4] `https://github.com/ankush-me/SynthText` last accessed 15 August 2017.

For each of these datasets, the annotation capturing strategies differed and, similarly, the annotation was either encoded as a binary MATLAB[5] file or serialised in plain text (either as XML or JSON). In the case where annotations were serialised in XML, a popular format was the PASCAL[6]-Visual Object Classes (VOC) [40] format. This format was made popular with the various PASCAL VOC Challenges [39–42], though custom annotation formats in XML are also present (e.g., ICDAR).

3.7.1 MS COCO and COCO-Text JSON Format

The Microsoft Common Objects in COntext (COCO) [98] dataset and its image captions extension [26] require all annotations to be labelled to one of 91 specified categories. This is done using a three-step annotation pipeline consisting of: (1) labelling all categories of objects in the image, (2) spotting all instances of these characters, and (3) segmenting the instances using a polygon. For our analysis, we focus only on the Object Keypoint and Image Caption Annotation challenges.

We mapped the annotation concepts from both MS COCO (and its text-reading equivalent COCO-Text) and mapped components to the ADF. As shown in Figure E.1, these high-level components are largely complete when expressed in ADF, and the mapping is powerful enough to compute required components of the COCO-based annotation formats to derived calculations within ADF.

Our metamodel did not consider using Run Length Encoding (RLE) to encode multiple boundaries (polygons), as is used when *Is Crowd* is set to 1 to indicate multiple instances in the image. However, we still capture this concept under the a form of a *Boundary*. Additionally, we found that the *Is Crowd* is the only mapping of all the formats that relates to an ADF *Image-Level Feature*.

3.7.2 PASCAL VOC XML Format

The PASCAL VOC XML format showed the most consistency with ADF's variant *Annotation* sub-types (Figure C.2), along with the ICDAR XML format. The PASCAL VOC format shows direct mappings to our components, such as *Bounding Box* to an *Annotation*'s equivalent, *Boundary*. Additionally, a collection of *Action*s (as given in the 2012 format to extend

[5]MATLAB is a registered trademark of The MathWorks, Inc.

[6]Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL)

the *Pose* of a person feature) can be used as a list of variant *Category* annotations. Further mappings can be seen in Figure E.3.

3.7.3 ICDAR XML Formats

The two ICDAR datasets (including the 2011 update) can also be captured within an ADF. We present this mapping in Figure E.2. Most notably, we map the *Text Line*, *Word*, *Atom* and *Text Parts* as ADF *Collection*s. The *Don't Care* component is a flag to notify if a line is illegible, which we can indicate using a *Boolean* annotation (a *Category*). The explicit *Color* type in the ICDAR 2011–2015 format can also be mapped to a *Label* (more specifically, the *Colour* label).

3.7.4 MATLAB-Encoded Binaries

To investigate the mapping of variant MATLAB binaries, we inspected the relevant guides on how to read the cell-arrays for both the SynthText 90k[7] and SVHN[8] datasets. Of the formats investigated, these were the simplest, and were the only ones to directly map an *Explicit Attribute* (for Synth90k). These are shown in further detail within Figure E.4.

[7] `http://www.robots.ox.ac.uk/~vgg/data/scenetext/readme.txt` last accessed 17 August 2017.
[8] `http://ufldl.stanford.edu/housenumbers/#downloads` last accessed 17 August 2017.

3.8 Summary

This chapter describes an inductive approach used for developing a generalisable metamodel from empirical iterations of systems development. We investigate the missing epistemology in the area of data capturing architectures. By developing a concrete implementation and reflecting upon the implementation, we were able to refine and discover a concrete metamodel. This metamodel alleviates the difficulties in maintaining data provenance when training AI models by making the serialisable, the therefore trackable with version control systems.

How may this metamodel be used elsewhere? It seems easily adaptable to similar areas in the field of computer vision processing, such as LPR or TSR. However, could we move this to beyond static images? Future works may extend our metamodel to multiple image frames, and therefore annotating videos frame-by-frame is within reach. Some current limitations of the metamodel (in its current form) is the difficulty to apply it to non-vision-related topics, such as Natural Language Processing (NLP), audio processing or sensor data processing. Some key concepts may be missing from our metamodel in these areas, and we expect that these can be closed by researchers with relevant domain expertise who can further extend it. We also propose that our approach could be used as a background for future works in metamodel development.

Additionally, we have shown that our metamodel is consistent with four popular existing data formats, and therefore is aligned to requirements of previous annotation systems. Hence, adapters can be used to map such formats, thereby unifying all datasets into one readable format which can be used for AI training and validation. We leave this open for future work.

The primary contributions of this chapter are:

- an approach to define prominence of marathon runners that is quantifiable,
- a metamodel that describes a schema to annotate large image datasets, and
- an exploratory methodological approach in designing a metamodel from a concrete system.

Minor contributions in this chapter include:

- a basic comparison of a metamodel against existing image-based datasets, and
- an overall pipeline in how to handle and deal with data provenance in AI models.

Chapter 4

Argus

Within this chapter, we present Argus[1], a partial implementation of our metamodel to label a marathon racing dataset. We discuss the overall design of Argus and how this fits into the data-capturing workflow (Section 3.5). Additionally, we discuss productivity and performance statistics gathered during the data gathering process, and assess Argus against other comparative tools. Further information related to Argus can be found at `http://deakin.edu.au/~ca/argus`, and source code is made publicly available at `http://github.com/alexcu/argus`.

4.1 Design

Argus is designed as a full-screen desktop application to utilise maximum screen space. The application is designed to be keyboard-driven, so keyboard shortcuts are displayed wherever possible. Additionally, the current instruction of the workflow is indicated on the top of the image in red to emphasise to annotators what input they are required to make. Figure 4.1 shows the main UI. Further segment-level features are captured in dialogs or user interactiondirectly on the image, as shown in Figures 4.2 and 4.3. We deployed Argus to data taggers remotely using the ClickOnce Deployment[2] strategy.

4.2 System Evaluation

We evaluated our implementation for throughput, biases on the Likelihood of Purchase (LoP), and the general quality of the annotations made by data taggers. We break this evaluation

[1] Inspired by the *all-seeing* giant of the same name from Greek mythology: "With his multiple sets of eyes, Argus could see nearly everything in his vicinity". See `http://www.loggia.com/myth/argus.html`.

[2] `https://msdn.microsoft.com/en-us/library/t71a733d.aspx` last accessed 11 August 2017.

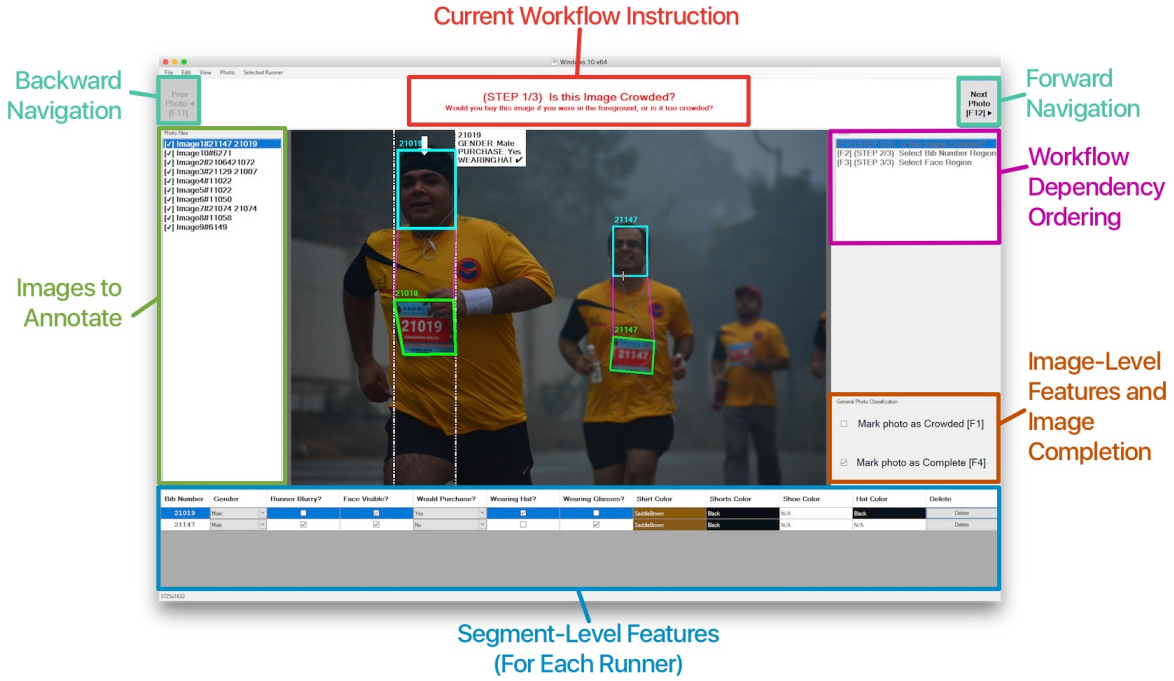


Figure 4.1: An overview of the Argus user interface.

under various metrics in the following sections.

4.2.1 Annotation Throughput

While Argus is running, we gather a number of metrics to determine what throughput is achievable in our 803 images, presented in Figure 4.4. We measure throughput as either image-level of segment-level features. A single image usually takes less than a minute to markup, and the longest feature to markup is the prominence and face features, especially as we gather most annotations here. These statistics are useful for productivity analysis on how long a given dataset may take to markup using Argus.

Another metric gathered were the number of mistakes made during annotation. We define 'mistake' as a rule violation, UI violation, or poor data entry. As shown in Table 4.1, adding construction rules is useful for our metamodel, as discovered during deployment trials of Argus. We needed to add the *FaceBounds* restriction (Figure 3.10), which was then violated a further 130 times during our annotation. UI restrictions automatically built into Argus also proved useful. These restrictions prevented users from dragging-and-dropping from bottom-right to top-left (inverse rectangle)[3] or checking if a polygon was being dragged-and-dropped instead of clicked (and vice versa for rectangles).

[3] Although, it would have been possible to automatically swap the x_1 and y_1 with x_2 and y_2 within Argus.

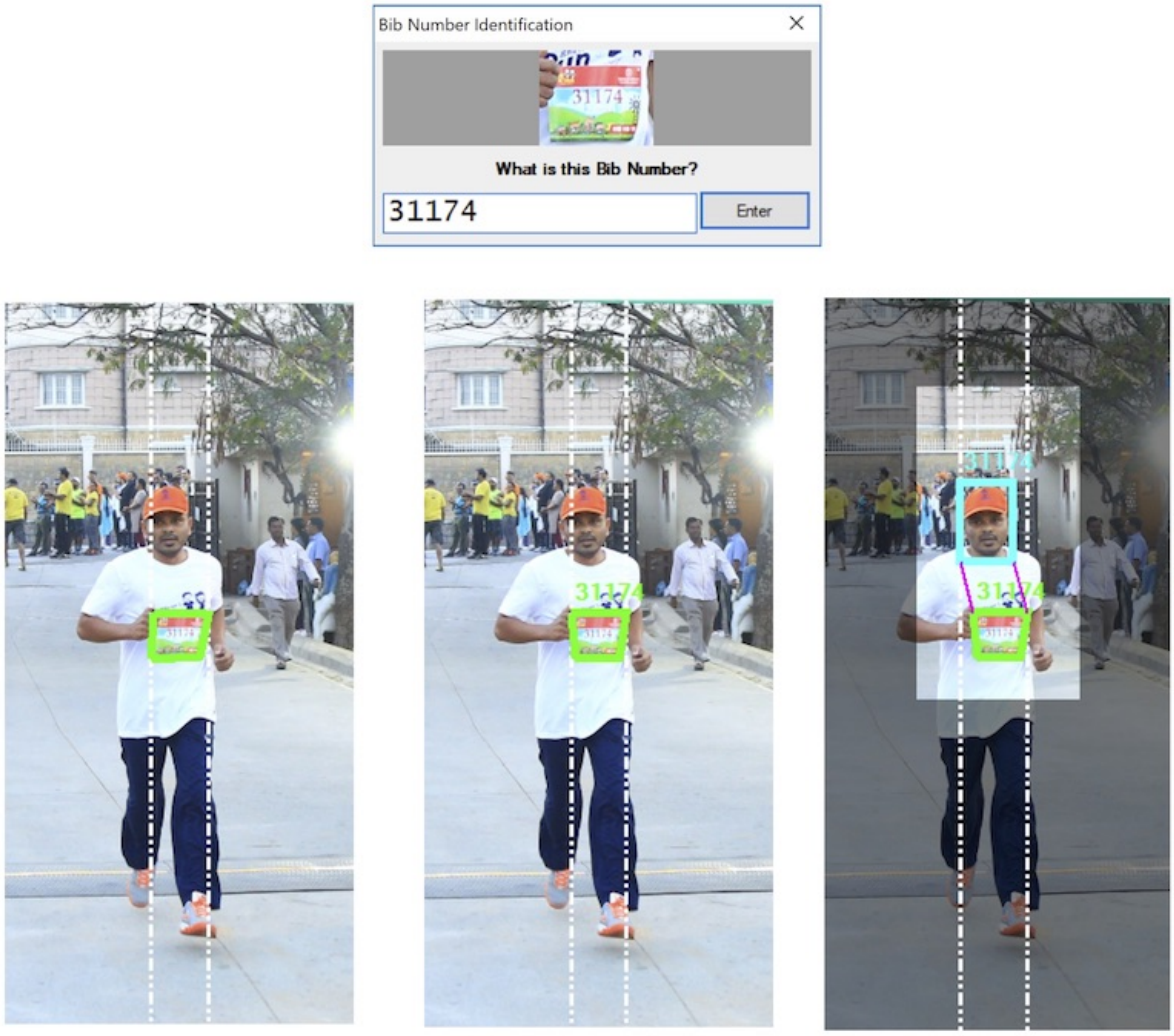


Figure 4.2: *Bib* and *Face* segment-level feature annotation. Users click four times around the bib to mark up the *BibSheet* region (left). A dialog asks the user to enter the *RBN* label (top). The RBN is annotated on the image (middle) and users can progress to drag-and-drop around the *Face* region within the restrictions set (see Section 3.4). Note the dependency ordering is present.

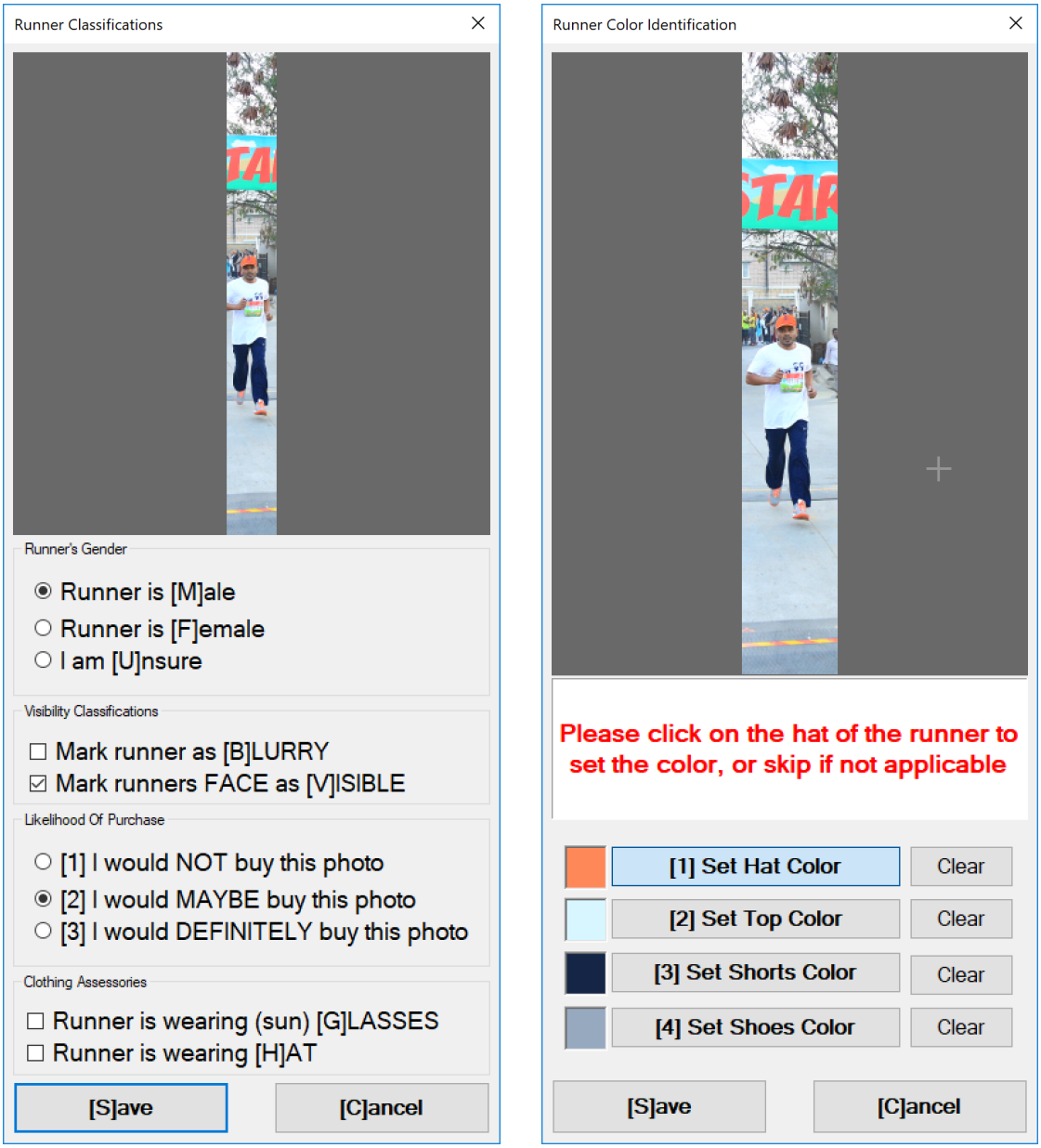
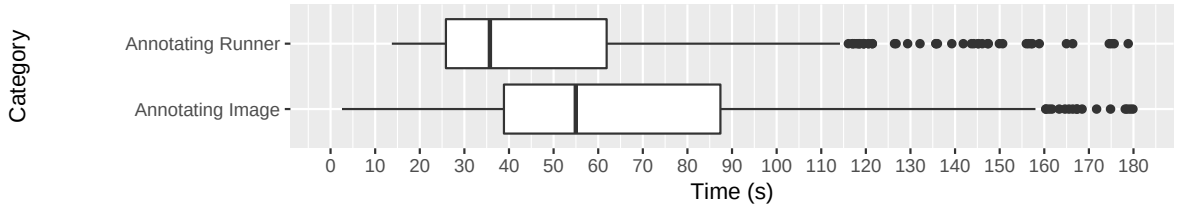
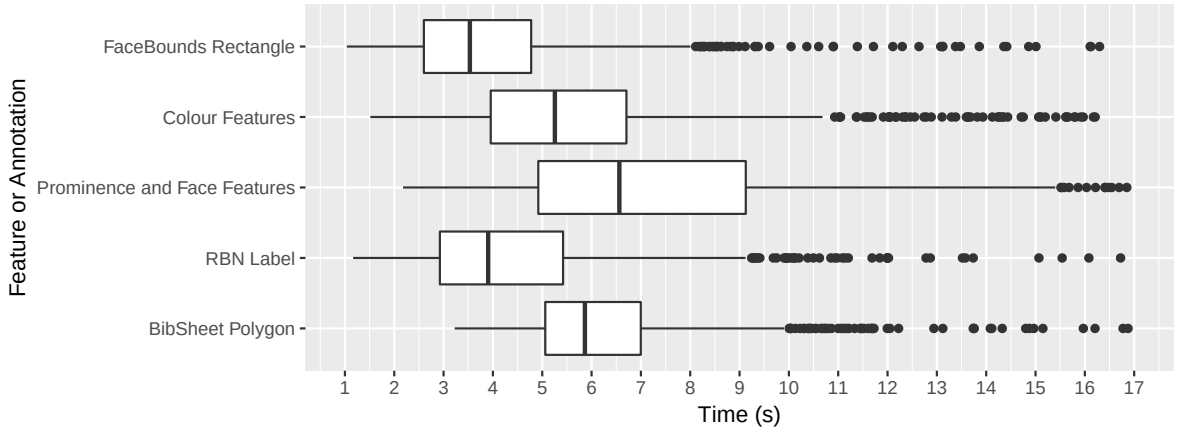


Figure 4.3: Annotation for the *Prominence* and *Colour* segment-level features.



(a) Overall timing per image.



(b) Segment-level features and annotations marked up.

Figure 4.4: Throughput of 803 images using Argus.

4.2.2 Annotation Likelihood of Purchase Bias

Upon inspection of our Likelihood of Purchase (LoP) metrics, we found that, of all runner's that were annotated (1,031), 77.1% were marked with a LoP of YES, 10.7% were marked as MAYBE and 12.2% were marked as NO. These values are important for prominence ranking training, as we we know there are far more prominent runners (with a higher LoP) to train an AI model. Furthermore, we want to reduce as many MAYBE LoP values as possible as these annotations are discarded in prominence training to reduce impartial bias when training a NN—and as shown our dataset was annotated with the MAYBE values at a minimum. Augmentation (Section 3.6.2) is required to increase the number of NO training samples.

4.2.3 Annotation Quality Evaluation

We took a subset of 260 randomly selected images (approximately one third of the entire dataset) tagged from the data tagging team and assessed the photos for quality assurance. Here, we inspected photos on three tiers: 'Good', 'Okay', and 'Bad' quality. We deemed photos as 'Good' if there were no flaws at all with the tagging, 'Okay' if there were minor

errors made that we can compensate with, and ‘Bad’ if there are significant flaws with the tagging that would affect the AI model. Refer to Figure 4.5 for samples of these images.

In Figure 4.5c, we see that the *FaceBounds* annotation is too small around the runner’s head, and that the *BibSheet* polygon has been marked as a square, rather than directly around the four corners of the bib sheet. This is considered ‘Okay’ as we are able to compensate with the extra padding around the *BibSheet*, and the *FaceBounds* could have padding to extend the area. In Figures 4.5a and 4.5b, however, we are given completely incorrect information. In the former, the *BibSheet* polygon is not at all reflective of the actual sheet, including part of the runner’s shirt in the annotation, and the *FaceRegion* is too small unless significant padding is added. In the latter, the *RBN* has been annotated with the value `100613`, confusing the first character as numeric when it is alphabetic (`I00613`). These instances show significant fallbacks to quality that may hinder training the NN.

In our sample set, we also evaluated whether the number of runner’s in an overall photo has a negative impact on the annotation quality (Figure 4.6), as well as how long taggers spend on annotating photos (Figure 4.7). Upon analysis, we can see that most runners are well-annotated—only in images containing one runner where there are more ‘Bad’ annotations than ‘Okay’ ones, and generally ‘Okay’ and ‘Good’ outcomes are largely consistent regardless of runner count. Therefore, likelihood of time spent on annotations is ‘Good’, and ‘Bad’ marking distribution is more biased towards short evaluation times where there are few runners in the image. Figure 4.7 confirms our hypothesis where, we can see, the longer the annotator spend time tagging, the better the quality of our data is: a median of 48 seconds is improved to 52 seconds between ‘Bad’ and ‘Okay’, and this jumps to a median of 55 seconds for ‘Good’. Therefore a difference in 7 seconds spend on tagging a photo shows a significant improvement on the quality of the tagging.

Table 4.1: Frequencies of mistakes made during the annotation process. These are separated into construction rule violations, UI violations, and poor data entry.

Statistic	**Frequency**
Marked as complete when *Face* region not tagged	20
Selected outside *FaceBounds* restriction	130
Selected *FaceBounds* below *BibSheet*	7
Drag-and-drop inversely for *FaceBounds*	4
Drag-and-drop made for *BibSheet* polygon	73
Clicked instead of drag-and-drop for *FaceBounds*	45
Undos made	70
Deleted runner annotation	46

(a) A ‘Bad’ annotation **(b)** A ‘Bad’ annotation **(c)** An ‘Okay’ annotation

Figure 4.5: ‘Okay’ and ‘Bad’ quality tagging tiers.

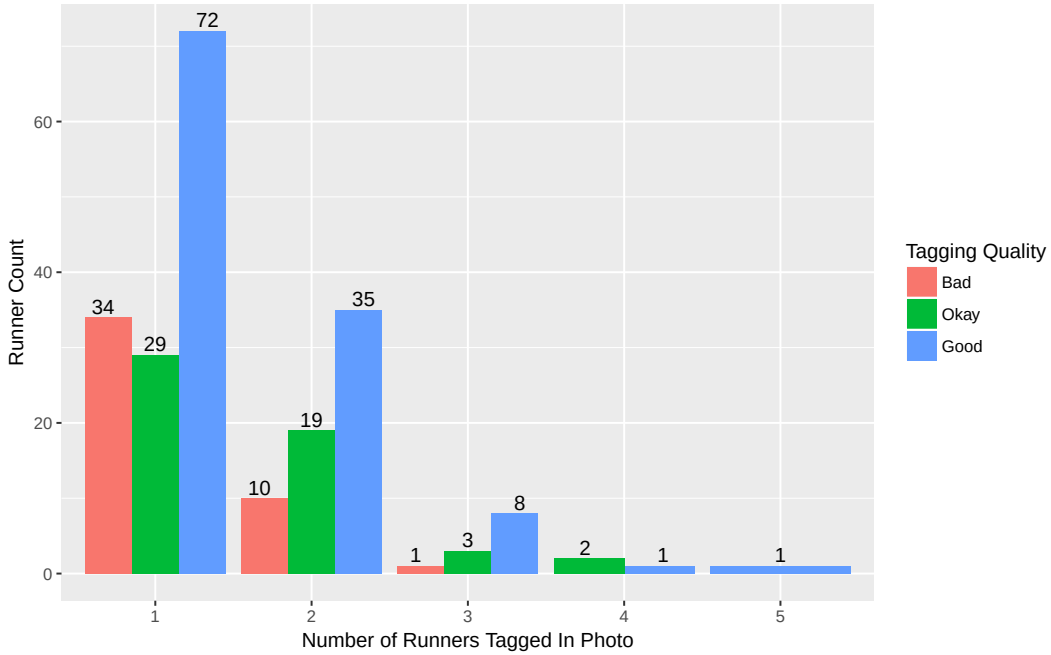


Figure 4.6: The quality of tagging against the number of runners per photo.

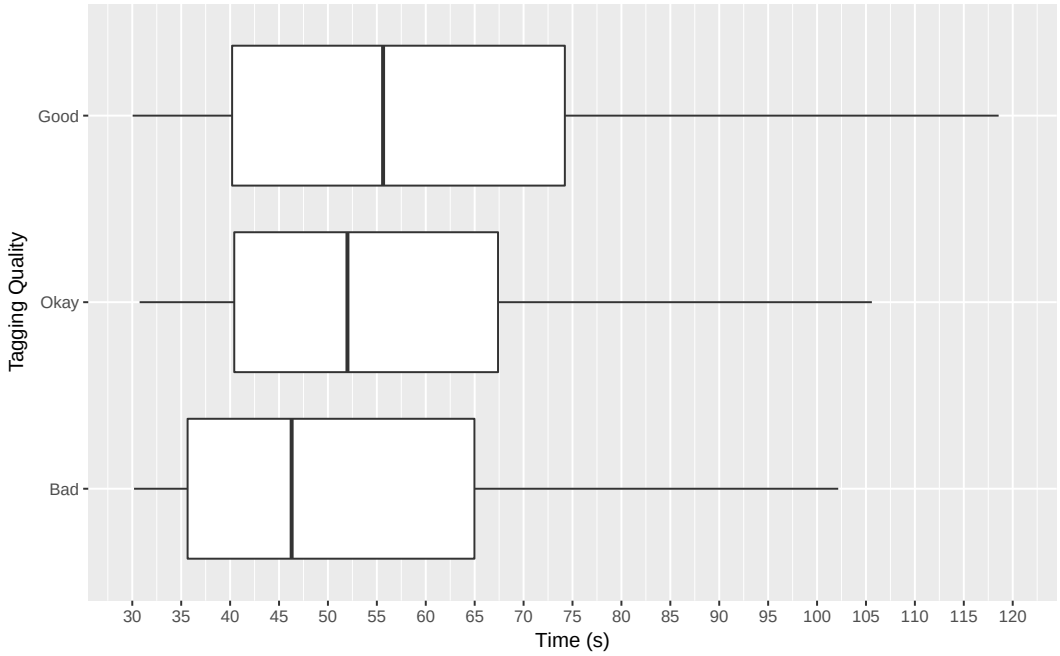


Figure 4.7: Quality of tagging against the time spent evaluating the photo.

4.3 Comparative Annotation Tools

In this section, we describe comparative annotation tools and services used to annotate similar datasets, and compare such tools with Argus.

4.3.1 LabelMe

LabelMe[4] (Figure 4.8), developed by Russell et al. [129], is a web-based application that requires annotators to set up the web service on their own system. It has been successfully used to markup the large Scene UNderstanding (SUN) dataset by novice users, though the developers note a number of reflective fallbacks and potential difficulties using the tool [5]. LabelMe does not contain a specific instructional workflow, unlike Argus, and therefore annotators are allowed to selectively choose whichever objects to annotate[5]. This makes the resulting dataset largely incomplete as (potentially) not all objects in each image are fully annotated. Additionally, there is no restriction to what types of objects are annotated in the image, as any text can be used to describe an object (rather from a hierarchical category, such as in [98]).

Figure 4.8: The LabelMe user interface [129].

[4] `http://labelme.csail.mit.edu/` last accessed 17 August 2017.

[5] Instructions are potentially vague: "Use your mouse to click around the boundary of *some* objects".

4.3.2 Annotation and Performance Evaluation Platform

Karatzas et al. [78] introduced the CVC Annotation and Performance Evaluation Platform (APEP)[6], made popular for use using the ICDAR 2011–2015 Robust Reading Competitions. Unlike our system, users can edit the per-pixel boundaries using a 'flood-fill' or 'magic-select' markup tool (with an adjustable tolerance), followed by a skeletonisation of the boundaries filled. This speeds up selection of specific text areas (Figure 4.9). While we did not incorporate a feature into Argus, a potential future version would benefit using such an annotation tool for a polygon, rather than clicking $n_{vertices}$ times if the vertices count is unknown. This could also be encoded using RLE as suggested from the COCO format in Section 3.7. Additionally, users are able to zoom in to get finer accuracies at the pixel level.

Figure 4.9: The APEP ground truth annotation tool, showing a hierachy of textual content and defined text parts (flood-filled areas and skeletons) [78].

4.3.3 Amazon Mechanical Turk

Amazon Mechanical Turk (AMT)[7] provides Software as a Service (SaaS) that does not require downloading or setting up (typically laborious for annotators). Annotation is achieved in the form of customised Human Intelligent Tasks (HITs), using a custom-built interface. Recently, this is typically the most popular choice of annotation outsourcing [26, 35, 98, 117, 154] due to flexibility in developing customised interfaces for the task at hand. The benefits of ensuring a user-friendly crowdsourcing annotation system are presented by Matera et al. [109]. Sorokin and Forsyth [142] discuss the utility of using AMT for annotation. The primary downside is that most interfaces need to be created from scratch, which can typically be laborious.

[6] `http://www.cvc.uab.es/apep/` last accessed 17 August 2017.
[7] `https://www.mturk.com/mturk/welcome` last accessed 11 August 2017.

4.3.4 VATIC

VATIC (the Video Annotation Tool from Irvine, California) is a a free web-based tool (Figure 4.10) developed by Vondrick, Patterson, and Ramanan [156], running on AMT. We note this tool specifically for the use of annotating video stills within an image: its intended purpose is for the sole use of object tracking in videos. There are only three annotations per object: the object name, the object is out of view, and the object is obstructed. It is useful in that not all frames have to be individually annotated, as the tool uses object tracking to estimate the movement between every 2-3 seconds. This may be useful to incorporate into future works of Argus.

Figure 4.10: The VATIC web-based user interface. Sourced from `https://github.com/cvondrick/vatic`. (Last viewed 11 August, 2017.)

4.3.5 ScaleAPI

ScaleAPI[8] is a recent SaaS that provides a web-based Application Programming Interface (API) to return human-marked annotation in realtime. Various boundary annotations can be made on an image. Listings 4.1 and 4.2 show the request and response for a simple instruction to draw bounding boxes around all pedestrians and cars in the image[9]. Additionally, object segmentation (on a per-pixel basis) is provided. While there are no need for workflows in a single request (as multiple requests can be made for multiple workflow steps), instructions to annotators are provided on a per-request basis. This said, a consistent annotator between requests is not guaranteed and this may have an impact on varying annotator quality.

[8] `http://www.scaleapi.com` last accessed 11 August 2017.

[9] `https://docs.scaleapi.com/?shell#bounding-box-annotation` last accessed 17 August 2017.

Listing 4.1: A sample ScaleAPI HTTP request made using cURL[10].

```
curl "https://api.scaleapi.com/v1/task/annotation" \
  -u "SCALE_API_KEY:" \
  -d callback_url="http://www.example.com/callback" \
  -d instruction="Draw a box around each **car** and **pedestrian**." \
  -d attachment_type=image \
  -d attachment="http://i.imgur.com/XOJbalC.jpg" \
  -d objects_to_annotate="car" \
  -d objects_to_annotate="pedestrian" \
  -d with_labels=true \
  -d min_width="30" \
  -d min_height="30"
```

Listing 4.2: Sample JSON response from the request made in Listing 4.1.

```
{
  "task_id": "5774cc78b01249ab09f089dd",
  "created_at": "2016-9-03T07:38:32.368Z",
  "callback_url": "http://www.example.com/callback",
  "type": "annotation",
  "status": "pending",
  "instruction": "Draw a box around each **car** and **pedestrian**",
  "urgency": "day",
  "params": {
    "with_labels": true,
    "min_width": 30,
    "min_height": 30,
    "objects_to_annotate": [
      "car",
      "pedestrian"
    ],
    "attachment_type": "image",
    "attachment": "http://i.imgur.com/XOJbalC.jpg"
  },
  "metadata": {}
}
```

[10] `https://curl.haxx.se/` last accessed 17 August 2017.

4.4 Summary

This chapter presents Argus, an implementation of our metamodel (in the context of marathon photography) shown in Chapter 3. We present the overall UI in Argus, and evaluate it against three key metrics: throughput, bias and quality. We also present comparative annotation tools that we assess against Argus to find specific overlaps between both.

The primary contributions of this chapter are:

- an open-source dataset tagging system for RBN and prominence extraction, *Argus*,
- an assessment of data tagger performance and throughput for specific phases in marathon-racing dataset tagging, and
- a methodology for quality assessment of annotations made in a dataset.

Minor contributions in this chapter include:

- a comparison of our proposed system against existing comparative annotation tools.

Chapter 5

Processing Pipeline

Chapters 3 and 4 discuss methods to curate and augment datasets. Using these datasets, we train a NN to develop a processing pipeline. This chapter describes this process and how we developed our pipeline to detect bib sheets 'in-the-wild' on any given marathon photo. Lastly, we explain our method of OCR to convert our RBN candidates into text. An overview of our pipeline is shown in Figure 5.1 and the source code is made available online[1].

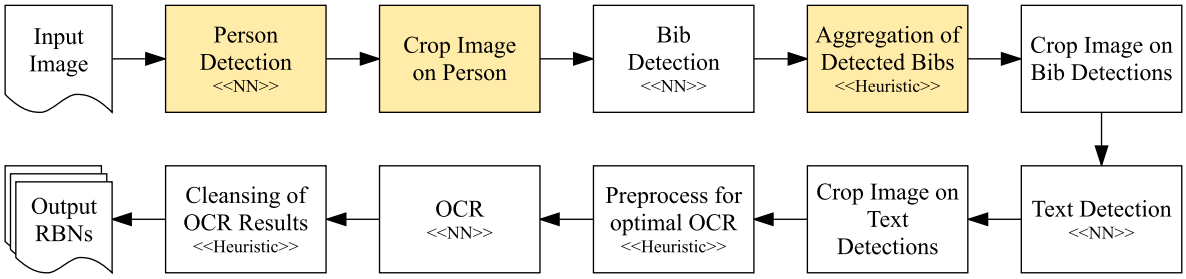


Figure 5.1: A high level overview of our processing pipeline. Highlighted components indicate optional steps that, when included, are proposed to increase accuracy. We use the stereotypes <<NN>> and <<Heuristic>> to differentiate between NN-based and heuristic-based components.

5.1 Bib Detection

Detecting bibs in an image is a image segmentation problem. There are multiple approaches that address this issue: typical approaches from literature in the area have used heuristic-based methods to segment an image, as highlighted in Section 2.1. As this research is part of a wider project under DSTIL, three heuristic-based approaches were investigated in addition to the AI approach proposed by this study.

[1] http://github.com/alexcu/hermes-bib-detect last accessed 25 September 2017.

5.1.1 Existing Approaches

The first was face detection and heuristics to find the bib–as presented in [9]—though as shown in Figure 2.14, this approach does not always work. To combat the issue of subjects leaning in photos, we looked at a two further heuristic-supported approaches: poselet detection to detect a human skeleton [118] and using a pre-trained CNN running the You Only Look Once (YOLO) object-detection system [126] on the Darknet framework [125]. We discuss the person detection in more detail within Section 5.1.3. In the latter, we only attempted to look at object detection of people. Both methods produced greater person detection candidate areas than using face areas alone, and we processed these results we used similar visual processing to that of Ben-ami et al. [9], albeit on a larger detection area.

But—as per the aims of this research—can a detection strategy be achieved without using heuristic-based approaches at all? We discuss our solution for solving this question using deep-learning NNs in the following section.

5.1.2 Deep-Learning Approaches

Section 2.1.2 introduced applications of deep-learning, specifically Convolutional Neural Networks (CNNs), that were successfully applied on object instance segmentation within images. We investigate similar approaches for recognising racer's bib sheets using the data collected from Argus (Chapter 4).

To implement our approach, we explored the viability of using region-based CNNs, known as R-CNNs. These were first introduced by Girshick et al. [50] in 2014, and more than doubled the accuracy of previous object detection methods such as HOG [34] and SIFT [103]. The R-CNN approach was improved upon with *Fast R-CNN* [51] that increased detection speed. We adopted the R-CNN approach that was improved upon in a recent 2017 study, *Faster R-CNN* [127], to predict bib regions. Faster R-CNN combines CNNs with state-of-the-art object detection networks (e.g., region proposal algorithms). The sharing of convolutional layers of a CNN with object detection networks produces an Region Proposal Network (RPN) that is utilised by Faster R-CNN.

We applied Faster R-CNN to the context of bib detection by the process of transfer learning [6, 19, 148]: this involves adjusting the final layer of a pre-trained NN and adding a context-specific layer (in our case, racing bibs). The concept of transfer learning mimics that of human learning in that humans do not learn tasks in isolation, but a sequence of training tasks over

a lifetime. (If one has learnt English and Latin, then learning French is easier due to transfer learning.) Transfer learning has successfully been applied to various contexts such as spam filtering [12] and text classification [36, 124].

To implement this detection pipeline, we chose a Keras [30] Python implementation of Faster R-CNN[2] that runs on top of the TensorFlow [1] library. To reduce false positives, we reject any candidate that is predicted with an accuracy of less than 50%. However, while we were able to extract a large amount of bibs from many images, there were cases of shoes, signs and hands being detected with accuracies above 90%. There were also some false negatives due to excessive noise in the photo. Figure 5.2 shows this in further detail. We decided to amend this by cropping each image down to detected people using the YOLO system.

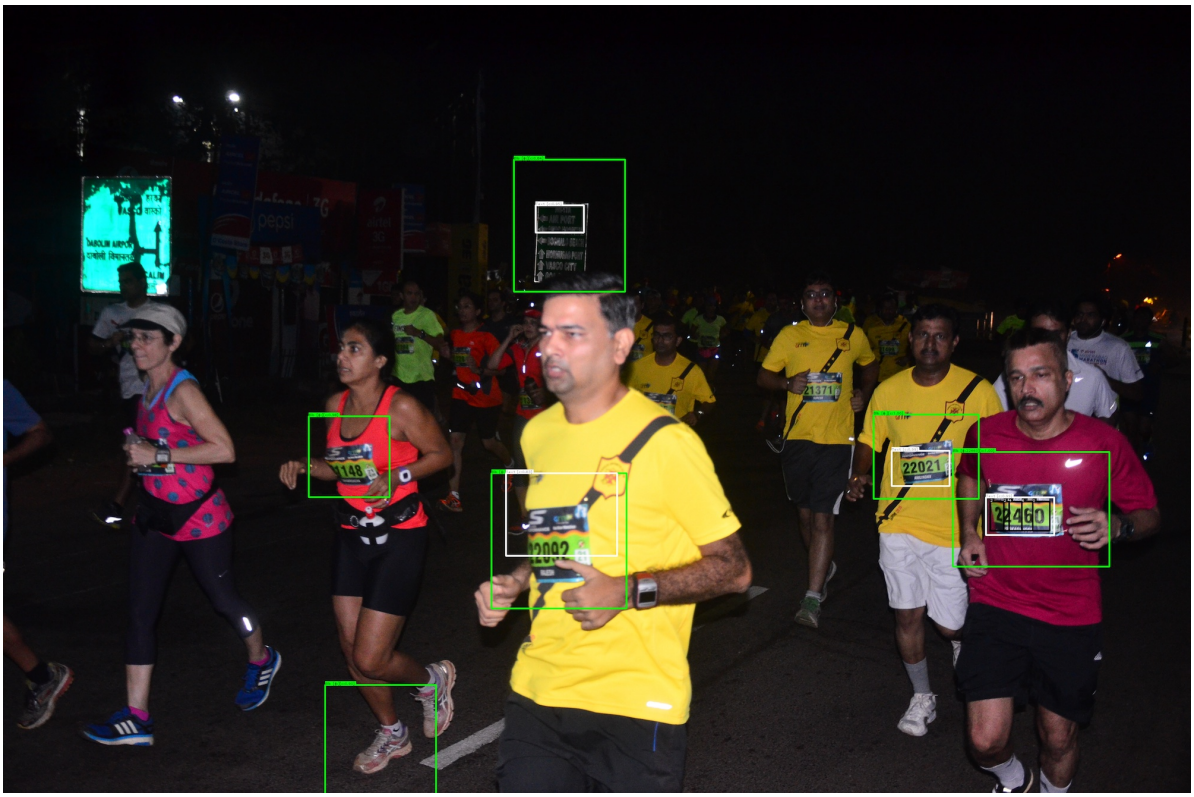


Figure 5.2: Results from using Faster R-CNN Keras implementation running on TensorFlow. Note the false negative for the third runner from the right and the false positive on the second runner's shoe from the left.

5.1.3 Person Filtering

As described in Section 5.1.1, a heuristic-based method explored was to apply visual transformation on a person detection to extract the RBN. The person detection component used

[2] `https://github.com/alexcu/keras-frcnn` last accessed 21 August 2017.

a pre-trained network built on the YOLO object detection system running on a customised implementation of Darknet[3].

We were able to utilise a similar method by extracting people from an image using YOLO (Figure 5.4a), and cropping each image to the respective bounding box of the detected areas. Using weights from the Tiny YOLO model[4], YOLO was able to detect the most prominent runners in the image, with which we found desirable considering the aims of the project. We then ran our Faster R-CNN prediction model on each cropped image (Figure 5.4b). As these images had reduced noise, accuracies of bib detections increased with fewer false positives.

The last stage of the bib detection was to aggregate all detections of separate runners back into one image (Figure 5.4c), and union overlapping predicted regions (Figure 5.4d). (We found that in some cases, two or more bibs would be detected in a single person crop due to runner's in the background.) In some cases, however, it was unwise to union *all* overlapping regions, such as that shown in Figure 5.3. Using $a(r)$ to denote the area of a region, we only formed the union of both regions, $r_1 \cup r_2$, if the area of the intersection region, $a(r_1 \cap r_2)$, was more than 75% of the area of one of the area of one of the regions.

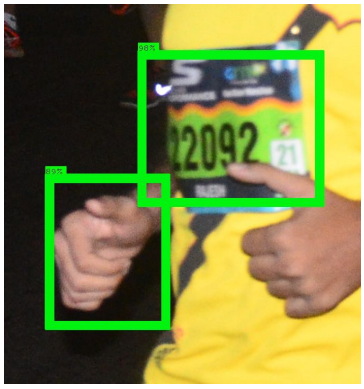


Figure 5.3: Invalid intersection to union both regions. The area of the intersection is not greater than 75% of the two regions.

Thus, given N detected runners in a single image with N cropped images, let R represent the union[5] of all regions r for the nth crop to produce Figure 5.4c. The final regions, R_f, of an image are all regions, R, and union regions, R_u but not intersection regions, R_i. Therefore, the

[3] `https://github.com/alexcu/darknet` last accessed 21 August 2017.
[4] `https://pjreddie.com/darknet/yolo/#tiny` last accessed 5 September 2017.
[5] Here, *union* refers to a union in set theory.

process of overlapping (Figure 5.4d) can formally be described in set notation:

$$A(r_1, r_2) = (\ a(r_1 \cap r_2) > 0.75 \times a(r_1)\ ) \ \vee \ (\ a(r_1 \cap r_2) > 0.75 \times a(r_2)\ )$$

$$R_u = \{\ r_1 \cup r_2 \mid (r_1, r_2) \in R \ \wedge \ r_1 \neq r_2 \ \wedge \ A(r_1, r_2)\ \}$$

$$R_i = \{\ r \mid (\exists r \in R)(\forall r_x \in R)[\ r \neq r_x \ \wedge \ A(r, r_x)\ ]\ \}$$

$$R_f = \{\ r \mid r \in R \wedge r \in R_u \wedge r \notin R_i\ \}$$

(a) Person detection **(b)** Bib detection **(c)** Aggregation **(d)** Overlapping

Figure 5.4: Improving accuracies of bib detection by cropping people. (a): Three bounding boxes detected for PEOPLE using YOLO. (b): Bib detections from Faster R-CNN (right) on each person crop (left). (c): Aggregated bib detections from individual person crops. Note the two overlapping blue and green regions. (d): The union of overlapped regions are merged into a single region, shown in white.

5.2 Text Detection

The penultimate stage of the detection pipeline is to crop each bib detection region and focus on the RBN itself, as shown in Figure 5.5. We attempted to train Faster R-CNN using both the COCO-Text [154] and SynthText In The Wild [55] datasets, following a similar process outlined in Section 5.1.2. However, we found that—due to the wide variance of typeface samples within the COCO-Text dataset—Faster R-CNN was only able to make reliable predictions on the SynthText dataset. Thus, we determined that Faster R-CNN is suitable for text extraction given that it is trained on a dataset where text samples similar (i.e., the consistency of the synthetic words generated in the SynthText dataset made this possible).

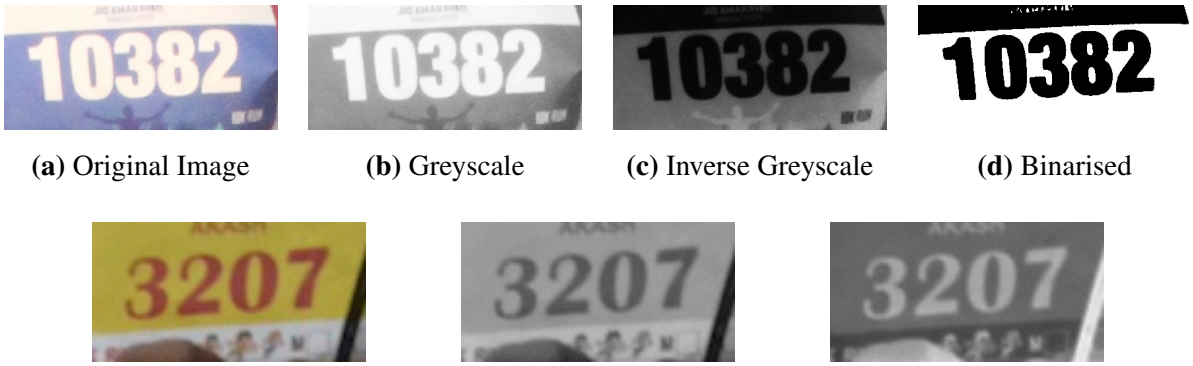
Text detection is used for two key purposes. Firstly, this helps to reduce false positives in the bib detection stage as if no text regions could be found, we assume the candidate region is not a bib image. Secondly, we feed the largest text region on the bib found into a text recognition engine. We explain this in the following section.

Figure 5.5: Text region detection. Detected bibs are cropped from the original image (left) and passed through a text detection pipeline (middle). We select the candidate with the largest area and crop it (right).

5.3 Text Recognition

Before we feed the text area into an OCR engine, we preprocess the extracted text images to optimise their recognition (Figure 5.6). We produce three preprocessed optimal candidates: greyscale, inverse greyscale, and threshold binary. For images with black and white numbers,

we use the statistical median of the image's colour range: where the median is greater than 128, we assume a white number is detected (i.e., white on a darker background, such as Figure 5.6a) and apply a binarised inverse threshold on the image. Otherwise, we assume a black number is detected, where we threshold without the inverse. Greyscale and inverse greyscale are most suitable for text regions where RBNs are not in a black or white typeface (Figure 5.6e).



(a) Original Image **(b)** Greyscale **(c)** Inverse Greyscale **(d)** Binarised

(e) Original Image **(f)** Greyscale **(g)** Inverse Greyscale

Figure 5.6: Image post-processing applied to optimise input into an OCR engine. *Top*: An RBN with white numbers can be applied binarised thresholding. *Bottom*: A RBN with red numbers cannot have binarised thresholding applied.

All preprocessed images are provided as input to Tesseract 4 alpha[6]. We use this OCR engine as it supports Long Short-Term Memory (LSTM) NN-based character recognition. From all output of our various candidates shown in Figure 5.6, multiple text candidates are extracted from Tesseract. To clean our output, we strip out characters that are not within our known character range of A–Z, 0–9 and figure dashes. Additionally, we only accept unique RBN sequences per bib region to prevent duplicates.

5.4 Runtime Performance

Our pipeline's runtime takes, on average, 23 seconds end-to-end per image (determined over all evaluations given in Chapter 6). OCR and person detection runs in less than a second. The algorithmic complexity of our pipeline bottlenecks typically at the bib and text detection stages: when cropping is introduced in the pipeline, bib detection increases by 2–10 extra seconds. Given the wide samples (fonts) used to train the text detection network, this is the

[6] `https://github.com/tesseract-ocr/tesseract/tree/4.00.00alpha` last accessed 5 September 2017.

slowest stage, at an average of 13 seconds. Further breakdowns of the runtime performance is given in Table F.4.

The engineering limitations of the Faster R-CNN implementation may be improved architecturally by running the system at scale on a cloud cluster such as Amazon Web Services (AWS). While memory complexity is not an issue, the time complexity is. We leave the improvements to our pipeline's runtime up to future works in systems architecture literature.

5.5 Summary

This chapter presents an end-to-end pipeline that solely uses NNs for object detection and text recognition to find the RBNs in a given marathon photo. We successfully utilise transfer-learning on Faster R-CNN to train it to detect both bibs sheets and text regions within those bib sheets. To reduce background noise, we use the YOLO object detection system and only check for bib sheets on a given crop of a detected human's bounding box region. We evaluate this pipeline in the following chapter.

The primary contributions of this chapter are:

- an end-to-end NN-based pipeline to detect and recognise RBNs in a natural scene, and
- a proposed method to increase accuracies by human extraction within photos.

Minor contributions in this chapter include:

- evolvement in the area of object detection and transfer learning,
- a methodology for improving the results from an OCR system given images without binarised text colours, and
- a methodology for removing false positive cases from an OCR system.

Chapter 6

Evaluation

This chapter discusses our evaluation strategies to test the performance of our pipeline. From an inspection of our results, we discuss the contributions and key limitations.

6.1 Evaluation Strategies

To apply a robust evaluation on our pipeline, we randomly sampled 200 images from our dataset that were not used in training or validation. Half of these evaluation images only contained one bib marked up in the image (100 RBNs), while the other half had more than one bib in the image (a total of 221 RBNs).

We used a two-stage pass on this set: the first pass applies to the first hundred images that we consider as an 'ideal' case for the pipeline to handle (i.e., to detect the only bib in the image). The second pass considers the 'realistic' case where there is more than one bib per image. Henceforth, we refer to our two subsets as 'ideal' and 'realistic'.

In addition to categorising the sets as ideal and realistic, we also considered what the minimal amount of training data required is to make the bib detection robust. We evaluated three additional trained models using Faster R-CNN for bib detection on 1, 100 and 500 training bibs, following the process outlined in Section 3.6.2. Each of these images are extracted from the same source dataset consisting of the 722 training samples[1] indicated in Table 3.3.

Lastly, we evaluated how the person filtering stage (Section 5.1.3) affects the overall performance of the pipeline. The permutations of these evaluation methods results in a total of sixteen different evaluations, shown in Table 6.1. We generated a dash-separated identifier to refer to each of these evaluations based on *I* or *R* for *ideal* and *realistic*, respectively, then the

[1] We refer to this as 'all' images in our training dataset.

number of training images (using *all* to refer to all training images) and lastly whether to crop on the image or not crop (*CR* for **cr**op, and *NC* for **n**o **c**ropping).

Table 6.1: Summary of Evaluations.

Evaluation ID	Evaluation Set	Training Bibs	Crop Human
I-1-CR	Ideal	1	Yes
I-1-NC	Ideal	1	No
I-100-CR	Ideal	100	Yes
I-100-NC	Ideal	100	No
I-500-CR	Ideal	500	Yes
I-500-NC	Ideal	500	No
I-ALL-CR	Ideal	722	Yes
I-ALL-NC	Ideal	722	No
R-1-CR	Realistic	1	Yes
R-1-NC	Realistic	1	No
R-100-CR	Realistic	100	Yes
R-100-NC	Realistic	100	No
R-500-CR	Realistic	500	Yes
R-500-NC	Realistic	500	No
R-ALL-CR	Realistic	722	Yes
R-ALL-NC	Realistic	722	No

6.2 Metrics

We breakdown the evaluation of our pipeline mapped its four stages=: (1) bib detection, (2) text detection within a bib crop, (3) OCR within a single bib crop against the ground truth, (4) overall performance on an entire image.

6.2.1 Bib Detection Performance

In addition to the standard precision, recall and f-score metrics defined in Section 2.3, we develop a secondary measure for our model performance based on the number of bib detections made in an image. For the ideal case, we assess the bib's detection accuracy assuming a binary scenario (the pipeline either finds the bib or it does not) and, where more than one detection is made, we limit its performance to 1. For the realistic case, the bib detection model performance is assessed as the number of estimated bibs found in the image divided by the total number of ground truth bibs. Thus, the resulting performance metric is in the range $[0, 1]$.

This means that, regardless of the number of bibs that are identified, as long as the number of estimated bibs is equal to or greater than the number of ground truth bibs, it is identified as a hit. Any more identified bibs are either classified as false positive estimates, or cases that were not considered for ground truth annotation.

Thus, given a set of ground truth bibs, T_{bib}, and estimates E_{bib}, we define the addition metric of bib performance, p_{bib}, as:

$$p_{bib} = \begin{cases} 0, & \text{if } |T_{bib}| = 0 \\ \min\left(1, \frac{|E_{bib}|}{|T_{bib}|}\right), & \text{otherwise} \end{cases}$$

Therefore, this performance metric captures *at least* all valid and correct bibs in the image, while ignoring the false alarms. As this stage, the metric is only dealing with bib *detection*. Our primary concern is with regards to *detecting* and not *recognising* bibs: having false positives detected in this stage is not an issue as we are likely to eliminate them in further stages should the estimated region not be a bib at all, thus maintaining our overall performance (Section 6.2.4).

6.2.2 Text Detection Performance

For text and character detection, we follow a similar process to that outlined in Section 6.2.1. Text detection works in a binarised matter: within a bib crop, the text detection model either accurately detects the ground truth RBN region, or it does not. For character segmentations, Tesseract's LSTM network either accurately detects all or partially detects characters within the RBN. The metric focus here is, once again, a partial match to ensure that *at least* all matches that can be made are made, and anything more is ignored.

6.2.3 Character Recognition

Character recognition introduces two additional metrics. For a *single* RBN, we define the recognition performance as the number of characters that match the characters within the ground truth RBN. Given a set of ground truth RBN sequences in an image, T_{rbn}, and recognition estimations made, E_{rbn}, a single element (either a ground truth or estimated RBN) in either set is defined as t and e, respectively.

Given that an RBN is a set of characters, t and e are therefore sets of characters, and thus, the character recognition performance on a single RBN, p_{rbn} is measured as:

$$p_{rbn} = \left\{ \max\left(0.5 \times \left(\frac{|e \cap t|}{|e|} + \frac{|e \cap t|}{|t|}\right)\right) \mid \forall e \in E, \forall t \in T \right\}$$

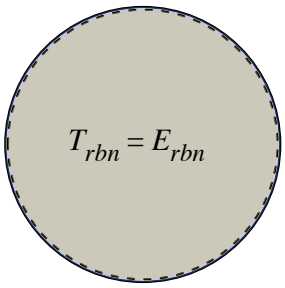
6.2.4 Overall Performance

We represent our pipeline as a way to narrow in on an image toward the true positive bib numbers that exist. Represented in set notation, the ground truth is a set of RBN sequences within the image, T_{rbn}, and our estimations of those RBNs are another set, E_{rbn}. Therefore, we formally describe that true positive matches are all estimations that fall within the ground truth ($E_{rbn} = T_{rbn}$, $E_{rbn} \subset T_{rbn}$), and all false positives are introduced when this is not the case ($T_{rbn} \subset E_{rbn}$, $E_{rbn} \not\subset T_{rbn}$, $E_{rbn} \cap T_{rbn} = \varnothing$).

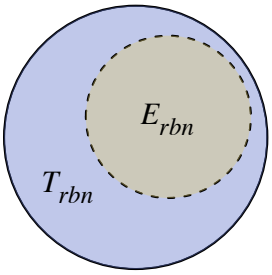
Represented as Euler diagrams (Figure 6.1), we visualise and describe such cases with context. Initially, our pipeline starts with Figure 6.1d—we narrow down our estimations using person, bib and text detection to eliminate false positives (that penalise the overall performance due to invalid recognitions), bringing us closer towards the ground truth.

If our pipeline is successful, we either end up with Figure 6.1b, or better, Figure 6.1a. With the former, we have recognised all RBNs that are within the ground truth, though are still some false negatives. Thus our estimation set is not complete, unlike Figure 6.1a where the estimation set matches the ground truth exactly. We consider Figure 6.1b as 'good' because, while it does not detect every RBN, we do not introduce false positives that do not exist in the image and introduce false data.

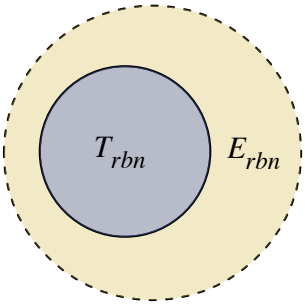
If our pipeline is not successful, we end up with Figures 6.1c to 6.1e. In Figure 6.1c, we see that—while the ground truths are within the subset of estimations—we produce false positives, and therefore mis-informed data that penalises performance. Similarly, in Figure 6.1d, there are still false positives, but—problematically—also false negatives as we are not recognising ground truth. (This must be penalised further as we introduce incorrect data and do not *fully* report correct data.) Lastly, when the ground truth and estimations are mutually exclusive (Figure 6.1e), the pipeline *only* reports false positives.



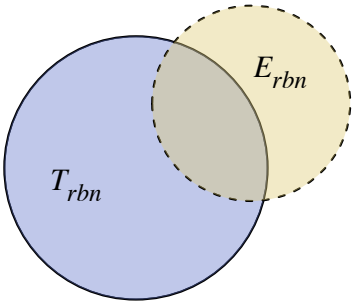
(a) Best performance as $E_{rbn} = T_{rbn}$.



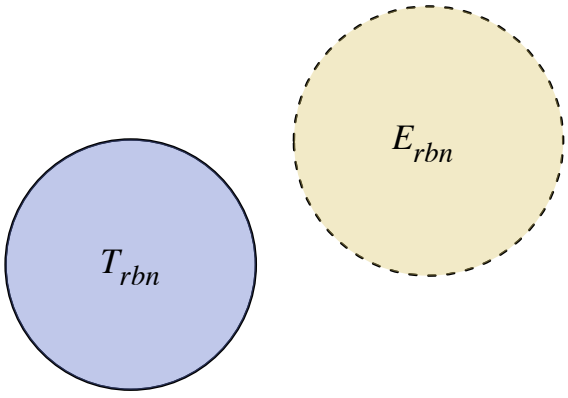
(b) Good performance as $E_{rbn} \subset T_{rbn}$.



(c) Moderate performance as $T_{rbn} \subset E_{rbn}$.



(d) Problematic performance as $T_{rbn} \not\subset E_{rbn}$.



(e) Worst performance as $E_{rbn} \cap T_{rbn} = \varnothing$.

Figure 6.1: Various scenarios of OCR performance. The order of subfigures indicate the degrading performance. Subfigures a and b show the successful cases, while non-successful cases are shown in c to e.

6.3 Results

In this section, we discuss the results of our findings from the evaluation strategies proposed in Section 6.1 using the metrics described in Sections 2.3 and 6.2. We also discuss limitations of the pipeline and possible mitigations.

6.3.1 Summarised Performance over all Evaluations

Presented in Figure 6.2 are the bib, text and character performance collated of all 16 evaluations. These values were calculated at a post-evaluation stage by comparing the output results with the ground truth. For further results, refer to Tables F.1 to F.3. For bib region accuracies detected by the pipeline, a mean f-score 0.14 is observed for ideal cases and 0.07 for realistic cases. A comparison between cropping and not cropping in both evaluation sets is presented in Table 6.2.

Generally, not cropping improves the f-score value in both evaluation sets, though our performance in the ideal cases are improved due to an increased false negative rate observed in the realistic cases (i.e., generally not all bibs are detected when there are more than one runner). This is most likely caused by the randomised dataset used to evaluate our pipeline as well as the quality of tagging made of ground truth accuracies on our pipeline: should a dataset of photos that have only been *sold* (and thus implicitly imply a greater prominence of the runner within each photo) with more fine-tuned tagging using Argus, we suggest that these detection rates would improve. We leave such evaluations up for future work.

Table 6.2: Comparison of various f-score values for cropping versus evaluation sets.

Evaluation Set	**Cropping**	
	Yes	No
Ideal	0.11	0.17
Realistic	0.03	0.10

We note the model performance is positively associated with the amount of training data supplied, and the significant improvement that *not* cropping on humans makes. Therefore, we conclude that this reduction is caused by the association that Faster R-CNN has made of a bib sheet to a human torso: it is likely that the network recognises these bib sheets best when the raw input is provided as this is the augmented dataset that we trained our network on (i.e., we did not train our network on cropped humans).

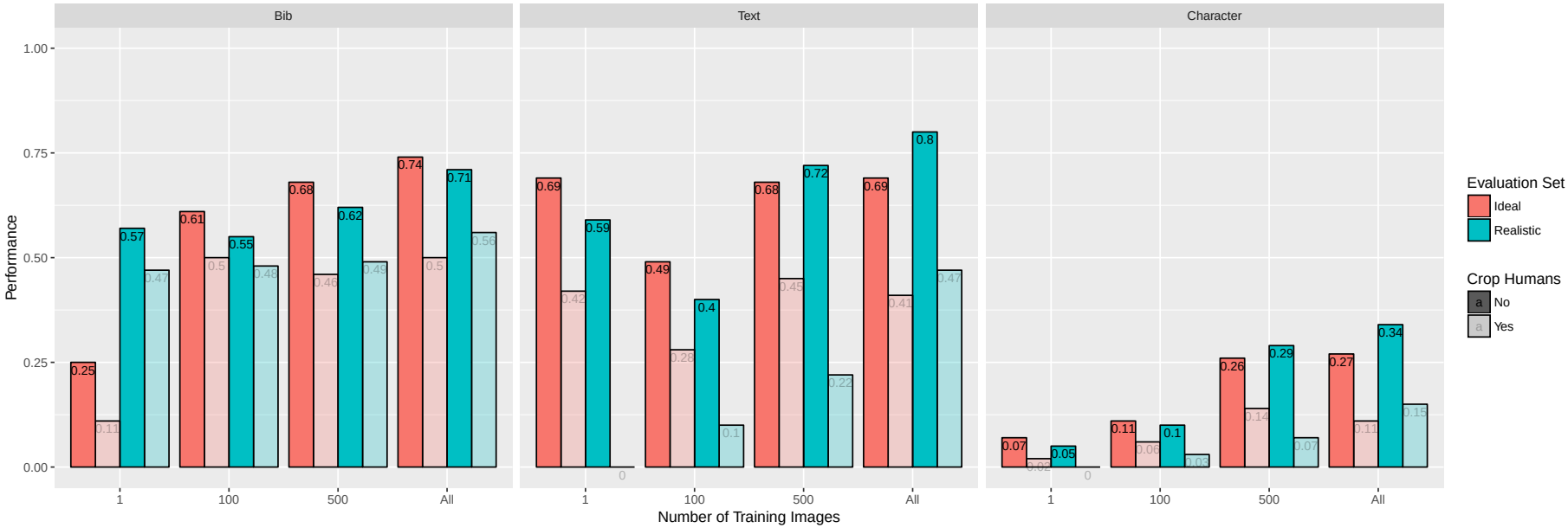


Figure 6.2: Bib, text and character performance of our pipeline over both realistic and ideal datasets using both cropping and non-cropping.

Thus, the association made by Faster R-CNN on the bib sheet to a human torso has a significant impact on the degrading accuracies. A 26% degrade in performance occurs for bibs which, in turn, has a hinderance on further stages of the pipeline; a 55% degrade in performance for text occurs and 65% for character recognition.

The significant decline between text detection and character detection prompted further investigation. We observe in Figure 6.2 that the highest bib, text and character performance is generally highest when all training data is used; as the automatic calculation of these stages were not manually verified by a human, we sampled 30 photos twice in two rounds of manual verification for a further evaluation. We discuss this evaluation in the following section.

6.3.2 Performance in Manual Evaluation

A total of 240 photos were manually inspected for both cropping and non-cropping over the ideal and realistic datasets in two rounds[2]. Results of both rounds have been averaged in Table F.5, which we present in Figure 6.3.

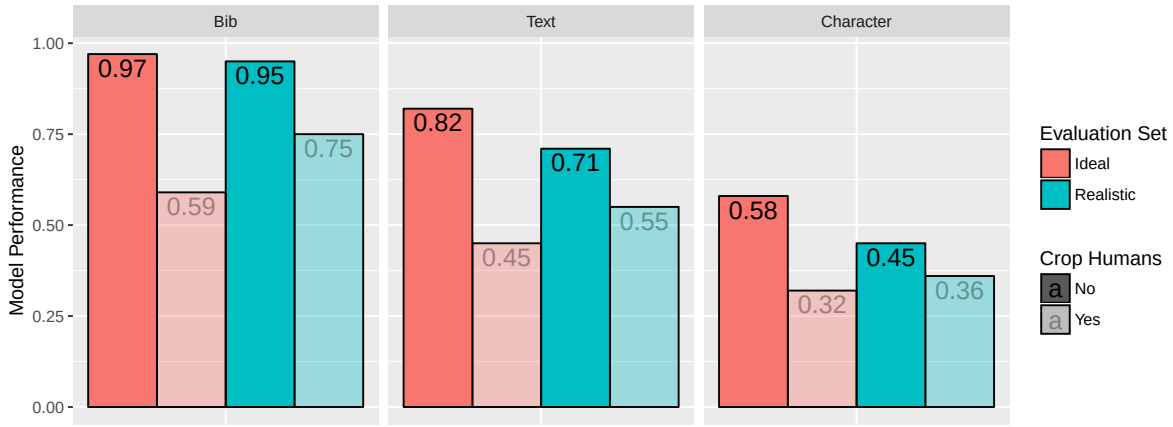


Figure 6.3: Results of bib, character and text performance manually inspected on a sample of all training images.

Similar to that of Figure 6.2, a negative association is shown as bib detection progresses to text and character detection. This negative association is clearer in the manual inspection: while bib detection is relatively high for both realistic and ideal cases, there is still a significant drop of performance for cropping and a gradual degrade for stages following bib detection. Character performance still acts as the biggest bottleneck to the pipeline.

In visualising the distribution of our manual evaluation, Wickham and Stryjewski [162] show that variations to a boxplot can introduce a richer distributional summary of the his-

[2]That is, using all training data, thus evaluation IDs I-ALL-CR, I-ALL-NC, R-ALL-CR, R-ALL-NC.

togram or density plot. This retains usefulness in comparing distributions across groups of data. Thus, a violin plot [60] is be useful to visualise the density of each individual photo we manually inspected. We present such information in Figure 6.4[3] and juxtapose both evaluation sets, comparing cropping versus non-cropping and further conditions of the photo, such as lighting conditions of the photo and the photo's blurriness.

Generally, we observe a wide dichotomy of the individual photo's performance with regards to bib performance. Within realistic non-cropping, we find that all only one photo has a miss-rate of 50%, with a large cluster toward 100%. Only two photos have false negatives in the ideal non-cropping case. Text and character detection preserves a similar dichotomy albeit with a larger distribution toward the median in the realistic cases. Most importantly, we see a reflection of our observation that performance in *all models* are improved whereby the mean performances increases from cropping to non-cropping. Generally, there is no bias toward photos taken at night or day as the distribution of both are consistent in all performance categories. Photos that are blurry reflect a poorer performance, and usually fall within the first 25% of our distribution within all datasets. However, there are some outliers in this case for realistic character performance whereby some blurry photos fall amongst all semi-interquartile ranges.

To follow up from the character detection performance, we assessed the character recognition performance for true positive cases following this manual evaluation, as described in the following section.

[3] Jitter has been added to this plot's y-axis to improve the plot's readability. The distance between data points in the y-axis bears no impact on results and is for aesthetics. Furthermore, the quartiles of each distribution are shown as vertical lines, and the mean of the data points themselves are indicated with the larger solid dot presented alongside with an error bar.

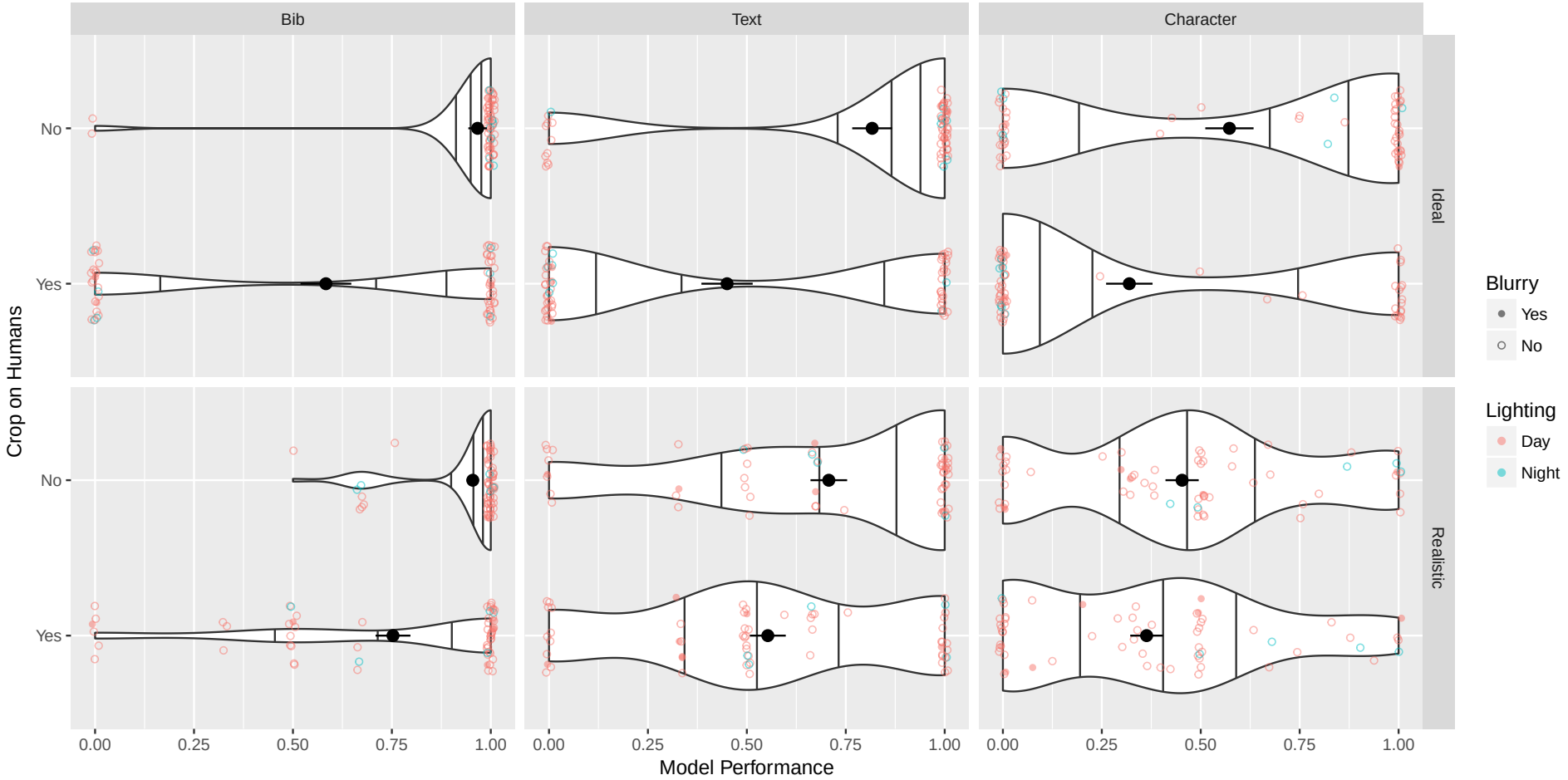
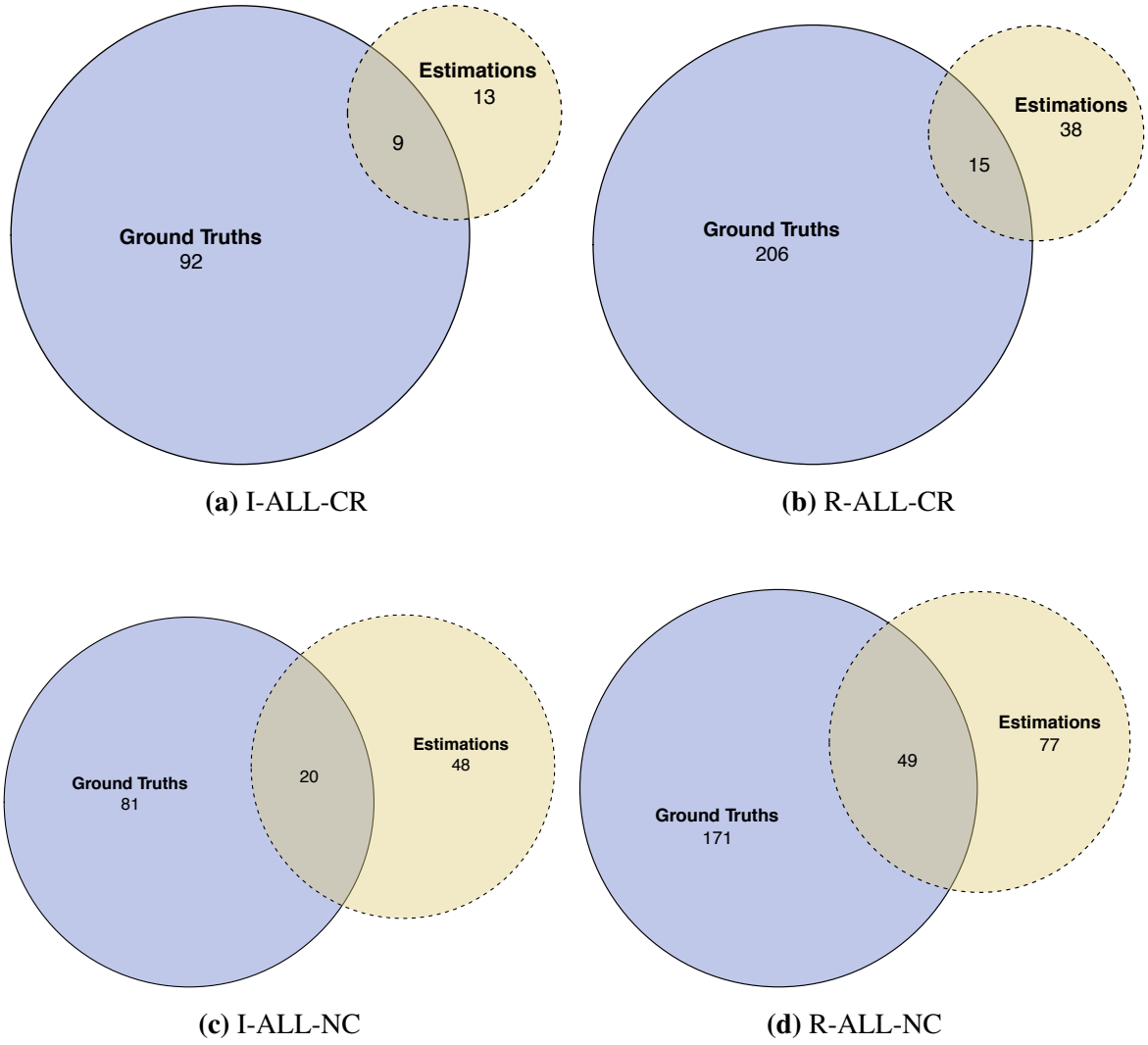


Figure 6.4: Distribution of manual evaluation of all 240 photos individually assessed for performance over three categories. We compare the ideal and realistic distribution in the vertical facets whilst also factoring in the photo's lighting and blur conditions.

6.3.3 Overall Performance

Presented in Figure 6.5 are the OCR performance metrics observed from photos trained with all training data. When compared to Figure 6.1, we observe problematic performance in all cases: whilst the pipeline returns true positive matches and some missing false negatives, false positives have been introduced. Cropping shows significicant disadvantages when compared to non-cropping, with a 21% improvement in true positive matches in realistic cases and 15% improvement in ideal.



(a) I-ALL-CR **(b)** R-ALL-CR

(c) I-ALL-NC **(d)** R-ALL-NC

Figure 6.5: OCR performance results of our pipeline on all training data. The first row (subfigures a and b) show ideal/realistic cropping, with a true positive rate of 9 and 7%, respectively. The second row (subfigures c and d) shows ideal/realistic without cropping, with a true positive rate of 24 and 28%, respectively.

6.4 Summary

In this chapter, we have presented an evaluation strategy of our pipeline that extends that of previous literature. Using previous evaluation strategies, discussed in Section 2.3, a comparison of a higher f-score metric of 0.59 by Ben-ami et al. [9] is noted, compared to our best of 0.17. This said, we suggest that this may be due to randomly sampled photos chosen from our dataset. An evaluation of the dataset and pipeline used by Ben-ami et al. against our own is not possible, since neither are publicly available. We cannot comment for the other metrics of text and character performance defined in this chapter as such metrics have not been applied in other works.

Bib detection in general is largely accurate, and we have successfully applied Faster R-CNN to recognise bib sheets with an accuracy larger than 95%. However, limitations showing a general decline in the performance of our pipeline following bib detection (71% for text; 45% for characters) can be mitigated through means that we discuss in the conclusion of this thesis.

The primary contributions of this chapter are:

- the development of metrics used to assess the performance of a alphanumeric sequence recognition pipeline,
- a demonstration that a CNN such as Faster R-CNN can detect bib sheets in a photo in-the-wild, and
- a demonstration that text regions can be detected using transfer learning with Faster R-CNN.

Minor contributions in this chapter include:

- a demonstration of a NN-based OCR's performance given a cropped alphanumeric sequence without the need to individually segment each character beforehand, and
- a visualisation technique of the distribution of individually assessed photos for performance quality for both detection and recognition.

Chapter 7

Conclusions

The unification of smartphones and cameras has sparked a vast amount of applications in image processing from natural scenes. Motivated by information extraction from the unstructured data within a natural image, many applications have sought to process images via heuristics, as opposed to deep-learning neural networks, for the purposes of object detection. Potential issues begin to arise when researchers rely on heuristics. In the context of marathon runners, facial detection is used in conjunction with a predefined ratio to determine the torso area [9]. In others—such as License Plate Recognition (LPR)—we find that distinct image properties (e.g., stroke, width and colour) are utilised, and thus as images become more complex, the performance is likely to degrade [92].

7.1 Primary Contributions

Overcoming the limitations of this Connected Component (CC)-based detection strategy has been investigated within this study, by only using deep-learning Neural Networks (NNs) for learning-based object detection. We find that detection with learning-based methods for object detection can achieve up to 97% accuracy using transfer-learning on pre-existing Region-based Convolutional Neural Network (R-CNN) (Faster R-CNN [127]) for visual processing given context of the desired object to be detected (i.e., racing bibs). We also find that transfer-learning can achieve partially similar results of up to 82% accuracy in detection of alphanumeric regions within those extracted features by changing the dataset and desired feature to text regions.

This thesis also aimed to tackle the issue of different Optical Character Recognition (OCR) strategies. We have shown that our pipeline is able to use NN-based OCR to recognise the alphanumeric sequences of a Racing Bib Number (RBN) without character segmentation. We

also show that preprocessing these images are a requirement for non-binarised text (i.e., red RBNs) into pre-existing OCR engines such as Tesseract.

We define an approach to define the prominence of a given runner within a photo using a data-collection system, *Argus*, that was used to curate labels for a large dataset in order to train the network. As a side research tangent of this thesis, we achieved a metamodel proposing how such applications can be used for further contexts, and not just marathon photography, as well as a metamodel development methodology to create metamodels using a quasi-experimental design and observational studies of a prototype systems.

7.2 Future Work

Our end-to-end NN pipeline shows significant fallbacks on human cropping—the proposed method of introducing cropping to improve accuracies has been shown to degrade our performance, rather than the initial intention to improve it. Overcoming this limitation may be as simple as re-training Faster R-CNN simply on cropped images with wider padding, rather than on raw images, though testing such a hypothesis is left open.

While we have developed a system to label the prominence of runners, we still leave the implementation of a classifier to understand what prominence is open to future work. By proposing a method by which all crowded photos are discarded and teaching a classifier on biased Likelihood of Purchase (LoP) (i.e., by removing the intermediary MAYBE candidates), such prominence ranking of subjects is within reach.

Similarly, the degrading text area detection may be improved by applying transfer learning to other R-CNN-based networks on a per-pixel level. As mentioned in Section 2.1.2, Mask-R-CNN, has recently appeared as a preprint in early 2017. Improvements to Argus to record data points on a per-pixel basis may allow for improvement in the areas of per-pixel detection, thereby extracting digits from a RBN within a bib even easier. An illustration of potential such applications are shown in Figure 7.1.

We propose to mitigate this bottleneck limitation of poor character detection by applying transfer learning to Faster R-CNN in a similar fashion to that described in Section 5.1.2. By producing a synthetic RBN dataset similar to that of Jaderberg et al. [67, 68] (with ground truth bounding boxes of each character known), we can augment these sample RBNs and therefore re-train Faster R-CNN to understand positions of characters. Each character from the detected regions can be extracted and piped into a CNN trained on recognising specific digits,

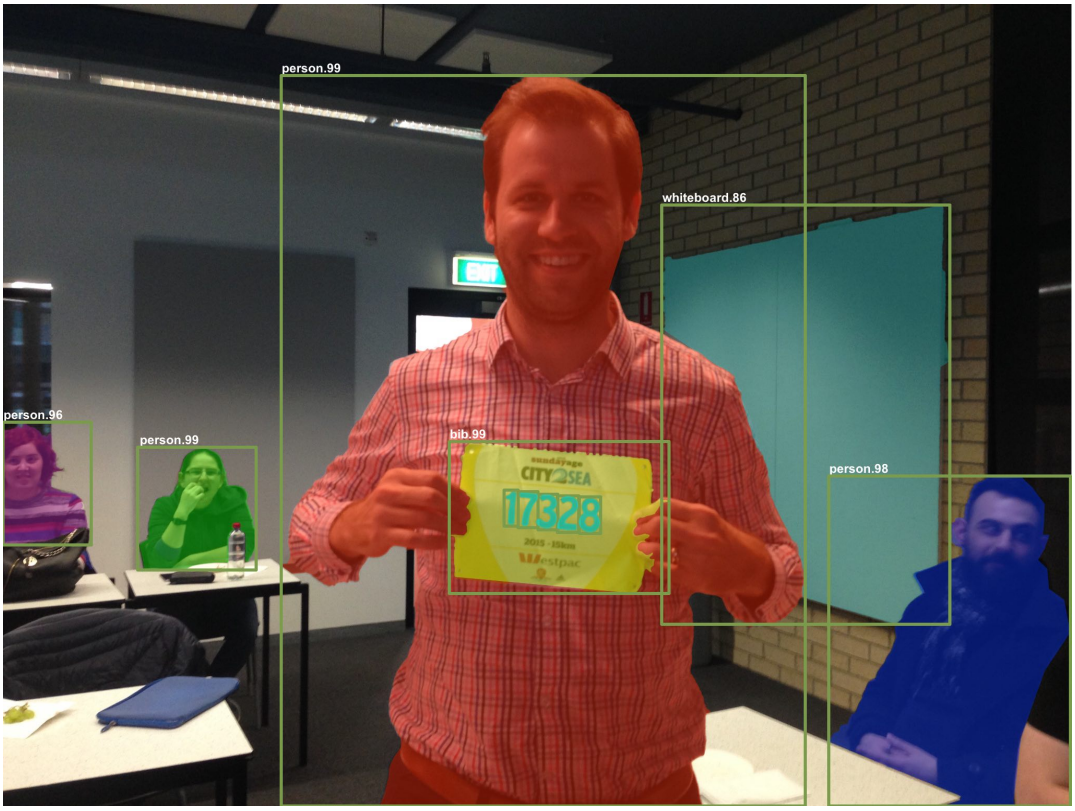


Figure 7.1: Potential use of Mask R-CNN to detect a bib sheet and its RBN on with per-pixel accuracy.

or perhaps—in addition to training Faster R-CNN to learn the positions of characters—the characters themselves can be trained in the same process. We leave such an implementation open to future works.

7.3 Wider Applicability

Alphanumeric character recognition goes far beyond RBN recognition. Other contexts in the area of self-driving vehicle literature, such as LPR and Traffic Sign Recognition (TSR), is imperative for the realisation of these vehicles. Similar areas include meter reading of electronic and water meters, or street sign numbers as discussed in Chapter 2.

The objective is to change the feature of interest: if one is interested in training a network about license plates, then train the network as we have done with racing bib numbers, but with an annotated dataset labelled with license plates. Additionally, training a text detection network with typeface specificity in mind (i.e., if one knows the typeface of a license plate or electronic meter), swapping out the text recognition portion of our pipeline with this trained network is achievable. This can be done to not only improve our OCR bottleneck, but to

improve context-specific cases.

7.4 Closing Remarks

In his 2017 book, Klaus Schwab [82], the founder and executive chairman of the World Economic Forum, comments on the impact on human society with the impending fourth industrial revolution, coined Industry 4.0 by Kagermann et al. [76]. Schwab discusses that such a revolution will fundamentally be underpinned by Artificial Intelligence (AI) systems. Thus, improvements in object detection, prominence ranking, and alphanumeric sequence recognition are just a slice of how the impact of automation may lessen the need for humanity to rely on heuristic-driven systems.

We anticipate that this thesis encompasses the body of how AI systems, like the pipeline developed within this work, show that we are only at the beginning for what is soon to come.

References

[1] Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[2] 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*.

[3] 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*.

[4] Baird, H. S. (1992). Document Image Defect Models. In *Erratum*, pp. 546–556. Berlin, Heidelberg: Springer Berlin Heidelberg.

[5] Barriuso, A. and A. Torralba (2012). Notes on image annotation. *CoRR abs/1210.3448*.

[6] Baxter, J. (1997). A Bayesian/Information Theoretic Model of Learning to Learn via Multiple Task Sampling. *Machine Learning 28*(1), 7–39.

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

[8] Belongie, S. J., J. Malik, and J. Puzicha (2001). Matching Shapes. *ICCV*.

[9] 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.

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

[11] Bézivin, J. (2006). Model Driven Engineering: An Emerging Technical Space. In *Computer Vision – ACCV 2010*, pp. 36–64. Berlin, Heidelberg: Springer Berlin Heidelberg.

[12] Bickel, S. (2006). Ecml-pkdd discovery challenge 2006 overview. *ECML-PKDD Discovery Challenge Workshop*.

[13] Bissacco, A., M. Cummins, Y. Netzer, and H. Neven (2013). PhotoOCR - Reading Text in Uncontrolled Conditions. *ICCV*.

[14] Buneman, P., S. Khanna, and W.-C. Tan (2000, November). Data Provenance: Some Basic Issues. In *Computer Vision – ACCV 2010*, pp. 87–93. Berlin, Heidelberg: Springer Berlin Heidelberg.

[15] Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov. 2*(2), 121–167.

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

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

[18] Cardelli, L. and P. Wegner (1985, December). On understanding types, data abstraction, and polymorphism. *ACM Computing Surveys (CSUR) 17*(4), 471–523.

[19] Caruana, R. (1997). Multitask Learning. *Machine Learning*.

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

[21] Chen, D., J.-M. Odobez, and H. Bourlard (2004a). Text detection, recognition in images and video frames. *Pattern Recognition*.

[22] Chen, D., J.-M. Odobez, and H. Bourlard (2004b). Text detection, recognition in images and video frames. *Pattern Recognition*.

[23] 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.*.

[24] 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*.

[25] Chen, T., M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang (2015, December). MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. *arXiv.org*.

[26] Chen, X., H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick (2015, April). Microsoft COCO Captions: Data Collection and Evaluation Server.

[27] 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.

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

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

[30] Chollet, F. et al. (2015). Keras. `https://github.com/fchollet/keras`.

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

[32] Cortes, C. and V. Vapnik (1995). Support-Vector Networks. *Machine Learning*.

[33] Cui, Y. and J. Widom (2003, May). Lineage tracing for general data warehouse transformations. *The VLDB Journal The International Journal on Very Large Data Bases 12*(1), 41–58.

[34] Dalal, N. and B. Triggs (2005). Histograms of Oriented Gradients for Human Detection. *CVPR 1*, 886–893.

[35] Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE.

[36] Do, C. B. and A. Y. Ng (2005). Transfer learning for text classification. *NIPS*.

[37] 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.

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

[39] Everingham, M., S. M. A. Eslami, L. J. Van Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman (2015). The Pascal Visual Object Classes Challenge - A Retrospective. *International Journal of Computer Vision*.

[40] Everingham, M., L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman (2009, September). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision 88*(2), 303–338.

[41] Everingham, M., L. J. Van Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman (2010). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*.

[42] Everingham, M., A. Zisserman, C. K. I. Williams, L. J. Van Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorkó, S. Duffner, J. Eichhorn, J. D. R. Farquhar, M. Fritz, C. Garcia, T. L. Griffiths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann, J. Shawe-Taylor, A. J. Storkey, S. Szedmák, B. Triggs, I. Ulusoy, V. Viitaniemi, and J. Zhang (2005). The 2005 PASCAL Visual Object Classes Challenge. *MLCW*.

[43] Faloutsos, C., R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz (1994, July). Efficient and effective Querying by Image Content. *Journal of Intelligent Information Systems 3*(3-4), 231–262.

[44] Freeman, W. T. and M. Roth (1995). Orientation histograms for hand gesture recognition. *International workshop on automatic ...*.

[45] Freund, Y. and R. E. Schapire (1996). Experiments with a New Boosting Algorithm. *ICML*.

[46] Friedman, J., R. Tibshirani, and T. Hastie (2000, April). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The annals of statistics 28*(2), 337–407.

[47] 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.

[48] Gatos, B., I. Pratikakis, and K. Kepene (2005). Text detection in indoor/outdoor scene images. *Proc First Workshop of* ....

[49] 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.

[50] Girshick, R., J. Donahue, T. Darrell, and J. Malik (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587. IEEE.

[51] Girshick, R. B. (2015). Fast R-CNN. *ICCV*.

[52] Gllavata, J., R. Ewerth, and B. Freisleben (2004). Text Detection in Images Based on Unsupervised Classification of High-Frequency Wavelet Coefficients. *ICPR*.

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

[54] Gray, D. E. (2013, November). *Doing Research in the Real World*. SAGE.

[55] Gupta, A., A. Vedaldi, and A. Zisserman (2016). Synthetic data for text localisation in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2315–2324. IEEE.

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

[57] Hanif, S. M., L. Prevost, and P. Negri (2008). A cascade detector for text detection in natural scene images. *ICPR*.

[58] He, K., G. Gkioxari, P. Dollár, and R. B. Girshick (2017). Mask R-CNN. *CoRR*.

[59] Heisele, B., T. Serre, S. Mukherjee, and T. A. Poggio (2001). Feature Reduction and Hierarchy of Classifiers for Fast Object Detection in Video Images. *CVPR*, 8.

[60] Hintze, J. L. and R. D. Nelson (1998, May). Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician 52*(2), 181.

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

[62] Hua, X.-S., L. Wenyin, and H. Zhang (2001). Automatic Performance Evaluation for Video Text Detection. *ICDAR*.

[63] 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.*.

[64] 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.

[65] Ikeda, R. and J. Widom (2009). Data lineage: A survey.

[66] Ivanov, I., J. Bzivin, and M. Aksit (2002, 10). Technological spaces: An initial appraisal. pp. 1–6. <http://www.cs.rmit.edu.au/fedconf/2002/program.html>.

[67] Jaderberg, M., K. Simonyan, A. Vedaldi, and A. Zisserman (2014, June). Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition. *arXiv.org*.

[68] Jaderberg, M., K. Simonyan, A. Vedaldi, and A. Zisserman (2016). Reading Text in the Wild with Convolutional Neural Networks. *International Journal of Computer Vision*.

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

[70] 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.

[71] 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.

[72] Jordan, P. W., B. Thomas, I. L. McClelland, and B. Weerdmeester (1996, June). *Usability Evaluation In Industry*. CRC Press.

[73] 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.

[74] 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.

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

[76] Kagermann, H., W.-D. Lukas, and W. Wahlster (2011). Industrie 4.0: Mit dem internet der dinge auf dem weg zur 4. industriellen revolution. *VDI nachrichten 13*, 11.

[77] Karatzas, D., L. Gomez-Bigorda, A. Nicolaou, S. K. Ghosh, A. D. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny (2015). ICDAR 2015 competition on Robust Reading. *ICDAR*.

[78] Karatzas, D., S. Robles, and L. Gomez (2014). An On-line Platform for Ground Truthing and Performance Evaluation of Text Extraction Systems. In *2014 11th IAPR International Workshop on Document Analysis Systems*, pp. 242–246. IEEE.

[79] Karatzas, D., F. Shafait, S. Uchida, M. Iwamura, L. G. i. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras (2013). ICDAR 2013 Robust Reading Competition. In *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1484–1493. IEEE.

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

[81] Kim, K. I., K. Jung, and J. H. Kim (2003). Texture-Based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.*.

[82] Klaus Schwab (2017, January). *The Fourth Industrial Revolution*. Penguin UK.

[83] Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). ImageNet Classification with Deep Convolutional Neural Networks. *NIPS*.

[84] Kumar, D., M. N. A. Prasad, and A. G. Ramakrishnan (2012, December). *MAPS: midline analysis and propagation of segmentation*. midline analysis and propagation of segmentation. New York, New York, USA: ACM.

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

[86] LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*(11), 2278–2324.

[87] 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.

[88] Lee, C.-Y. and S. Osindero (2016, March). Recursive Recurrent Nets with Attention Modeling for OCR in the Wild. pp. 1–10.

[89] 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.

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

[91] Li, H., D. S. Doermann, and O. E. Kia (2000). Automatic text detection and tracking in digital video. *IEEE Trans. Image Processing*.

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

[93] Li, Y., H. Qi, J. Dai, X. Ji, and Y. Wei (2016). Fully convolutional instance-aware semantic segmentation. *arXiv.org*.

[94] 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.

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

[96] Lienhart, R. and J. Maydt (2002). An extended set of Haar-like features for rapid object detection. *ICIP*.

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

[98] Lin, T.-Y., M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár (2014, May). Microsoft COCO: Common Objects in Context. *arXiv.org*.

[99] Liu, C. L., M. Koga, and H. Fujisawa (2005). Gabor feature extraction for character recognition: comparison with gradient feature. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pp. 121–125 Vol. 1. IEEE.

[100] Liu, H. and X. Ding (2005). Handwritten Character Recognition Using Gradient Feature and Quadratic Classifier with Multiple Discrimination Schemes. *ICDAR*.

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

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

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

[104] 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.

[105] 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.

[106] Lucas, S. M., A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, H. Miyao, J. Zhu, W. Ou, C. Wolf, J.-M. Jolion, L. Todoran,

M. Worring, and X. Lin (2005, July). ICDAR 2003 robust reading competitions: entries, results, and future directions. *IJDAR 7*(2-3), 105–122.

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

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

[109] Matera, T., J. Jakes, and M. Cheng (2014). A user friendly crowdsourcing task manager. . . . *on Computer Vision* . . . .

[110] McConnell, R. (1986, January 28). Method of and apparatus for pattern recognition. US Patent 4,567,610.

[111] 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*.

[112] Minetto, R., N. Thome, M. Cord, J. Fabrizio, and B. Marcotegui (2010). Snoopertext - A multiresolution system for text detection in complex visual scenes. *ICIP*.

[113] Mohan, A., C. Papageorgiou, and T. A. Poggio (2001). Example-Based Object Detection in Images by Components. *IEEE Trans. Pattern Anal. Mach. Intell.*.

[114] Moody, D. L. (2009). The "Physics" of Notations - Toward a Scientific Basis for Constructing Visual Notations in Software Engineering. *IEEE Trans. Software Eng.*.

[115] Mottaghi, R., X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille (2014). The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, pp. 891–898. IEEE.

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

[117] 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*.

[118] Nguyen, D. T., M.-K. Tran, and S.-K. Yeung (2015). An MRF-Poselets Model for Detecting Highly Articulated Humans. *ICCV*.

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

[120] Ojala, T., M. Pietikainen, and D. Harwood (1994). Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *12th International Conference on Pattern Recognition*, pp. 582–585. IEEE Comput. Soc. Press.

[121] Oxford English Dictionary (2015). 5th Edition. Retrieved 13 July 2017, <http://www.oed.com/view/Entry/196665>.

[122] Pan, Y.-F., C.-L. Liu, and X. Hou (2010). Fast scene text localization by learning-based filtering and verification. In *2010 17th IEEE International Conference on Image Processing (ICIP 2010)*, pp. 2269–2272. IEEE.

[123] Phan, T. Q., P. Shivakumara, and C. L. Tan (2009). A Laplacian Method for Video Text Detection. In *2009 10th International Conference on Document Analysis and Recognition*, pp. 66–70. IEEE.

[124] Raina, R., A. Y. Ng, and D. Koller (2006). Constructing informative priors using transfer learning. *ICML*.

[125] Redmon, J. (2013–2016). Darknet: Open source neural networks in c. http://pjreddie.com/darknet/.

[126] Redmon, J. and A. Farhadi (2016). Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*.

[127] Ren, S., K. He, R. B. Girshick, and J. S. 0001 (2017). Faster R-CNN - Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.*.

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

[129] Russell, B. C., A. T. 0001, K. P. Murphy, and W. T. Freeman (2008). LabelMe - A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*.

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

[131] 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.

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

[133] 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.

[134] 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*.

[135] 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.

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

[137] 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.*.

[138] 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.

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

[140] Sobottka, K., H. Bunke, and H. Kronenberg (1999). Identification of Text on Colored Book and Journal Covers. *ICDAR*.

[141] Sochman, J. and J. Matas (2005). WaldBoost ? Learning for Time Constrained Sequential Detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 150–156. IEEE.

[142] Sorokin, A. and D. A. Forsyth (2008). Utility data annotation with Amazon Mechanical Turk. *CVPR Workshops*.

[143] 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.

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

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

[146] Sung, K. K. and T. A. Poggio (1998). Example-Based Learning for View-Based Human Face Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*.

[147] 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*.

[148] Thrun, S. (1996). Is learning the n-th thing any easier than learning the first? *Advances in neural information processing systems*.

[149] 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.

[150] Trochim, W. and J. P. Donnelly (2001). Research methods knowledge base.

[151] 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*.

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

[153] Vapnik, V. (1999, November). *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

[154] Veit, A., T. Matera, L. Neumann, J. Matas, and S. J. Belongie (2016). COCO-Text - Dataset and Benchmark for Text Detection and Recognition in Natural Images. *CoRR*.

[155] Viola, P. A., M. J. Jones, and D. Snow (2003). Detecting Pedestrians Using Patterns of Motion and Appearance. *ICCV*.

[156] Vondrick, C., D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, 1–21. 10.1007/s11263-012-0564-1.

[157] Wang, K., B. Babenko, and S. J. Belongie (2011). End-to-end scene text recognition. *ICCV*.

[158] Wang, X., T. X. Han, and S. Yan (2009). An HOG-LBP human detector with partial occlusion handling. *ICCV*.

[159] Wang, X., L. Huang, and C. Liu (2009). A New Block Partitioned Text Feature for Text Verification. In *2009 10th International Conference on Document Analysis and Recognition*, pp. 366–370. IEEE.

[160] Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*.

[161] Wickham, H. (2010, January). A Layered Grammar of Graphics. *Journal of Computational and Graphical Statistics 19*(1), 3–28.

[162] Wickham, H. and L. Stryjewski (2012). 40 years of boxplots. Technical report, had.co.nz.

[163] Wolf, C. and J.-M. Jolion (2006, April). Object count/area graphs for the evaluation of object detection and segmentation algorithms. *IJDAR 8*(4), 280–296.

[164] Wong, S. C., A. Gatt, V. Stamatescu, and M. D. McDonnell (2016). Understanding data augmentation for classification - when to warp? *CoRR cs.CV*, arXiv:1609.08764.

[165] Xiao, J., J. Hays, K. A. Ehinger, A. Oliva, and A. T. 0001 (2010). SUN database - Large-scale scene recognition from abbey to zoo. *CVPR*.

[166] Yaeger, L. S., R. F. Lyon, and B. J. Webb (1996). Effective Training of a Neural Network Character Classifier for Word Recognition. *NIPS*.

[167] Ye, Q. and D. Doermann (2014, December). Text Detection and Recognition in Imagery: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*(7), 1480–1500.

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

[169] 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.

[170] 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.

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

[172] 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.

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

Appendix A

Ethics Clearance

Professor Rajesh Vasa
Associate Professor Andrew Cain
Mr Alex Cummaudo
School of Information Technology
Burwood Campus

9 August 2017

Dear Rajesh, Andrew and Alex

STEC-49-2017-CUMMAUDO titled *"Recognition and Prominence Ranking of Alphanumeric Number Sequences in Images."*

Thank you for submitting the above project for consideration by the Faculty Human Ethics Advisory Group (HEAG). The HEAG recognised that the project complies with the National Statement on Ethical Conduct in Human Research (2007) and has approved it. You may commence the project upon receipt of this communication.

The approval period is for three years. It is your responsibility to contact the Faculty HEAG immediately should any of the following occur:

- Serious or unexpected adverse effects on the participants
- Any proposed changes in the protocol, including extensions of time
- Any changes to the research team or changes to contact details
- Any events which might affect the continuing ethical acceptability of the project
- The project is discontinued before the expected date of completion.

You will be required to submit an annual report giving details of the progress of your research. Please forward your first annual report on 9/8/18 Failure to do so may result in the termination of the project. Once the project is completed, you will be required to submit a final report informing the HEAG of its completion.

Please ensure that the Deakin logo is on the Plain Language Statement and Consent Forms. You should also ensure that the project ID is inserted in the complaints clause on the Plain Language Statement, and be reminded that the project number must always be quoted in any communication with the HEAG to avoid delays. All communication should be directed to sciethic@deakin.edu.au

The Faculty HEAG and/or Deakin University Human Research Ethics Committee (HREC) may need to audit this project as part of the requirements for monitoring set out in the National Statement on Ethical Conduct in Human Research (2007).

If you have any queries in the future, please do not hesitate to contact me.

We wish you well with your research.

Kind regards

Teresa Treffry

Teresa Treffry
Secretary, Human Ethics Advisory Group (HEAG)
Faculty of Science Engineering & Built Environment

Appendix B

Usability Notes

B.1 Notes

Suggestions from participants that were implemented:

- Remove hat color option if not selected in color wizard.
- Otherwise if selected have Hat as option number 1.
- Prevent completion of bib numbers being drawn to screen if face not drawn.
- Note that steps are not updating for the images (retains previous image's step).
- Color wizard missing in main menu.
- Reword 'is image crowded' step to 'foreground crowded'.
- Cancel the color wizard should take you back to the base class wizard in first tag mode.
- Spit out a JSON file per photo rather than one dump JSON.
- Save on navigate to prevent data loss.
- Add big "Foreground is Crowded" over image if crowded is true.
- Open class dialogs to the side (left of centre)—that way we can see the runner.
- Add missing shortcut to delete runner.
- When going back an image, ask if you are finished tagging that image.
- Ask if image is crowded after 1.5 second delay—that way we can see the runner.
- Update bib number to be set in face mode (instruction text).
- Once tagged a person, automatically move to next person (in face mode).

B.2 System Usability Scale Scores

Table B.1: System Usability Scale (SUS) scores retrieved from usability testing.

Question	P1	P2	P3
1. I think that I would like to use this system frequently	4	4	4
2. I found the system unnecessarily complex	2	2	1
3. I thought the system was easy to use	4	4	5
4. I think that I would need the support of a technical person to be able to use this system	1	1	1
5. I found the various functions in this system were well integrated	5	4	5
6. I thought there was too much inconsistency in this system	1	1	2
7. I would imagine that most people would learn to use this system very quickly	5	4	4
8. I found the system very cumbersome to use	2	1	1
9. I felt very confident using the system	5	4	4
10. I needed to learn a lot of things before I could get going with this system	1	1	1
SUS Score	75	65	70
Average SUS Score			70

Appendix C

Metamodel Class Diagrams

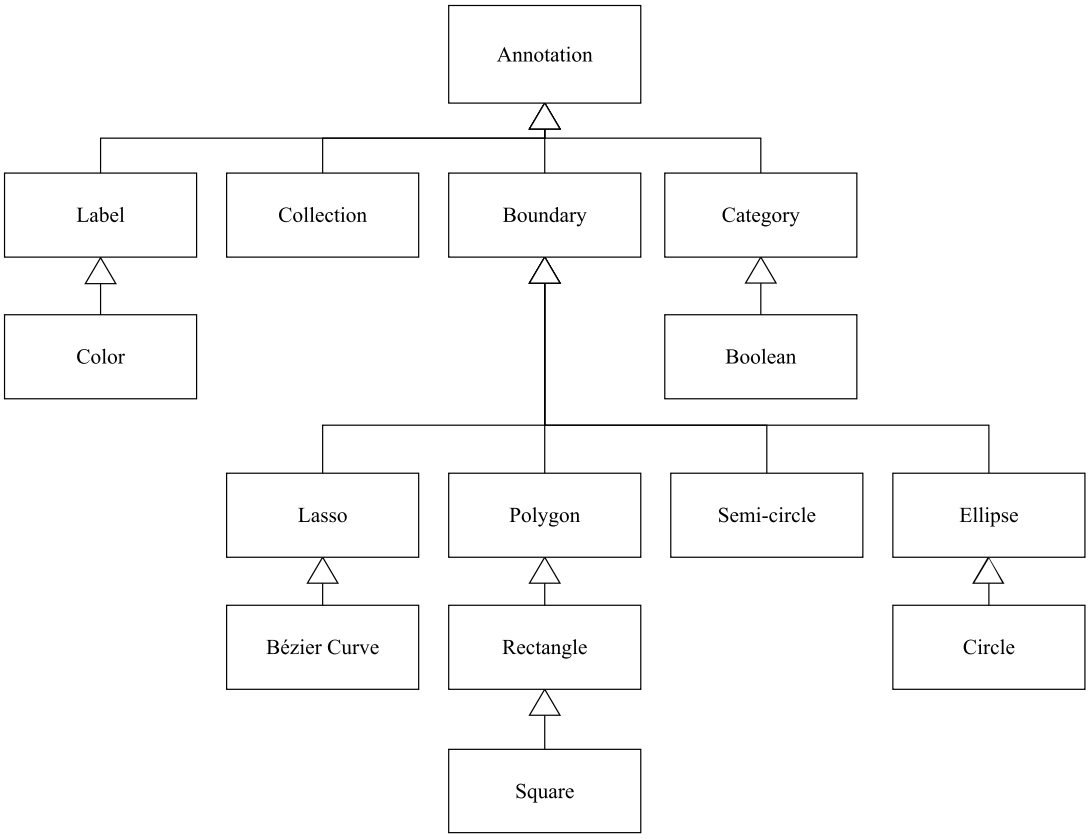


Figure C.1: Type class hierarchy of annotations

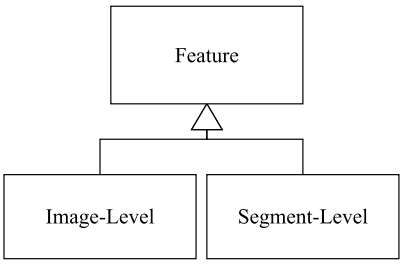


Figure C.2: Type class hierarchy of features

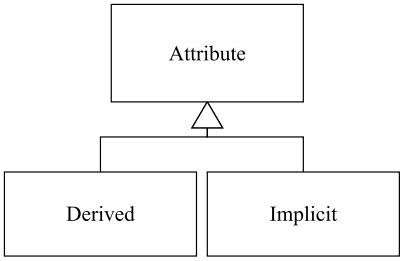


Figure C.3: Type class hierarchy of attributes

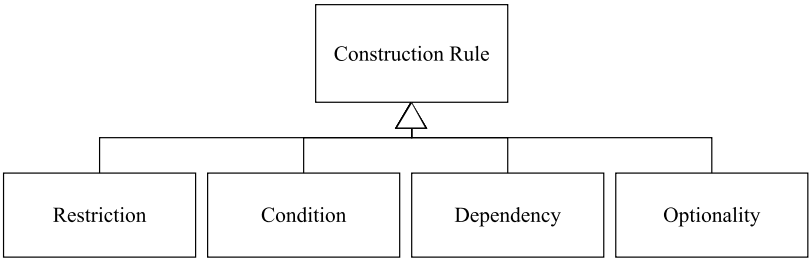


Figure C.4: Type class hierarchy of construction rules

Appendix D

Dataset Schemas

Listing D.1: MS COCO JavaScript Object Notation (JSON) Schema, sourced from `http://mscoco.org/dataset/#download`.

```
{
  "info" : info,
  "images" : [image],
  "annotations" : [annotation],
  "licenses" : [license],
}

info {
  "year" : int,
  "version" : str,
  "description" : str,
  "contributor" : str,
  "url" : str,
  "date_created" : datetime,
}

image {
  "id" : int,
  "width" : int,
  "height" : int,
  "file_name" : str,
  "license" : int,
  "flickr_url" : str,
  "coco_url" : str,
  "date_captured" : datetime,
}

license {
  "id" : int,
  "name" : str,
  "url" : str,
}

// Object Instance Annotations

annotation {
  "id" : int,
  "image_id" : int,
  "category_id" : int,
  "segmentation" : RLE or [polygon],
```

```
  "area" : float,
  "bbox" : [x,y,width,height],
  "iscrowd" : 0 or 1,
}

categories [{
  "id" : int,
  "name" : str,
  "supercategory" : str,
}]

// Object Keypoint Annotations

annotation {
  "keypoints" : [x1,y1,v1,...],
  "num_keypoints" : int,
  "[cloned]" : ...,
}

categories [{
  "keypoints" : [str],
  "skeleton" : [edge],
  "[cloned]" : ...,
}]

// Image Caption Annotations

annotation {
  "id" : int,
  "image_id" : int,
  "caption" : str,
}
```

Listing D.2: COCO-Text JSON Schema, sourced from `https://vision.cornell.edu/se3/coco-text-2/`.

```
{
  "info" : info,
  "imgs" : [image],
  "anns" : [annotation]
}

info {
  "version" : str,
  "description" : str,
  "author" : str,
  "url" : str,
  "date_created" : datetime
}

image {
  "id" : int,
  "file_name" : str,
  "width" : int,
  "height" : int,
  "set" : str // train or val
}

annotation {
  "id" : int,
  "image_id" : int,
  "class" : str, // machine printed or handwritten or others
  "legibility" : str, // legible or illegible
  "language" : str, // english or not english or na
  "area" : float,
  "bbox" : [x,y,width,height],
  "utf8_string" : str,
  "polygon" : []
}
```

Listing D.3: XML Schema Definition (XSD) of the ICDAR 2011–2015 annotation schema, sourced within `http://www.cvc.uab.es/apep/downloads.php`.

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="textGt">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="textLine" maxOccurs="unbounded" />
      </xs:sequence>
      <xs:attribute name="width" type="xs:integer" use="required" />
      <xs:attribute name="height" type="xs:integer" use="required" />
      <xs:attribute name="imageFileName" type="xs:string" use="required"
        ↪ />
      <xs:attribute name="areaImageFilename" type="xs:string" use="
        ↪ optional" />
      <xs:attribute name="skeletonImageFilename" type="xs:string" use="
        ↪ optional" />
      <xs:attribute name="docVersion" type="xs:integer" use="required"
        ↪ fixed="3" />
      <xs:attribute name="author" type="xs:string" use="required" />
      <xs:attribute name="comments" type="xs:string" use="optional" />
    </xs:complexType>
  </xs:element>
  <xs:element name="textLine">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="word" minOccurs="0" maxOccurs="unbounded" />
        <xs:element ref="atom" minOccurs="0" maxOccurs="unbounded" />
      </xs:sequence>
      <xs:attribute name="id" type="xs:integer" use="required" />
      <xs:attribute name="transcription" type="xs:string" use="required"
        ↪ />
      <xs:attribute name="xmin" type="xs:integer" use="required" />
      <xs:attribute name="xmax" type="xs:integer" use="required" />
      <xs:attribute name="ymin" type="xs:integer" use="required" />
      <xs:attribute name="ymax" type="xs:integer" use="required" />
      <xs:attribute name="dontCare" type="xs:boolean" use="required" />
    </xs:complexType>
  </xs:element>
  <xs:element name="word">
    <xs:complexType>
      <xs:sequence>
```

```
      <xs:element ref="atom" minOccurs="0" maxOccurs="unbounded" />
    </xs:sequence>
    <xs:attribute name="id" type="xs:integer" use="required" />
    <xs:attribute name="transcription" type="xs:string" use="required" />
    <xs:attribute name="xmin" type="xs:integer" use="required" />
    <xs:attribute name="xmax" type="xs:integer" use="required" />
    <xs:attribute name="ymin" type="xs:integer" use="required" />
    <xs:attribute name="ymax" type="xs:integer" use="required" />
    <xs:attribute name="dontCare" type="xs:boolean" use="required" />
  </xs:complexType>
 </xs:element>
 <xs:simpleType name="colorHex">
  <xs:restriction base="xs:string">
   <xs:pattern value="#[0-9a-fA-F]{6}"/>
  </xs:restriction>
 </xs:simpleType>
 <xs:element name="atom">
  <xs:complexType>
   <xs:attribute name="id" type="xs:integer" use="required" />
   <xs:attribute name="transcription" type="xs:string" use="required" />
   <xs:attribute name="splitByDesign" type="xs:boolean" use="required" />
   <xs:attribute name="mergedByDesign" type="xs:boolean" use="required " />
   <xs:attribute name="color" type="colorHex" use="required" />
   <xs:attribute name="dontCare" type="xs:boolean" use="required" />
   <xs:attribute name="textParts" type="xs:integer" use="required" />
   <xs:attribute name="xmin" type="xs:integer" use="required" />
   <xs:attribute name="xmax" type="xs:integer" use="required" />
   <xs:attribute name="ymin" type="xs:integer" use="required" />
   <xs:attribute name="ymax" type="xs:integer" use="required" />
  </xs:complexType>
 </xs:element>
</xs:schema>
```

Listing D.4: Sample annotation file of the ICDAR 2003–2011 annotation schema, from [106].

```
<tagset>
  <image>
    <imageName>scene/ComputerScienceSmall.jpg</imageName>
    <resolution x="338" y="255" />
    <taggedRectangles>
      <taggedRectangle x="99" y="94" width="128" height="20" offset="0"
          ↪ rotation="0">
      <tag>Department</tag>
      <segmentation>
        <xOff>16</xOff>
        <xOff>29</xOff>
        <xOff>43</xOff>
        <xOff>54</xOff>
        <xOff>64</xOff>
        <xOff>74</xOff>
        <xOff>93</xOff>
        <xOff>106</xOff>
        <xOff>117</xOff>
      </segmentation>
    </taggedRectangle>
    ...
  </image>
...
</tagset>
```

Listing D.5: Sample annotation file of the PASCAL VOC 2007 annotation schema, from `http://host.robots.ox.ac.uk/pascal/VOC/voc2007`.

```
<annotation>
  <folder>VOC2012</folder>
  <filename>2007_000032.jpg</filename>
  <source>
    <database>The VOC2007 Database</database>
    <annotation>PASCAL VOC2007</annotation>
    <image>flickr</image>
  </source>
  <size>
    <width>500</width>
    <height>281</height>
    <depth>3</depth>
  </size>
  <segmented>1</segmented>
  <object>
    <name>aeroplane</name>
    <pose>Frontal</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <bndbox>
      <xmin>104</xmin>
      <ymin>78</ymin>
      <xmax>375</xmax>
      <ymax>183</ymax>
    </bndbox>
  </object>
  <object>
    <name>aeroplane</name>
    <pose>Left</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <bndbox>
      <xmin>133</xmin>
      <ymin>88</ymin>
      <xmax>197</xmax>
      <ymax>123</ymax>
    </bndbox>
  </object>
  <object>
    <name>person</name>
```

```
    <pose>Rear</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <bndbox>
      <xmin>195</xmin>
      <ymin>180</ymin>
      <xmax>213</xmax>
      <ymax>229</ymax>
    </bndbox>
  </object>
  <object>
    <name>person</name>
    <pose>Rear</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <bndbox>
      <xmin>26</xmin>
      <ymin>189</ymin>
      <xmax>44</xmax>
      <ymax>238</ymax>
    </bndbox>
  </object>
</annotation>
```

Listing D.6: Sample annotation file of the PASCAL VOC 2012 annotation schema (extended from 2007), from `http://host.robots.ox.ac.uk/pascal/VOC/voc2012`.

```
<annotation>
  <filename>2012_004317.jpg</filename>
  <folder>VOC2012</folder>
  <object>
    <name>person</name>
    <actions>
      <jumping>0</jumping>
      <other>0</other>
      <phoning>0</phoning>
      <playinginstrument>0</playinginstrument>
      <reading>0</reading>
      <ridingbike>0</ridingbike>
      <ridinghorse>0</ridinghorse>
      <running>0</running>
      <takingphoto>0</takingphoto>
      <usingcomputer>1</usingcomputer>
      <walking>0</walking>
    </actions>
    <bndbox>
      <xmax>170</xmax>
      <xmin>64</xmin>
      <ymax>310</ymax>
      <ymin>131</ymin>
    </bndbox>
    <difficult>0</difficult>
    <pose>Unspecified</pose>
    <point>
      <x>109</x>
      <y>193</y>
    </point>
  </object>
  <segmented>0</segmented>
  <size>
    <depth>3</depth>
    <height>500</height>
    <width>375</width>
  </size>
  <source>
    <annotation>PASCAL VOC2012</annotation>
    <database>The VOC2012 Database</database>
```

```
41      <image>flickr</image>
42    </source>
43  </annotation>
```

Appendix E

Dataset Mappings

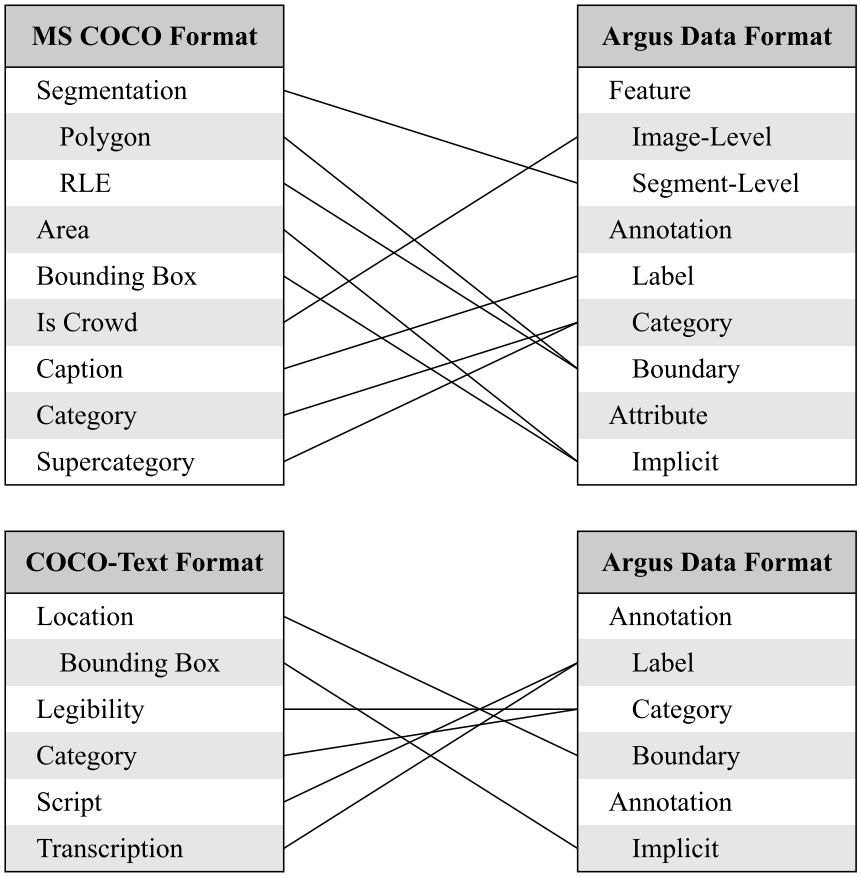


Figure E.1: Annotation components of both MS-COCO and COCO-text (left) and ADF (right).

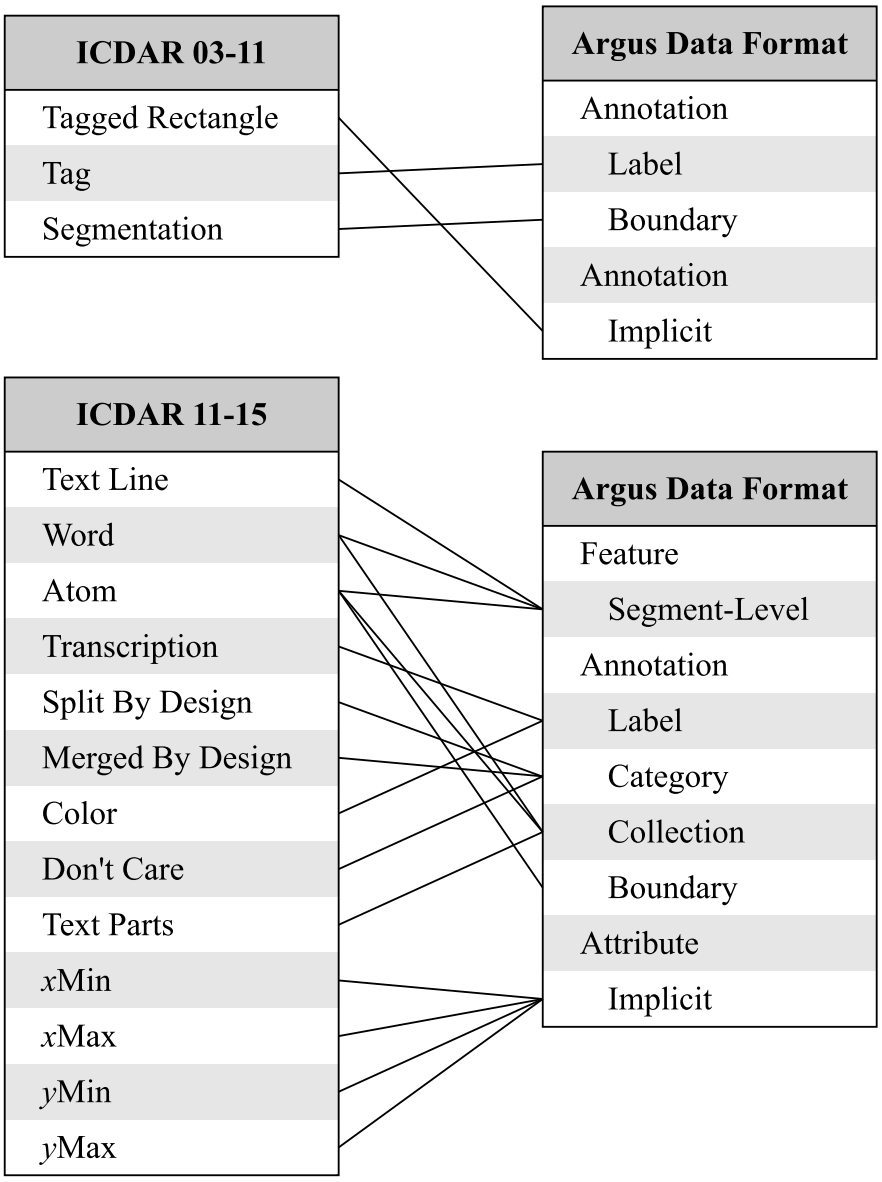


Figure E.2: Annotation components of the ICDAR 03–11 and 11–15 formats (left) with ADF (right).

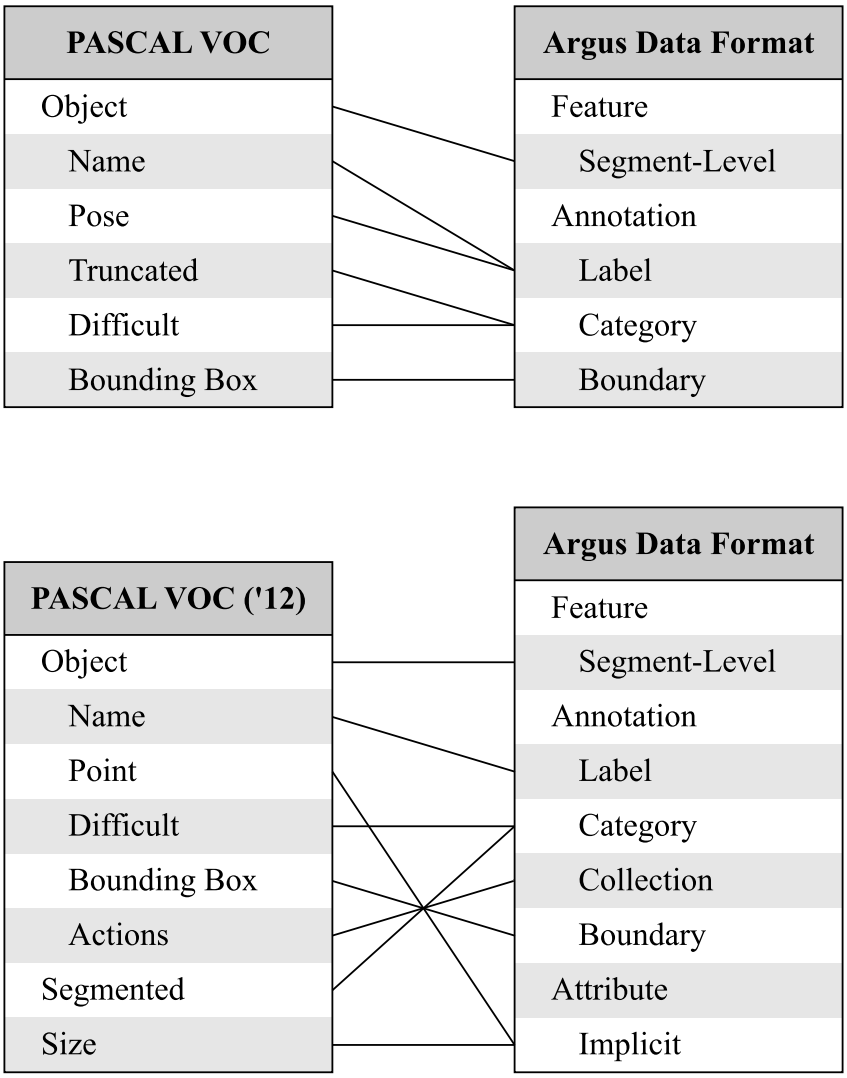


Figure E.3: Annotation components of the PASCAL VOC formats (left) with ADF (right).

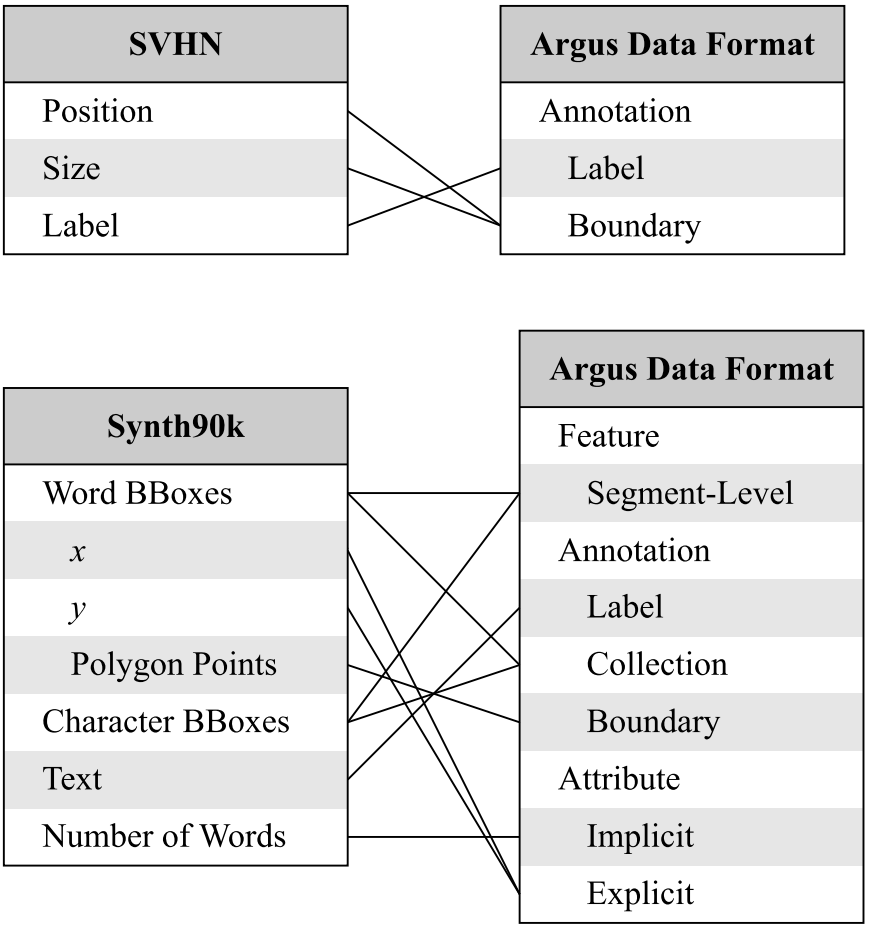


Figure E.4: Annotation components of binary-encoded MATLAB formats of the SVHN and Synth90k datasets (left) with ADF (right).

Appendix F

Supplementary Data

Table F.1: Summary of Bib Detection

Evaluation ID	Model Performance	Confidence	Precision	Recall	f-score
I-1-CR	0.11	0.99	0.15	0.15	0.15
I-1-NC	0.25	1.00	0.17	0.15	0.16
I-100-CR	0.50	0.96	0.08	0.08	0.08
I-100-NC	0.61	0.98	0.16	0.15	0.15
I-500-CR	0.46	0.95	0.17	0.14	0.15
I-500-NC	0.68	0.97	0.26	0.24	0.24
I-ALL-CR	0.50	0.96	0.07	0.06	0.06
I-ALL-NC	0.74	0.97	0.14	0.11	0.12
R-1-CR	0.47	0.96	0.04	0.02	0.03
R-1-NC	0.57	0.97	0.08	0.04	0.05
R-100-CR	0.48	0.96	0.03	0.01	0.02
R-100-NC	0.55	0.97	0.12	0.06	0.07
R-500-CR	0.49	0.95	0.07	0.04	0.04
R-500-NC	0.62	0.95	0.22	0.12	0.15
R-ALL-CR	0.56	0.95	0.06	0.04	0.04
R-ALL-NC	0.71	0.96	0.18	0.10	0.12

Table F.2: Summary of Text Detection

Evaluation ID	Number of Regions	Mean Performance
I-1-CR	12	0.42
I-1-NC	32	0.69
I-100-CR	57	0.28
I-100-NC	72	0.49
I-500-CR	49	0.45
I-500-NC	77	0.68
I-ALL-CR	54	0.41
I-ALL-NC	99	0.69
R-1-CR	28	0.00
R-1-NC	39	0.59
R-100-CR	90	0.10
R-100-NC	105	0.40
R-500-CR	91	0.22
R-500-NC	138	0.72
R-ALL-CR	118	0.47
R-ALL-NC	161	0.80

Table F.3: Summary of OCR

Evaluation ID	Model Performance	TP Rate	FN Rate	FP Rate
I-1-CR	0.02	0.01	0.99	0.04
I-1-NC	0.07	0.03	0.97	0.19
I-100-CR	0.06	0.06	0.94	0.10
I-100-NC	0.11	0.06	0.94	0.29
I-500-CR	0.14	0.11	0.89	0.11
I-500-NC	0.26	0.22	0.78	0.30
I-ALL-CR	0.11	0.09	0.91	0.13
I-ALL-NC	0.27	0.20	0.80	0.48
R-1-CR	0.00	0.00	1.00	0.00
R-1-NC	0.05	0.02	0.99	0.08
R-100-CR	0.03	0.01	0.99	0.03
R-100-NC	0.10	0.05	0.95	0.14
R-500-CR	0.07	0.01	0.99	0.08
R-500-NC	0.29	0.15	0.85	0.32
R-ALL-CR	0.15	0.07	0.93	0.19
R-ALL-NC	0.34	0.23	0.77	0.36

Table F.4: Summary of Runtime

Evaluation ID	Person Runtime	Bib Runtime	Text Runtime	OCR Runtime	Total Runtime
I-1-CR	1.08	8.99	0.13	8.08	18.27
I-1-NC	N/A	6.04	0.16	10.53	16.73
I-100-CR	0.95	10.49	0.19	11.42	23.05
I-100-NC	N/A	6.08	0.25	11.15	17.48
I-500-CR	0.94	10.05	0.15	9.94	21.07
I-500-NC	N/A	6.05	0.15	9.40	15.60
I-ALL-CR	0.95	10.43	0.18	10.29	21.85
I-ALL-NC	N/A	6.07	0.22	8.24	14.52
R-1-CR	0.90	15.30	0.19	13.62	30.01
R-1-NC	N/A	6.11	0.16	11.36	17.63
R-100-CR	0.96	16.75	0.30	22.27	40.28
R-100-NC	N/A	6.06	0.29	16.58	22.94
R-500-CR	0.96	15.81	0.27	17.82	34.86
R-500-NC	N/A	6.07	0.23	15.23	21.53
R-ALL-CR	0.96	15.85	0.26	16.95	34.02
R-ALL-NC	N/A	6.05	0.27	15.80	22.12

Table F.5: Summary of Manual Inspections

Evaluation ID	Bib Performance	Text Performance	Character Performance
R-ALL-NC	0.94	0.72	0.46
R-ALL-NC	0.97	0.69	0.45
R-ALL-CR	0.73	0.59	0.37
R-ALL-CR	0.77	0.52	0.35
I-ALL-NC	0.94	0.71	0.49
I-ALL-NC	1.00	0.93	0.66
I-ALL-CR	0.52	0.42	0.30
I-ALL-CR	0.66	0.48	0.34