Computer Vision and Machine Learning for Biomechanics Applications

Human Detection, Pose and Shape Estimation and Tracking in Unconstrained Environment From Uncalibrated Images, Videos and Depth

Ami Drory

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University

September 2017

Research School of Engineering
College of Engineering and Computer Science
The Australian National University

Final Copy – 26 September 2017

Final Copy – 26 September 2017

This PhD thesis has been conducted under the supervision of Prof. Richard Hartley and Prof. Hongdong Li at the Australian National University. Except where otherwise indicated, this thesis is my own original work. Most of the results in this thesis have been submitted to or published at international conferences and journals or are currently being prepared for publication. Some of these results have been achieved in collaboration with other researchers as indicated.

These results include:

Peer reviewed journal papers:

- Drory A., Li, H. & Hartley, R. (2017). A Learning-based Markerless Approach for Full-body Kinematics Estimation *In-Natura* from a Single Image. *Journal of Biomechanics* 55:1-10
- Drory A., Zhu, G., Li, H. & Hartley, R. (2017). Automated Detection and Tracking of Slalom Paddlers from Broadcast Image Sequences using Cascade Classifiers and Discriminative Correlation Filters. *Computer Vision and Image Understanding* 159:116-127.
- Drory A., Li, H. & Hartley, R. (2017). Estimating the Projected Frontal Surface Area of Cyclists from Images using a Variational Framework and Statistical Shape and Appearance Models. *Proc IMechE, Part P: J Sports Engineering and Technology*. 231(3):169-183
- Drory A, Yanagisawa M. (2012). Predictive Mathematical Model of Time Saved on Descents in Road Cycling Achieved Through Reduction in Aerodynamic Drag Area. *Proc IMechE, Part P: J Sports Engineering and Technology*. 226(2):152-160.

Peer reviewed conference papers:

- Drory, A., Li, H., Hartley, R. (2016). Markerless Sagittal Skeletal Kinematics Estimation from Uncalibrated Images Using Mixture of Parts Classification. 10^{th} Australasian Biomechanics Conference. Melbourne.
- Drory, A., Hartley, R., Li, H. (2014). Cyclist Detection in Images for Pose Estimation Using Cascade Deformable Part Based Model over Histogram of Oriented Gradients. Proceedings of the 12^{th} international symposium on Computer Methods in Biomechanics and Biomedical engineering. Amsterdam
- Drory, A., Hartley, R., Li, H. (2014). Statistical Shape Model of Cyclists in Level Set Formulation for Pose Estimation Towards Quantification of Aerodynamic Drag Area. Proceedings of the 12^{th} international symposium on Computer Methods in Biomechanics and Biomedical engineering. Amsterdam.
- Drory A, Yanagisawa M. (2011). A Mathematical Model For Estimation Of Time Saved In Road Cycling Descents Through Reduction In Aerodynamic Drag

Area Achieved Via Modifications To Static Rider Position. 8th Australasian Biomechanics Conference. Canberra

Book chapter:

- Drory A. (2012). Bike fit and Aerodynamics. In Hopker, J., Jobson, S. (Eds.). Performance Cycling: The Science of Success. London: Bloomsbury; p. 104-22.

Ami Drory
26 September 2017

I dedicate this work to Antje and Hugo without whose support this work would have never materialised. To Yitzchak Rosenberg, my late grandfather, a quiet and humble man whose personal qualities and integrity remain a beacon to behold.

To my supervisors, Richard Hartley and Hongdong Li, mostly for their patience, but also for their guidance, dedication and enthusiasm. To Fatih Porikli for his advice. To Anoop Cherian, Gao Zhu and Masahiro Yanagisawa for their assistance. To Alireza and Moshen for robust political debates.

To those who have come before me and whose seminal work has inspired mine and my interest in the biomechanical foundation of human motion. In particular, a tip of the akubra to Borelli, Muybridge, Hill, Van Ingen Scheneau, Bobert and Stacov.

To Shaul Ladany, whom I have never met, but whose life story of accomplishments and perseverance in the face of adversity through sport continue to inspired me.

Finally, to Winnie the Pooh. For everything.

> "It is not necessary that you leave the house. Remain at your table and listen. Do not even listen, only wait. Do not even wait, be wholly still and alone. The world will present itself to you for its unmasking, it can do no other, in ecstasy it will writhe at your feet."
>
> "It is, after all, not necessary to fly right into the middle of the sun, but it is necessary to crawl to a clean little spot on Earth where the sun sometimes shines and one can warm oneself a little."
>
> (Franz Kafka)

Final Copy – 26 September 2017

Acknowledgements

I wish to express my gratitude to the following collaborators, who contributed to the respective chapters and recognised as co-authors of the respective manuscripts;

Richard Hartley and Hongdong Li have contributed to all chapters. Anoop Cherian contributed to chapter 3. Masahiro Yanagisawa contributed to chapter 6. Gao Zhu contributed to chapter 7.

I am also grateful to the University of Washington's wind tunnel and the San Diego Wind tunnel for sharing some of their images.

Abstract

Motivation In Biomechanics, musculoskeletal models yield information that cannot be non-invasively obtained by direct measurement based on skeletal kinematics. Unsatisfactorily, obtaining accurate skeletal kinematics is limited to either user manual labelling or marker-based motion capture systems (MoCaps) that are limited by expansive infrastructure, environmental conditions, obtrusive markers causing movement impediment and occlusion errors. Moreover, they cannot yield surface geometry that is critical for many biomechanical applications. To advance the state of knowledge, real-time user-free acquisition of individualised pose and surface geometry is currently needed, and motivates our work on this thesis.

Aims The goals of this dissertation are; 1) to explore how advances in computer vision and machine learning algorithms can be levered to provide the necessary framework for *in-natura* acquisition of skeletal kinematics, 2) in a challenge to the traditional biomechanics modelling reliance on skeletal pose only, explore how computer vision algorithms can be used to develop shape recovery framework, 3) to demonstrate the potential of human detection, tracking, pose estimation and surface recovery techniques to address open problems in biomechanics.

Contributions We demonstrate skeletal pose estimation from monocular images in challenging environments under a discriminative pictorial structure framework. We extend the flexible part based approach to explicitly model human-object interaction. Our empirical performance results show that our proposed extension to the technique improves pose estimation. Further, we develop a hybrid framework for human detection and shape recovery using a discriminative deformable part based model for detection with a learnt shape and appearance model priors[1] for shape recovery from monocular images. We also develop a real time framework for simultaneous activity recognition, pose estimation and shape recovery using information from a structured light sensor. For a demonstrator application, we develop a theoretical model that uses the recovered shape to solve downstream open questions in biomechanics. Finally, we develop object detection and tracking in a particularly challenging environment from image sequences that include rapid shot and view transition using complementary trained discriminative classifiers. We apply our techniques to the human ambulatory modalities of cycling and kayaking because they are common in both the clinical and sports biomechanics settings, but are rarely studied because they present

[1] In Bayesian statistical inference, a *prior* refers to a probability distribution that represents the belief about an unknown quantity before some evidence is considered. A prior can be determined from past information, such as previous experiments. Consistent with the relevant literature, we use simply *prior* throughout the text.

unique challenges. Specifically, many applied problems relating to those modalities remain open due to absence of robust markerless motion capture that can recover skeletal kinematics and surface geometry *in-natura*.

Impact The developed methods can subsequently provide new insights into open applied problems, such as enhance the understanding of bluff bodies, specifically cycling, aerodynamics, and kayaking performance. More importantly, we believe that from a higher level standpoint, our full-body human shape modelling and surface recovery represents a significant paradigm shift in biomechanical modelling, which traditionally relies on skeletal pose only. The knowledge gained is intended to form the foundation for the development of evidence-based decision support tools for diagnosis and treatment through enhanced understanding of human motion. We envision that these methods will have a transformative effect on the field of biomechanics, analogously to the effect of medical imaging on the field of medicine.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background to the problem

1.1.1 Biomechanics

Human musculoskeletal modelling in biomechanics. Characterising the non-linear behaviour of human motion enhances the understanding of neuromuscular coordination patterns and dysfunction. Musculoskeletal computer models and simulations yield information that cannot be obtained non-invasively by direct measurement. Using inverse dynamics or dynamic optimisation, resultant compressive and shear tissue loads and muscle contributions to segment and joint accelerations can be estimated based on the measured kinetics, inertial properties and skeletal kinematics (fig. 1.1).

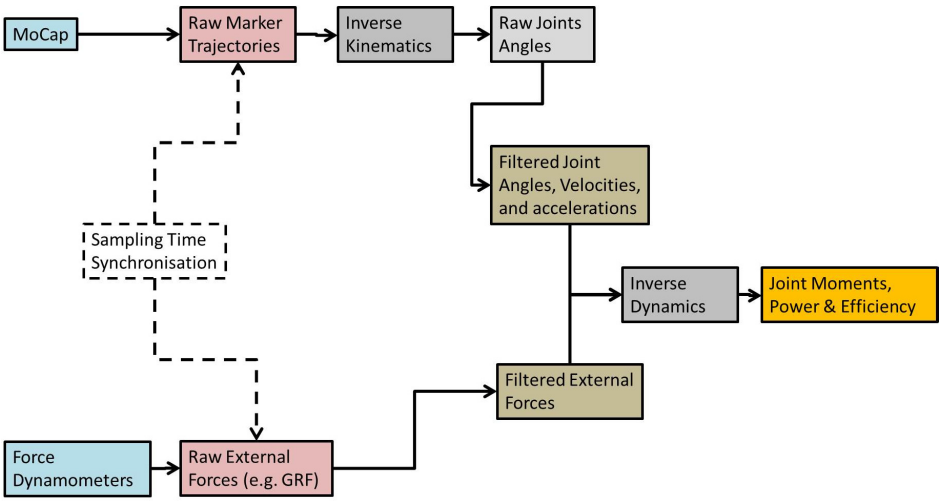
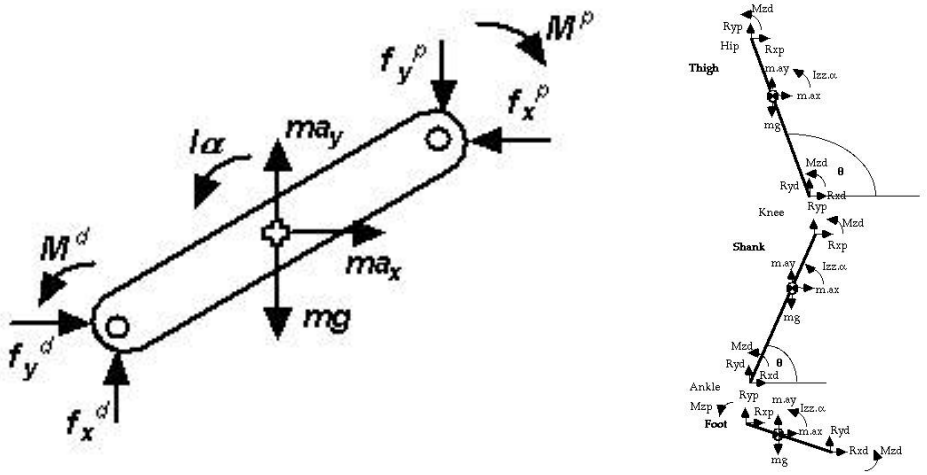


Figure 1.1: Overview of Inverse dynamics for biomechanics.

Inverse dynamics Inverse dynamics derives the minimum forces and moments responsible for motion by solving Newton-Euler equations of motion from measurement of skeletal kinematics and external forces under a set of assumptions [143](fig. 1.2).



(a) Single segment

(b) multi segment - lower leg

Figure 1.2: Inverse dynamics models for biomechanics. (source: clinicalgaitanalysis.com). For a single body segment (a), the shear and compressive forces and net muscle torque at the proximal joint $(\mathbf{f}_x^p, \mathbf{f}_y^p, \mathbf{M}^p)$ can be estimated from the estimated segment's inertial properties $(\mathbf{m}, \mathbf{I})$, measured external forces and moments $(\mathbf{f}_x^d, \mathbf{f}_y^d, \mathbf{M}^d)$ acting at the distal end, and measured kinematics $(\mathbf{a}_x, \mathbf{a}_y, \alpha)$. For a multi segment model (b), the approach can be extended to estimate the kinetics at each joint of the linkage system by solving a simultaneous system of equations.

The approach has a number of limitations including; frictional forces within joints, individual muscle, tendon, and ligament forces all remain unknown as are muscle co-contractions, forces are attenuated since the segments are not completely rigid, calculations are very sensitive to errors in external forces [210], centre of pressure measurements [169, 211], segment inertial properties, joint centre estimates [85, 204, 205, 224, 225], and segment accelerations [177].

Despite its limitations the approach has been successfully applied to the study of human gait [19, 205, 207]. In the clinical setting, it is extensively used for finding characteristic patterns to predict whether artificial joints or pathological conditions will have sufficient support (e.g. when walking or running), for estimating the compressive and shear loads in joints (e.g. loads on the spine during lifting and lower back pain), patient pre- and post-operative assessment of mobility, implant design, and neuromuscular disorder diagnosis and monitoring. In sport biomechanics the approach is used for technique optimisation, injury prevention and performance modelling.

1.1.2 Statement of the Problem

With respect to inverse dynamics for biomechanics, this dissertation is only concerned with development of markerless techniques for the acquisition of skeletal

kinematics from images and videos.

State-of-the-art Obtaining skeletal kinematics is currently limited mostly to marker-based Motion Capture (MoCap) systems. This is unsatisfactory because the approach is constrained by expansive laboratory infrastructure with camera array, control of lighting and environmental conditions, laborious and time consuming post processing, and the obtrusive use of markers requiring palpation. Inherent to the use of surface mounted markers are output errors caused by a critical reliance on a strong assumption of rigid linkage skeletal system and ignoring surface deformation [34, 40, 116]. To achieve a person-specific model, the parameters of a generic model are adjusted to an individual's anthropometry by inverse kinematics. Inverse kinematics solves a weighted least square problem to minimise the discrepancy between experimental and virtual markers [155]. In particular, the effect of Soft Tissue Artefacts (STA) causing movement impediment has received extensive attention in the literature [5, 30, 31, 56, 107, 144, 149, 170, 186, 195, 196], as has the precision of anatomical landmark determination [68, 85, 155, 212, 224, 225, 236, 239] (see Fig. 1.3). Consequently, the development of unconstrained and remote evidence-based decision support tools for diagnosis and treatment is inhibited. To advance the state of knowledge, real-time acquisition of skeletal kinematics is currently needed.

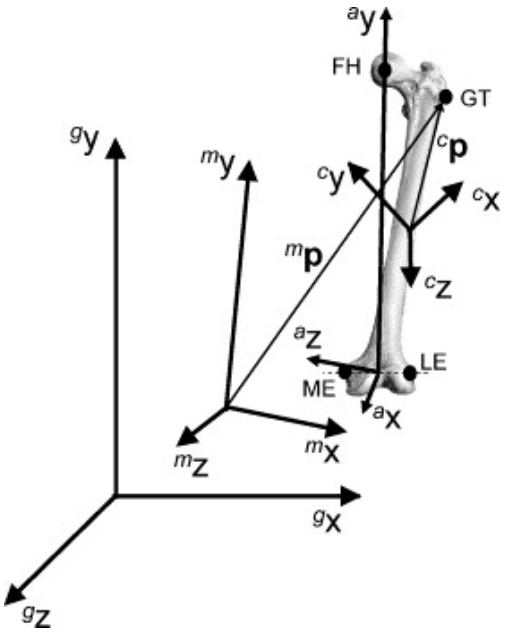


Figure 1.3: In traditional marker-based MoCap capture of kinematics, soft tissue artefact refers to the relative movement between a surface mounted cluster of markers ${}^{m}P = ({}^{m}X, {}^{m}Y, {}^{m}Z)$ and the underlying bone $({}^{c}X, {}^{c}Y, {}^{c}Z)$. Inertial effects, skin deformation and sliding, which occur mainly in areas closer to the joints, and deformation caused by muscle contractions, contribute independently to STA (source: [34]).

1.1.3 Alternate approaches to motion capture

The following introduction to alternate approaches is presented here to provide a broad context to the work of this thesis with respect to pose estimation. However, these approaches were not explored nor discussed further in detail in this dissertation.

1.1.3.1 Intertial-based MoCap

Alternate inertial-based MoCap approaches to motion capture use a set of small Inertial Measurement Units (IMUs) attached to the segments of the body to obtain skeletal kinematics. Compared to marker-based MoCap systems, these are unconstrained by the need for camera arrays, lighting conditions or space. Further, they are not affected by occlusions and can be used underneath clothing. Thus, they overcome some of the limitations of marker-based systems, and are suitable for use *in-natura*.

However, since these approaches require double integration of the sensors' acceleration signal, they suffer from low accuracy due to signal drift and inherent lack of fixed global reference frame [111, 158, 163, 208]. Furthermore, since the problem is under-constrained, a large number of segmental sensors are typically used [158], which remains unsatisfactorily intrusive.

Recent advances describe minimally intrusive solutions, which use a sparse set of IMUs on the distal body's extremities (wrists, lower legs and head) and the body's root at the pelvis [223, 232, 233]. The sparse IMUs only provide weak constraints, which is insufficient to recover the full pose. Tautges et al. [223] reconstruct the pose via retrieval of poses with similar segmental accelerations from a pre-recorded dataset of motions obtained from a marker-based MoCap. This solution suffers from the noisy acceleration signal. Further, the very large space of possible segmental acceleration makes learning very difficult. In contrast, von Marcard et al. [232, 233] use a learned statistical body model [154] to provide kinematic and anthropometric constraints, and fit the poses of the body model to the orientation and acceleration measurements over multiple frames. In the context of this dissertation however, this approach cannot be readily generelised to describe human-object interaction. Likewise, it cannot be extended to recover both pose estimate and body geometry.

1.1.3.2 Convolutional neural networks

A neural network consists of a very large number of densely interconnected processing nodes that are typically organized into layers. Each node might be connected to several nodes in the layer beneath it, from which it receives data, and several nodes in the layer above it, to which it sends data (feed forward). To each of its incoming connections, a node will assign a weight. Patterns are presented to the network via an input layer, which communicates to one or more 'hidden' layers where the processing is done via a system of weighted connections. A learning rule modifies the weights of the connections according to the input patterns that it is presented with. During training of a neural network, the weights and thresholds are continually adjusted

until training data with the same labels consistently yield similar outputs. Learning can be supervised with backpropagation, where in each cycle the network is presented with a new input pattern through a forward activation flow of outputs, and the backwards error propagation of weight adjustments. The neural network then makes an appropriate adjustment to its connection weight based on a discrepancy measure between its answer and the true value.

Convolutional Neural Networks (CNNs) are a category of neural networks that have proven very effective for image recognition and classification. CNNs can be incorporated into a pose machine framework [190] for learning image features and image-dependent spatial models for the task of human pose estimation. Convolutional Pose Machine [235], is a a sequential multi-layered modular network architecture composed of convolutional networks that directly operate on belief maps from previous stages. It uses message passing to predict a confidence of the location for each variable (body part), iteratively improving its estimates for part location at each stage, without the need for explicit graphical model-style inference. Compared to a graphical models, for which inference is difficult and inexact in all but the most simple models, such as a tree-structured or star-structured models, it incorporates richer interactions among multiple variables at a time. Furthermore, it models long-range dependencies between variables, and learns an expressive image-dependent spatial models of the relationships between parts directly from the data without the need for designing part-specific classifiers, or specifying the parametric form of the potential functions.

Recently, frameworks that use CNNs achieved state-of-the-art performance on standard human pose estimation benchmarks (MPII, LSP, and FLIC datasets), and outperform competing methods including the graphical model-based Deformable Parts Model (DPM) framework that we adopted in part I [101, 174, 235]. Generally, DPMs and CNNs are viewed as distinct approaches to pose estimation. However, Girshick et al. [101] elegantly showed that a DPM can be formulated as a CNN. Their approach involves mapping each DPM inference step to an equivalent CNN layer, and replaced the standard image features used in DPM with a learned feature extractor.

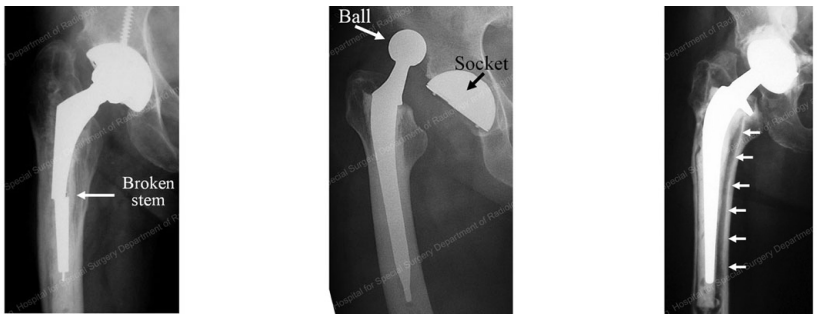
Notwithstanding, we opt to use the classical DPM approach here for three main reasons; 1) the CNN approach remains largely a 'black-box' approach, whilst the the spatial relations of a DPM is intuitive to the applied practitioner, 2) the CNN approach requires very large datasets of training samples, and remains prohibitively time consuming in terms of obtaining very large amount of annotated data that is necessary to achieve well-trained models, and 3) the CNN approach cannot readily handle the aggregation of multiple scenarios, such as in the case for human-object interaction.

1.1.4 Beyond inverse dynamics - recovery of both skeletal pose and body geometry

While the inverse dynamics approach is standard practice that has served the biomechanics community well for the study of human movement for several decades since Borrelli in the 17th century, we argue that the skeletal linkage system at its model's core (fig. 1.2), is an over-simplification that fails to explain human motion in many important cases. Most importantly, the approach suffers from high sensitivity to errors in skeletal kinematics estimation [36, 104, 206], the consequences of which can be catastrophic (Fig. 1.4).

Moreover, previous approaches to determination of the individualised segmental inertial properties that are required for an inverse dynamics approach [193, 238], used *static* medical scanning techniques (MRI, DEXA and CT) [12, 44, 84, 166, 185, 198, 217]. These are constrained by expensive infrastructure, radiation exposure and lengthy exposure time that can only yield static geometry. Alternate generic approaches used regression models over cadaver information that is lacking gender, age and racial diversity and unrepresentative of the wider population [61, 257], or indirectly using the immersion in water method under the assumption of homogenous tissue density [74]. Evidently, a method that allows dynamic estimation of segment volume from which segment tissue composition and subsequent inertial properties can be determined, is currently needed.

Critically, current marker based MoCaps approaches use a rigid marker triad cluster to represent a body segment's surface from which the underlying anatomical structure is inferred under the strong rigidity and surface non-deformity assumption (see Fig. 1.3). This strong assumption contributes significantly to STA and subsequent estimation errors. In recovering the body's surface geometry directly, this assumption can be relaxed. Subsequently, surface deformation caused by muscle contraction and inertial effects can be modelled and understood.



(a) broken stem (b) dislocation (c) looseness

Figure 1.4: Hip prosthesis failures. (source: [173]). Failure to accurately predict joint loads and surface geometry may result in inferior implant design and insertion with severe consequences.

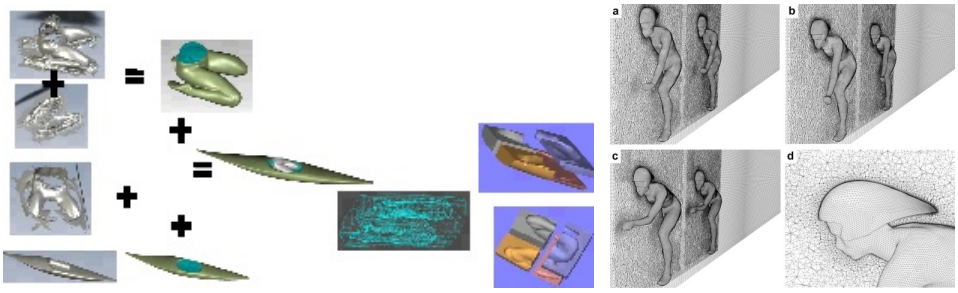
The dimensionality-reduced modelled system is appropriate when the system is

subjected to external forces that can be realistically approximated as acting at a point on the system. Alas, this cannot be justified when the external forces are acting on the entire object's surface. Simultaneous reconstruction of skeletal pose and body shape and geometry is paramount for the understanding of common human ambulatory modalities and many applications (e.g. orthotic and prosthetic design in Biomedical Engineering, or bluff bodies flow simulation in fluid dynamics). Importantly, existing commercially available MoCap systems are limited to recovered skeletal pose and are not capable of providing 3D surface and geometry information required for important applications. Alternate scan and meshing techniques are limited to recovery of *static* shape only, which is inadequate for dynamic human actions ([219, 237]), and typically cannot provide real-time solution. With respect to shape models for biomechanics, this dissertation is concerned with development of techniques for the modelling and extraction of human geometry recovered from images and videos.

Motivating example 1: Orthotic and prosthetic systems Current orthotic and prosthetic (O&P) systems are passive devices whose characteristics are set to maximize mobility during level walking. However, O&P users perform a large variety of ambulatory tasks throughout the course of daily living, including stair climbing, ramp walking and cycling. In order to maximize user's mobility, O&P systems should adapt to the tasks the user performs. This requires accurate characterisation of pose and surface deformation during a range of activities. Current approaches to O&P design are limited to computer simulation of bone remodelling and material properties response models performed on generic bone models coupled with skeletal pose obtained from MoCap systems. This is unsatisfactory because skeleton pose cannot yield an accurate 3D surface model that is required for optimal design of O&Ps. Alternate CAD based approach is used to bridge static scan data with simulation. This methodology is problematic due to a need to mesh more than one domain (see Fig. 1.5a). Hence, investigations are limited to either cadaver or isolated single limb or organ studies.

Motivating example 2: bluff bodies aerodynamics Similarly to O&P design, in the domain of sport performance the need for accurate 3D surface and geometry information of highly dynamic human motion is manifested in research efforts to understand flow characteristics around dynamic bluff bodies, for instance in the study of aerodynamic drag in cycling. Detailed review of related work is provided in chapter 5 as appeared in the corresponding manuscript (fig. 1.5b).

Dissertation focus This dissertation focuses on recovery of shape and pose characteristics of cyclists from images and videos for the study of bluff bodies aerodynamics, and the *in-natura* detection and tracking of canoe/kayak paddlers from videos for the study of obstacle negotiation. Essentially, the problems are formulated as an inference process that infers human pose and motion from observations. Observations in the context of this work refers to either the image data or features extracted



(a) Individualised human-instrument interface (b) Cyclist models for CFD

Figure 1.5: Motivating applications in sport biomechanics that can benefit from reconstruction of surface geometry; a) Illustration of an optimal individualised human-instrument interface design applied to the paddler-canoe interface. This project is problematic due to a need to mesh more than one domain. This project does not form part of this dissertation. b) Computational Fluid Dynamics (CFD) studies of aerodynamics drag in cycling is currently limited to static scanned cyclist's models (source: [17]).

from the image.

Human motion can be naturally described as a joint spatial and temporal representation of the human body. We simplify human motion to describe a sequence of poses in time, modelled by an articulated skeleton determined by a pose vector comprised of a position, an orientation and joint angles.

1.1.5 Challenges

3D Reconstruction remains a fundamental problem in computer vision. Robust solutions for rigid structures from motion have been achieved [261] due to the rigidity assumption [246] because of known multi-view geometric relations [115]. Yet, state-of-the-art techniques offer no similar solutions for dynamically deforming non-rigid objects. The solution resides in a very high dimensional parameter space, compounded by the requirement to estimate surface coordinates at each pixel or voxel. Tracking and matching algorithms are further complicated by uniform colour and texture [162]. The articulated body motion also induces severe occlusions, which present critical challenges [95, 148]. Most solutions require restrictive use of calibrated cameras arrays [123] and controlled conditions [98].

For the reconstruction of surface geometry, shape-from-silhouette has been achieved using local convex approximation of the human's surface geometry in a calibrated camera array setting by intersecting generalised cones via back projection of the object's multi-view silhouette and camera parameters [51–53]. This requires considerable control over lighting and environmental conditions, and unsuitable for estimation of kinematics in realistic natural environments. Moreover, the visual hull approach tends to overestimate the volume of the subject and fails to reconstruct cavities in the subject's surface. Critically, because the pose estimation is an ill-

conditioned problem, the visual hull does not uniquely determine a single pose. Rather, the under-determined system may result in a silhouette that corresponds to multiple postures. In the context of this work, the use of structured lighting from multimodal images (see Chapter 4) provides additional constraints to reduce the pose search space and assist in making the pose unambiguous.

1.1.6 Applications

The applications serve as demonstrators to the exploitation of advances in computer vision and machine learning techniques to investigate open problems in biomechanics. To advance the current state of knowledge on these problems, acquisition of individualised real time surface geometry and motion information in natural environment is needed. Many human activities involve interaction with objects in their environment. Therefore, an additional underlying theme of the thesis is concerned with the sub-problem of the interaction between human and objects as a cue for improved pose estimation, detection and tracking, with the view that the kinematics of the interactions will form an important component for studying the human motion itself.

1.2 Impact

Finding sustainable solutions for health management presents a major and persisting challenge to society and policy makers. In coupling technological opportunities with maximising health impacts, the research objectives will create opportunities for unimpeded patient *in-natura* monitoring of postural control, ambulatory activities and assisted living. In particular, markerless MoCap will enable remote networked patient assessment and monitoring.

1.3 Thesis Outline

This dissertation addresses the problem of human skeletal pose estimation and shape recovery introduced in 1.1 through the development of alternate frameworks that utilise computer vision and machine learning techniques. These are further discussed in detail in the corresponding chapters. The dissertation is divided into three parts.

Markerless human detection and pose estimation Part I uses a supervised learning approach to learn an activity-specific discriminative parts-based models for detection and pose estimation based on 2D pictorial structure framework. In this setup, the human is an object category whose instances need to be detected in an image. The framework described in chapter 2 returns detections in the form of skeletal pose. Chapter 3 models human-object interactions as a special case of a more general framework that models object-object interactions. It proposes a principled way of modelling spatial relationships between objects to facilitate their respective detection

and pose estimation in images. For example, people tend to sit on bicycles with feet on the pedals and hand on the handlebars. In contrast to previous approaches, the spatial relationships are not learnt from training data and are used in the energy function to penalise inappropriate detection proposals. We show empirical results demonstrating this framework outperforms alternative approaches. Demonstrating the connection between the geometry of an interacting object on the one hand and the appearance of its individual parts on the other, is one of the key insights of this thesis. This chapter demonstrates the benefit of modelling activity-specific appearances for simultaneous detection and pose estimation. The ideas described in chapter 2 were submitted for publication. A generalisation of chapter 3 is currently being prepared for submission.

Markerless recovery of surface geometry Whereas the part I focuses on skeletal pose estimation as a necessary pre-requisite to inverse dynamics (1.1.2), part II presents two alternate approaches for the recovery of human *geometrical shape*. The approaches are applied to cycling as a demonstrator application that requires the human shape for the study of the movement modality as described in section 1.1.4 . In contrast to part I, which advocates for modelling the human-object interactions, both approaches ignore the inter-relations between the human and object. Chapter 4 takes a supervised learning approach to learn an activity-specific model for activity recognition based on skeletal kinematic features reconstructed from depth information, and simultaneously recovers the human surface geometry. The developed framework outputs estimated pose and recovered shape in real-time. The ideas described in this chapter are currently being prepared for publication.

Chapter 5 takes a hybrid approach, which uses a pictorial structures discriminative approach to human detection and a generative approach to shape recovery using an active contour framework regularized by a learnt shape and appearance models prior. The detection framework was first published in Drory et al. [78]. The shape model framework was first published in Drory et al. [79]. Finally, the joint framework including the active contour component has been submitted for publication.

To demonstrate the potential downstream practical use of the recovered pose and geometry, chapter 6 provides theoretical foundation to the study of the relationship between recovered human geometry and human motion using a motivating example of cycling performance and aerodynamics. To improve our understanding of this human ambulatory modality, both skeletal pose and recovered shape are necessary. The position of a cyclist on a bicycle is key to optimal performance and injury prevention in both optimising power delivery and minimising aerodynamics resistance. While skeletal pose is required to estimate the former, modelling the flow characteristics around a bluff-body such as a cyclist, requires dynamically recovered shape. This part demonstrates how levering the techniques developed in this dissertation can provide information not previously available to answer research questions in allied fields. The ideas described in this chapter were first published in Drory [75], Drory and Yanagisawa [82, 83].

Markerless Human Detection and Tracking Whereas previous parts propose techniques for pose estimation and shape recovery in a common movement modality including the interaction with objects, part III investigates techniques for understanding human motion in a less common modality (slalom kayaking), where the task is detection and spatial tracking of the human-object complex in a challenging natural environment.

Related Work associated with each of these sub-problems has been incorporated inside the corresponding chapter that addresses it.

Part I

Detection and Pose Estimation from Monocular Images Using Flexible Mixture of Parts

Chapter 2

A Learning-based Markerless Approach for Full-body Kinematics Estimation *In-Natura* from a Single Image

Abstract

We present a supervised machine learning approach for markerless estimation of human full-body kinematics for a cyclist from an unconstrained color image. This approach is motivated by the limitations of existing marker-based approaches restricted by infrastructure, environmental conditions, and obtrusive markers. By using a discriminatively learned mixture-of-parts model, we construct a probabilistic tree representation to model the configuration and appearance of human body joints. During the learning stage, a Structured Support Vector Machine (SSVM) learns body parts appearance and spatial relations. In the testing stage, the learned models are employed to recover body pose via searching in a test image over a pyramid structure. We focus on the movement modality of cycling to demonstrate the efficacy of our approach. *In natura* estimation of cycling kinematics using images is challenging because of human interaction with a bicycle causing frequent occlusions. We make no assumptions in relation to the kinematic constraints of the model, nor the appearance of the scene. Our technique finds multiple quality hypotheses for the pose. We evaluate the precision of our method on two new datasets using loss functions. Our method achieves a score of 91.1 and 69.3 on mean Probability of Correct Keypoint (PCK) measure and 88.7 and 66.1 on the Average Precision of Keypoints (APK) measure for the frontal and sagittal datasets respectively. We conclude that our method opens new vistas to robust user-interaction free estimation of full body kinematics, a prerequisite to motion analysis.

2.1 Introduction

2.1.1 Motivation - Deficiencies of Marker Based Mocap

Characterizing the non-linear behaviour of human motion enhances the understanding of neuromuscular coordination patterns and dysfunction. Using inverse dynamics or dynamic optimisation, resultant compressive and shear loads and muscle contributions to segment and joint accelerations can be estimated based on the measured kinetics, inertial properties and skeletal kinematics. Obtaining skeletal kinematics is currently limited mostly to marker-based motion capture systems. This is unsatisfactory because the approach is constrained by expansive laboratory infrastructure with camera array, control of lighting and environmental conditions, and the obtrusive use of markers requiring palpation. Inherent to the use of surface mounted markers are output errors caused by a critical reliance on a strong assumption of rigid linkage skeletal system and ignoring surface deformation ([34, 40, 116]). In particular, the effect of Soft Tissue Artefacts (STA) causing movement impediment has received extensive attention in the literature ([5, 30, 31, 56, 107, 144, 149, 170, 186, 195, 196]), as has the precision of anatomical landmark determination ([68, 85, 156, 224, 225]). Consequently, the development of evidence-based decision support tools for diagnosis and treatment is inhibited. Hence, the development of a markerless solution for acquisition of full body kinematics has attracted significant research efforts.

2.1.2 Previous Work

2.1.2.1 Kinematics Estimation from Images

Estimation of the full body human kinematics from monocular images remains an open problem. The difficulties stem from background clutter, scene illumination and the weak local appearance support, which is further hindered by out-of-plane motion and severe occlusions caused by the motion of the articulated body ([109]). Since 2D intensity images remain the most readily obtainable for capture of unrestricted motion *in-natura*, feature tracking via direct manual digitization has formed the most common form of analysis. Krosshaug and Bahr [136] reconstructed motion kinematics from uncalibrated images using manual annotation of anatomical landmark locations that was matched across camera views and applied to a subject-specific scaled anatomical model with joint constraints. Likewise, Sanders et al. [202] have shown high repeatability of manual 3D marker trajectories digitised from multi view swimming images. Magalhaes et al. [161] attempted to automatically track surface mounted markers underwater using optical flow with limited success. Using textured clothing to replace surface mounted markers approach Lerasle et al. [146] tracked low level image features of a cycling leg using a Kalman filter. Similarly, Sandau et al. [201] used a texture enhanced clothing aided by background subtraction to achieve point correspondences for surface reconstruction in a calibrated multi-view camera setup. They fitted an articulated model to the 3D surface reconstruction using a patch matching technique, which enforces local photometric consistency and

global visibility constraints.

2.1.2.2 Computer Vision and Machine Learning Approaches

In generative approaches, pose estimation is formulated as an optimisation problem whose objective function is a discrepancy between a parametric prior body model and the input observation ([8, 90, 200] (for review, see [247])). This approach, however, suffers from local minima and solution multiplicity due to its often highly non-convex nature. For instance, Corazza et al. [53] fitted prior articulated model to a 3D surface visual hull reconstruction using patch matching with high accuracy. They used body part segmentation and least-squares optimisation to identify the location of joint centres under the assumption of rigid links connected by pivot joints ([52]) and to estimate the centre of mass ([51]). The same method was modified to use adaptive Gaussian mixture models to enhance background subtraction for the pose estimation in a water environment ([37]). Notably, the visual hull approach tends to overestimate the volume of the subject and fails to reconstruct cavities in the subject's surface. Whilst less obtrusive than marker-based methods, the method critically relies on background subtraction and a constrained capture space. This requires considerable control over lighting and environmental conditions, and remains unsuitable for estimation of kinematics in realistic natural environments.

In contrast, discriminative approaches seek a mapping from image observation space to a set of body pose parameters space, from which the kinematics can be estimated ([1]). The pictorial structures framework uses a probabilistic graph model to model the appearance and configuration of body parts. Pose estimation can then be formulated as a statistical inference problem, where the model parameters are learned from training examples using maximum likelihood estimation ([92]). This powerful framework allows for efficient inference and captures large variations in posture and appearance. The inter-part relative deformation term makes this framework invariant to some global transformation. Additionally, the overall decision is made with no assumptions being made about the initial location of parts. For these reasons, the approach has been popular for simultaneous human detection and pose estimation tasks ([7, 45, 87, 187, 220, 249]).

2.1.2.3 Deformable Part-Based Methods

Variants of the approach have been proven to outperform single object templates in detecting humans in images. In Felzenszwalb and Huttenlocher [92] a discriminatively trained, multiscale Deformable Parts Model (DPM) approach is introduced for pedestrian detection. The DPM model consists of a coarse root filter, a mixture of body parts filters, and part deformation relative to the root model to represent a person. The models are trained offline on a positive and negative image set using Support Vector Machines (SVM). In inference, the learned model is used for object search in a new image over a pyramid of image features, for instance, an appearance representation based on Histogram of Oriented Gradients (HOG) features ([57]). An

object proposal is calculated from a unary data term representing the scores of each appearance filter at their respective locations and a deformation cost that depends on the position of each part with respect to the root.

Recently, approaches that use Convolutional Neural Networks (CNNs) have outperformed pictorial structures in pose estimation tasks ([42, 47]). However, CNNs require prohibitively large datasets for training, or risk overfitting a model to the data. Consequently, the approach also requires extensive computing resources and training time. Furthermore, due to its intractable nature, a CNN remains largely a 'black box' approach, which provides little insight or intuition to its performance. These limitations justify our decision to adopt the pictorial structures framework.

2.2 Method

2.2.1 Problem Scope and Contributions

Motivated by the limitations of existing approaches, we address in this paper the problem of estimating full-body kinematics from challenging monocular images that contain severe occlusions in unconstrained environments. We opt for a discriminative part-based approach that requires an offline learning of a model that recovers pose estimates from observable image metrics. To demonstrate the efficacy of our approach, we focus our experiments on the movement modality of cycling. Our motivation stems from the observation that this movement modality is especially challenging due to the human interaction with an object (i.e the bicycle), which induces severe occlusions, the similarity of the posture in the frontal plane to normal human gait, and the severely occluded sagittal plane posture, for which a pose estimation method was not found in the literature. We use images captured in natural environment and a variety of resolutions. Importantly, We make no assumptions about the anthropometric proportions nor the kinematic constraints of the human model, nor the appearance of the scene. Our technique finds multiple good hypotheses for the human posture rather than just a single best solution. This is advantageous for cases where imprecision in the model may result in the desired match not being the one with the minimum energy.

2.2.2 Method Overview

In this section we introduce our framework for the estimation of a cyclist's posture from unconstrained images. Given a monocular image with one or more cyclists, we aim to simultaneously detect and estimate the cyclists' posture characterised by the joints' spatial locations and limbs' orientations in the image. Our method learns disparate appearance and geometry models of a cyclist offline, and estimates the human posture in a new image. Specifically, our work builds on the deformable mixture of parts framework of Yang and Ramanan [249] and Desai and Ramanan [69], who used local part *mixtures* that capture spatial relations between parts and

local appearance. We provide a diagrammatic overview of our learning and inference frameworks in figures 2.1 and 2.2 respectively.

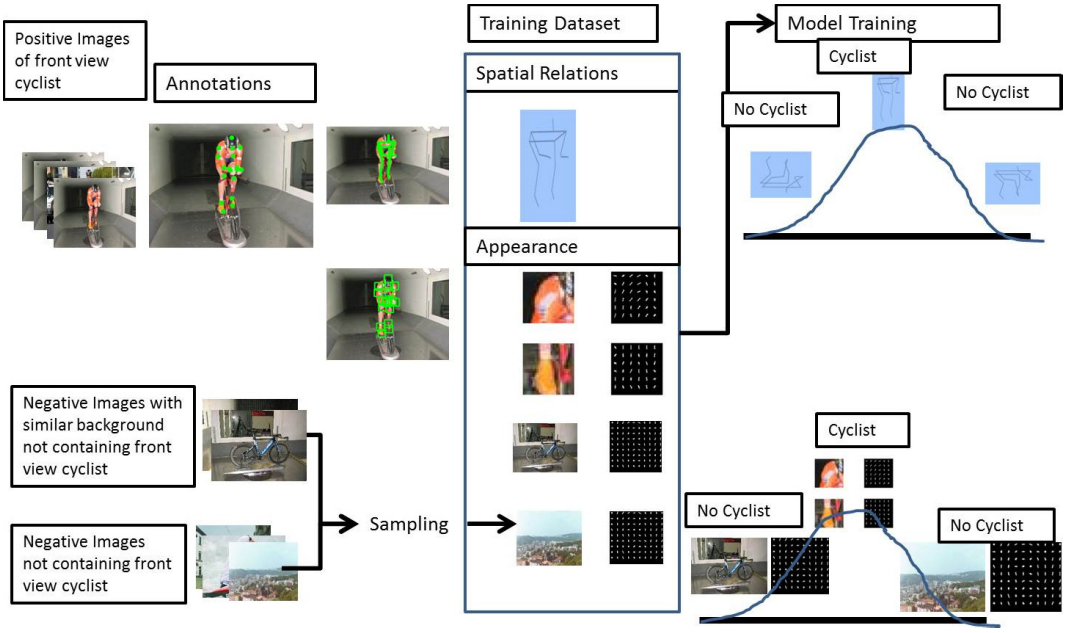


Figure 2.1: An overall learning framework of the flexible mixture of parts approach applied to a cycling model. Appearance and spatial relations features are extracted from both positive (object is present) and negative (object is not present) images to train an object model.

2.2.3 Mixture of Parts Human Model

We model the human body as a collection of the body's articulations (joints) whose spatial location is represented as a point in the 2D plane, and local appearance filters. We model the articulations as ball-and-socket joints expressed in Joint Coordinate System (JCS) following Wu et al. [242, 243]. We express a human model as a tree-structured undirected graph $G = (V, E)$, where the vertices $V = \{v_1, \ldots, v_n\}$ correspond to n body joints, and an edge $(v_i, v_j) \in E$ for each pair of connected body joints v_i and v_j corresponding to the body's segments (Fig. 2.3). An instance of a full body in the image I is given by a configuration of body parts $\boldsymbol{L} = \{\mathbf{l}_1, \ldots, \mathbf{l}_n\} \in \mathbf{R}^{n \times 2}$, where $\mathbf{l}_i = (x_i, y_i)$ denotes the location of part v_i.

Appearance Model We represent the appearance of body joint v_i by a concatenated HOG ([57]) feature vector $\phi(I, \mathbf{l}_i) \in \mathbb{R}^{5 \times 5 \times 32}$. HOG is an edge orientation histogram based feature descriptor, which we compute on a dense grid of uniformly spaced cells over an image patch of size 32×32 pixels centred at $\mathbf{l}_i$. In training, we learn a full body appearance model $\boldsymbol{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ where $\mathbf{w}_i \in \mathbb{R}^{800}$ is the template

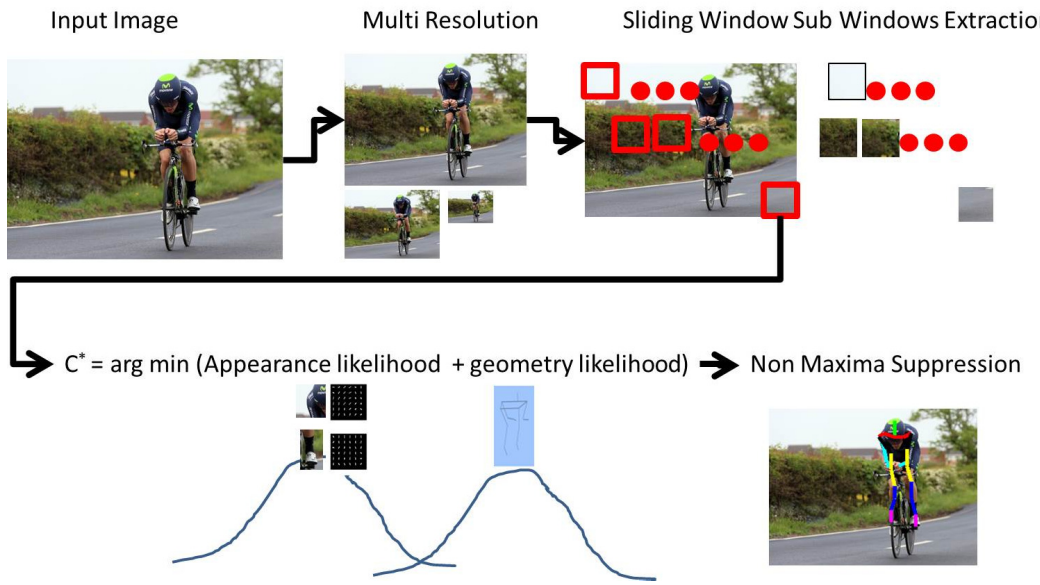
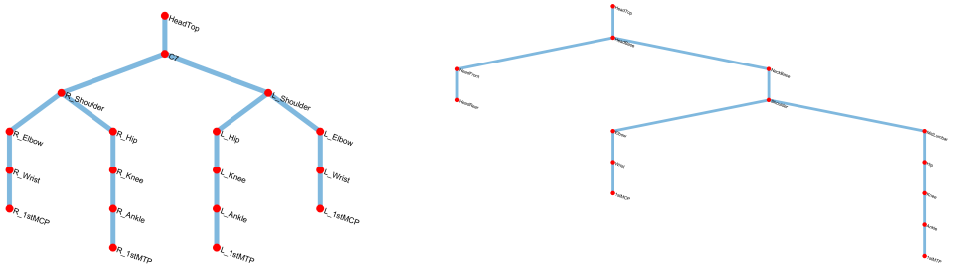


Figure 2.2: An overall inference framework of the flexible mixture of parts approach. Sub windows are extracted from a new image at multiple resolutions using a sliding window scheme. A probabilistic object hypothesis is tested against the appearance and geometry models for each sub window resulting in a likelihood score. Finally, a non maxima suppression is applied to overlapping object proposals.



(a) (b)

Figure 2.3: Tree structured graph models of a cyclist in frontal (a) and sagittal (b) views. Note that for the frontal view the model is an acyclical approximation of a natural representation that links the two hip nodes with a pelvic edge. The approximation simplifies the graph model and enables exact inference.

feature vector for body joint v_i (see a part visualisation in Fig. 2.5 and a full body visualisation in Fig. 2.4a). To arrive at the final feature length of 800, an image patch of size 32×32 pixels is divided into 5×5 overlapping cells, such that cell (1,1) contains pixels (1:16,1:16), cell (1,2) contains pixels (1:16,5:20), and so forth. Each

cell is comprised of 2×2 blocks of 8×8 pixels each. Each block is encoded by a histogram with 8 bins representing edge orientation in 45 degree step size.

(a) (b) (c)

Figure 2.4: a) A visualisation of the learnt full body frontal cyclist model, where parts' appearance models characterised by HOG ([57]) filters and their relative spatial location relative to a root filter. Since our model also learns co-occurrence probabilities for part mixture components, the displayed model shows one example of such co-occurrence. In inference, a feature representation of a new image (b) is computed at multiple image resolutions. c) shows an example feature representation of the image at one resolution level. The feature vector is then convolved with each of the learnt appearance model parts' filters (a) to yield a local response.

Spatial Relations Model To encode the spatial relations between adjacent joints, we represent the spatial relations for an edge (v_i, v_j) by a quadratic deformation vector $\psi(\mathbf{l}_i, \mathbf{l}_j) = [dx, dy, dx^2, dy^2]^T$ from the relative position of the connected joints v_i and v_j. This term is often interpreted as a negative spring energy resulting from pulling body part j from a relative position with respect to body part i ([92])[1]. In training, we learn the spatial relations model $\mathbf{w}_{ij} \in \mathbb{R}^4$ for each edge (v_i, v_j). This parameter can be viewed as indicating the spring's position at equilibrium and rigidity. It also encodes implicit relations to distal parts through connected edges.

Thus, a score S associated with a particular configuration of body parts in an image I is a function of the parts' appearance and deformation and can be written as

$$S(I, \boldsymbol{L}) = \sum_{i=1}^{n} m_i(I, \mathbf{l}_i) + \sum_{(v_i, v_j) \in E} d_{ij}(\mathbf{l}_i, \mathbf{l}_j) \tag{2.1}$$

where, $m_i(I, \mathbf{l}_i) = \langle \mathbf{w}_i, \phi(I, \mathbf{l}_i) \rangle$ is a unary scalar term measuring the appearance discrepancy for each part v_i at location $\mathbf{l}_i$ in the image I with the local template $\mathbf{w}_i \in \mathbb{R}^{800}$, and is based on convolving the image I with a family of underlying linear local templates $\boldsymbol{W}$, and $\langle \cdot, \cdot \rangle$ is the inner product operator. Similarly, $d_{ij}(\mathbf{l}_i, \mathbf{l}_j) = \langle \mathbf{w}_{ij}, \psi(\mathbf{l}_i, \mathbf{l}_j) \rangle$ is a pairwise scalar term measuring the deformation cost for a given pair of connected parts, that is of part v_i at location $\mathbf{l}_i$ and part v_j at location $\mathbf{l}_j$. A low negative d_{ij} score indicates that a body part's location and orientation with respect to its parent (proximal body segment) is close to the learnt prior spatial relations model.

[1] A.1 provides further interpretation and generalisation of the parts deformation cost.

For clarity, the terms are summarised as follows:

- $\mathbf{w}_i \in \mathbb{R}^{800}$ is the learnt HOG feature appearance template for joint v_i
- $\phi(I, \mathbf{l}_i) \in \mathbb{R}^{800}$ is the concatenated HOG feature descriptor of a 32×32 pixels sized patch of image I centred at $\mathbf{l}_i$
- $m_i(I, \mathbf{l}_i) = \langle \mathbf{w}_i, \phi(I, \mathbf{l}_i) \rangle$ is a scalar measuring how well the feature at the image patch centred at $\mathbf{l}_i$ matches the template of v_i
- $\mathbf{w}_{ij} \in \mathbb{R}^4$ is the learnt spatial deformation model of joint v_j with respect to its parent v_i
- $\psi(\mathbf{l}_i, \mathbf{l}_j) \in \mathbb{R}^4$ is the spatial deformation between two points $\mathbf{l}_i$ and $\mathbf{l}_j$ in the image I
- $d_{ij}(\mathbf{l}_i, \mathbf{l}_j) = \langle \mathbf{w}_{ij}, \psi(\mathbf{l}_i, \mathbf{l}_j) \rangle$ is a scalar measuring how well the deformation between two points in the image match the learnt deformation model for joints v_u and v_j

Mixture of Parts Notwithstanding the activity-specific application, the appearance of body parts is highly variable. For instance, an appearance patch for the first metatarsophalangeal (1$^{\text{st}}$MTP) joint looks different at the top-dead-centre of the cycling stroke than at the bottom-dead-centre. Likewise, the helmet and the hand position varies between a road cyclist, a sprint track cyclist and a mountain bike rider. Therefore, to encode a richer family of appearances for each body part, we model the appearance of each body joint $v_i$by a mixture of templates, instead of a single fixed appearance template. We write $t^i \in \{1, \ldots, T\}$ for a latent variable denoting an appearance mixture component for part v_i, and model the appearance of each body *joint* v_i by a mixture $\mathbf{\Phi_i} = \{\phi_{i1}, \ldots, \phi_{iT}\} \in \mathbb{R}^{800 \times T}$, where $\phi_{it}(I, \mathbf{l}_i)$ is a feature vector component t^i centred at $\mathbf{l}_i$ following Desai and Ramanan [69]. An assignment of mixtures for a full body model can then be denoted by $\mathbf{t} = \{t_1, \ldots, t_n\}$. In training, we learn a unary term $b_i(t^i)$that supports a particular mixture component assignment for the body joint v_i (see a visualisation in Fig. 2.5).

Co-occurrence Model Our intuition is that support for a part's particular mixture component assignment depends somewhat on the full-body pose. That is, two joints connected by a rigid limb are likely to present consistent pairing of appearance representations based on the limb's global orientation, for example the elbow and shoulder joints connected by the upper arm when elevated versus externally rotated. To capture the dependency of global pose on local appearance variations, we also learn a pairwise term $b_{ij}(t^i, t^j)$ that supports a particular mixture components co-assignment for the body joints v_i and v_j.

We write $\mathbf{z}_i = (\mathbf{l}_i, t^i) \in \mathbb{R}^3$ for the pixel location and mixture component for part v_i. We can then write the full score S associated with a particular configuration of

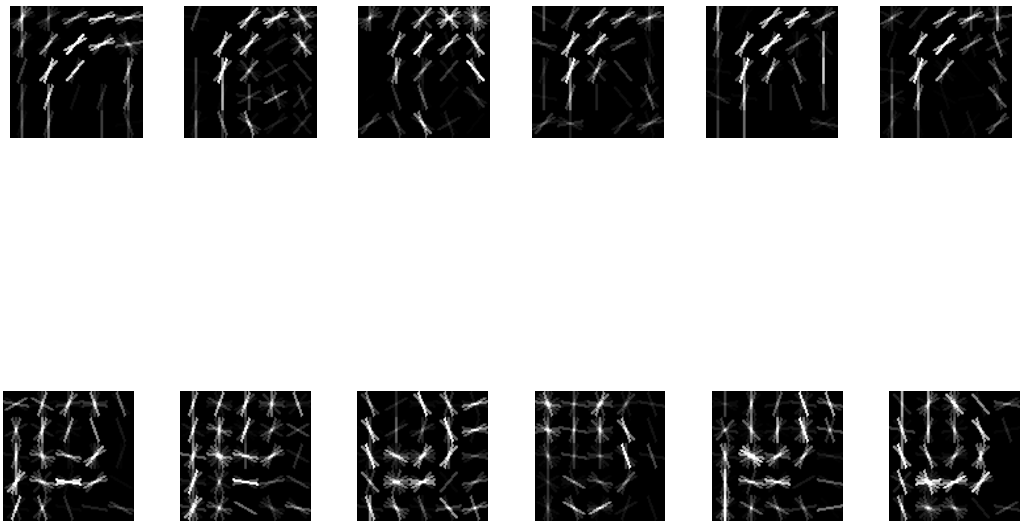


Figure 2.5: Each joint is modelled by 6 possible templates learnt during the training phase and represent alternative HOG filter appearance representations of an image patch of size 32×32 pixels centred at the joint. The filters capture variance in part appearance due to small changes in view point, background, shape, illumination and texture (e.g. changes induced by the pedalling cycle). Above is a visualisation of the filter mixtures for the right shoulder (top) and left 1stMTP joints (bottom) in the frontal view. Predictably, in cycling the appearance variability at the foot is higher than at the shoulder. A particular full body model has one mixture component for each joint, for instance the second component for the shoulder and the fifth component for the 1stMTP.

body parts in an image I as

$$S(I, \mathbf{z}) = \sum_{i=1}^{n} m_i(I, \mathbf{z}_i) + \sum_{(v_i, v_j) \in E} d_{ij}(\mathbf{z}_i, \mathbf{z}_j) \tag{2.2}$$

where

$$\boldsymbol{M}_i(I, \mathbf{z}_i) = \langle \mathbf{w}_i(t^i), \phi(I, \mathbf{l}_i) \rangle + b_i(t^i)$$
$$d_{ij}(\mathbf{z}_i, \mathbf{z}_j) = \langle \mathbf{w}_{ij}(t^i, t^j), \psi(\mathbf{l}_i, \mathbf{l}_j) \rangle + b_{ij}(t^i, t^j)$$

The pairwise term $b_{ij}(t^i, t^j)$ favours consistent co-occurence of mixture components for the corresponding parts v_i and v_j, such that a positive $b_{ij}(t^i, t^j)$ score reflects consistent pose assignments, and a negative score reflects the alternative. Thus, an

instance of a human in the image I indicates which mixture component from each body joint is used and its relative location.

2.2.4 Inference

Our goal is to detect and estimate the posture of a cyclist from test images. In inference, we produce candidate pose proposals by using a sliding window detection scheme over an image pyramid. The optimal match of a model to an image is found by maximising (2.2) over $\mathbf{z}$

$$C(z) = \max_{z}(\sum_{i=1}^{n} m_i(\mathbf{z}_i) + \sum_{(v_i,v_j)\in E} d_{ij}(\mathbf{z}_i, \mathbf{z}_j)). \tag{2.3}$$

Conveniently, the tree graph structure leads to an efficient and tractable inference such as sampling or belief propagation. We use our models to compute the score for each part v_i, at every pixel location of image I, and for all appearance mixture components $\mathbf{t}^i$, which includes messages from the children nodes of v_i by

$$s_i(\mathbf{z}_i) = m_i(I, \mathbf{z}_i) + r_i \tag{2.4}$$

where r_i is the sum of messages passed by the children of v_i. We provide a score heatmap visualisation for a representative part v_i for all its mixture components $\mathbf{t}^i$ in Fig. 2.6. A message from a child part to its parent computes the best location and mixture component for the child part.

Upon arrival of all messages at the root node, its score represents the optimal pose at its location. Retaining the indices of the best scoring part proposal, it is then possible to track back to find the location and mixture of each body part that is optimal for the pose. The principle that underpins this approach is that a collection of weak classifiers, collectively creates a strong class classifier. We retain the q best-scoring candidate cyclist poses using a threshold and apply non maxima suppression to prune overlapping proposals and, for each, select the one with top cyclist score. This enable the retention of multiple instances of cyclists in the image (Fig. 2.7).

2.2.5 Learning

We adopt the supervised learning paradigm of Kumar et al. [137] and train a part based detector for the human (cyclist) using manually annotated positive examples of joint locations and negative examples. We separately trained classifiers for the frontal and saggital view on 141 and 144 sagittal images of cyclists, respectively, whose pose was manually annotated for each of our models' body joint keypoints (see Figure 2.3). Our negative set contains 1217 images of people but not cyclists, and background scenery images.

The learning problem can be cast as obtaining a weight vector $\gamma = (\mathbf{w}_i, \mathbf{w}_{i,j}, b_i, b_{i,j})$ and scalar bias ξ, such that the learnt model parameters are able to discriminate between positive and negative examples in terms of their energy value. Learning

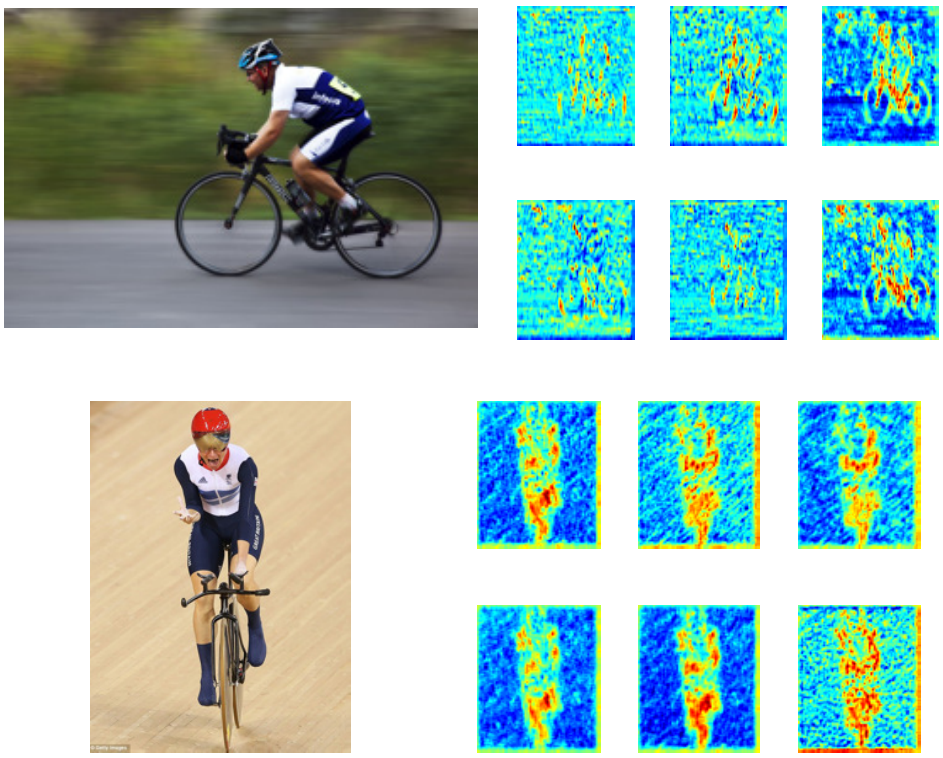


Figure 2.6: In inference, a feature representation of a new image is convolved with the learnt model's mixture parts appearance filters. The computation yields a local response as part appearance scores represented as heatmaps for the ankle body part for each mixture component (example from one indicative pyramid level). A hot colour represents high score / low discrepancy signal with respect to the body part appearance mixture component model.

the most discriminative model parameters is equivalent to solving the optimisation problem

$$(\gamma^\star, \xi^\star) = \underset{\gamma, \xi \geq 0}{\operatorname{argmin}} \frac{1}{2}\|\gamma\|^2 + C(\sum_n \xi_n) \tag{2.5}$$

such that

$$\begin{aligned} \gamma \cdot \phi(I_n, \mathbf{z}_n) &\geq 1 - \xi_n \;\; \forall n \in \text{positive images}, \\ \gamma \cdot \phi(I_n, \mathbf{z}) &\leq -1 + \xi_n \;\; \forall \mathbf{z}, \forall n \in \text{negative images} \end{aligned}$$

where the $\|\cdot\|$ operator represents the the standard l_2-norm, n the number of images, and $C \geq 0$ is a constant, which specifies the trade-off between accuracy and regularisation of the weights vector. The constraints ensure that positive samples score better than 1, and the negative samples score less than -1, violations of which are penalised by the objective function using the slack variables ξ_n.

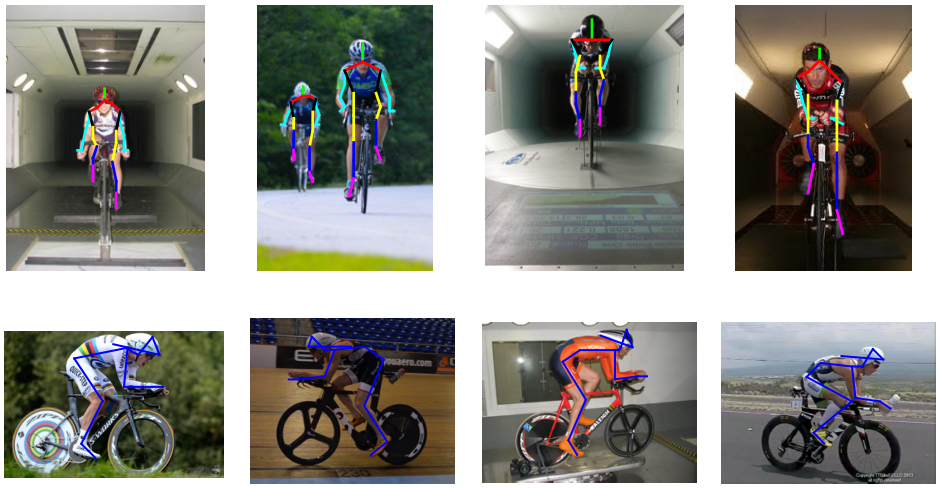


Figure 2.7: Our method yields pose estimation from a monocular image of a cyclist in the frontal and sagittal views and is robust to changes in scene, illumination and can yield multiple instances in an image.

Whilst convex, this learning problem cannot be solved efficiently due to the number of constraints. Kumar et al. [137] have reduced this problem to an equivalent problem with a polynomial number of constraints, from which an optimal solution can be reached. Known as a structural SVM, this learning problem has many efficient solvers. We use the dual-coordinate-descent quadratic programming solver of Yang and Ramanan [249] (see visualisation in Fig. 2.4a).

Part mixture components determination. We assume that the appearance of a body joint depends on its position relative to its parent (proximal body segment) in E. Therefore, we can use this relative position as criterion to cluster the part's appearance instances in our training data into consistent relative orientation. We define the mixture label for a part based on cluster membership achieved via K-means with $K = T$.

2.2.6 Implementation

We represent the human pose using a tree-structured graph with 26 nodes comprised of 18 and 14 keypoint nodes representing joints and 8 and 12 secondary mid-limb nodes for the frontal and sagittal models respectively, with the base of the neck at joint *C7* as its root and the limbs and head as its extremities (Fig. 2.1). We justify supplementing the number of joint keypoints by secondary mid-limb nodes by the experiments of Yang and Ramanan [249], who showed that 26 nodes provide a good trade-off of performance vs. computation. They experimented with varying the number of mixture types (14 to 26) and the number of parts (1 to 6), as well as with a star graph model. When published, their framework reduced error by 25%

and computation speed improvement of order of magnitude over the previously published state-of-the-art methods. In the sagittal case, we consider the base, top, front and rear edges of the bicycle helmet as key points to enable helmet orientation characterisation relative to the cyclist's trunk for future analysis.

A sagittal view of a cyclist presents severe occlusion of most or all of the limbs on the far side. Thus, it presents a large and significant view change that the framework of Yang and Ramanan [249], who impose a fixed number of keypoint nodes to model *small* changes in foreshortening, is unable to handle. Instead, in agreement with Felzenszwalb and Huttenlocher [92] we handle the severe occlusions induced by these rotations by explicitly encoding out-of-plane rotations by using a separate model with a different number of keypoint nodes for each view.

2.3 Results

To evaluate the performance of our pose estimation method, we apply our method to a set of challenging test images comprised of frontal and sagittal views of human cycling in unconstrained environment. In this section we report on both qualitative and quantitative results.

2.3.1 Quantitative Results

We conducted experiments on new task-specific datasets containing 141 frontal and 144 sagittal images of cyclists, whose pose was manually annotated for each of our models' body joint keypoints. The datasets contain images that have been either taken by the authors, downloaded from on-line repositories with a licence search criteria set to creative common (Flickr), labelled for reuse (Google Images), or provided by the University of Washington's windtunnel.

To train our models, we have split the datasets into a standard 60%, 20% and 20% for model learning, cross-validation and test sets respectively. Our negative set contains 1217 images comprised of the positive and negative training images from the INRIA Person ([57]) and Parse ([191]) datasets. In both datasets, the positive training images contain images of people with images of cyclists removed, whilst the negative sets contain mostly background scenery images. Using images that contain people in our negative set ensures that our cyclist model discriminates well between people in normal gait and cyclists for our specific task.

We measure the pose prediction of our method using loss functions, the Probability of Correct Keypoint (PCK) and Average Precision of Keypoints (APK) ([249]). A prediction is considered correct if it resides within a small distance from the annotated 'ground truth' point. For a given part at the annotated location i_*, the loss for prediction $\hat{i}$ is defined by

$$\triangle^p(i_*, \hat{i}) = I\left(\|i_* - \hat{i}\| > \alpha \ \max(h, w)\right), \tag{2.6}$$

where I is the indicator function, and h and w are the vertical and horizontal dis-

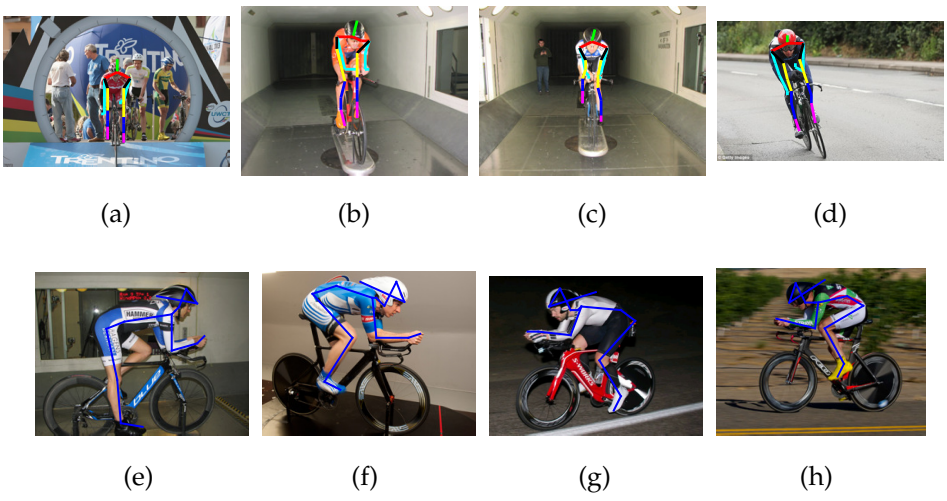
Table 2.1: Probability of Correct Keypoint (PCK) and Average Precision of Keypoints (APK) Results.

Dataset	Mean PCK	Mean APK
Frontal view	91.1	88.7
Sagittal view	69.3	66.1

tances respectively, and α is a detection region threshold parameter. The results are presented in table 2.1.

2.3.2 Qualitative Results

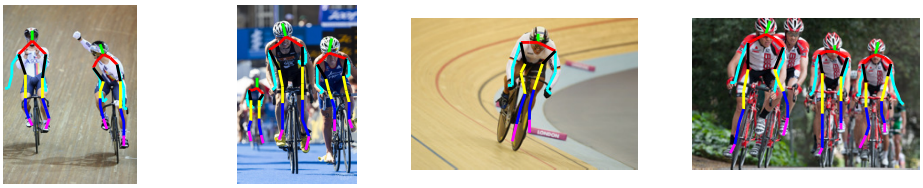
In contrast to the test images datasets, which contain images of a single cyclist with no other human objects in the scene, here we aim to qualitatively investigate the performance of our methods on unannotated images that contain multiple cyclists, severe occlusions and additional objects and humans in the scene (see results in Fig. 2.8). Whilst our approach is robust to small changes in orientation, view point and scale of the human object in images, the pictorial structure framework assumes that all body parts have a fixed scale. Hence, it may suffer local failure in such images that present cases where certain body parts experience severe fore-shortening effect due to change in view point, where the problem becomes ill-posed. Fig. 2.9 presents such cases.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2.8: Qualitative results in the frontal (top) and sagittal (bottom) planes. The discriminative power of the classifier is demonstrated by rejection of cyclist present in the scene but not in a riding position in (a), presence of people in (a) and (c), and robustness to small changes in object's view (b,f,g) and orientation (d,e).



(a) (b) (c) (d)

Figure 2.9: Examples of local failures; a) shows local failure at the wrists due to part unlearned spatial location. b) shows failure at the ankle of the cyclist on the right due to unlearnt part appearance (barefoot). c) shows failure at the legs due to hip occlusion. d) shows failure at the left ankle of the lead cyclist due to part appearance similarity of the second cyclist's ankle.

2.4 Discussion

The presented results demonstrate the utility of estimating human kinematics via a fully supervised learning approach to reconstruction of human posture from unconstrained images as an alternative to marker-based motion capture. Our method is underpinned by a spatial relations model and parts' appearance and co-occurrence models. Therefore, like other markerless approaches it does not suffer from the deficiencies of marker-based system. For instance, our method is not constrained to an expensive laboratory setting with controlled lighting, reflectance and other environmental conditions, nor does it have any direct sensitivity to STA. likewise, it does not suffer the inter-trial and inter-tester repeatability and reliability of direct digitisation approaches ([136, 202]).

For our task, we applied the flexible mixture of parts method ([249]) to the estimation of cyclists' posture and achieved robust results with a task-specific trained classifier. It is, however, impossible to design a discriminative classifier for the general case because of the high variability in the appearance of human ambulation. The performance of the approach critically relies on time consuming and costly process and the availability of a large quantity of training samples. Further, the generalisation of the approach can only be achieved through the addition of adequate training samples, as its adaptability to unseen body postures is low, as typically manifested by poor performance were occlusions exist. Thus, to adapt our method to a different movement modality, a task-specific classifier needs to be trained on an appropriate training set.

With respect to our specific cycling task, the coupling of the human and object reduces the search space of likely poses represented by the spatial relations between parts in our model as a result, for instance, of the predictable position of the hands on the handlebars and feet on the pedals. This may be viewed as an advantage of our task of choice when compared to unconstrained activity such as running. On the other hand, the human-object interaction, in fact, complicates the search for a body part's appearance. This is due to increased incidence of body part occlusions and

appearance similarity or ambiguity. Therefore, a trade-off exists between improved spatial relations and less discriminative appearance. Since the apriori location of the human is unknown in a new image, the search space for parts' appearance is an image pyramid over the entire input image. Hence, overall it is less challenging to estimate the pose in running than when a human-object interaction exist. Normal gait and running have been addressed often previously ([6, 7, 86, 87], and others). Typically, these approaches are aided by segmentation and background modelling algorithms that separate the human (foreground) and background scene prior to pose estimation by reducing the image search space to the foreground alone. We avoid pre processing steps, and directly estimate the pose from the parts' appearance and spatial relations simultaneously.

This is consistent with our aforestated motivation, which stems from the observation that the cycling is particularly challenging due to the human interaction with the bicycle, which induces severe occlusions, the similarity of the posture in the frontal plane to normal human gait, which presents a discrimination challenge, and the severely occluded sagittal plane posture, for which a pose estimation method was not found in the literature. In follow-up work that is outside the scope of this paper, we will investigate whether modelling the human-object interaction explicitly improves the pose estimation.

A principal limitation of the pictorial structures framework is that exact inference is only possible with tree structured graph model. However, in certain situations, such as temporal and occlusion models, it is advantageous to have non-tree models. Kiefel and Gehler [131] used a binary random variable to model occlusions at every possible location, scale and orientation and graph topology that is not restricted to a tree structure, with modest improved performance. Cherian et al. [45] generated pose candidates in each time frame by enforcing temporal constancy between instances of a tree model. They decomposed the body parts to generate temporally smoothed body part sequences, followed by re-composition of the body pose. A non-tree extension to our work that models occlusions and allows for temporal constancy to be enforced would make the approach very attractive for a variety of applications.

Despite significant progress, accurate inference remains an extremely challenging problem principally due to occlusions and self-occlusions in the image. Consequently, in our framework, we model the human body only for the visible body parts for each view. This results in separate models for the frontal and sagittal views. The challenge is further compounded by inter-object interaction, such as the interaction between a cyclist and the bicycle. Therefore, a natural extension to our approach would exploit advances in modelling the human-object interaction within the framework. Moreover, in this work we avoided imposing explicit kinematic constraints on the pose proposals. Introducing such constraints will significantly reduce the search space of probable poses and improve the accuracy.

We note that whilst manual annotation of point location remains the standard ground truth for performance evaluation of pose estimation techniques in the computer vision domain, a higher level of accuracy is often expected in the biomechanics domain. The reason stems from error propagation with subsequent calculations and

double differentiation required for inverse dynamics optimisation. The accuracy required is commonly achieved through the obtrusive use of surface mounted reflective markers. Nevertheless, our choice to validate our model estimation against manual annotation is justified since the use of reflective markers would contaminate the model in learning and the image data in inference. It will result in appearance models that are tuned for the presence of a marker in the image patch. Consequently, performance evaluation would be grossly overestimated. Furthermore, marker-based systems require control of scene, lighting and environmental conditions, which would unfairly penalise our method.

The estimation of pose using our method can be used for the reconstruction of the human body's geometry. Using our cycling task as an example, the geometric shape of the cyclist can be used to enhance the understanding of the relationship between a cyclist's posture and aerodynamic drag. The task then becomes a problem of extracting the cyclist's shape from the background in an image in a segmentation pipeline. Thus, the extracted skeletal pose can be used as a necessary prior foreground seed for segmentation techniques such as graph cuts ([21, 229]), and the exterior to a convex hull that contains all part patches as its background seed.

2.5 Conclusions

In this paper, we investigated the challenging problem of markerless estimation of a human full body kinematics from monocular images. We proposed a discriminative part-based approach that develops a probabilistic prior model based on learned measurements. In learning, a structured SVM solver learns spatial relations of skeletal segment orientation and co-occurrence relation between parts appearance. In inference, the model detects the human in the image and recovers pose estimates. We applied our approach to images of cyclists captured in natural environment with no assumptions in relation to kinematic constraints, nor the appearance of the scene. Our technique finds multiple good hypotheses for the human pose rather than just a single best solution. This is advantageous for cases where imprecision in the model resulting in the desired match not being the one with the minimum energy. Our method yields a robust user-interaction-free approach for estimation of full body kinematics, which serves as a crucial evidence base pre-requisite to motion analysis. Based on skeletal kinematics, joint forces, torques, power and efficiency of motion can then be determined. Furthermore, our pose estimate can be viewed as a necessary first step in a segmentation pipeline aimed at characterising the geometry of human motion.

Chapter 3

Modelling Human-Object Interactions for Improved Human Pose Estimation

Abstract

Estimating human poses in real-world monocular images under occlusions continue to be a challenging problem in computer vision. In this chapter, we present a technique for addressing this problem by explicitly modelling the interactions of humans with complex objects. Going by the state of the art approaches for pose estimation, we model the human pose and the object by separate probabilistic tree-structured graph models. The interactions of the human with the object is captured by imposing constraints between the models; each interaction point contributing to a proximity constraint between the respective human joint and the object part. We propose an algorithm for efficient inference over the combined model. Next, we provide an application of this framework for estimating the pose of a cyclist. Specifically, based on skeletal pose kinematics, joint forces, torques, power, and efficiency of cycling can be determined using inverse dynamics optimization. In the absence of public datasets for the task, we propose a novel human-bicycle interaction dataset, experiments over which are provided and demonstrates significant improvements against alternative schemes.

3.1 Introduction

Problem Scope In this chapter we consider the challenging problem of simultaneous human detection and pose estimation from monocular images. The task serves as a crucial evidence base pre-requisite to analysis of technique and optimisation of performance as well as training anticipatory skill in sport.

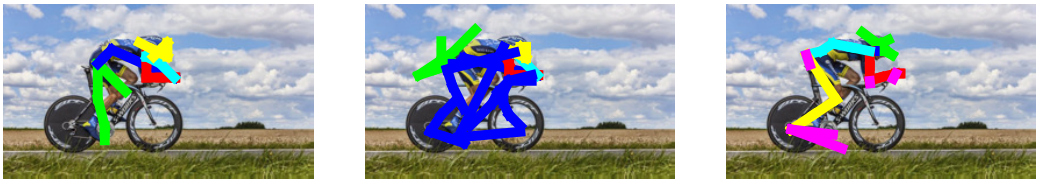


Figure 3.1: Example demonstrating difference between methods. Detection and human pose estimation comparison between cyclist model only (left) implicit (centre - bike included) and explicit (right - inter-object relations marked in magenta) method of modelling human-object relations.

Figure 3.2: Example where there was no difference. Detection and human pose estimation comparison between cyclist model only (left) implicit (centre - bike included) and explicit (right - inter-object relations marked in magenta) method of modelling human-object relations.

3.1.1 Approaches to inter-object relations

The challenge of pose estimation is compounded by interactions between humans and ambulatory devices, instruments and accessories that cause occlusion and changes to appearance, but are common in many human ambulatory activities. Accurate model of the human and object interaction often yields loopy graph representation, which significantly hinders inference. Instead, most methods use an approximate tree graph representations that enables exact inference [45, 249]. However, this approximation results in unrepresented edges, which may contain important information about the direct influence of adjacent nodes. Consequently, outcome of inference on approximated trees is sensitive to the choice made on removing edges to avoid loopy graphs.

In its generic form, the connections between parts of the pictorial structure framework can be used to encode generic relationships between objects [92] or geometrical constraints. Nevertheless, the majority of human pose estimation methods treat the human in isolation with respect to its interaction with the environment, except for the environment information surrounding the object that is invariably captured in

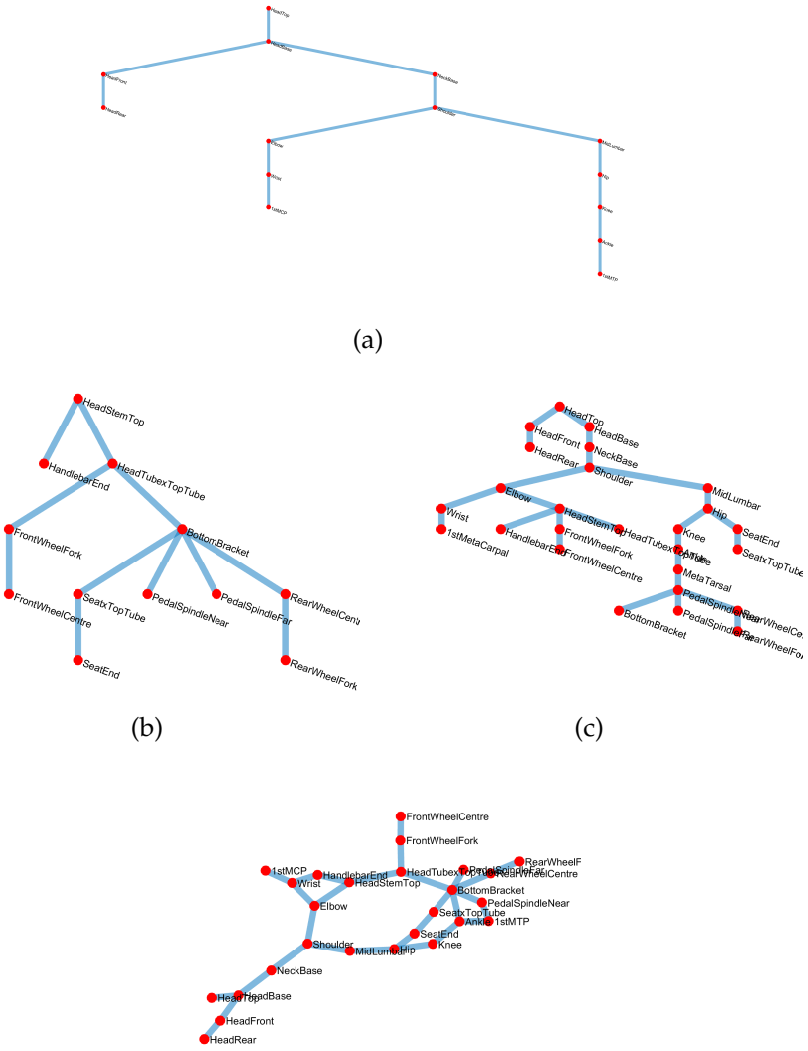


Figure 3.3: Graph representation of the experimental conditions; a) is a tree graph model of the cyclist in side view. b) is a tree model of the bicycle in side view. A natural representation of the cyclist-bike complex is depicted in d, composed of the objects in a and b that are linked by the highlighted interaction edges. This, however, is a cyclic graph, which makes inference difficult. c) is an acyclic approximate representation of the cyclist-bike complex. It has the same number of nodes whilst retaining a tree structure enabling exact inference. In this paper we considers inference over models a, c and d, where in d inference is performed separately on its decomposed sub-trees a and b.

the training of an appearance model. Hence, these methods do not explicitly model

the objects themselves or the human-object interaction under the strong assumption that in inference, the learnt human-only model will treat the object and the interaction as noise, and therefore, ignored. This optimistic assumption often results in failed detection and pose estimation (see fig. 3.2).

Several methods for activity recognition *do* exploit human-object interaction [109, 110], where the object label is used to constrain the activity semantic. Delaitre et al. [67] introduced person-object interaction features based on spatial co-occurrences of individual body parts and objects. A discriminative model of the interaction is learnt offline to improve action classification. However, they do not attempt to estimate the human pose, nor do the learnt features necessarily correspond to meaningful configuration or body pose. In fact, their method of learning class-specific object and body part interaction is predicated on avoiding the challenging task of estimating the full pose. Likewise, Sadeghi and Farhadi [199] avoid estimating pose and treat the activity recognition as a classification problem by learning "visual phrases" for interaction. Similarly, Prest et al. [189] introduced a weakly supervised learning approach to interactions that yields a probabilistic model of the spatial interaction between the human and the object. Our approach significantly differs from these methods as we exploit interaction to improve pose estimation.

Alternate approaches focus on the interaction with the ground plane. Rosenhahn et al. [197] integrate the floor plane into the pose estimation framework by imposing constraints that penalise pose proposals with body parts that intersect the ground plane. Likewise, Amamoto and Agishita [4] use kinematic constraints from the scene to reduce the degrees of freedom of a skier model in a tracking task. This approach, however, is not effective in human movement modalities that take place at a distance from the ground (e.g. cycling) or performed through fluid environment (e.g. rowing , swimming), where the motion intersects the medium, or where the floor plane is absent from the scene. In contrast, we detect the object in the scene and use the spatial relationship with the human to infer the human pose.

Whilst objects were commonly considered occluding interferers in pose estimation tasks, Hamer et al. [112] exploit contact points and object geometry to learn an object dependent hand pose prior to assist tracking. In agreement with Hamer et al. [112], Kjellstrom et al. [134], Singh et al. [216], we argue that if an object can be detected, then the apriori knowledge of the context of the pose in terms of its interaction with an object can be used to reduce the search space and subsequently the overall degrees of freedom of the pose. For instance, it can be reasonably inferred that a cyclist holding onto a bicycle handlebar has the forearms internally pronated. Our work is inspired by the hand pose estimation work of Hamer et al. [112], Oikonomidis et al. [180] and conceptually similar to Bangpeng and Li [10], Yao and Fei-Fei [252] in explicitly modelling the human and object separately and using inter-object interaction to infer pose proposals that facilitate pose estimation, but differs from Desai and Ramanan [69], Sadeghi and Farhadi [199], who have implicitly jointly modelled the human with the object . However, Bangpeng and Li [10] as do Gupta et al. [110] use a fully supervised setting that requires training with annotated human and object locations. This requires the construction of a large dictionary of human poses,

they term 'atomic poses', that correspond to particular configurations of body parts. Their mutual context model outperforms traditional methods in human pose estimation. In contrast to Kjellstrom et al. [134], who use a generative framework, we use a discriminative part based model framework for pose estimation. Furthermore, Kjellstrom et al. [134] use a complicated human model with 40 pose parameters relative to global coordinate system and 34 Euler angles defining relative pose of the limbs. We use a simple and efficient tree model with a task specific inter-relations model.

Singh et al. [216] introduced pose context tree, which adds an object node to the tree structured human graph model. They make a strong assumption that objects are only handled at the extremities and at a single point of contact. The approach is unsuitable for our task due to the multiple points of contact with the object including at a common root note (pelvic-bike seat interaction). Further, the multiple contacts will create cyclical graph models. These can be viewed as a Markov chain, which necessitate solving for the stationary state to compute various probabilities, for example with loopy belief propagation. In which case, the solution may not converge or if it does, the solution is an acceptable approximation [250]. For cyclic models with larger graphs the situation becomes even more complex and computationally infeasible.

It is out of scope of this chapter to develop a pose estimation method that applies to all situations. Instead, without loss of generality, we focus on pose estimation of a person interacting with a large complex object with multiple contact points in cycling.

A further insight to our approach is the observation that objects are generally rigid. They do not display the complexity of deformable objects such as human motion. Hence, rigid-object pose reconstruction is easier to detect reliably in agreement with Urtasun et al. [228], who tracked a golf swing. Likewise, Gupta et al. [109] exploit tennis ball kinematics used as constraints for forearm reconstruction. The majority of these methods use a simple object with a single point of contact. In contrast, we use a complex object with multiple points of interaction.

3.2 Inter-Relations Pose Estimation Framework

Given an image of a cyclist in an unconstrained environment, our goal is to estimate the human pose (shown in Fig. 3.6). In the case of a human interacting with an object, pose estimation using a human-only classifier often fails as a result of occlusions, self-occlusions or the varied appearance induced by the articulated human as demonstrated in Figure 3.2. In contrast, rigid objects can be more reliably detected (Fig. 3.2). In this section, we introduce our framework, which exploits a successfully applied discriminative part based pose estimation technique to independently detect both the human pose and the pose of an associated rigid object. Both human and object classifiers are learnt offline. We introduce an inter-relations model, which defines the spatial relationships between a human part and an object part (Fig. 3.4. Our model deals with the situations where the human interacts with an object at multiple

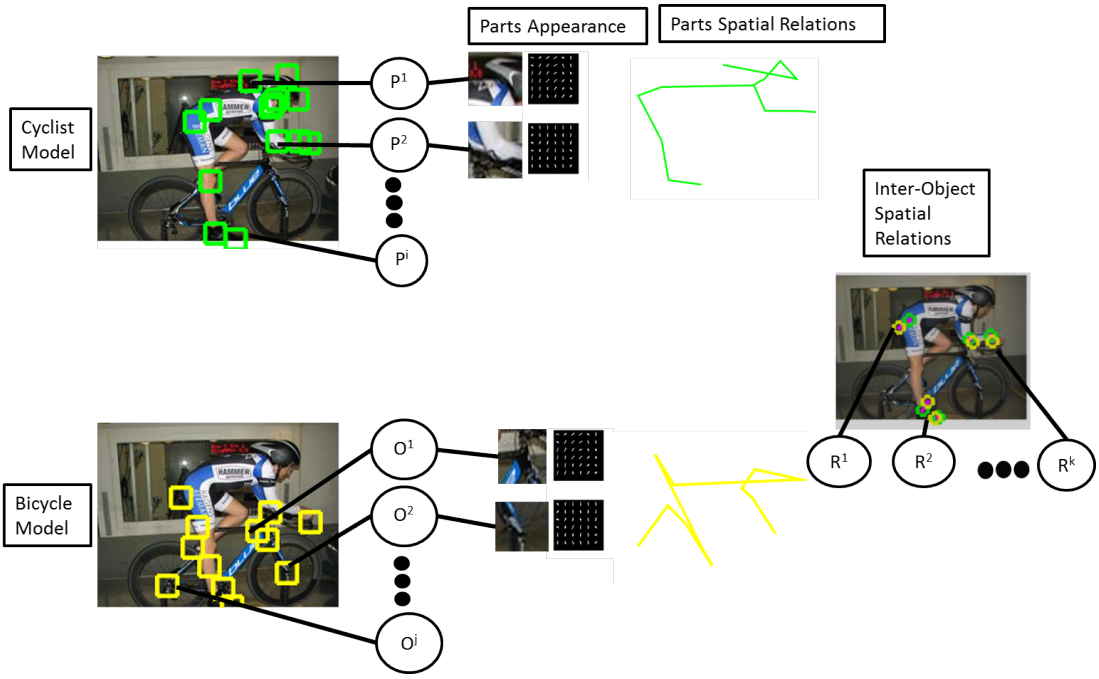


Figure 3.4: Inter-object spatial relations model learning. Our framework learns a person model from person parts appearance and spatial relations (**P**), and object parts based appearance and spatial relations model (**O**), and an explicit person-object relations model (**R**).

points. We make no assumption in relation to the number of interaction points nor on the nature of the interaction, i.e. whether the interaction denotes a contact or proximity of the object's part to the human part, nor its uniqueness. We introduce the part based classification framework in section 3.2.1 followed by our formulation of the inter relations model in section 3.2.2.

3.2.1 Part Based Body and Object Models

The standard pictorial structures model as described in chapter 2, is expressed in terms of an undirected graph $G = (V, E)$, where the vertices $V = \{v_1, \ldots, v_n\}$ correspond to n parts, and an edge $(v_i, v_j) \in E$ for each pair of connected parts v_i and v_j. An instance of the object in the image I is given by a joint configuration of body parts $L = (\mathbf{l}_1, \ldots, \mathbf{l}_n)$, where $\mathbf{l}_i = (x_i, y_i)$ denotes the location of part v_i.

A tree-structured graph model is designed whose scores are provided by body part templates that are learnt from training data [7]. Conveniently, this model structure leads to an efficient and tractable inference as described in detail in chapter 2. The optimal match of a model to an image is found by minimizing an energy function that measures both discrepancy for each part and a deformation cost for each pair of connected parts and is defined by eq (2.3).

3.2.2 Inter-Relations Model

For our task, a composite human and object model would be an accurate natural representation. However, this would result in loopy graph representation, which significantly complicates inference. Instead, we decompose the problem into a separate parts based models for the human (a cyclist) and the rigid object (a bicycle) and model the interaction between the models separately. Thus, our approach can be viewed as a sparse bipartite graph of two (or more) disjoint sets U and $U^\star$, where U and $U^\star$ are independent sets, such that some edges in U connect to some edges in $U^\star$. Our approach, whilst an approximation of the composite problem, yields a tree representation of each object enabling exact inference on each.

We adopt a supervised learning paradigm and train a separate part based detector for the human (cyclist) and the object (bicycle) following section 2.2.5 using manually annotated positive examples of keypoint locations and negative examples. In our task, a side view of a cyclist presents severe occlusion of most or all of the limbs on the far side, whilst the majority of the complete set of body parts is visible in the front view. However, the approach in Yang and Ramanan [249] imposes a fixed number of nodes. Therefore, we use a different representation for each view point T.

3.2.3 learning and Inference

For our side view model, we represent the human pose using a tree-structured graph with 26 nodes comprised of 14 keypoint nodes representing joints and 8 secondary mid-limb nodes with the base of the neck at joint *C7* as its root and the four limbs and head as its extremities (Fig. 3.3). We justify supplementing the number of joint keypoints by secondary mid-limb nodes by the results of Yang and Ramanan [249]. Similarly, we represent the bicycle object using a tree-structured graph model with 14 nodes comprised of frame joints, extremity points or wheel centres with the bottom bracket as its root. Thus we denote the joints of the human body $V^h = \{v_1^h, \ldots, v_n^h\}$ and the keypoints of the bicycle object $V^b = \{v_1^b, \ldots, v_m^b\}$.

Given an image I, our models report best scoring pose configurations $\mathbf{l}_i^h$ and $\mathbf{l}_k^b$ by minimising (2.3) for the cyclist and bicycle respectively. To model the interaction between the cyclist and the bicycle, we write $(v_i^h, v_k^b) \in H$ to denote an edge for each pair of associated inter-object parts, where $H \in \{1, \ldots, T\}$. The inference problem than becomes

$$C(z) = \max_z \Big(\sum_{i=1}^{n} m_i(\mathbf{z}_i) + \sum_{(v_i^h, v_j^h) \in E} d_{ij}(\mathbf{z}_i, \mathbf{z}_j)\Big) + \delta \sum_{(v_i^h, v_k^b) \in H} g_{ik}(\mathbf{l}_i^h, \mathbf{l}_k^b)\Big), \tag{3.1}$$

where δ is a scalar tuning parameter and $g_{ik}(\mathbf{l}_i^h, \mathbf{l}_k^b)$ is a function measuring the proximity for a given pair of inter-object related parts, that is of a human part v_i^h at location $\mathbf{l}_i^h$ and bicycle object part v_k^b at location $\mathbf{l}_k^b$. Note that in an absence of an object or interaction between the human and an object, the final term disappears and our model becomes the human-only model described in section 2.2.5.

Our goal is to detect and estimate the pose of a cyclist from front and side test

images. In inference, we produce candidate pose proposals by using a sliding window detection scheme over an image pyramid. We use our models to generate Q number of best scoring candidate bicycle poses. We then generate S number of best scoring candidate cyclist poses. Of these candidates, we select the R best candidate poses that minimise the inter-object relation model in (3.1). We apply non maxima suppression to prune overlapping proposals and elect the one with top cyclist score.

3.3 Experiments

In this section we report on quantitative experiments to evaluate the performance of our pose estimation method. We apply our method to new datasets of challenging task-specific test images, which involve front and side views of cyclists in an unconstrained environment. We test our model against a human-only model and with an implicit composite human-object model, to which we apply the flexible articulated model of Yang and Ramanan [249].

The FV and SV datasets contain 141 and 144 front and side view pose-annotated images of a cyclists respectively. Both datasets include a standard train and test split. To train our models, as our negative set we use both the positive and negative training images from both the INRIA Person [57] and Parse [191] datasets. In both datasets, the positive training images contain images of people with images of cyclists removed, whilst the negative sets contain mostly background scenery images. Using images that contain people in our negative set ensures that our cyclist model discriminates well between people and cyclists for our specific task (see right image in Fig. 3.5).

Figure 3.5: Top likely pose proposals returned by a cyclist-only model. Some likely poses are not associated with an object defining the activity. For clarity, the side view pose limbs are marked in yellow for the lower limbs, in cyan for the trunk, in red for the upper limbs and in green for the head/helmet complex. For the front view the lower limbs are marked in magenta, blue, and yellow, in black for the trunk, cyan for the upper limbs, red for the shoulders and green for the head/helmet.

Qualitative results of our pose estimation compared to implicit composite person-

object and person-only models are presented in Fig 3.2. The figure demonstrates that our inter-object relations model outperforms both alternate models. Notably, at worst our model performs as well as the person-only model following (3.1). The implicit composite model appears to consistently perform worse than the inter-object and human-only models. A possible explanation relates to the size of the graph model that is required to jointly represent both human and object. A large model yields an exponential number of pose configurations that are difficult to satisfy with image evidence.

3.3.1 Quantitative Results

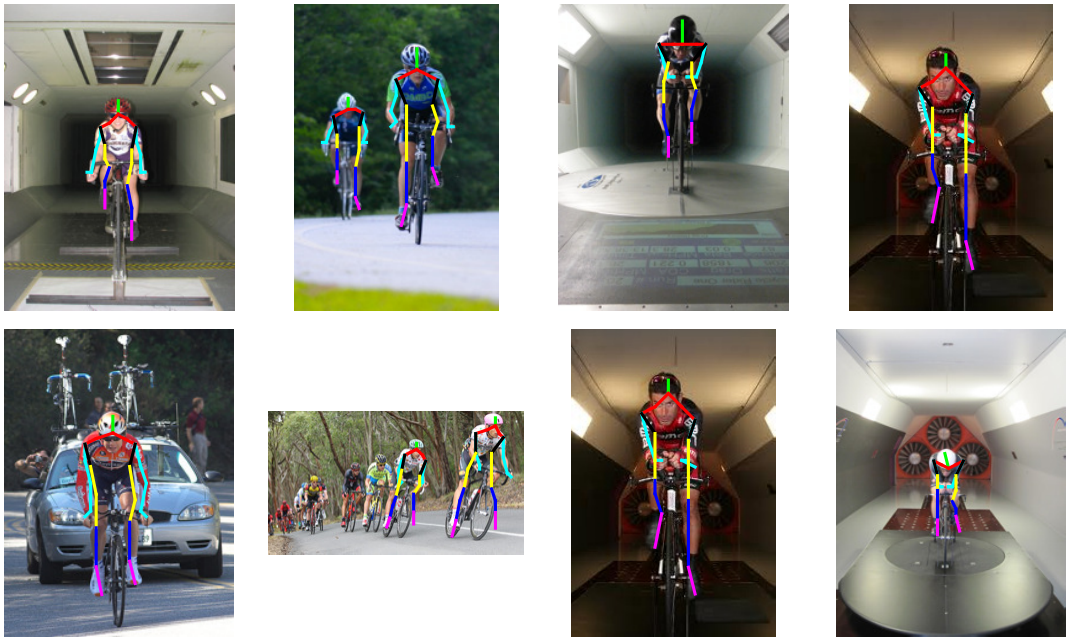


Figure 3.6: Detection and pose estimation of cyclists in challenging environments.

We quantitatively measure the performance of model estimation using loss functions that represent the desired output of the system. For the performance of body pose estimation, we use the Probability of Correct Keypoint (PCK) and Average Precision of Keypoints (APK) [249]. A prediction is considered a true positive if it resides within a small distance from the annotated ground-truth keypoint. For a given part at the annotated location i^*, the loss for prediction $\hat{i}$ is defined by

$$\triangle^p(i^*, \hat{i}) = I(\|i^* - \hat{i}\| > \alpha \max(h, w)), \tag{3.2}$$

where I is the indicator function, and h and w are the horizontal and vertical distance respectively, and α is a detection region threshold parameter. The APK formulation is not only a measure of true positives, but also penalises both false positives and

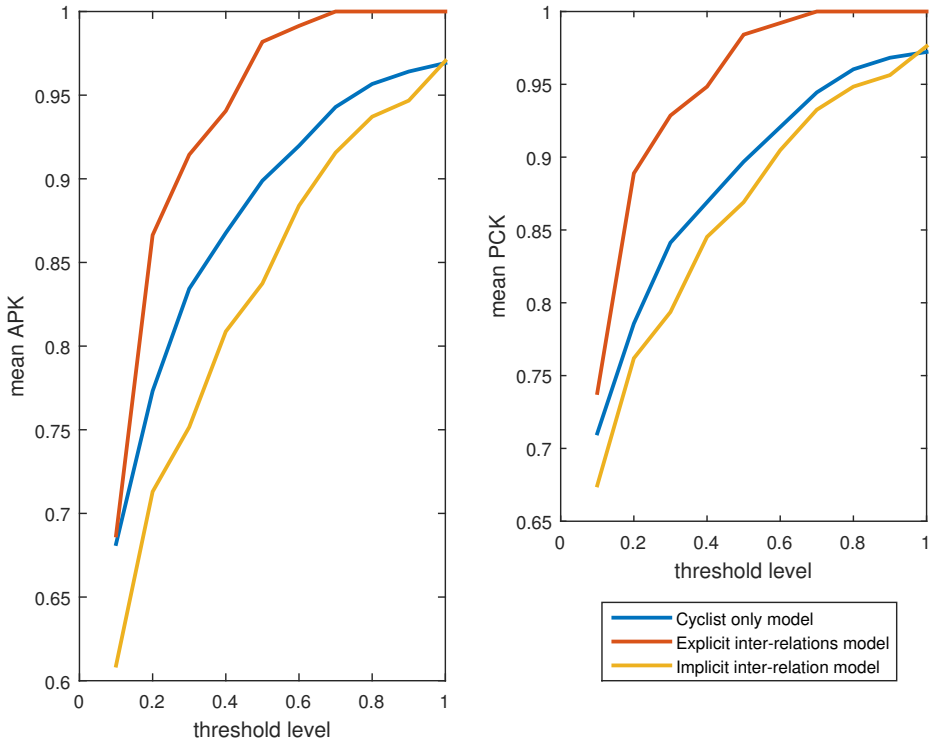


Figure 3.7: mean key points APK (left) and PCK (right) as threshold level changes for the three inter-object relations conditions

false negatives.

The direct comparison using PCK and APK is reported in Table 3.1. The table indicates that our method outperforms the composite person-object model on both PCK and APK. Further, we consider the effect of varying λ, a threshold tuning parameter imposed on the location of pose candidates through the score of the root node l_1. Figure 3.7 shows that our method performs better than the two alternatives consistently across all levels of λ. The composite person-object performing significantly worse than both.

Moreover, we explore whether the effect is manifested stronger at body part nodes distant from the root node. Interestingly, the wrist body part, a body part furthest from the root displays similar performance for all models. In contrast, the toe body part, an equally distant body part from the root, does not. Both body parts have an assigned inter-object relation edge. We reason that this stems from the role of those body parts during the target activity. The wrist is quasi-static and expected to be in contact with the handlebars during cycling. Hence, its appearance and spatial relation models would have low variability and would perform equally well in the three experimental conditions tested here. Indeed, we observe this behaviour for all the upper body parts in our task. In contrast, the toe is the most dynamic body part during cycling. Therefore, it manifests highly varied appearance and spatial relations

Table 3.1: Probability of Correct Keypoint (PCK) and Average Precision of Keypoints (APK) Results.

Method	Mean PCK	Mean APK
Integrated Bicycle-Cyclist Model	67.5	60.9
Inter-Object Relations Model	71.0	68.2

throughout the pedalling cycle that is evident in the training set. Thus, anchor constraint via human-object interaction model is likely to benefit this body part most. This observation provides supporting evidence that our approach may be particularly beneficial to highly dynamic motions and where highly varied appearance is expected. Furthermore, this also indicates anchor constraint of a body part to an object benefits the pose estimation problem for the entire pose rather than locally at the body part nodes. Examples of body part specific comparisons are given in Figure 3.8.

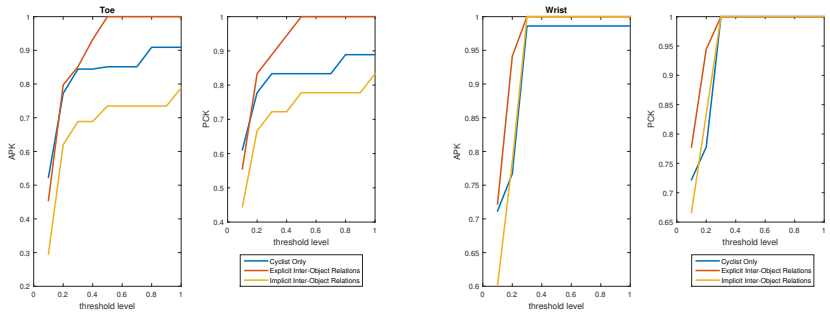


Figure 3.8: APK (left) and PCK (right) for the Wrist and Toe joints as threshold level changes at each of the three inter-object relations conditions

3.4 Conclusions

In this chapter, we investigated the challenging problem of simultaneous human detection and pose estimation from monocular images where human-object interactions induce severe occlusions. We proposed an extension to the discriminative approach for concurrent detection and pose estimation using flexible mixture of parts. The approach develops a probabilistic prior model based on learnt measurements to be used for inference, but cannot be easily generalised to un-trained scenarios such as when human-object interactions create occlusions and self-occlusions. We argue that global composite models are unnecessary. Instead, our framework uses an object model that is levered to prune human object pose proposals. We propose an inter-relations model to define the interaction between a human and an object models.

For each model in our method, a structured SVM solver learns spatial relations of skeletal segment orientation and co-occurrence relation between parts appearance. In inference, inter-object spatial relation model helps prune unlikely detections by penalising unlikely pose proposals. Our experiments demonstrated that our explicit inter-relations model outperforms implicit composite human-object and human only models on a new task-specific dataset. This pose estimation task serves as a crucial evidence base pre-requisite to inverse dynamics optimization.

Limitations and Future Work A number of challenges remain that need addressing for enhanced robustness of the proposed method. For instance, tree based models may produce pose candidates that co-locate symmetric limbs. Due to the similarity in limbs appearance, the model may score both limbs in the same part of the image. Whilst this may result in a plausible pose, it would be advantageous to penalise such a pose proposal.

Occlusions and self-occlusions present the main difficulty in pose estimation. To address this challenge, an extension to our approach could encode the presence or absence of a part in an image with a binary random variable. This results in a significantly increased inference problem. However, Kiefel and Gehler [131] demonstrated an efficient approximate solution.

Whilst tree graph models allow for efficient inference procedure, cyclical models are often desired, for example in temporal tracking of pose or multiple views of a person. The former can be addressed by generating pose candidates in each time frame followed by a decomposition of the body parts to generated temporally smoothed body part sequences and pose re-composition Cherian et al. [45]. The latter can be addressed by introducing a latent mixture model representing the view. The inference problem than discomposes into a tree constraint. Conceptually, our approach can be considered similar to Cherian et al. [45] in the sense that they enforces constancy between instances of the same tree graph model in the temporal domain, whilst we enforce constancy between distinct tree graph models in the spatial domain. Conveniently, our approach does not require decomposition of the body parts.

As noted in chapter 2, manual annotation of point location remains the standard ground truth for performance evaluation of pose estimation in the computer vision domain, but higher level of accuracy is expected in the biomechanics domain. Nevertheless, our choice to validate our model estimation against manual annotation is justified since the use of reflective markers would contaminate the model in learning and image data in inference. This will result in an appearance models that are tuned for the presence of a marker in the image patch. Consequently, performance evaluation would be grossly overestimated.

Part II

Activity Recognition and Full Body Shape Recovery

This part presents two techniques that yields surface geometry in addition to skeletal pose to be used for biomechanical analysis. We demonstrate the approaches for the estimation of projected Frontal Surface Area (pFSA) of cyclists to be used for the investigation of bluff bodies aerodynamics and optimising cycling performance. This presents a significant paradigm shift in biomechanical analysis of human movement,which traditionally uses skeletal pose only.

Structured light sensors capture 2.5D scene geometry at video frame rate, which provide data that allows for more reliable pose reconstruction from a single viewpoint. Previous work have shown successful reconstruction of poses based on the depth signal. Chapter 4 presents a real-time activity-specific framework. It takes a supervised learning approach to learn a model for activity recognition based on skeletal kinematic features reconstructed from depth information, and simultaneously recovers the human surface geometry. The developed framework outputs estimated pose and recovered shape in real-time.

The problem of real-time reconstruction of complex human motions from monocular intensity images is considerably under-constrained and remains open. Non-trivial inference or optimisation tasks are required in combination with strong priors in order to have a chance to reconstruct human movements. Chapter 5 develops an alternate technique to pFSA estimation from monocular intensity stream images because these are the most readily available signal sources despite advances in multimodal and stereo imagery systems. It takes a hybrid approach, which uses a pictorial structures discriminative approach to human detection and a generative approach to shape recovery using an active contour framework regularized by a learnt shape and appearance models prior.

This part demonstrates how levering the techniques developed in this dissertation can provide information not previously available to answer research questions in allied fields. Chapter 6 provides theoretical foundation to the study of the relationship between recovered human geometry and human motion using a motivating example of cycling performance and aerodynamics. To improve our understanding of this human ambulatory modality, both skeletal pose and recovered shape are necessary.

Chapter 4

Real-Time Periodic Motion Activity Shape Reconstruction Using Random Regression Forest over Kinematic Derivatives from Depth

abstract

This chapter addresses the challenging problem of continuous activity recognition and shape estimation of dynamic articulated human performing periodic motion from depth in real-time. Modelling the non-linear behaviour of human motion requires detailed geometrical shape in addition to skeletal kinematics. In this work, we perform detection and recognition of unstructured human periodic motion activity in unstructured environments from noisy depth information by a structured light sensor. We cast the activity recognition as a labelling problem and learn a statistical model over a set of features that are computed from the estimated human skeletal pose, motion and point cloud information. The key components of our framework include kinematic features that encode the relative spatio-temporal characteristics of the target activity. Our spatial relative joint angle features provide invariance to sensor view and differences in individual proportions. The temporal derivative features capture the salient characteristics of the dynamic activity of interest. Our trained boosted classifier is robust for detecting the target activity and allows the extraction of estimated geometrical parameters of the subject efficiently. We test our algorithm on detecting and recognizing cyclists and achieve good performance even when the person was not seen before in the training set. Our activity classifier achieves an average accuracy of 98.3% for approximately 360 seconds recordings at 30Hz. Following a successful recognition of the activity we output the object's geometrical shape by a silhouette-from-skeleton approach, which infers pixel or voxel labels by segment-group membership that corresponds to the object's shape.

4.1 Introduction

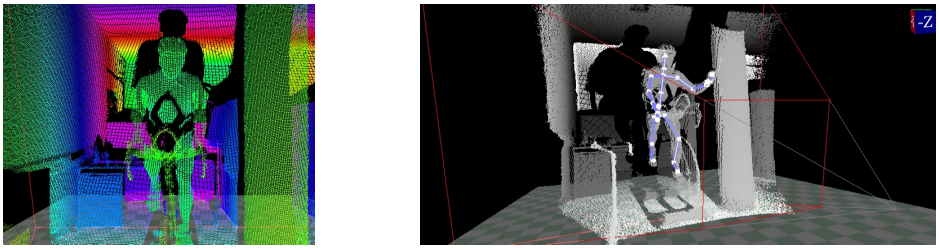


Figure 4.1: Surface geometry (surface normal) and estimated skeletal pose (right) are computed from the point cloud information (left) that is obtained from a single structured light sensor.

In this chapter, we address the challenging problem of continuous activity recognition and geometrical shape estimation of a human performing periodic motion from depth in real-time. Modelling the non-linear behaviour of human motion requires detailed geometrical shape in addition to skeletal kinematics. Simultaneous reconstruction of body kinematics and geometric shape is critical for many diagnostic and clinical applications. The *person-specific* body's geometry influences the initiation and progression of musculoskeletal injury and disease, and joint and limb geometries improve load predictions [99, 214, 259]. Whilst the inverse dynamics approach is standard practice that has served the biomechanics community well for the study of human motion for several decades, we argue that the skeletal linkage system at its model's core is an over-simplification that fails to explain full-body human motion in many cases. The dimensionality-reduced modelled system is appropriate when the system is subjected to external forces that can be realistically approximated as acting at a point on the system. Alas, this cannot be justified when the external forces are acting on the entire object's geometrical shape. To advance the state of knowledge, real-time acquisition of individualised full-body geometrical and motion is needed.

In this chapter, we perform detection and recognition of unstructured human periodic motion activity from depth maps that enables a subsequent surface geometry estimation. We cast the activity recognition as a labelling problem, the key components of which, include spatio-temporal kinematic features computed from a human pose and motion and point cloud information obtained from noisy depth information from a structured light sensor, and a trained boosted activity classifier. Once detected, we are able to extract the surface geometry from the data stream based on the estimated skeletal pose, facilitating further analysis of the motion.

Human activities are often considered to be primitive actions consisting of one or few atomic body movement or poses, such as gait, sitting, standing , kicking, punching or jumping. More complex actions often involve interactions with objects and other people, and can occur over longer periods. In particular, repetitive periodic

actions such as cycling, rowing or kayaking present a critical challenge to detection and pose estimation tasks, in which the human-object interactions may induce severe occlusions of the human limbs. Consequently, the characteristics of repetitive periodic activities that involve human-object interactions, whilst common in real-life, are rarely studied.

4.1.1 Motivating Application: Real-time Estimation of Projected Surface Area of Cyclists

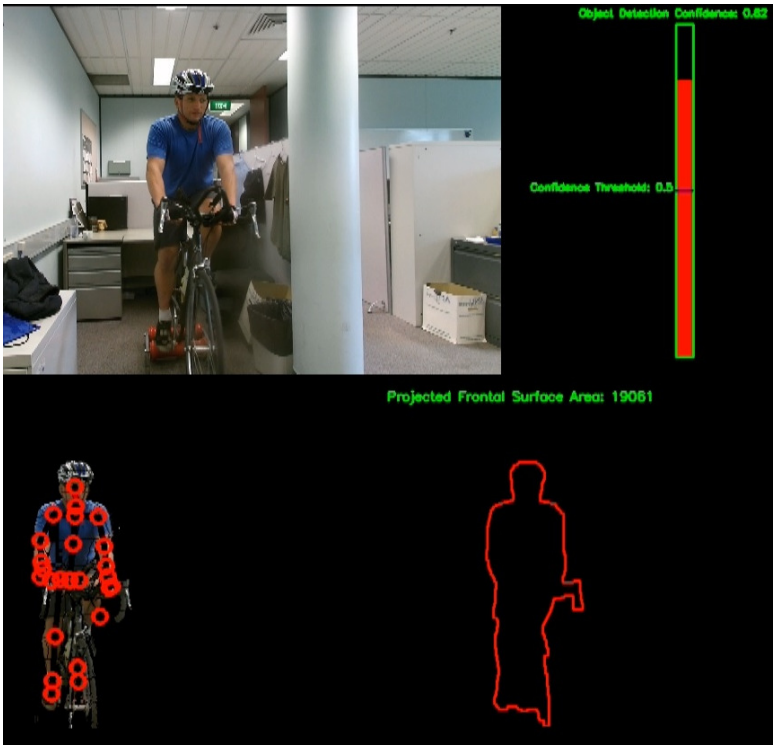


Figure 4.2: Our framework takes RGBD input and outputs a human skeletal pose represented by the joint locations (bottom left) and geometrical shape (bottom right) in real-time. An offline leant classifier over skeletal kinematics enables an activity-specific operation conditioned on a heuristic detection confidence level (top right). Our framework is applied to the estimation of projected surface area of cyclists.

Cycling is a common human activity in rehabilitation and sport, and thus deserves special focus. In competitive cycling, projected Frontal Surface Area (pFSA) is widely accepted to predict aerodynamic drag C_d in cyclists resulting in increased velocity. Large reductions in C_d have been reported as a result of modifications to cycling posture in both wind tunnel testing [63, 83, 97, 106] and Computational Fluid

Dynamics (CFD) simulations [63]. However, some studies reported contradictory results. Defraeye et al. [66] found no correlation between pFSA and C_d in CFD simulations. Drory et al. [76] reported only a weak correlation between pFSA and C_d on 17 elite cyclists tested in a wind tunnel. They demonstrated that reduction in pFSA may in fact be accompanied by an increase in C_d. Whilst reductions in C_d can be found for most cyclists with trial and error modification of posture accompanied by empirical testing of C_d in a wind tunnel, the mechanisms that lead to reduction in drag are still poorly understood, because flow characteristics of *dynamic* bluff bodies remains an open problem. This is amplified by lack of an approach that enables comparison of precise surface geometry at varied experimental conditions. Therefore, development of a method for real-time automatic estimation of surface geometry is needed to advance the current state-of-knowledge to enhance our understanding of bluff-bodies aerodynamics and reduction of aerodynamic resistance on cyclists.

4.1.2 Contributions

In this chapter, our objective is to develop an approach for real-time user-free periodic activity recognition and estimation of geometrical shape using depth information. By exploiting recent advance in inexpensive structured light sensing and pose estimation algorithms, our framework uses kinematic features extracted from the temporal pose to infer activity. To address the objective, we learn a statistical model, which encodes the spatio-temporal characteristics of the target activity, and then uses the model to classify the input skeletal motion. A key to the overall framework performance is a small set of skeletal kinematic features comprised of scalar time series that can be robustly estimated from a noisy input, yet retain the salient characteristics of the motion. Specifically, our spatial relative joint angle features provide invariance to sensor view and differences in individual proportions. The temporal derivative features capture the unique dynamic characteristics of the activity of interest. The features' quaternion representation are compact and stable representation.

We train a classifier for a specific activity recognition based on labelled training data comprised of pose kinematics. The classifier is used for mapping new instances in real-time and outputs an activity class for each detected skeleton. For our demonstrator application, we then use a silhouette-from-skeleton approach, the inverse of the skeletal pose estimation, to yields pixel or voxel labels for a cyclist class. In our example, the object's boundary represents the pFSA of the cyclist.

4.2 Related Work

4.2.1 Shape Reconstruction

Our task requires a concurrent activity recognition and pose estimation. Reconstruction of articulated human motion and human activity recognition from monocular intensity image sequences remains a challenge. Most techniques operate on intensity images, but suffer from cluttered image scene and varied lighting conditions and the

weak local appearance support, which is further hindered by out-of-plane motion and severe occlusions and self-occlusions caused by the motion of the articulated body [95, 148]. Additionally, visual data from regular image acquisition systems can only capture projective information of the real world. Human motion resides in a very high-dimensional parameter space, which is further compounded by the requirement of recovering surface coordinates at every image pixel or shape voxel. Moreover, the uniform human skin colour, as well as repetitive cloth textures, have further complicated feature matching algorithms. Most solutions require restrictive use of calibrated cameras arrays [60, 123] and controlled conditions [98, 241]. For the reconstruction of surface geometry, shape-from-silhouette has been achieved using local convex approximation of the human's surface geometry in a calibrated camera array setting by intersecting generalised cones via back projection of the object's multi-view silhouette and camera parameters [51–53]. The method, however, overestimates the volume of the subject, fails to reconstruct cavities in the subject's surface, and cannot uniquely determine a single pose, since the problem is ill-conditioned. Solutions were also achieved by using an array of synchronised cameras in combination with inertial sensors [188, 232], but are generally complex, restrictive and expensive.

Compared with conventional intensity images, emerging consumer range sensors take advantage of data capture, which provides light-invariant metric that is no longer limited to projective measurements of the geometry. These devices yield depth maps that reflect shape and geometry cues, and provide colour and texture invariance and reduced sensitivity to lighting conditions, which are favourable for segmentation and subsequent object detection. Whilst a single sensor can only provide information on one side of the object in view, 3D (often referred to as 2.5D) body parts were successfully extracted based on a per-pixel classification and mean-shift clustering, from which joint locations could be predicted [215]. Conveniently, skeletal joint tracking algorithms have been provided with consumer devices and can parse a depth-map stream to estimate in real-time the positions of predefined points that constitute a skeleton of a subject (see Fig. 4.1).

4.2.2 Activity Recognition

A typical pipeline of activity recognition involves feature extraction followed by classification. However, direct application of effective and efficient feature descriptors designed for intensity images does not provide satisfactory results [183]. Depth-map features may contain severe occlusions , which make the global features unstable. Further, extracting reliable temporal correspondences by using local differential operators (e.g. gradients) on depth maps is difficult as they are often too noisy due to the absence of texture.

Since there exist a natural correspondence of skeletal pose across time, which is difficult to establish for visual and depth information, skeleton-based features are an intuitive representation for activity recognition from depth-map sequences. A space-time approach extracts global features from the volume (e.g. features from

3D silhouettes), without explicit modelling of temporal dynamics, and then uses a discriminative classifier for recognition. For instance Wang et al. [234] used the skeleton tracking algorithm in Shotton et al. [215] to obtain the joints of the skeleton as reliable interest points for tracking at each frame. Their feature representation captured, in addition to the joint position, the shape of the area surrounding the joint using a local occupancy pattern and a pairwise distance feature. Subsequently, the temporal variation of the features is described by Fourier transform coefficients of the features.

Sequential approaches, In contrast, explicitly model the activity's temporal dynamics by constructing a statistical model over extracted local features from each time instance. Compared to the features-from-silhouettes approach, the skeletal joint features are camera-view and subject-appearance invariant. The pairwise joint position difference is an intuitive compact spatial representation of the skeletal pose in a single frame. Most commonly for activity recognition tasks, the difference is taken with respect to a key or neutral pose (e.g. T-pose) in object coordinates [167, 248]. Other researchers compute planes from joints and measure joint-to-plane distance and motion as features. Yun et al. [254] compute the relationship between a joint and a plane spanned by three joints. Characteristic temporal information about an activity can also be obtained by computing the difference of the joint motion between the two frames. Zhang and Tian [258] computes the temporal difference with respect to a global reference (head-floor distance) for falls detection. Xia et al. [245] builds a statistical posture representation by constructing histogram of 3D joints location by casting the joint positions into 3D cone bins.

The relative pairwise joint orientation, computed from the skeletal joint locations, is a feature that it is invariant to the individual anthropometric differences. Sempena et al. [213] built a feature vector from joint orientation along time series and apply dynamic time warping onto the feature vector for action recognition. In order to recognise gaming actions, Bloom et al. [18] concatenates pairwise joint position difference, joint velocity, velocity magnitude, joint angle velocity, and a 3D joint angle between three distinct joints.

A class of dynamic human activity can be naturally represented by the human pose and its temporal change, as captured by a set of features. The instantaneous skeletal pose at a given time encodes characteristic information about an activity and is commonly used in activity recognition tasks [170, 194, 245, 248]. The pose derivatives further encode information about the dynamics of the activity that has been shown to induce significant improvement over the use of pose-only features in action recognition tasks [209, 256]. Therefore, unlike Oreifej and Liu [183], who bypassed the skeleton tracker, we use the skeleton tracker as basis for our skeleton-based motion features.

4.3 Our Approach

We focus on recognizing a periodic activity using a compact pose representation extracted from sequences of depth maps. Our task presents unique challenges to activity recognition algorithms as a result of severe occlusions (e.g. handlebars), self-occlusions (e.g. hip joint) and segmentation ambiguity due to human-object interaction, which are present in our cycling task. Given a sequence of depth images $\{I_1, I_2, \ldots, I_Q\}$ from a single structured light sensor containing multiple people performing a variety of activities, our goal is to compute a global descriptor which is able to discriminate the class of the target action to other actions being performed.

Despite its limitations, the available skeleton tracker provides a convenient and fast starting point for our task. Our instinct is that within a small range of views around the frontal pose with respect to the sensor, the skeletal tracker is sufficiently robust to detect or infer the loci of the skeletal joints of a cyclist in each frame. The challenge then becomes to construct effective features to train a classifier that can efficiently discriminate between a cyclist and a human performing a different activity. Inspired by Sempena et al. [213] we base our features on spatio-temporal relative pairwise joint differences and their derivatives that are computed from the estimated skeletal joint locations at each frame. Essentially, we convert the points representing the nodes of the skeleton to spatio-temporal joint angle representation. Our intuition is that this feature construction provides not only invariance to individual anthropometric differences, global translation and sensor view, but importantly retains the unique dynamic characteristics of the activity for classification in a concise representation. We then use a supervised learning approach in which we collected ground-truth labelled data for training our classifier that is used to infer the activity. An overview of our method is provided in Fig. 4.3.

4.3.1 Skeletal Kinematics Representation

In order to compute the human pose features, we model the human body as a collection of the body's articulations (joints) whose spatial location is represented as a point in the 3D space. We express a human model as a tree-structured undirected graph $G = (V, E)$, where the vertices $V = \{v_1, \ldots, v_n\}$ correspond to $n \in \{1, \ldots, N\}$ body joints, and the edges $E = \{e_1, \ldots, e_l\}$ corresponding to $l \in \{1, \ldots, L\}$ body's segments, such that $e_l = (v_i, v_j) \in E$ for each pair of connected body joints v_i and v_j (Fig. 4.4). An instance of a skeleton representation of a pose in each frame I_q of a sequence is given by the 3D positions of the skeletal joints $\boldsymbol{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$, where $i \in \{1, \ldots, N\}$ in global Cartesian coordinates. Specifically, we have the 3D Euclidean coordinates $\mathbf{p}_i \in \mathbb{R}^3$ of each joint v_i with respect to the sensor. We extract this skeleton using the tracking system described in Shotton et al. [215] at a rate of 30fps for each person recognised in the data.

From this pose representation, our goal is to infer a set of features that enable the recognition of a periodic activity. Our kinematic descriptor encodes the spatio-temporal relations between joints leading to a low-dimensional descriptor. Zanfir

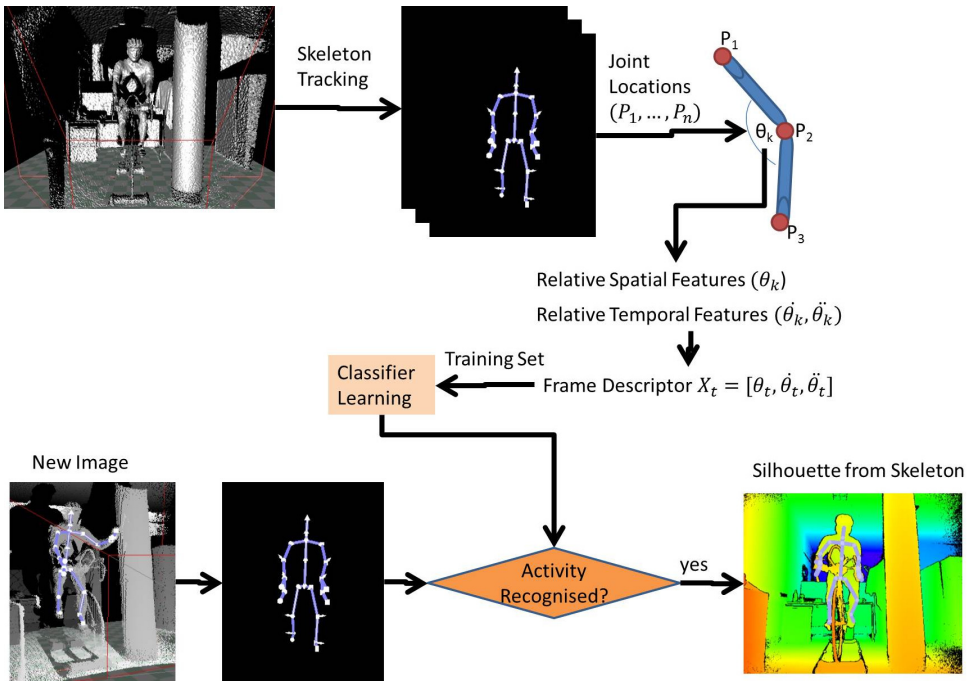


Figure 4.3: Method Overview. Our method takes a surface normal information obtained by a structured light sensor, from which the skeleton is estimated. Spatial and temporal features are estimated and then scored by a trained classifier. If the target activity is recognised, the framework extract the silhouette of the person based on the tracked skeleton.

et al. [256] normalise the pose and its derivatives so they have a unit norm to compensate for noise in the input pose, to account for variations in body anthropometrics, and to remove absolute speed and acceleration in order to preserve relative distributions of different joints. In contrast, we address variations in body anthropometrics by considering the joint's *relative angles* and their time derivatives instead of the pose. This also provides invariance to camera view. In using joint angles we avoid the necessity to normalise segment lengths for the individuals in our dataset.

From the 3D joint positions $\boldsymbol{P}$, the skeleton pose can be described by the lengths of the body's segments $\boldsymbol{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_l\}$, where $\mathbf{s}_l$ is the length and rotation matrix in a Cartesian coordinate system of edge $e_l = (v_i, v_j) \in E$ computed from $\mathbf{p}_i$ and $\mathbf{p}_j$. Typically, the segment angles of the proximal segment are computed with respect to a reference T-pose and expressed in local coordinate system . We convert each rotation matrix to half-space quaternions in order to compactly represent a segmentâĂŹs orientation. Whilst Euler angles enable a more compact representation, they suffer from a mathematical singularity (gimbal lock) and order dependency, a problem that is generally manifested more often in the upper than the lower extremities. Thus, we represent the orientation of each segment s_l by a quaternion vector $\mathbf{s}_l = [q_1, q_2, q_3, q_4]$ with unit norm.

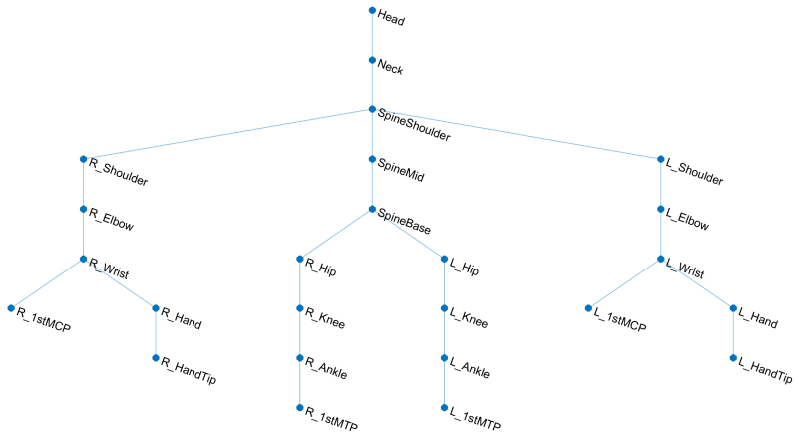


Figure 4.4: Tree structured graph model of a cyclist in frontal view. Note that the model is an acyclical approximation of a natural representation that links the two hip nodes with a pelvic edge. The approximation simplifies the graph model and enables exact inference.

The dynamic characteristics of a periodic motion that is comprised of cyclical actions, such as cycling, is highly repeatable. Characterising the action via differential quantities, such as the relative position, velocity and acceleration of the joints can be expected to be invariant to the anthropometric characteristics of the person performing the action, and be more stable than their actual 3D locations. The pairwise relative position of the joints are more discriminative than the joint positions themselves. Similarly Yao et al. [251] encoded the skeleton motion by relational pose features that describe geometric relation between specific joints in a single pose or a short sequence of poses. Therefore, for each joint $u_k \in U \subset V$ (excluding the skeleton's end points) we compute the *relative* angle $\theta_k \in \mathbb{R}^4$ between its two connected segments e_r and e_s corresponding to the edges (v_i, v_k) and (v_k, v_j) respectively in Joint Coordinate System (JCS). The pose is than represented by $\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_k\}$.

For instance, in order to compute the knee angle θ_k, we calculate the angle between the thigh segment $\mathbf{s}_T$ and shank segment $\mathbf{s}_S$, which were determined from the thigh, knee and ankle skeletal coordinates in $\boldsymbol{P}$, such that

$$\theta_K = \arccos\left(\frac{\mathbf{s}_T \cdot \mathbf{s}_S}{\|\mathbf{s}_T\| \|\mathbf{s}_S\|}\right) \tag{4.1}$$

Arms kinematics generally play an important role in activity classification. Notwithstanding, in our specific activity task of cycling, the arms play no role. The hands are expected to be quasi-staticly in contact with the handlebars for the majority of the time. However, we relaxed the requirement in order to obtain a flexible model that will capture cycling whilst the arms are in the air, by excluding the pose features that

are related to arm kinematics from our training data.

An activity can be represented as a collection of time series describing the change in joint angles over time, or alternatively as a sequence of frame descriptors ordered in time. The first requires dynamic time warping of the joint positions or angles to match a template for a nearest-neighbour classification [150, 171] or a Fourier pyramid [170]. However, for a periodic motion, particularly one that is performed with a large variability of cadence, this is likely to produce significant temporal misalignment that will critically compromise the performance of the activity classifier. Instead, we opt for the second approach, but in contrast with Miranda et al. [170], Nowozin and Shotton [178], Sempena et al. [213] and others, we do not seek to assign key poses or action points. Given that the nature of our target activity is periodic, we consider each frame instance of the motion equally. We deduce that since our feature vector captures temporal characteristics of the activity, our classifier will be able to recognise the activity.

We view the pose as a continuous and differentiable function of the joint angles. Further, we argue that due to the inertial properties of human limbs and the latency in muscle actuation, a human activity can be well approximated by a quadratic function of the segmental motion, expressed in terms of the first and second derivatives of the joint angles with respect to time. Thus, its second-order Taylor approximation in a window around the current time step t_0 would be defined by expanding around the current pose $\boldsymbol{\Theta}(t_0)$ based on the first and second order derivatives, $\dot{\boldsymbol{\Theta}}(t_o)$ and $\ddot{\boldsymbol{\Theta}}(t_0)$ such that

$$\boldsymbol{\Theta}(t) \approx \boldsymbol{\Theta}(t_0) + \dot{\boldsymbol{\Theta}}(t_0)(t - t_0) + \frac{1}{2}\ddot{\boldsymbol{\Theta}}(t_0)(t - t_0)^2. \tag{4.2}$$

For each frame we compute a feature descriptor as a concatenation of the 3D pose $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_n]$ and its first and second derivatives $\dot{\boldsymbol{\Theta}}(t_0)$ and $\ddot{\boldsymbol{\Theta}}(t_0)$. The derivatives are estimated such that $\dot{\boldsymbol{\Theta}}(t_0) = \frac{1}{2}\mathbf{w}(t_0) * \boldsymbol{\Theta}(t_0)$ and $\ddot{\boldsymbol{\Theta}}(t_0) = \frac{1}{2}(\dot{\mathbf{w}}(t_0) * \boldsymbol{\Theta}(t_0) + \mathbf{w}(t_0) * \dot{\boldsymbol{\Theta}}(t_0))$, where $\mathbf{w}(t_0)$ and $\dot{\mathbf{w}}(t_0) \in \mathbb{R}^3$ are the angular velocity and acceleration, respectively, estimated numerically by using a temporal window of 3 and 5 frames centred at the current by $\mathbf{w}(t_0) \approx \mathbf{w}(t_1) - \mathbf{w}(t_{-1})$, and $\dot{\mathbf{w}}(t_0) \approx \mathbf{w}(t_{+2}) + \mathbf{w}(t_{-2}) - 2\mathbf{w}(t_0)$ respectively, and $*$ represents the quaternion multiplication operator. Our final kinematic descriptor $\boldsymbol{X}_t$ for frame at time t is obtained by concatenating the joint angles and their derivatives over time, such that $\boldsymbol{X}_t = [\boldsymbol{\Theta}_t, \dot{\boldsymbol{\Theta}}_t, \ddot{\boldsymbol{\Theta}}_t] \in \mathbb{R}^{1\times(12K)}$.

We assume that all 3d joint positions of the human body are available in each frame. Whilst occluded or substantially noisy joint positions are identified by the skeletal tracking algorithm as 'inferred', we consider them as if they were 'tracked' under the assumption that tracking recovery will eventuate throughout the cycle and be treated as noise.

In summary, our features convert 3D point trajectories to a set of time series that represent the relative motion of the body segments that are robust to noise and invariant to sensor view and individual anthropometric proportions, and are well suited as input to a classifier.

4.3.2 Training

By defining the feature vectors of a periodic activity, we train a boosted classifier with the computed features as input, whose objective is to accurately classify new input to the correct class. We build a model of the periodic activity by using a training set that constitutes of positive and negative samples of participants cycling.

4.3.3 Boosted Random Forest Classifier

Decision trees [23] are tree-structured learning methods that map complex input spaces into discrete output spaces by splitting the original problem into smaller ones, solvable with simple predictors. A decision tree is comprised of split and leaf nodes. A split node in a tree consists of a split function which tests the input feature for information gain. The result directs a data sample towards one of its child nodes. During training, the tests are chosen in order to group the training data in clusters where simple models achieve good predictions. Such models are stored at the leaves and computed from the annotated data which reached each leaf at train time.

Given a test instance U with description $F = \{f^k\}$, where k is the index of the feature, the split function assigns U to the left split if $f^k < \tau$, or to the right split otherwise, where τ is a decision threshold. Therefore, training the forest is equivalent to optimising the parameters τ and k at each node of each tree to achieve the best split, i.e. minimal error. Our goal then is given the training set S, to produce a classifier $\hat{C}(X) \in \{1, -1\}$ and estimate the probability of the class labels $P(Y = +1|X)$.

Standard decision trees alone suffer from overfitting. However, a collection of decorrelated randomly trained decision trees has been shown to have high generalisation power [22]. In particular, individually each of the features may only provide weak support, but collectively, using bagging, they constitute a strong classifier. Thus, To bag our forest of tree classifiers $C(S, x)$, we draw bootstrap samples $(S^1, \ldots, S^B)$ each of size R with replacement from the training data. Then $\hat{C}(x) = MajorityVote\{(S^{*b}, x)\}_{b=1}^{B}$.

Rooted in bagging, which seeks to average noise and unbiased models to create a model with low variance [94], Adaptive Boosting (AdaBoost) assumes that most of the trees can provide correct classification prediction for most of the data. AdaBoost is thus an ensemble that consists of multiple trees (learners) that are employed to build a stronger classifier via iterative improvement by accounting for the incorrectly classified examples in the training set. It assigns an equal weight α to each weak classifier c_t, where $t = \{1, \ldots, T\}$ from the pool. At each iteration, it then calculates confidence α_t for c_t, where $\alpha_t = 0.5 \times \log(\frac{1-\epsilon_t}{\epsilon_t})$, and ϵ_t is the error value of learner t, and increases the weights of the incorrectly classified examples. The output is a strong classifier $C(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t c_t(x))$.

The resultant strong classifier can be thought of as a taylor series, where the number of terms chosen results in a computation versus accuracy tradeoff. This is done by keeping the strongest week classifiers and truncating the weak classifiers. AdaBoost has shown to be very effective mutliclass classifier due to its capability to

handle large training sets, high generalization power, fast computation, and ease of implementation. For our task, we use the OpenCV AdaBoost implementation for binary classification of the target activity, but our method can readily be expanded to multiclass activity classification.

4.3.4 Silhouette from Skeleton

Motivated by our specific target task, our objective is to develop an approach for real-time user-free periodic activity recognition and estimation of geometrical shape. Our activity classification is underpinned by features extracted from the estimated skeletal tracking algorithm, which can be viewed as a skeleton-from-silhouette approach. Following a successful recognition of the activity using the trained classifier, we want to output the object's geometrical shape. This can be viewed a silhouette-from-skeleton approach, the inverse of the skeletal pose estimation, and is already available in real-time from the previous step. Therefore, we can use the skeleton location in the segmented image to infer pixel or voxel labels by segment-group membership, which corresponds to the object's shape. In our example, the object's boundary represents the pFSA of the cyclist (see Fig. 4.5).

We note that whilst we output the pFSA as an area in 2D pixel units, it is trivial to convert those units to real world units by either a simple calibration routine or by averaging the sensor depth information for the voxels corresponding to the subject. We do not perform this conversion for two reasons; 1) our motivating example is to subsequently study the relationship between aerodynamic drag of cyclists and pFSA, which is often debated in the literature, and 2) in the study of this relationship, the absolute values of pFSA are of no interest. Rather, it is the *relative* change in pFSA in response to a change in a cyclist's pose that is of interest. Therefore, for our purpose pixel area values suffice.

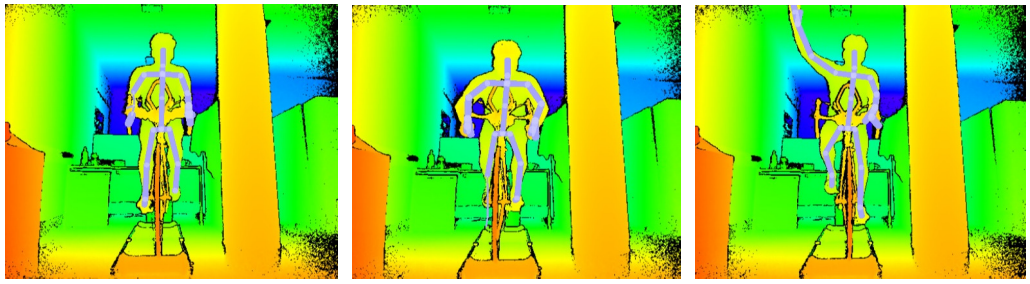
4.4 Experimental Results

In the absence of publicly available dataset for our target task, we evaluate our approach on a new dataset described below, for which we provide annotation of activity labels. In this section we describe the dataset and results of our experiments.

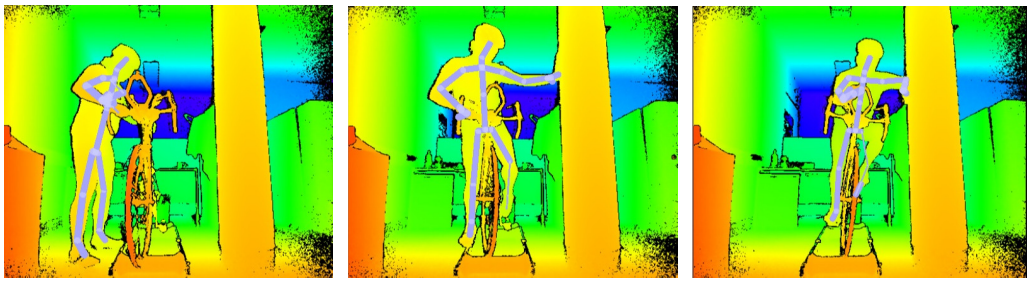
4.4.1 Dataset

The majority of publicly available datasets of multimodal imagery of dynamic human motion consists of human gait and activities with little or no object interaction. In addition, the activities that are found in those datasets can be described as a discrete non-periodic gestures. We did not find a dataset of a repeated periodic activity or cycling specifically. Whilst our learning framework is suitable for multi-class classification, we focus our experiments on the binary classification of and activity as 'cycling' or 'non-cycling'. Therefore, we use multimodal image sequences of cycling, which were performed on cycling rollers. Cycling on rollers, whilst enabling

convenient image capture in a constrained space, retains the natural motion of the bicycle and cyclist. In contrast to studies of cycling where the bicycle is constrained by a fixed ergometer or trainer, our approach ensures that the cyclist's motion most closely mimics realistic outdoor cycling. The capture was performed by both a fixed and moving sensor within a small range of sensor view change and distance in the frontal plane. Therefore, our dataset contains varied appearance and depth information of bicycle and cyclist from a range of orientations. This, in addition to the constant motion of the bike on the rollers, ensures that our model is invariant to small changes of sensor view.



(a) $+1$



(b) -1

Figure 4.5: 2D segmentation and pose estimation from depth. Images in a depict examples of positively labelled $(+1)$ instances of cycling from our training dataset. Images in b depict examples of non-cycling, i.e. negatively labelled (-1) instances of cycling.

Our dataset consists of 10 image sequences totalling $N(29306)$ frames of RGBD images of 3 participants that were hand labelled for activity class. The labelled training samples include 20842 positive cycling samples and 8464 negative samples, such that $S = (x_n, y_n)_{n=1}^{N}, x \in X$, where $x =$ skeleton feature data, $y \in Y$, where $Y = \{-1, 1\}$.

The image sequences include a person mounting and dismounting a bicycle, stationary sitting on a bicycle, and even falling from the bicycle and the rollers. All frames were labelled as 'cycling' or 'non-cycling' (see Fig. 4.5). To enhance the specificity of our activity recognition model, our criteria for labelling a frame as 'cycling' is whether with respect to its nearest temporal neighbour, it belongs to an image sequence of cycling at a cadence greater than 10rpm. This ensures that the

temporal dynamics of the activity are captured by our classifier. Hence, frames belonging to a sequence that describes stationary sitting on a bicycle, mounting and dismounting, pedalling at a cadence lower than 10rpm, or falling have been labelled as 'non-cycling'. All image sequences include cycling at a variety of cadences from 10 to 120rpm to ensure that our classifier captures activity characteristics that change with the speed of execution. For instance, variability in cycling mechanics may be the result of change to muscle recruitment pattern as the cadence increases. The dataset includes images of a person mounting and dismounting the bicycle, stationary and moving. Some involve hands off the handlebars.

We evaluate the performance of the activity classification following a k-fold cross validation scheme, where $k = 8$. At each fold, the dataset was randomly partitioned following a standard 70/30% for training and testing respectively. Note that all kinematic features have been computed for all frames in our dataset prior to partitioning. This is necessary because of the dependency of the angle derivative features on the frame's immediate temporal neighbours. Using a detection threshold of 0.039 , our classifier achieves 98.3% accuracy on positive samples and 0% false positives on our training data.

Finally, we wish to get an insight into which weak classifiers contributed most to the construction of the strong classifier. Table 4.1 shows the top 10 contributing weak classifiers to our final activity classifier. Not surprising the feature derivatives from the knee, ankle and hip are the prominent contributors, as the lower limbs perform the majority of the motion in the target activity of cycling. Likewise, the neck and spine angles, whilst not participating in the motion, are stable characteristics that can be used to distinct the activity from other actions.

Table 4.1: Top contributing weak classifiers

Feature	Associated skeletal joints	α
$L\_\dot{K}nee$	(L_Hip,L_Knee,L_Ankle)	1.197646
$Neck$	(Head,Neck,SpineShoulder)	0.659523
$L\_\ddot{K}nee$	(L_Hip,L_Knee,L_Ankle)	0.487340
$R\_\ddot{K}nee$	(R_Hip,R_Knee,R_Ankle)	0.432138
$L\_\ddot{A}nkle$	(L_Knee,L_Ankle,L_1stMTP)	0.424726
$R\_\ddot{K}nee$	(R_Hip,R_Knee,R_Ankle)	0.419117
$R\_\dot{A}nkle$	(R_Knee,R_Ankle,R_1stMTP)	0.396936
$L\_\dot{A}nkle$	(L_Knee,L_Ankle,L_1stMTP)	0.381479
$L\_\dot{H}ip$	(SpineBase,L_Hip,L_Knee)	0.365908
$SpineMid$	(SpineShoulder,SpineMid,SpineBase)	0.340869

4.4.2 Qualitative Results

In order to qualitatively test our framework, we performed additional tests that included the presence of additional persons in the capture space, which resulted in various levels of occlusions. We also experimented with various heuristic activity recognition confidence threshold values. Our experiments show that a threshold value ≥ 0.4 yields robust activity recognition and subsequent shape extraction in a

variety of sensor view and depth, lighting and subject appearance (e.g. apparel, helmet, glasses etc.) conditions. An example qualitative output is provided in figure 4.2 and associated demo image sequences.

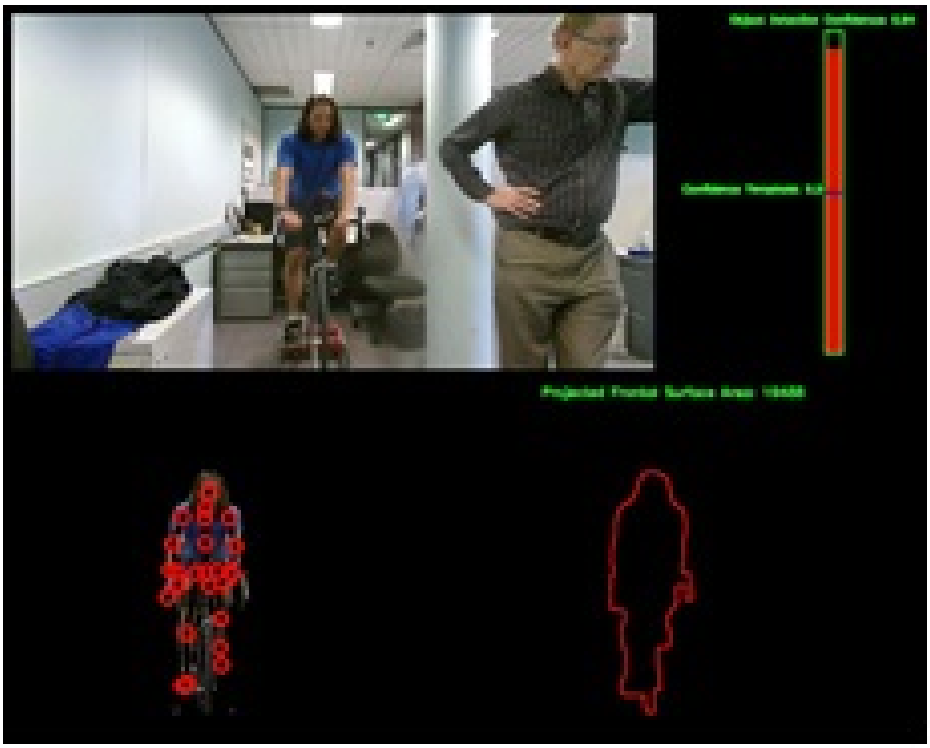


Figure 4.6: A sample result from qualitative testing of our framework for activity recognition and shape estimation. Our method successfully identifies the person performing the target activity and ignores individuals performing other actions. It then outputs the estimated geometrical shape of the object

4.5 Discussion

In this chapter, we presented a real-time framework for activity recognition and shape estimation of a human performing a periodic motion that involves human-object interaction from depth information. The key components of our framework include spatio-temporal kinematic features computed from a human pose and motion and point cloud information obtained from noisy depth information by a structured light sensor, and boosted forest activity classifier. The results we presented demonstrate that the framework is robust to noise, invariant to differences in individual proportions, lighting conditions and sensor view, and can robustly discriminate the target activity from a human performing alternate actions.

4.5.1 Limitations and Future Work

We represent an activity as a sequence of frame descriptors ordered in time by the rational of section 4.3. An alternative approach would represent the activity as a collection of time series describing the change in joint angles over time. This, however, requires dynamic time warping of the joint angles to match a template [150, 171], else a significant temporal misalignment may be produced that will compromise the performance of the activity classifier. An alternative classification approach could use regression decision tree that enables mapping of a non-linear complex input space to *continuous* output parameters by dividing the problem into a set of simpler sub-problems. Thus the approach is particularly suitable for time-series data and has been successfully employed for object detection, pose estimation and tracking [55, 89, 95, 102, 234]. The feature vector in this case would represent a sequence of joint angle observations over a time period $t_1, \ldots, t_q$ that is defined as a joint angle trajectory $T_k = \{X^k_{t_1}, \ldots, X^k_{t_q}\}$, and $T = \{T_1, \ldots, T_k\}$ as the set of all k joint angle trajectories in an image sequence. Nevertheless, the challenge remains of classification of vector space curves into action categories because it suffers from rate variations, temporal misalignment and noise.

The framework also suffers from a number of limitations; due to its ready availability, the majority of current approaches use features computed from the skeleton tracking that is provided with consumer sensors. However, the skeletal tracking algorithm has been trained on a dataset of humans in a frontal view to the sensor [215]. Consequently, the algorithm does not generalises well to unseen scenarios, and only works well in an occlusion-free frontal plane conditions, when a human subject is facing the sensor in an upright position. The reliability rapidly diminishes with occluded body parts, and partial or changed view and pose. Such scenarios include a side sensor view, which inevitably induces severe self-occlusions, result in less reliable body part labelling and subsequently a contaminated feature training dataset for our activity classifier. The original skeletal tracker [215] was trained on a million images consisting of real but mostly syntheticly rendered poses on a 1000 core cluster. Evidently, the cost of training a robust system for each scenario is prohibitively high. Therefore, past research focused on simple mostly discrete human actions, such as gait and grasping.

One possible solution for this problem is to use intensity image based pose estimation techniques to generate training samples on the corresponding depth information by labelling body parts based on parts-based appearance. Drory et al. [80] proposed a markerless approach for pose estimation of cyclists in the sagittal plane from uncalibrated intensity images using mixture of parts classification. The approach can be used to label body parts in the sagittal plane from which depth skeleton features can be trained. Conceptually, this is similar to the hybrid approach of Lei et al. [145], who used depth to extract hands and objects and intensity images to track hands. However, the disadvantage of the approach is that it does not explicitly handles occluded body parts. Nevertheless, correct labelling of one side of the body may impose sufficient constraints on the search space of possible poses, to force

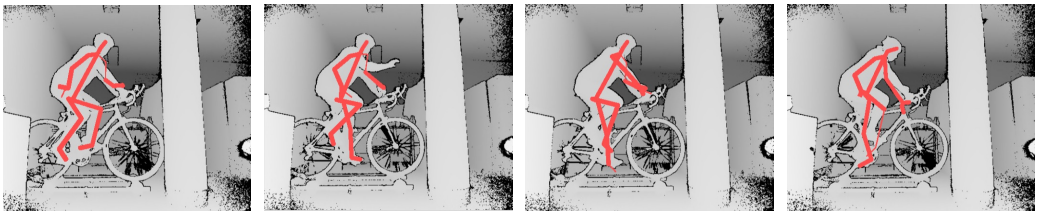


Figure 4.7: Examples of sagittal view skeletal tracking failures. The skeletal tracking algorithm of Shotton et al. [215] was trained on a million real and synthetic samples of upright humans performing actions in the frontal view relative to the sensor. Thus, it has low adaptability to unseen scenarios.

a correct 3D skeletal pose from the depth information. For instance, to infer the overall direction in which the subject is facing with respect to the sensor, we can compute the cross product between the shoulders distance vector (R_Shoulder - L_Shoulder) and the trunk length vector (SpineShoulder - SpineBase).

An alternate hybrid solution will incorporate a learnt statistical activity shape and appearance prior to the skeletal tracking algorithm to encourage it to arrive at the desired solution. Drory et al. [79] has proposed a statistical shape models of cyclists in frontal and sagittal views. The models along with appearance models were used to regulate an evolving curve to estimate pFSA from intensity images in Drory et al. 81(Submitted for publication).

Similarly, due to the periodic nature of our task, it is possible to incorporate a learnt statistical model of our kinematic features in our framework, and is intended for future work. This requires a robust segmentation and alignment of motion features for all cycles in a manner that is similar to the work of Ormoneit et al. [184] on gait data.

Handling occluded body parts remains a persistent challenge for pose estimation approaches. A more robust solution will use a synchronised multi-view sensor array that ensures that the depth information of all body parts is captured. Appealingly, this will also enable a 3D shape reconstruction as opposed to the 2D or 2.5D output from our framework, and is intended for future work. However, this presents a challenge of handling reflections created by multiple structured light streams.

Additional persistent challenge for skeletal tracking algorithms based on depth information is correct body part classification when a human is interacting with an object. The skeletal features do not capture information about the surrounding objects. Due to lack of texture and appearance information and its reliance on segmentation at its core, body part misclassification is frequent as a result of segmentation ambiguity. Whilst it did not manifest as a problem in our frontal view case, sagittal view data demonstrates the problem.

When modelling person-object interaction, object detection and tracking have to be combined. A hybrid intensity-depth approach, which can be used to address this problem, incorporates an intensity-based object detector and impose an explicit or

learnt human-object interaction model to reduce the search space of possible body parts and pose configuration. For example, Lei et al. [145] use depth to extract hands and objects from a tabletop and intensity images to track hands. Then they combine a global feature using PCA on hand trajectories, and a local feature using bag-of-words of trajectory gradients to recognize kitchen activities from an overhead multimodal sensor. Similarly, Koppula et al. [135] explicitly consider human-object interactions by training discriminative object detection classifiers from the intensity information that are tracked with the human skeleton from depth. An explicit human-object interaction approach was used by Drory et al. [77] (Submitted for publication) to improve pose estimation of cyclists from intensity images. An alternate approach was used by Wang et al. [234], who used local occupancy pattern features based on the point cloud information at a particular joint, for example, when a person grasps a cup.

The skeleton tracking algorithm infers skeletal pose separately at each image frame. No temporal constancy constraints are imposed on the classification of body parts. Therefore, the method is sensitive to successful estimation of the skeletal pose. Thus, our method, whilst capturing the dynamic characteristics of the activity through the construction of spatio-temporal kinematic features, remains sensitive to the performance of the skeletal tracker, on which it is underpinned. A significant consequence of the approach is that the skeletal linkage rigidity assumption is regularly violated. It is common to observe significant changes in body part lengths across consecutive frames as a result of the upright pose assumption. Important future work would impose a rigid or semi-rigid linkage skeleton, for example with respect to a reference pose (e.g. T-pose), to constrain the search space of likely poses.

Because of the high variability of human ambulation, it is impossible to design a discriminative classification approach for the general case of activity recognition. The performance of the approach critically relies on time consuming and costly process and the availability of a large quantity of training samples. Further, the generalisation of the approach can only be achieved through the addition of adequate training samples, as its adaptability to unseen dynamic postures is low and is typically manifested by poor performance were occlusions exist. Notwithstanding, Our framework can be trivially extended to recognise other activities. The features can be extended to model, for instance, person-to-person interaction, if the skeletal joint loci of multiple persons are known [254]. This can be achieved by using our classifier for multi-class classification as opposed to binary as was used here.

Training a classifier using a boosted forest requires very large and varied training dataset to avoid the risk of overfitting a model to the data. Therefore, a further risk associated with our method is that the relatively small number of participants and capture conditions may have biased our classifier. The performance of our qualitative testing on image sequences that are not in our dataset indicate that our classifier is robust. Nontheless, our framework would benefit from supplementary samples of additional participants in a varied capture environment.

4.6 Conclusions

In this chapter, we described a real time framework for periodic activity recognition and geometrical shape estimation using depth information from a structured light sensor. We learn a statistical model over a set of features that were extracted from the estimated skeletal pose. The features encode the relative spatio-temporal characteristics of the target activity. Our spatial relative joint angle features provide invariance to sensor view and differences in individual proportions. The temporal derivative features capture the salient characteristics of the dynamic activity of interest. Our trained boosted classifier is robust for detecting the target activity and allows the extraction of estimated geometrical parameters of the subject efficiently.

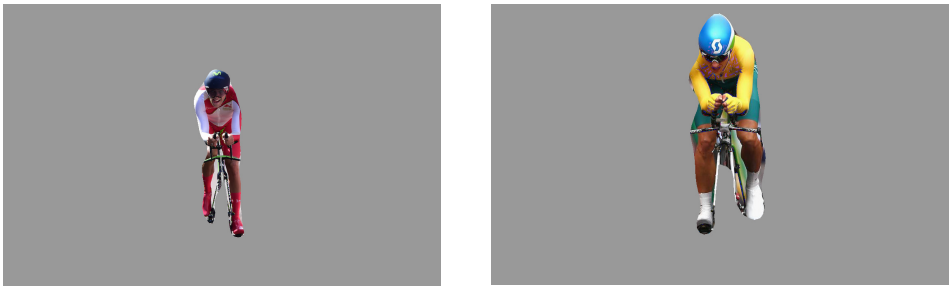
Chapter 5

Estimating the Projected Frontal Surface Area of Cyclists from Images using a Variational Framework and Statistical Shape and Appearance Models

Abstract

We present an approach to estimating the projected Frontal Surface Area (pFSA) of cyclists from unconstrained images. Wind-tunnel studies show reduction in cyclists' aerodynamic drag through manipulation of the cyclist's pose. Whilst the mechanism by which reduction is achieved remains unknown, it is widely accepted in the literature that the drag is proportional to the cyclist's pFSA. In this chapter we aim to develop a repeatable automatic method for pFSA estimation for the study of its relationship with aerodynamic drag in cyclists. The proposed approach is based on finding object boundaries in images. An initialised curve dynamically evolves in the image to minimise an energy function designed to force the curve to gravitate towards image features. To overcome occlusions and pose variation, we use a statistical cyclist shape and appearance models as priors to encourage the evolving curve to arrive at the desired solution. Contour initialisation is achieved using a discriminative object detection method based on offline supervised learning that yields a cyclist classifier. Once an instance of a cyclist is detected in an image and segmented, the pFSA is calculated from the area of the final curve. We demonstrate that our method is successfully applied to cyclist images. We discuss the performance of our method under occlusion, orientation, and pose conditions. We show that our method successfully estimates pFSA in cyclists and opens new vistas for exploration of the relationship between pFSA and aerodynamic drag.

5.1 Introduction



(a) (b)

Figure 5.1: The curve gravitates toward image features and arrives at equilibrium. This results in an object segmentation that provides an estimation of the projected frontal surface area of a cyclist. Top right image shows some failure with partially occluded second object.

Cyclist-bike systems' aerodynamic properties greatly affect cycling performance measured by the time to completion of a race. A cyclist's velocity is dependent on the total resistive forces acting on the system. At racing speeds greater than $14m/s$, aerodynamic drag force C_d accounts for more than 90% of the total resistive forces acting on the cyclist-bike system [141, 159], of which up to 72% are attributed to the cyclist alone [63, 108, 141].

Recent Computational Fluid Dynamic (CFD) simulations showed that aerodynamic drag of a cyclist, as typical to bluff bodies, is dominated primarily by form drag associated with the geometric shape of the cyclist-bike system and the vortical contribution of wake flow [17, 65, 66, 160]. This is in contrast to highly streamlined airfoil bodies, which are dominated by the viscous drag component due to velocity slowdown in the boundary layers and associated skin friction.

The form drag can be reduced principally by modifications to the shape of the cyclist-bike system through changes to the riding posture [63, 66, 97, 106, 227]. Given the potential for improved cycling performance, past research has focused on reduction of aerodynamic drag through modification to riding posture [63, 83, 93, 97, 227] or equipment [97]. However, despite the growing interest in cycling aerodynamics, current approaches rely on *ad-hoc* manipulation of cyclists' posture that minimise average aerodynamic drag as measured in wind tunnel testing [26, 63]. While reductions in aerodynamic drag can be found for most cyclists, the mechanisms that lead to reduction in drag are still poorly understood, because flow characteristics of *dynamic* bluff bodies remain an open problem.

Large reductions in aerodynamic drag have been reported as a result of modifications to riding posture in both wind tunnel testing [63, 83, 97, 106] and CFD simulations [63]. In those studies, however, the modifications to rider postures were achieved through manipulation of peripheral equipment (e.g., bicycle type, handlebar height and elbow pads location) or an ambiguous subjective posture description

making precise and repeatable cyclist's geometrical shape characterisation impossible(e.g., 'upright', 'dropped' postures). It is uncertain if the various studies describe similar geometrical structures or whether the different cyclist poses lie on a geometrical continuum that would enable interpolation or extrapolation to other poses. This highlights the necessity for a development of a shape modelling approach that enables precise characterisation and comparison of experimental conditions to be adopted in future research. In the following section we provide a brief review of existing approaches and techniques that have been developed to characterise cyclist postures. In this chapter we focus on a method that directly measures or predicts the cyclist's pFSA to enhance the understanding of its relationship with aerodynamic drag F_D.

5.1.1 Contributions

In this work, we seek to exploit advances in object detection and segmentation algorithms to develop a method for pFSA estimation in cyclists. We propose a new approach based on the variational framework that includes shape and appearance prior. The initialisation is achieved through object detection using a fully supervised learning approach based on robust image features. We describe our algorithm in section 5.3, and present our experiments in section 5.4. The potential benefits of the approach include:

1. near real-time, repeatable, user-independent direct geometric characterisation of a cyclist's posture.

2. it enables direct comparison of pFSA and Drag Area (CdA) in empirical studies, and facilitates the study of the dependency of Cd on the Reynolds number.

3. it enables field estimation of pFSA for monitoring posture deviations from lab-prescribed posture.

4. extended to 3D, it will enable dynamic CFD simulation of cycling.

5.2 Technical background

5.2.1 Indirect Characterisation of Cyclist Position

Past studies focused on the differences between upright, dropped and time-trial riding postures [63, 97, 227]. Unsurprisingly, they showed moderate agreement between the reduced torso angle and aerodynamic drag from upright to time-trial posture (see Defraeye et al. [63] for partial summary). We reason that if a cyclist's torso is approximated by a cylinder with a large longitudinal to transverse axial ratio, it is evident that it would cause greater aerodynamic drag in its upright position than its horizontal position [175, 176, 253]. Having discovered this relationship some decades ago, when optimising performance of importance, competitive cyclists ride in subtle

variants of the time-trial posture. Consequently, the upright posture is no longer relevant in the context of minimisation of aerodynamic drag. However, no studies were found in the literature that were successful in finding a relationship between aerodynamic drag and these subtle variants modelled using torso angle. It is noted that the dropped posture remains relevant in competitive cycling sprints. Furthermore, by modelling a cyclist's posture by the angle of the torso's longitudinal axis alone, the relationship of the principal geometrical dimension that causes drag, the surface normal to the approach wind, is lost and cannot be recovered.

Additional simplified and indirect characterisations of the geometric changes to the cyclist-bike system were used to explain reduction in aerodynamic drag. García-López et al. [97] used kinematic variables including profile length and height and arm-torso and arm-forearm angles to distinguish between experimental conditions relating to posture changes. These, however, showed large participant and trial specific variations and are neither consistent nor reproducible. Drory et al. [76] showed anthropometric measures of segment lengths and girths that characterise body proportions were not correlated with aerodynamic drag. Defraeye et al. [63] attempted to control the variability in rider postures using a physical constraints positioning system. They studied three rider posture (upright, dropped and time-trial positions) and attempted to characterise the wake flow field using CFD, but only achieved moderate agreement with wind tunnel empirical data. Evidently, indirect shape characterisation methods are inadequate to describe the geometric shape of a cyclist and its relationship to aerodynamic drag.

5.2.2 Direct Geometric Characterisation of Cyclist Position

5.2.2.1 Cyclist Geometry for CFD

CFD can provide high-resolution simulated flow field and drag information on a cyclist as an alternative to empirical ad-hoc wind tunnel testing. To achieve valid results, recent studies obtained high resolution digital 3D models of cyclists using laser scans [15–17, 63, 65, 66, 72, 160]. Using the scanning approach enables the capture of a cyclist's pose whilst holding a static posture only. Consequently, only flow field and drag data CFD simulations of static cycling have been reported to date. In practice, however, large object segments are rotating in the direction of motion, the entire object oscillates in the plane transverse to the motion, and changes to the overall geometry take place and disturb the airstream. It is yet to be demonstrated that the study of static cycling (both empirical and simulated) is a valid approximation of dynamic cycling. This inadequate physical model representation results in a significant deficiency in CFD cycling models making their findings uncertain. In addition, the high cost, infrastructure and post-processes that are required preclude the use of laser scanning from near real-time and in-field use. An alternative geometry capture and modelling approach that facilitates CFD simulation of dynamic cycling is yet to be reported in the literature and constitutes a necessary step to establish the validity of investigation of cycling in wind tunnels and CFD simulation.

5.2.2.2 Estimation of Projected Frontal Surface Area

Aerodynamic drag is quantified by the relationship between the drag force $F_D(N)$, the object's Projected Frontal Surface Area (pFSA) $A(m^2)$ and a dimensionless drag coefficient C_D, such that $F_D = C_D A(\rho U^2/2)$, where ρ is the air density (kg/m^2) and U is the approach wind speed (m/s). Two techniques have been reported for direct estimation of pFSA; weighing cyclists' photograph cut-outs [33, 97, 126, 181, 182, 222], and manual digitization of coordinate data [62]. Both methods are achieved through laborious and time consuming post-processing. Furthermore, Jensen et al. [127] reported significant intra- and inter-tester differences in manual digitization of pFSA as well as test repeatability differences the authors associated with the learning effect. Consequently, CdA is typically reported instead of Cd, as it avoids explicit determination of the pFSA. This hinders exploration of the dependency of aerodynamic drag on pFSA and Cd. Moreover, it imposes an impediment for pFSA estimation to be adopted as valuable tool to inform decision making in wind tunnel aerodynamic testing, training environment monitoring and competition applications where near real-time performance is required.

A few attempts have been made to explain the reduction in aerodynamic drag via changes to pFSA, the cyclist's projected surface normal to the fluid displacement and a direct geometric shape characteristic [33, 70, 126, 181, 182, 222]. This has led to wide acceptance that pFSA is directly proportional to aerodynamic drag under constant drag coefficient conditions. In contrast, Drory et al. [76] reported only a weak correlation between pFSA and aerodynamic drag on 17 elite cyclists. They demonstrated that reduction in pFSA may in fact be accompanied by an increase in CdA. Defraeye et al. [66] found no correlation between pFSA and aerodynamic drag in CFD simulations. Nevertheless, the prevailing view supports reduction in pFSA as a measure to reduce aerodynamic drag.

Despite the uncertainty about the relationship between pFSA and CdA, some researchers suggested that cyclists' physiological parameters normalised to pFSA are good predictors of time trial cycling performance [33, 168, 182]. Moreover, some researchers have attempted to estimate pFSA indirectly from the ratio of anthropometric parameters including body dimensions and mass [97, 117, 182, 222] and Helmet length and angle [11]. However, these were shown to have weak correlation with pFSA [29, 97, 221]. Considering the conflicting views it is therefore necessary to establish the nature of the relationship between pFSA and aerodynamic drag.

The utility of the indirect methods remains limited to static laboratory use. Thus, near real-time in-field monitoring of cyclists' adherence to a lab-prescribed position is not yet possible. This highlights the need for a pose estimation method that can be used in the daily training environment and out-of-laboratory applications. Computer vision techniques for object detection, shape reconstruction and image segmentation, present an attractive alternative to the above techniques for repeatable automatic near real-time pFSA estimation extracted from images.

5.2.3 Geometric Active Contours

Object segmentation remains a fundamental problem in the computer vision domain that aims at extraction of structures of interest from images. Geometric Active Contours (GAC) is a segmentation approach for finding object boundaries in images that has been popularised in medical imaging. In GAC, initialised curves dynamically evolve in the image to minimise an energy function. The energy function is designed to force the curve to gravitate towards an image feature such as an intensity edge [35, 130]. The curve becomes stationary when it coincides with the structure boundary and the energy function is at its minimum [130]. However, boundary finding algorithms do not guarantee arrival at the desired structure due to noise, occlusion or boundary discontinuities [172]. These methods are generally local and are therefore sensitive to the position of the initialised curve [41, 147] and 'leakage' through the boundary of the object if the feature is not prominent enough [43, 54].

GAC sensitivity to curve initialisation is commonly addressed by either semi-automatic 'user in-the-loop' approaches that require user input of appropriate initialisation [157, 255], or by using object detection algorithms which provide the Region of Interest (ROI) of the desired object [133]. In the context of near real time pFSA estimation in testing, training and competition environments, the first approach is unlikely to be adequate or adopted by the user. Therefore, while more challenging, only the later approach is considered further in this work (see further details in section 5.2.5).

5.2.4 Statistical Shape Model

GAC's tendency to 'leak' through weak underlying boundary features can be mitigated by apriori knowledge of the expected shape or appearance of the desired object [43, 54, 147]. Gaps in the boundary can then be linked in a supervised semantically meaningful manner. In order to incorporate prior shape knowledge into a GAC, a geometric shape model is required. Active Shape Models (ASM) is a technique in which a statistical shape variation model is developed from a set of n manually labelled corresponding boundary points in training images [50]. The training shapes are aligned with respect to similarity transformation into a common coordinate frame using an iterative technique which minimizes the least squared error between the points, until convergence. The resultant set of landmark points can be viewed as a point cloud that lies on a manifold in a $2n$-D space forming a Gaussian distribution of likely correlated vectors. The statistical shape model is obtained via Principal Component Analysis (PCA) performed in a high dimensional vector space of the sampled landmarks. The technique is used extensively in the medical imaging domain for the segmentation of known anatomical structures [133, 157, 255].

Statistical shape models have been previously introduced into GAC implementation. Leventon et al. [147] used a Bayesian framework in which a shape prior is embedded as the zero level set of a higher dimensional surface. They evolve the surface towards the maximum a posteriori estimate until convergence. Chen et al. [43] modified the energy functional of the GAC so it also depends on a shape prior. As

a consequence, their GAC was able to find boundaries in the presence of gaps and occlusions. Bresson et al. [24] extended Chen's and Leventon's work to introduce a geometric shape prior into the Mumford-Shah functional. They have successfully applied this approach in segmentation of synthetic and medical images. Cremers et al. [54] incorporated a low dimensional statistical shape prior in explicit parametrisation into a modified Mumford-Shah functional, thereby minimising a single energy functional. Cremers et al. [54] used this approach for segmentation of noisy or occluded images of hands. The advances in segmentation algorithms with shape priors as high level regularisation present a technique that is attractive to be applied to modelling the cyclist's shape from a set of annotated images.

5.2.5 Object Detection

The utility of GAC for object segmentation has been hindered by its sensitivity to curve initialisation. This is due to its local energy minimisation characteristics. In applications that seek a user-free solution, object detection algorithms are often used to provide the ROI of the desired object [133]. The object detection literature is vast and we do not seek to review it in detail. Instead we focus on recent advances to multi-scale Deformable Part based Models (DPM) approach to object detection.

Dalal and Triggs [57] constructed a filter on Histogram of Oriented Gradients (HOG) features to represent object categories. They used a multi scale sliding window approach to score an instance of the object at given scale and position. Their approach has shown to be particularly effective for pedestrian detection [71, 260]. Felzenszwalb et al. [91] introduced a discriminatively trained, multi-scale DPM for object detection. In their work, each model consists of a mixture of a coarse root filter, a mixture of parts filters, commonly HOG features,and part deformation relative to the root model to represent an object category. The object class specific models are trained using Support Vector Machines (SVM). The learnt model can then be used for object search in a new image. The DPM approach of Felzenszwalb et al. [91] has proved to be effective for detection of a large number of classes including bicycles, cars and animals. Many recent state-of-art object detection algorithms are built on extensions to this approach. For this reason, the approach is suitable and enticing for exploitation for detection of cyclists in images as a pre-process step for GAC initialisation.

5.3 pFSA Estimation Framework

This section provides an overview of our method for pFSA estimation of cyclists from images. The method is based on segmentation of the object boundary. This is achieved through curve evolution towards energy minima assumed to lie at the boundary of the cyclist's pFSA. The curve is initialised subsequent to cyclist detection using a trained DPM [78] and is regularised using statistical shape [79] and appearance models of a cyclist. As the cyclist detection and construction of the shape

and appearance statistical models are pre-requisites for segmentation, we describe those first.

5.3.1 Cyclist Detection

In order to initialise the GAC to segment the boundary of a cyclist, a detection of the position and scale of a cyclist is required. We use a discriminatively trained, multi-scale DPM [91] on HOG features [57] to do so. An early version of the cyclist detection algorithm was published in Drory et al. [78].

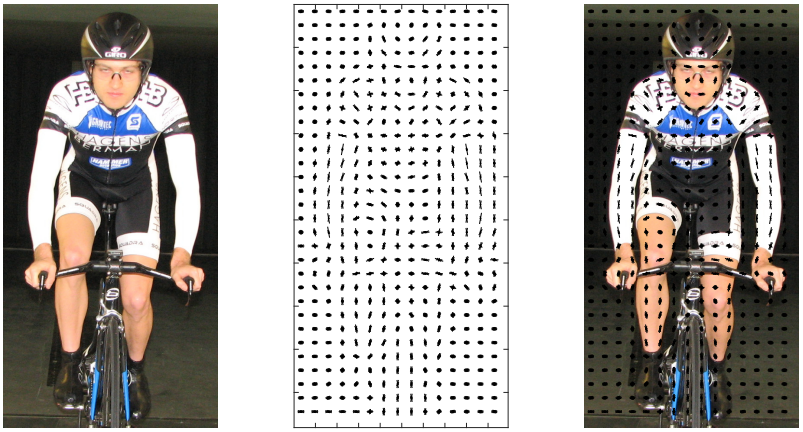


Figure 5.2: a cyclist's image (left its HOG feature representation (middle) and a superimposed of HOG visualisation on the original image (right).

HOG features represent an image window by the spatial distribution of locally normalised intensity gradients sampled on a dense grid of overlapping spaced cells to form a feature map. The histogram of oriented gradients of each cell is computed, where the gradient f of a function $f(x_1, x_2, \ldots, x_n)$ is a vector of partial derivatives of f. The distribution of the gradient orientations is depicted through a histogram over $p \in N$ discrete bins weighted by the gradient magnitude. The cells are then normalised with neighbouring cells to form spatial blocks. The final Descriptor is the concatenated feature blocks (fig. 5.2). The result is a descriptor that is invariant to small object deformation, image noise and bias. For pedestrian detection, Dalal and Triggs [57] found that a 36 dimensional feature vector consisting of 2×2 overlapping blocks of 6 to 8 cells representing 9 unsigned gradient orientation bins worked best. For cyclist detection, we found that a 31 dimensional feature vector that combines both signed and unsigned orientations results in superior performance (Fig. 5.2).

Following Felzenszwalb et al. [91] we construct a deformable part model that consists of a coarse whole object root filter and several high resolution small part filters sampled at twice the root's resolution. The relative position of a part to the root filter is assigned deformation weights.

A filter score at a position (x, y) in the feature map is defined by

$$\sum_{x',y'} \langle F(x', y'), G(x + x', y + y') \rangle,$$

where $\langle F, G \rangle$ denotes the inner product of filter F and a feature map G using a sliding window propagation. In practice the filter F is applied on a multiscale feature pyramid H at $\mathbf{p} = (x, y, l)$, where l denotes the l-th pyramid level of position (x, y), to minimise its sensitivity to object scale in an image.

To deal with pose variations, Felzenszwalb et al. [91] used a mixture model M with m components $(M_1, \ldots, M_m)$, where each component M_i is a deformable part model for a particular pose of the object class (fig. 5.3). Formally, they define the model by a $(n + 2)$-tuple $(F_0, P_1, \ldots, P_n, b)$, where F_0 is the root filter, P_i is a model for the i-th part and b is a bias value. A part P_i consists of (F_i, v_i, d_i), where F_i is the i-th part filter, v_i is its anchor position relative to the root filter and d_i is a four dimensional vector containing its deformation costs.

Figure 5.3: A single component HOG based cyclist front view model. The model is defined by coarse root filter (left), several high resolution part filters (middle) and a spatial weight for each part model's location relative to the root model (right) [91].

For model learning, we use Linear SVM. The technique takes a training set of feature maps with labelled positive and negative class examples and produces a model which predicts the label of test data, such that

$$D = (\langle x_1, y_1 \rangle, \ldots, \langle x_n, y_n \rangle | x_i \in X, y_i \in \{-1, 1\}),$$

where x_i is the feature vector and y_i is the class label (1 for positive and -1 for negative examples).

For inference of a new image, we compute a feature pyramid of the target image. The root filter is applied at coarse levels followed by application of the part filter at

fine pyramid levels. We calculate a score of an object hypothesis from a data term representing the scores of each filter at their respective locations and a deformation cost that depends on the position of each part with respect to the root plus the bias,

$$s(p_0, \ldots, p_n) = \sum_{i=1}^{n} \langle F_i, \Phi(H, p_i) \rangle - \sum_{i=1}^{n} d_i \cdot \phi(dx_i, dy_i) + b,$$

where $(dx_i, dy_i) = (x_i, y_i) - (2(x_0, y_0) + v_i)$ is the displacement of the i-th part with respect to its anchor position, and $\phi_d(dx_i, dy_i) = (dx, dy, dx^2, dy^2)$ are deformation features. The bias term b is required when multiple models are combined into a mixture model. High scoring root locations define detections while the location of the parts that yields a high scoring root location defines an object hypothesis. Finally, non-maximum suppression is applied to eliminate detections that are overlapping by more than 50%. For example, when a cyclist is occluding another by more than 50%, only the first will be detected (see fig 5.4b and d). Calculation of overall score for each root location allows detection of multiple object instances (see fig 5.4c).

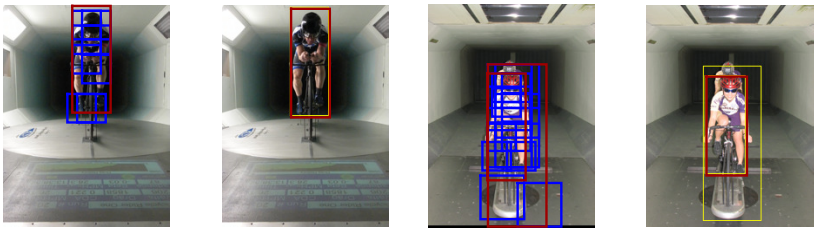
Our method is implemented to work on the VOC2009 challenge dataset for model training and evaluation as described in Drory et al. [78]. We use the code published by Felzenszwalb et al. [91], Girshick et al. [103] release 4 to train and apply our models. Our method only considered the frontal pose of a cyclist, hence we used mixture model M with $m = 1$. This, however, can be extended to additional poses.

5.3.2 Cyclists Statistical Shape Model

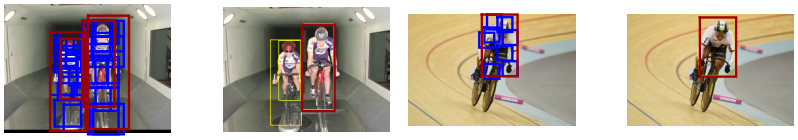
In order to incorporate high-level regularisation into the GAC framework, a statistical model of a cyclist needs to be developed. We construct a statistical shape variation model from a set of corresponding points across training images similarly to Cootes and Taylor [49]. Suppose $\mathbf{x} = (x_1, \ldots, x_n, y_1, \ldots, y_n)$ is a $2n$ element vector where (x_i, y_i) are sampled boundary landmarks for a 2D image $\mathbf{x}_j$ of s training images. The set of training shapes need to be aligned with respect to a similarity transformation into a common coordinate frame in order to construct a statistical model of a cyclist (fig 5.5a) such that $D = \sum |x_i - x|^2$ is the distance of each shape to the mean $\mathbf{x}$ is minimised. We use 97 boundary landmark points to represent a cyclist in each image following Drory et al. [79].

5.3.2.1 Point Distribution Model, Procrustes Analysis

The minimisation is ill defined unless some constraints are placed on the alignment of at least one of the shapes. Generalised Procrustes Analysis (GPA) is an iterative technique to superimpose a set of objects, typically centred on the origin, have a mean scale of unity and a fixed arbitrary orientation until convergence. To construct our statistical shape model we first rigidly align the 57 manually segmented training images with respect to similarity transform following Gower [105]. This results in the shape model that consists of the aligned shapes (fig 5.5b [79]).



(a) (b)



(c) (d)

Figure 5.4: Cyclist detection using deformable part based model showing parts model in left image and cyclist ROI on right depicting invariance to a) unilateral illumination (image: property of San Diego Wind Tunnel. Used with permission), b) posterior obstruction, c) multiple instances (images b and c: property of University of Washington Wind Tunnel. Used with permission)., and d) obstruction and angled orientation on track (image: Drory, A.).

5.3.2.2 Principal Component Analysis

The set of landmark points of the shape model can be viewed as a point cloud that lies on a manifold in a $2n - D$ space to form a Gaussian distribution of likely correlated vectors. Using PCA it is possible to project those vectors onto an uncorrelated coordinate frame and compute the main axes of this cloud. This can be computed efficiently using singular value decomposition on the covariance matrix Σ of the data.

This operation does not alter the original point cloud and it can be recovered from the new coordinate frame. Typically, it emerges that only few major axes are responsible for the majority of the model's variance. It is therefore desirable to reduce the dimensionality of the data to a more manageable size. In our case, 13 principal components were found to represent 98% of the variance in our data (fig 5.5c). Thus, we were able to discard the remaining data and reduce the dimensionality from $194 - D$ space to a manageable $13 - D$ space (Fig. 5.5).

Each of the training shapes can then be approximated using

$$\mathbf{x} \approx \bar{\mathbf{x}} + \boldsymbol{P}\mathbf{b},$$

where $\boldsymbol{P} = (\mathbf{p}_1^T, \ldots, \mathbf{p}_t^T)$ contains t eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$ and $\mathbf{b}$ is a t dimensional vector given by

$$\mathbf{b} = \boldsymbol{P}^T(\mathbf{x} - \bar{\mathbf{x}}). \tag{5.1}$$

In figure 5.5d we demonstrate reconstruction of 10 training shapes from the dimensionality reduced model.

We define our final shape model $M(\bar{\mathbf{x}}, \boldsymbol{P}, \mathbf{b})$, where $\bar{\mathbf{x}}^{2n}$ is the mean shape, $\boldsymbol{P}^{2n \times t}$ is the matrix containing the principal components and $\mathbf{b}^t$ is the vector of eigenvalues corresponding to the principal components with t the number of retained components that explain at least 98% of the variance in our data.

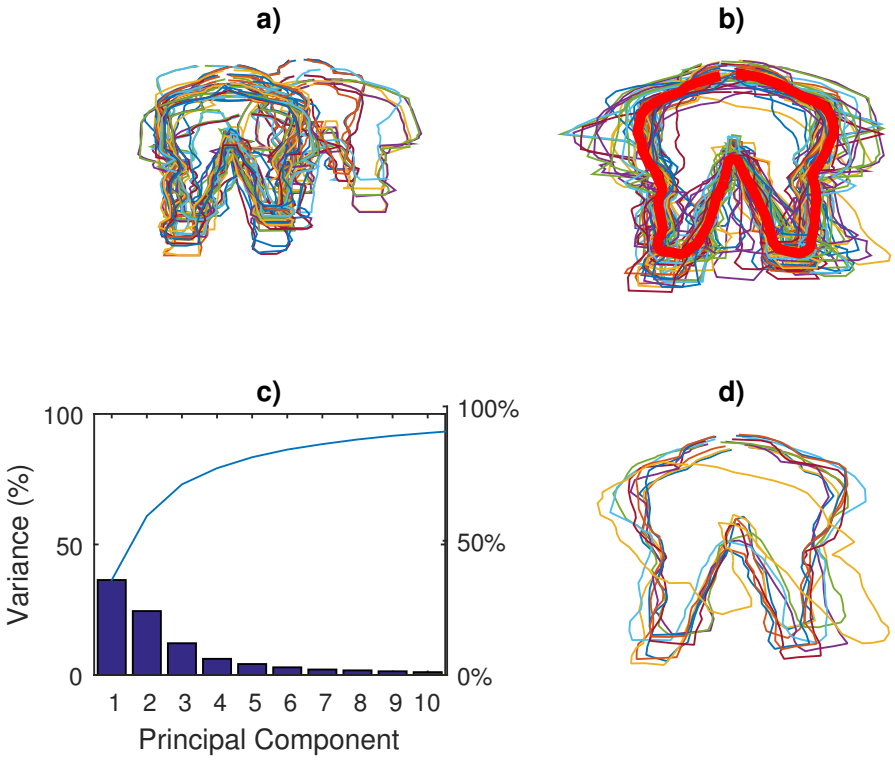


Figure 5.5: Statistical shape model of a cyclist. a) boundaries of cyclists in training images. b) Procrustes analysis of training shapes achieves co-alignment and scaling. Our final model M is rendered in red. c) The eigenvalues corresponding to the top 10 principal components. The solid line depicts the cumulative percentage of total variance explained by the principal components. d) Training shapes can be recovered using a dimensionality reduced model (using PCA). Here this consists of only the top 13 principal components accounting for 98% of the variance.

5.3.2.3 Effect of modifying principal components

The vector $\mathbf{b}$ defines a set of parameters of a deformable model. This allows the creation of synthetic shapes that are not in the training set simply by varying $\mathbf{b}$ in

equation(5.1). By constraining the variance of b_i to $\pm\sqrt{\lambda_i}$, where λ_i is the variance of the i-th parameter, we ensure that the generated shape is within the statistical probability of the model. In figure 5.6 we demonstrate varying of parameters to create likely synthetic shapes. The significance of this capacity is discussed in section 5.5.

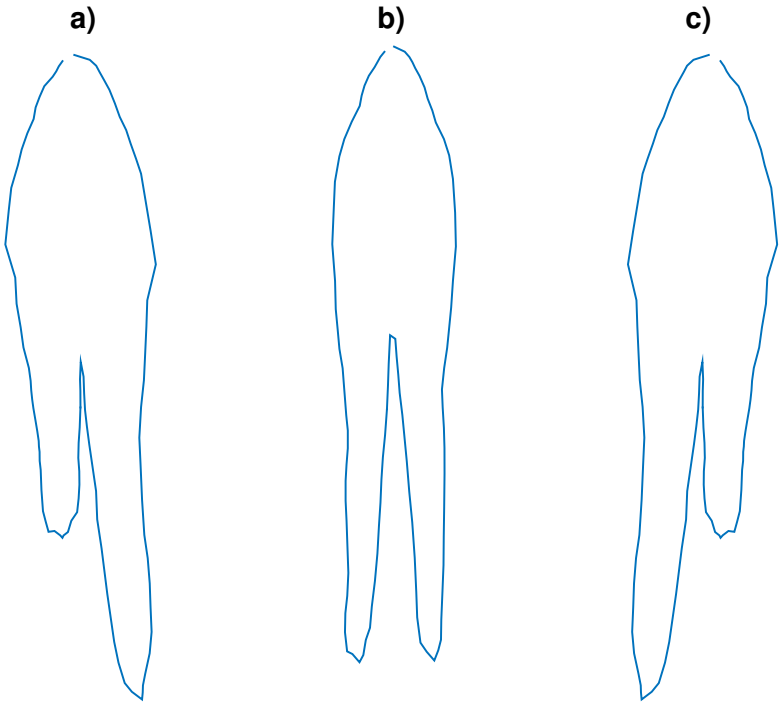


Figure 5.6: The statistical model of a cyclist can be manipulated to create likely synthetic shapes that are within a determined standard deviation from the mean simply by changing the eigenvalue that corresponds to a principal component. here b) depicts the mean shape. a) depicts change to the first principal component corresponding to 3 standard deviations from the mean. This corresponds to the leg position of the cyclist (pedal at top dead centre) and height of the cyclist. c) depicts change to the second principal component, which corresponds to foot position in the pedalling cycle.

5.3.2.4 Level Set Formulation

To facilitate modelling changes in shape topology, multiple cyclists and shape 'holes' such as underarm gaps, we adopt an alternative formulation of the shape model using level set construction following Leventon et al. [147]. In this formulation the object boundary is embedded as the zero level set of a higher dimensional signed distance transform where each sample encodes the distance to the nearest point on the curve (fig 5.7a). The effect of modifying the model's principal components in

level set formulation is demonstrated in fig 5.9.

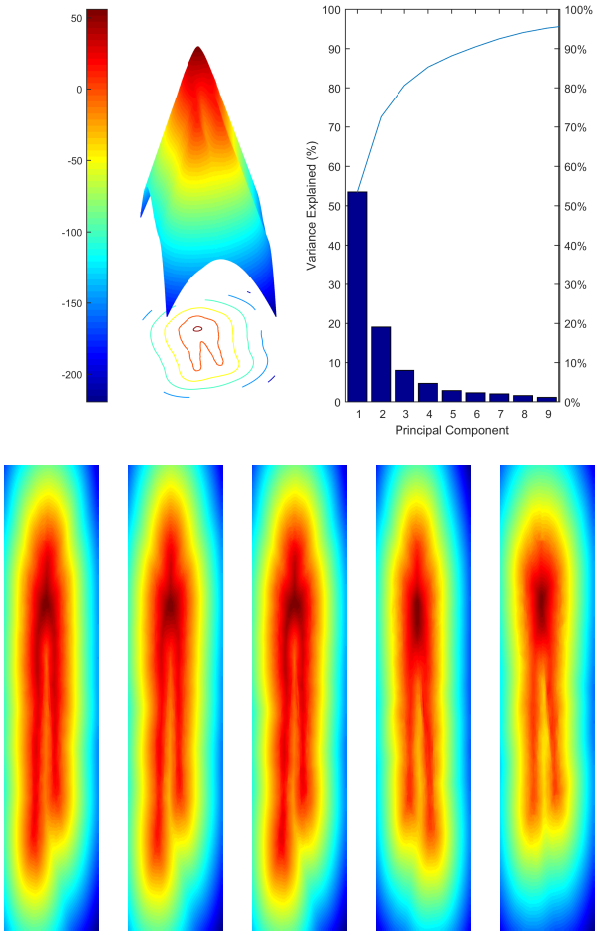


Figure 5.7: Statistical shape model of cyclists in level set formulation following Drory et al. [79]. a) mean shape of cyclists in training images with the mean curve represented at the zero level set and projected onto $2D$. b) The eigenvalues corresponding to the top 10 principal components. The solid line depicts the cumulative percentage of total variance explained by the principal components. c) The training shapes can be recovered using a dimensionality reduced model (using PCA). In this case, this consists of only the top 16 principal components accounting for 98% of the variance.

5.3.3 Cyclists Statistical Appearance Model

To further enhance the regularisation of our prior models, we construct a local appearance model from our training images at each keypoint. Our local appearance feature is based on the profile vector $\mathbf{v}$ of length r of the normal to the boundary at each keypoint and its derivative ∂v, with the keypoint at the feature vector's centre (Fig.5.10).

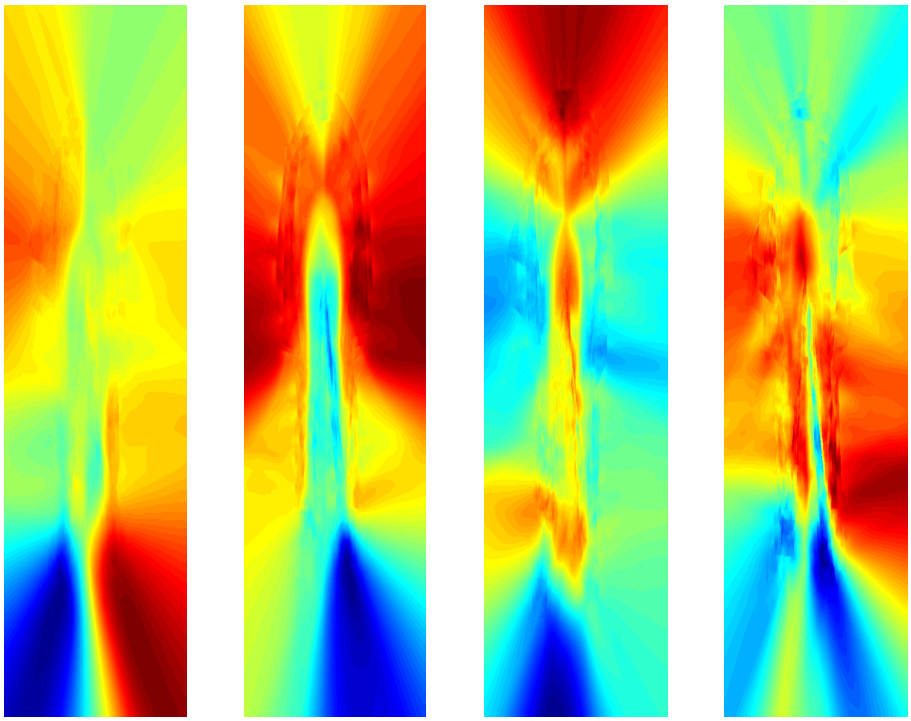


Figure 5.8: A visualisation of the first 4 principal components in level set formulation [79].

We define our final appearance model $\boldsymbol{L}(\bar{\Phi}, \sigma)$, where $\bar{\boldsymbol{\Phi}}^{n\times 2r}$ is a matrix containing the mean appearance feature vector $\mathbf{l}_i^{2r} = [v_i^r, \partial v_i^r]$ of the i-th keypoint, and σ_i^r is its variance.

At each iteration of the GAC, we ensure that the proposed move remains within the statistical probability of the appearance model $\boldsymbol{L}$ by constraining the variance of σ_i to $\pm\sqrt{\lambda_i}$, where λ_i is the variance of $\mathbf{l}_i$ at keypoint i.

5.3.4 Geometric Active Contour

The basic GAC technique defines an energy function $E(C)$ over a curve $C(s) = [x(s), y(s)], s \in [0,1]$ as the sum of internal and external energies of the curve. The curve evolves to minimise an energy functional

$$E(C,t) = \int E_{internal}(C(s))\mathrm{d}s + \int E_{external}(C(s))\mathrm{d}s, \tag{5.2}$$

where the internal energy $E_{internal}$ represents the curve's tension and stiffness, which serves to impose the piecewise smoothness constraint, and the external energy $E_{external}$ represents salient image features such as edges and constraints from high-level sources such as user initialisation (Fig. 5.11).

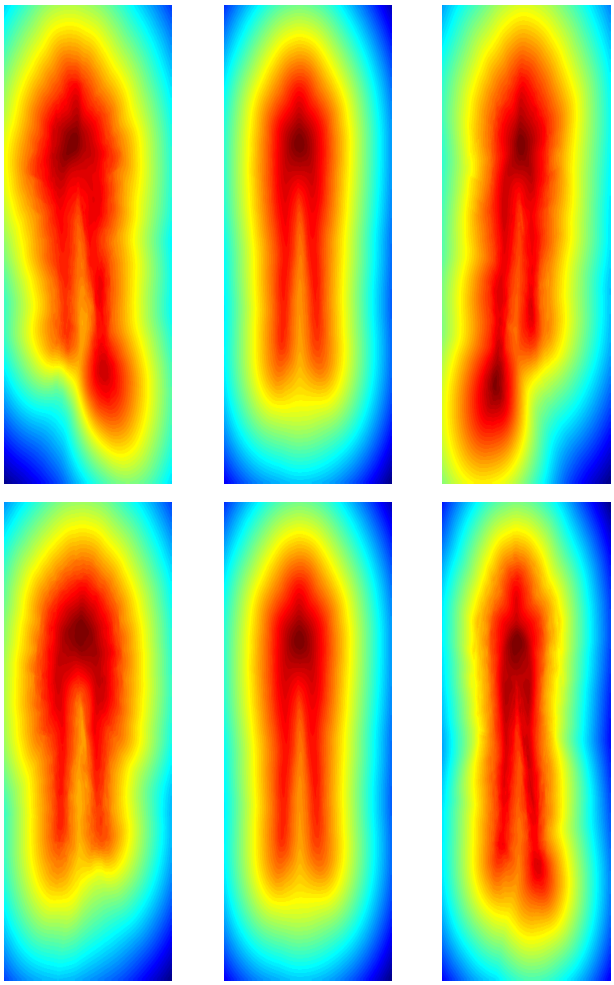


Figure 5.9: The effect of changing the eigenvalues that correspond to principal components can create new likely synthetic images that are not in the training data from the mean shape (centre image top and bottom) [79]. In the top row the first principal component was changed by ± 2 standard variations (left and right in top row respectively). This principal component captures the leg position (simulates pedalling). In the bottom row the second principal component was changed by ± 2 standard variations (left and right in bottom row respectively). This principal component captures the cyclistâĂŹs size variations.

In keeping with Kass et al. [130] we define the internal energy

$$E_{internal} = \int_0^1 (\alpha(s)|c_s(s)|^2 + \beta(s)|c_{ss}(s)|^2), \tag{5.3}$$

Where $c_s(s)$ is the first order term representing the curve's elasticity,which makes the spline act like a membrane, and $c_{ss}(s)$ is the second order term representing the curve's stiffness, which makes the spline act like a thin-plate. The parameters α and

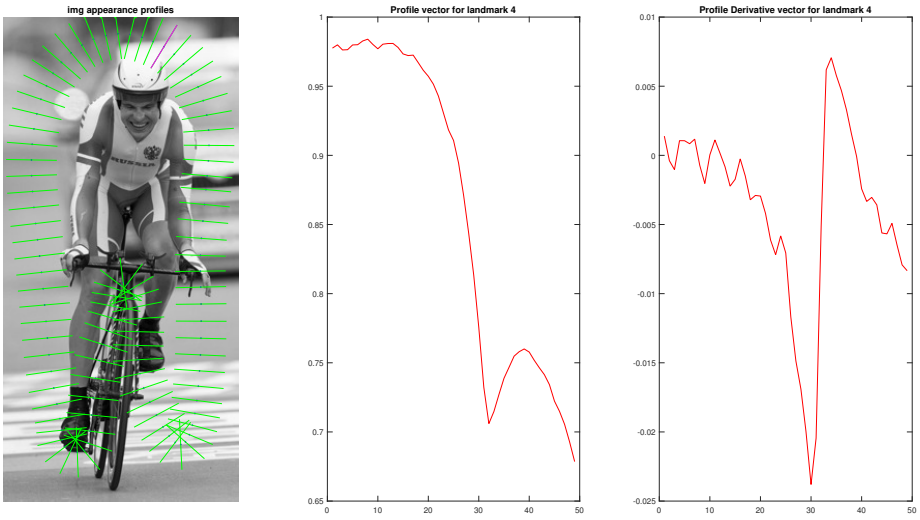


Figure 5.10: Appearance model of a cyclist captures the normalised appearance profile along the normal to the cyclist's contour at each key point (left image). The centre plot depicts and example appearance profile at keypoint 4 (highlighted in magenta on the left image). The plot on the right is the profile derivative at keypoint 4. The bottom row shows partial occlusion of second object.

β control the relative contribution of the curve's elasticity and stiffness respectively. Setting $\beta(s) = 0$ at a point allows the snake to become second-order discontinuous and develop a corner.

The external energy $E_{external}$ produces a force that gravitates the curve towards salient image features. In our implementation we define

$$E_{external} = -\int_0^1 |\nabla G_\sigma(s) * I(s)|^2, \tag{5.4}$$

Where $*$ and ∇ are the convolution and gradient operators respectively and $G_\sigma(s)$ is a two dimensional Gaussian function with standard deviation σ.

Using variational calculus we arrive at the Euler-Lagrange energy gradient equation

$$C_t(s,t) = \alpha c_{ss}(s,t) - \beta c_{ssss}(s,t) - \nabla E_{external}, \tag{5.5}$$

where the curve is made dynamic by addition of dependency with respect to time t. The energy function is minimised when $C_t(s,t) = 0$. In practice equation (5.5) is discretised and solved iteratively using semi-implicit relaxation methods.

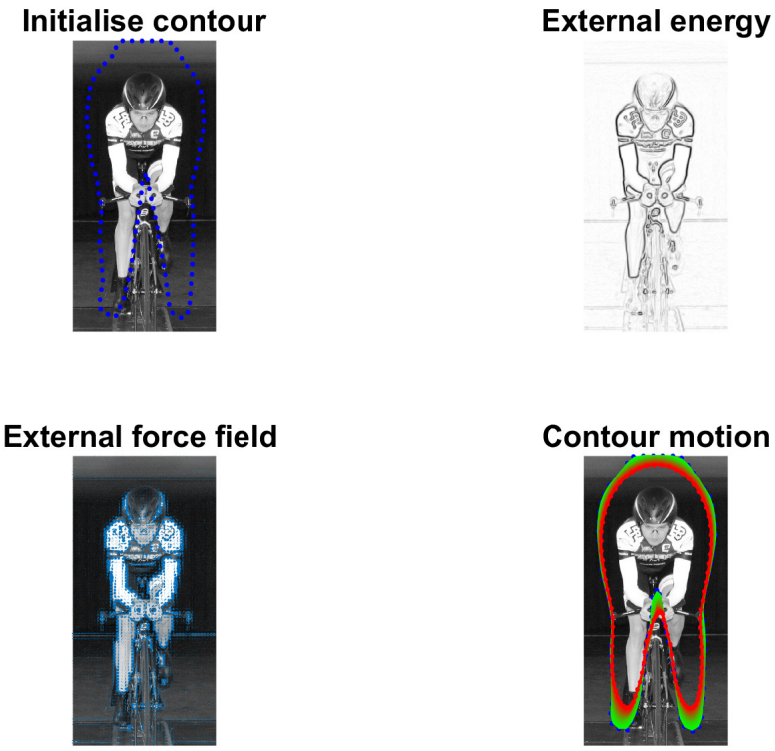


Figure 5.11: The contour is initialised by a curve matching the expectation of the statistical shape model at a location provided by the output of the DPM object detection algorithm (top left). The energy of the GAC is comprised of an external energy (top right), and a force field (bottom left). The curve gravitates towards image features with each iteration. (bottom right - progress in time is depicted by green to red motion (50 iterations)).

5.4 Experiments and Evaluation

5.4.1 Datasets

In this section, we report on a set of quantitative experiments performed to evaluate our method. The experiments were conducted on two challenging datasets containing images of cyclists. The WT dataset contains images of a single cyclist in a controlled environment,typically during wind tunnel testing. In this dataset, there is a predictable, relatively smooth background, no occlusions and no other person in the scene. On the other hand, in this dataset, the cyclists present varied appearance in their pose, garments, accessories (e.g. helmets), lighting conditions or colours that present a significant challenge to the classifier's model. The UCI dataset contains images of multiple cyclists *in natura*, in an uncontrolled environment, typically during road races. In this dataset, there are many instances of occlusions, varied appearance and multitude of other objects such as crowds, cars and motorcycles that present a

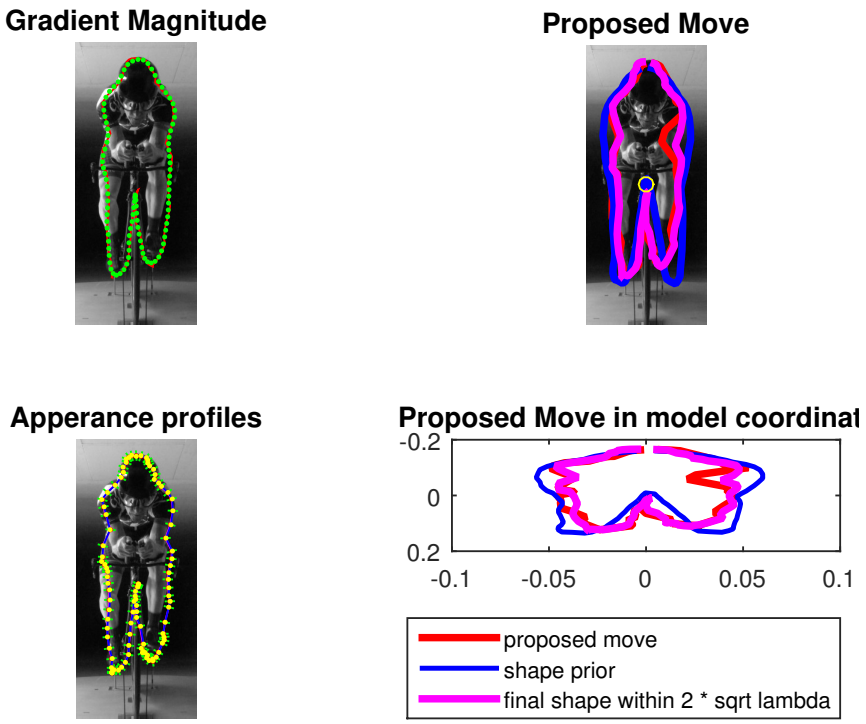


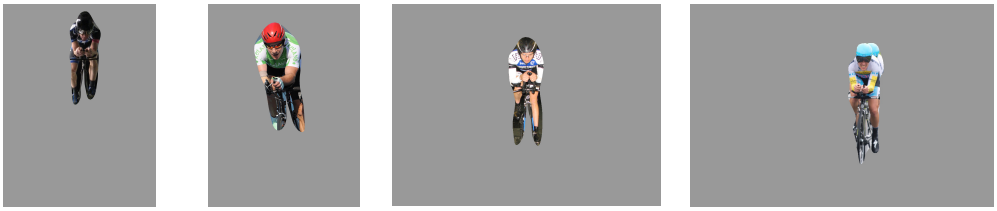
Figure 5.12: Visualisation of a contour move decision framework at a single GAC iteration. The sum of the contour's gradient (Top Left) and local appearance profiles (bottom left) probabilities at each keypoint results in a proposed contour move (right images - solid red line). The proposal is tested against our statistical shape model M, whose expectation is depicted by the solid blue line on the right images. By constraining the variance of b_i to $\pm\sqrt{\lambda_i}$, where λ_i is the variance of the i-th parameter, we ensure that the generated shape is within the statistical probability of the model (right images - solid magenta line). For clarity, the three curves were projected onto the model coordinate system in the bottom right image.

different challenge than the WT dataset.

Using these datasets, section 5.4.2 reports on the performance of our cyclist detector. This is followed by reports on the segmentation performance and the estimation of pFSA in section 5.4.3.

5.4.2 Cyclist Detection

To evaluate the performance of our cyclist detection we use standard metrics and calculate precision and recall using $p_t/(p_t + p_f)$ and $p_t/(p_t + n_f)$ respectively, where p_t is true positive, p_f is false positive and n_f is false negative. However, in contrast to standard evaluation techniques in the object detection research community, which score detection based on their centroid distance and overlap ratio, we use a far stricter definition of p_t. We consider a detection to be a p_t, if and only if the cyclist in the



(a) (b) (c) (d)

Figure 5.13: The curve gravitates toward image features and arrives at equilibrium. This results in an object segmentation that provides an estimation of the projected frontal surface area of a cyclist. The image in (a) shows some sub-optimal result at the legs due to shoes' colour similarity with the background. The image in (d) demonstrates sub-optimal result due to compromised curve initialisation due to partially occluded second object with colour similarity.

image is entirely contained within the detection bounding box. We justify this definition by the downstream effect that partial detection has on the curve initialisation for segmentation. The performance of the GAC critically relies on good initialisation and many segmentation techniques assume that such initialisation is given. Instead, we exclude partial detections from our evaluation and reported separately. Our results (see table 5.1) indicate that this algorithm is very effective in cyclist detections from images both in controlled and uncontrolled environments. Interestingly, our detector performs similarly well in precision on both the WT and UCI datasets, but far better in recall on the UCI dataset. We reason that the challenging lighting conditions, varied garment appearance and often unusual equipment use (e.g. helmets) in wind tunnel testing result in poorer detector performance on the WT dataset. We also note that multiple cyclists may be present in an image of the UCI dataset. Only cyclists, who were visible by more than 50%, were considered for the performance evaluation of the detector.

Table 5.1: Cyclist Detection Results

dataset	#images evaluated	$\#p_t$	$\#p_f$	$\#n_f$	#Partial	Precision	Recall
WT	171	120	0	57	4	1	0.68
UCI	130	129	5	26	10	0.96	0.83

5.4.3 Cyclist Segmentation and pFSA Estimation

Our framework is designed such that a cyclist is detected in an image, and that at final curve evolution of the GAC at equilibrium rests on the boundary of the cyclist's image. This provides a segmentation of the cyclist from the background of the image scene, the area of which we consider to represent the pFSA of the cyclist.

To evaluate the performance of our cyclist segmentation and pFSA estimation we use a standard segmentation evaluation metric, the mean Dice Similarity Coef-

ficient (DSC) over 40 and 20 p_t images from the challenging WT and UCI datasets respectively. Given G a set of pixels annotated as a ground truth cyclist and S a set of pixels segmented using our GAC framework, DSC is defined as $\text{DSC}(G, S) = (2|G \cap S|/(|G| + |S|)$. Our segmentation framework achieved mean DSC scores of 0.88 and 0.92 on our test images from the WT and UCI datasets respectively. We present representative results in figures 5.15.13.

As can be seen in figure 5.13 sub-optimal result occur generally around and between the leg segments of the cyclists, or in the presence of like objects with similar appearance.

5.5 Discussion

In this chapter, we investigated the challenging problem of estimating the pFSA of cyclists from monocular images. We introduced a repeatable automatic method for pFSA estimation for the study of its relationship with aerodynamic drag in cyclists. Our approach is based on detection of a cyclist object class in an image using a discriminatively trained DPM model, followed by finding a cyclist's boundary in the image. An initialised curve dynamically evolves in the image to minimise an Euler-Lagrangian energy function designed to force the curve to gravitate towards image features. To overcome occlusions and pose variation, we use statistical cyclist shape and appearance models as priors to encourage the evolving curve to arrive at the desired solution. Once an instance of a cyclist is detected in an image and segmented, the pFSA is calculated from the area of the final curve. Our experiments demonstrated that our framework is successfully applied to cyclist images of two challenging datasets. We discuss the performance of our method under occlusion, orientation, and pose conditions. We show that our method successfully estimates pFSA in cyclists and allows exploration of the relationship between pFSA and aerodynamic drag. The output of our framework forms a critical component and a crucial evidence base pre-requisite to the study of the relationship between a cyclist's riding position and reduction of aerodynamic drag.

Limitations and Future Work A number of challenges remain that need addressing for enhanced robustness of our framework. For instance, for cyclist detection we use a standard DPM framework. Recent advances in part-based simultaneous detection and human pose estimation shown superior performance over DPM [249]. An extension of our framework would use Flexible Mixture of Parts (FMP) approach for cyclist detection, which exploits the spatial relations of the parts. Moreover, the output of a FMP framework provides a pose estimate of the object in addition to bounding boxes for each part. Attractively, this skeletal linkage model can be used as a foreground seed for a graph-cut approach to segmentation with the parts' bounding boxes as its background seed. More recently, Convolutional Neural Network (CNN) approaches showed state-of-the-art performance on object detection datasets. It would be interesting to test the performance of our framework with CNN detector as an alternative

to DPM in our framework.

A natural extension to our method will use a multi view approach to achieve a comprehensive deformable 3D dynamic cyclist model. This, however, requires pose estimation algorithms that better handle occlusions and self-occlusions than has so far been achieved. This is necessary as the current approach cannot handle missing nodes (Fig. 5.14.

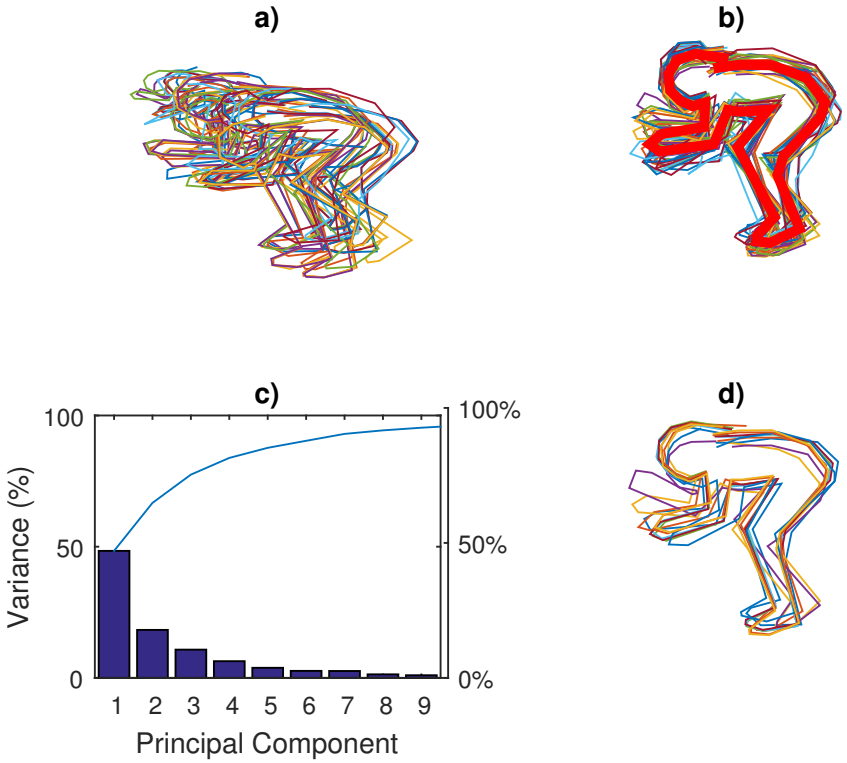


Figure 5.14: A side view statistical shape model of a cyclist can be used to extend our approach to 3D. The underlying graphical model, however, currently has nodes representing one side of the body only.

5.5.1 Acknowledgement

We wish to express our gratitude to the San Diego Wind Tunnel and the Wind Tunnel at the University of Washington for the sharing of data.

Chapter 6

Predictive Model of Time Saved on Descents in Road Cycling Achieved Through Reduction in Aerodynamic Drag Area

Abstract

This chapter presents simulation results of descent time to completion in road cycling. A mathematical model is formulated to predict time saved on road cycling descents where a cyclist's position is static through manipulation of Aerodynamic Drag Area, system parameters and initial conditions. Road cyclists often adopt drastic static riding positions on long descents in order to minimise aerodynamic drag and optimise performance measured as race time to completion. In those riding positions bicycle control is compromised and the risk of fall and injury increases. The aims of this study were to investigate the effect of the difference on time to completion of road descents associated with the 'Top-Tube' descending riding position compared to 'Normal' descending riding position. Based on Newtonian-Lagrangian equations of motion of the cyclist-bicycle system, an analogue mathematical model as a non-linear Riccati ordinary differential equation was developed to enable prediction of velocity and time to completion of a road cycling course descent of known length and gradient as measures of performance for an athlete of known mass and drag area in a descending position. Previously proposed models of cycling performance have been based on physiological, anthropometric, and mechanical power output. No general closed form time to completion mathematical model for cycling was found in the literature. Analytical solutions allow for a concise investigation of a dynamical system model's behaviour that is not as readily available with a numerical solution. The analytical solution to the non-linear Riccati differential equation showed large time savings as a result of reduction of drag area can be made on road cycling race descents. In the example scenario simulated here, 29.2sec may be saved on a 5km descent of 10% gradient with 25% reduction in aerodynamic drag area (CdA). We conclude that the 'Top-Tube' riding position is associated with large

reduction in aerodynamic drag area in road descents compared to conventional descending riding position. Our model enables the prediction of time to completion on descents. This may assist cyclists to assess the trade off between undertaking increased risk associated with drastic rider descending position and the potential for improved performance in the context of race tactics and strategy.

6.1 Introduction

The performance of cyclists in road racing as measured by the time to completion of a race course is significantly affected by the cyclist's aerodynamic properties [140]. Three types of external forces act on the bicycle/cyclist couple, being the system's weight, aerodynamic drag force and contact forces between the road and the bicycle tyres [38] At racing speeds greater than 14m/s, aerodynamic drag force account for more than 90% of the total resistive forces acting on the cyclist/bike system [106, 140, 141], of which up to 70% are attributed to the cyclist [39, 64, 141]. The aerodynamic drag of a cyclist, as typical to bluff bodies, is dominated primarily by form drag associated with the geometrical shape of the bike-cyclist system and the vortical wake contribution [48]. This is in contrast to highly streamlined airfoil bodies, the aerodynamic properties of which are dominated by viscous drag component due to the velocity slow down in the boundary layers and associated skin friction [25]. The form drag can be reduced mainly by modifications to the geometrical shape of the bike-cyclist system through changes to cyclist's riding position [11, 25]. Given the large potential of improved cycling performance, past research focused on reduction of aerodynamic drag through modification to riding position [48, 62, 64, 97, 106, 128, 139, 179, 226] or equipment [3, 11, 14, 25, 25, 28, 38, 138, 142]. Grappe et al. [106] reported up to 27% reduction in drag area in the O'bree position compared to upright, dropped and aero positions, which indicate the possibility for rider position optimisation for reduction in aerodynamic drag. Similarly, Garcia-Lopez14 reported reference values of aerodynamic drag of professional cyclists as measured in a wind tunnel in five riding positions under static and dynamic conditions. They showed that modifications to riding position can reduce the drag area of the system by up to 14%. However, the modifications in rider positions were achieved through manipulation of peripheral equipment (bicycle type, handlebar height and elbow pads location). Hence, the kinematic variables used to describe each position showed large variations and are participant and trial specific and are not consistent or reproducible. Defraeye et al. [64] attempted to control for the variability in rider position using a physical constraints positioning system. They studied three rider positions and attempted to characterise the wake flow field using CFD, but only achieve moderate agreement with wind tunnel empirical data. The majority of past research focused on aerodynamic optimisation for flat surface or quasi-flat surface cycling such as time trial or track pursuit events [106, 226]. Grappe et al. [106] showed 7.8% and 12.4% reduction in drag area (CdA) in the dropped and aero positions respectively relative to an upright position. Underwood and Jermy [226] showed a small

(0.6+/-0.4N) reduction in aerodynamic drag for arrow-style hand position with no loss in power output. This also corresponded with reduction in estimated time to completion of a 250m lap of individual pursuit effort at assumed zero wind and gradient conditions. Chowdhury et al. [46] showed the CdA in a time trial position was lower than dropped position or the upright position, but failed to control the shape of the cyclist or the kinematic characteristics of the cyclist's position. Despite the increased proportional effect of aerodynamic drag on total resistive forces acting on the cyclist-bicycle system, which exceeds 90% at 14m/s [140], no studies were found that measured or reported aerodynamic drag in cycling descents where the velocity of the cyclist may exceed 25m/s. This chapter will present first reference empirical values for aerodynamic drag of cyclists in descending riding position. The effect of reduced aerodynamic drag through modification of riding position on cycling performance is difficult to determine. Previous studies report a trade off between aerodynamic drag reduction achieved through adoption of aerodynamic favourable position and the ability of the cyclist to deliver the same power to the system. Savelberg et al. [203] showed that changes to trunk angle influences muscle recruitment pattern and limb kinematics. This may be due to reasons relating to either the ability of the neuro-muscular system to produce muscular effort in those positions [13] or the motor pattern specificity in trained participants, who are unable to execute a modified pattern in a position different to that they are trained in. Irrespective of the reason, this limits the ability to infer cause and effect relationship between reduction of aerodynamic drag achieved through modification of riding position and cycling performance. In cycling descents, however, cyclists regularly assume static position with the aim of increasing velocity through reduction of aerodynamic drag, in which case there is no cyclist's muscular effort input that influences propulsion. In this specific case where propulsion is dominated by gravity, the direct effect of reduction in aerodynamic drag through changes in riding position on cycling performance can be studied. For this reason, this chapter will use cycling descent to estimate the effect of reduced aerodynamic drag achieved through changes to riding position on simulated cycling performance.

6.1.1 Mathematical modelling of cycling performance

Mathematical cycling models enable prediction of the effect of variations to system parameters and environmental conditions that are difficult to measure, control or isolate on cycling performance. Several predictive mathematical models have been previously proposed [70, 140, 181]. Majority of past research presented approximate numerical solutions to cycling models. Olds et al. [181] presented a cycling energy demand and supply model to predict performance time of a 4,000m individual pursuit in 18 elite cyclists on quasi-flat track. They reported good correlation (0.8) and 3% mean absolute performance time difference. They have extended this model to predict flat road time-trial performance [182], with similar mean absolute performance time difference (3.9%). Ahmad et al. [2] used third order Runge-Kutta methods to derive at numerical solution for estimation of cycling velocity. Marqués-

Bruna and Grimshaw [164] used numerical solution to linear regression model [165] to simulate the effect of changes to road gradient and environmental conditions on aerodynamic drag and resulting velocity and time to completion of a time trial. Numerical methods such as the Runge-Kutta method provide only an approximation over a brief time period. Never the less, they can provide a very good approximation under certain circumstances. Analytical models allow for a concise investigation of a model's behaviour and insight to a dynamic system that is not as readily available with a numerical solution. The use of analytical model is more robust and provides confidence that results are transparent irrespective of the platform they have been implemented on. A general expression for prediction time to completion may play important role in race strategy and tactics. A general analytical closed-form expression for predictive time to completion model was not found in the literature. For this reason, this chapter will present a general closed-form method. The non-linear equations for cycling are formulated and expressed in terms of velocity, initial conditions and system parameters. The integration of these equations is conducted symbolically and is used to compute the time to completion value and time history of speed for various cases. Likewise, a mathematical model that predicts descent time in road cycling was not found. Yet, field observations show that in the course of road races, cyclists often adopt radical static positions on long descents with the aim of maximising velocity through reduction of aerodynamic drag. This chapter will present experimental observations in support of this hypothesis through CdA measurements from wind tunnel lab testing, which demonstrate the reduction in CdA that can be found for these static descending positions. Specifically, the 'Top-Tube' descending riding position has emerged in recent years and is used by some riders. The difference in drag area associated with this position compared to traditional or 'Normal' descending riding position will be quantified using laboratory testing. The ability to predict time saved on descents can play an important role in race tactics and strategy. The following example demonstrates the often overlooked importance of time saved on descent in road racing. In the 2011 edition of the Tour de France, Cadel Evans won the three week tour by 94 seconds from Andy Shleck, who led the race by 57 seconds entering the individual time-trial stage on the penultimate day of racing. While much of the post race analysis focused on the time Evans gained over Shleck at the time-trial to claim the overall win, less focus was granted to the 69 seconds Evans gained over Shleck on the 9.5km descent of mean 5.5% gradient 16th stage to Gap. Had Evans not made this time gain, he may have started the penultimate time-trial stage with a deficit of 2:06 minutes, which may have influenced the final outcome. It is worth noting that the Tour de France was previously won by time margins as small as 8 seconds (1989, LeMond from Fignon) and Evans twice lost the overall win by margins of less than a minute (23 seconds to Contador in 2007; 58 seconds to Sastre in 2008). While this anecdotal example demonstrates the potential impact of time gained on descents on overall win, it is not suggested here that the reason for the time Evans gained over Shleck on stage 16th was caused by reduction of drag area through adoption of the riding positions studied in this chapter. Other factors including handling skills are likely at play. Never the less, it demonstrates

that significant gains can be made if a cyclist is prepared to undertake increased risk on descent. This chapter will present simulation results for time to completion of road cycling descent with reduction in drag area.

6.2 Method

6.2.1 The Mathematical Model

A mathematical model is proposed to enable prediction of time saved on a descent of known length and gradient for an athlete of known mass and laboratory determined percent reduction in CdA in a particular position, changes to cyclist's parameters (mass), changes to initial conditions (initial velocity) and system parameters (gradient, descent length, and rolling friction). Newtonian âĂŞ Lagrangian equations of motion of the cyclist-bicycle system have been described previously [2, 70]. Consistent with previous literature, here a coordinate system that is parallel to the slope of the ground is used. From Cyclist free body diagram [2, 179, 226], sum of the forces in the y direction (perpendicular to the slope) is

$$\sum F_y : N - mg\cos\theta = 0, \tag{6.1}$$

where N is the normal component of the ground reaction force applied on the bike, m is the combined mass of the cyclist and bike system, g is the gravitational acceleration, and θ is the slope's angle. From Cyclist free body diagram, sum of the forces in the x direction (parallel to the slope) is

$$\sum F_x : mg\sin\theta - F_{RR} - D = ma, \tag{6.2}$$

where F_{RR} is the component of the ground reaction force acting parallel to the slope. Here, if is assumed that it is caused predominantly by bike tyres rolling resistance force, D is the Aerodynamic drag force acting on the cyclist and bike system, and a is the acceleration of the cyclist and bike system along the slope. To obtain an autonomous differential equation for this system, detailed structures into equation (2) are introduced. In general, a relationship between $F_R R$ and the normal force N is given by

$$F_{RR} = \mu N, \tag{6.3}$$

where μ is the coefficient of rolling resistance. The slope's angle typically derived from the slope's gradient as

$$\theta = \tan^{-1}(G_R), \tag{6.4}$$

where G_R is the slope's gradient. The aerodynamic force is given by

$$D = \frac{1}{2}\rho C_D A V^2, \tag{6.5}$$

where ρ is the air density, C_D is the coefficient of drag, A is the projected surface area, and V is the bike-rider velocity. Substituting eqs (6.3),(6.4), and (6.5), into eq

(6.2), results in

$$a = g[\sin(\tan^{-1}(G_R)) - \mu\cos(\tan^{-1}(G_R))] - \frac{1}{2}\frac{1}{m}\rho C_D A V^2. \tag{6.6}$$

This can be written as

$$\frac{\partial v}{\partial t} = -Kv^2 + B, \tag{6.7}$$

where two positive constants are defined as

$$K = \frac{1}{2}\frac{1}{m}\rho C_D A \tag{6.8}$$

and,

$$B = g[\sin(\tan^{-1}(G_R)) - \mu\cos(tan^{-1}(G_R))]. \tag{6.9}$$

These two constants can be calculated from the known parameters $G_R, C_D A, \mu, \rho, m$.

Of prime interest is calculating the amount of time the cyclist needs to ride to reach a given distance on the slope. To this end, the differential equation (6.7) is first solved. Note that the terminal velocity is given by

$$\begin{aligned} 0 &= -Kv^2(\infty) + B \\ v(\infty) &= \sqrt{B/K}, \end{aligned}$$

where K, B are positive. Suppose that the initial velocity is faster than the terminal velocity

$$v(0) \equiv v_0 > \sqrt{\frac{B}{K}},$$

equation (6.7) can be rewritten as

$$\left[\frac{1}{v - \sqrt{\frac{B}{K}}} - \frac{1}{v + \sqrt{\frac{B}{K}}}\right] dv = -2K\sqrt{\frac{B}{K}}dt.$$

Making a use of

$$\int \frac{f'(v)}{f(v)}dv = \ln|f(v)| + constant$$

results in

$$\ln\frac{v - \sqrt{\frac{B}{K}}}{v + \sqrt{\frac{B}{K}}} = -2K\sqrt{\frac{B}{K}}t + constant.$$

A solution to the differential equation (6.7) is therefore given by

$$v(t) = \sqrt{\frac{B}{K}}\frac{1 + Je^{-2K\sqrt{\frac{B}{K}}t}}{1 - Je^{-2K\sqrt{\frac{B}{K}}t}}, \tag{6.10}$$

J is a constant to be determined from the initial conditions. For $t = 0$,

$$v_0 = \sqrt{\frac{B}{K}} \frac{1+J}{1-J}. \tag{6.11}$$

Therefore,

$$J = \frac{v_0 - \sqrt{\frac{B}{K}}}{v_0 + \sqrt{\frac{B}{K}}} \tag{6.12}$$

This results in

$$v(t) = \sqrt{\frac{B}{K}} \frac{1 + \frac{v_0 - \sqrt{\frac{B}{K}}}{v_0 + \sqrt{\frac{B}{K}}} e^{-2K\sqrt{\frac{B}{K}}t}}{1 - \frac{v_0 - \sqrt{\frac{B}{K}}}{v_0 + \sqrt{\frac{B}{K}}} e^{-2K\sqrt{\frac{B}{K}}t}}. \tag{6.13}$$

If the initial velocity is slower than the terminal velocity,

$$v_0 < \sqrt{\frac{B}{K}}$$

then,

$$v(t) = \sqrt{\frac{B}{K}} \frac{1 - Je^{-2K\sqrt{\frac{B}{K}}}}{1 + Je^{-2K\sqrt{\frac{B}{K}}}}. \tag{6.14}$$

Likewise,

$$J = -\frac{v_0 - \sqrt{\frac{B}{K}}}{v_0 + \sqrt{\frac{B}{K}}} \tag{6.15}$$

Substituting (6.15) into (6.14) demonstrate that the solution is identical to (6.13). Thus, (6.14) is the general solution for any initial conditions.

6.2.2 Relationship Between Time and Distance

Assume that the cyclist rides for a distance L for a time T. Then,

$$L(T) = \int_0^T v(t)\partial t = T\sqrt{\frac{B}{K}} + \frac{1}{K} \ln \frac{1 - Je^{-2K\sqrt{\frac{B}{K}}T}}{1 - J}. \tag{6.16}$$

Partially differentiating both sides with respect to v_0 , results in

$$\frac{\partial T}{\partial v_0} = -\frac{1 - e^{-2K\sqrt{\frac{B}{K}}T}}{2v_0 K\sqrt{\frac{B}{K}} + B(1 - e^{-2K\sqrt{\frac{B}{K}}T})} < 0. \tag{6.17}$$

This indicates that faster initial velocity results in a shorter time to reach L. This is almost trivial because v is monotonic. Equation (6.16) can be rewritten as

$$Je^{-2K\sqrt{\frac{B}{K}}T} + (1-J)e^{KL}e^{-K\sqrt{\frac{B}{K}}T} - 1 = 0. \tag{6.18}$$

This takes the form of a general second order equation of the form $AX^2 + BX + C = 0$,, where $X = e^{-K\sqrt{\frac{B}{K}}T}$ Solving this second order equation, results in

$$e^{-K\sqrt{\frac{B}{K}}T} = \frac{1}{2J}\sqrt{(1-J)^2e^{2KL} + 4J} - (1-J)e^{KL} \tag{6.19}$$

$$T(L) = -\sqrt{\frac{1}{KB}}\ln\left[\frac{1}{2J}\sqrt{(1-J)^2e^{2KL} + 4J} - (1-J)e^{KL}\right]. \tag{6.20}$$

The numerical results for $T(L)$ using Eq. (6.20) are shown.

6.2.3 Mathematical Model Implementation

The analytical solution to the mathematical model was implemented in code using Matlab 2011a (The Mathworks Inc., Massachusetts, USA). The range of drag area values that were measured experimentally for the two descending positions were used for the model simulation. A set of system parameters were used in simulation of the implemented analytical model to assess their effect on the model's output, descent time to completion. In order to assess the effect of coefficient of rolling resistance on a descent's time to completion, the range of $0.0042 - 0.0047$ was used for Coefficient of rolling resistance. This is consistent with the range of values reported in the literature [32, 118, 152]. The range represents a change in tyre air pressure from 130 to 100psi for a 700C rims and 23 mm 'clincher' tires. This was calculated using the relationship between air pressure and rolling coefficient described by Lim et al. [152]. The mass of the lightest of the participants tested and bicycle was used for athlete-bicycle system mass in simulation The effect of an addition of 2kg to the system mass on the difference in time to completion between the two riding positions was also simulated. A gradient slope of 10% and a descent of length 5km were used in the simulation. These values are consistent but not all encompassing with the range of descents typical in road racing. For example, the mostly straight descent from Cote de Pra'Martino (km 171) to Pinerolo during stage 17 of the 2011 Tour de France from Gap to Pinerolo was of 5.1km length of average gradient of 9.01%. An initial velocity of 50kmh was used as system's initial condition to represent the velocity in which cyclists assume one of the static descent riding positions. This is based on field observations of cycling races and the participants' self-report. Air Density of 1.1839 was used. This represents air density at sea level and temperature of 25^o. Considering all competing cyclists experience the same environmental conditions, the value of air density is not important for consideration and only used here for reference to standardised measures. For this reason, the effects of altitude, humidity and temperature on air density were not considered.

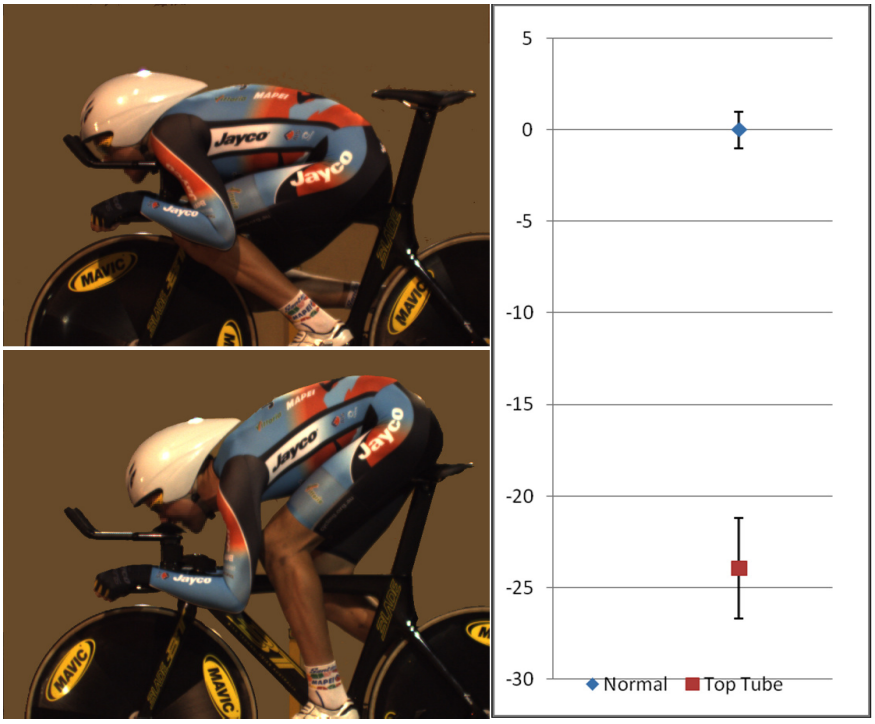


Figure 6.1: Elite Time Trial Cyclist in his 'normal' descending riding position (bottom), and in the 'top-tube' descending riding position (top) [75].

6.3 Results

6.3.1 Simultaion Results

The values of drag area values used for simulation of the implemented analytical model were 0.160-0.220m^2. This is consistent with the range of experimental values obtained in Drory and Yanagisawa [83]. Figure 6.2 shows the results of the model simulation for the time gap and difference in velocity for a cyclist using either the 'Normal' descending position with a CdA of 0.220m^2, or the 'Top-Tube' position with a CdA of 0.160m^2. The simulation was performed for a 5km descent of 10% gradient for a rider-bicycle with starting velocity of 50kmh and combined mass of 75kg, with environmental conditions of air density equivalent to sea level at 25^{0}C (1.1839) and rolling resistance coefficient of 0.0045.The figure shows the difference in velocity and time gap attributable to the difference in CdA between the two riding positions in 1km intervals. Under those conditions, the difference in time to completion is predicted to be reduced by 4.1sec after 1km and 29.2sec after 5km descent. The velocity difference is predicted to be 3.4(m/s) after the 5km descent. The simulated effect of 2kg added system mass to a cyclist in the 'Top-Tube' riding position on the difference in time to completion is indicated by the vertical bars in figure 6.2. The model predicts that additional 2.1sec can be saved after 5km descent by addition of this mass under the simulated conditions.

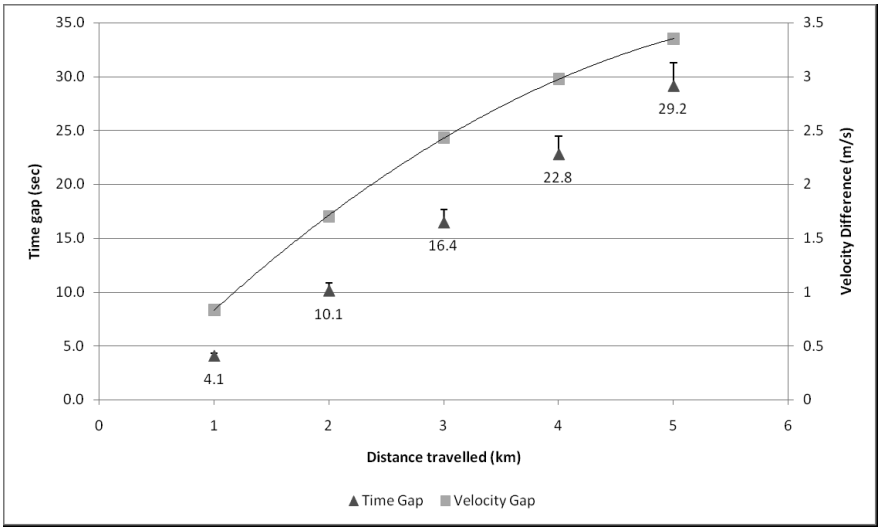


Figure 6.2: Simulated difference in velocity (square) and time to completion gap (triangle) between a cyclist with either a CdA of 0.160, which corresponds to the 'Top-Tube' position, and a CdA of 0.220, which corresponds to the 'Normal' position. The vertical error bars indicate the effect of addition of 2kg mass to the cyclist on the time gap.

Table 6.1 shows the model prediction of difference in time to completion where the rolling resistance is either 0.0042 or 0.0047 these values correspond to a reduced tyre pressure from 130psi to 100psi based on a cyclist with a CdA of $0.160m^2$ associated with the 'Top-Tube' descending position under the same course condition simulated above. The model predicts a negligible difference in time to completion of 0.48sec after 5km descent.

6.4 Discussion

The main findings of this study were that a static descending riding position can be found that reduces cyclist-bicycle system drag area by up to 25%, and that the saving in time for completion of a specific descent by a specific cyclist under known conditions can be estimated using a closed form solution of a non linear Ricatti type equation derived from equations of motion. The finding presented here are benchmark for riding positions for road descents. Reports of wind tunnel testing of the descending riding positions empirically tested here were not found in the literature. The values for drag area measured in this study (table 1) are considerably lower than drag area values reported for cycling riding positions reported previously [64, 97]. Defraeye et al. [64] reported values of 0.211, 0.243 and $0.270m^2$ for the time trial, dropped and upright riding positions respectively. García-López et al. [97] tested five professional cyclists in the time trial position. They reported drag area value range of $0.260 \pm 0.024m^2$ for static tests. However, considering there are no previous reports of wind tunnel testing of cyclists in descending riding positions, comparing

Distance (km)	Difference in time to completion (sec)
1	0.09
2	0.19
3	0.28
4	0.38
5	0.48

Table 6.1: Drag area wind tunnel measurement results for the 'Normal' and 'Top Tube' conditions

the magnitude of the measured drag area is not appropriate, and the use of the experimental values as input to the mathematical model (table 1) is justified. The time savings for the simulated conditions and parameters are only presented here as example use of the mathematical model. However, the authors consider the theoretical circumstance to be a realistic scenario for descents in road racing. Never the less, it demonstrates that significant gains can be made if a cyclist is prepared to undertake increased risk on descent. This highlights the importance of time saved on descent in road racing may often be overlooked, and the potential impact of time gained on descents on overall win. For this reason, the ability to apply the general mathematical model presented here to any descent racing situation may provide a powerful capability that enables the planning of racing strategy for a particular racing course and the evaluation of the likelihood of a strategy to succeed. It is important to acknowledge that assuming the 'Top-Tube' riding position on steep descents is far from trivial. It is likely that a trade off with bicycle handling and control exist, which increases the risk of fall and injury. The ability to predict the likelihood of success of a descending strategy may increase the ability of a cyclist to assess whether the increased risk is worthwhile. In a similar manner to reduction in aerodynamic drag, increased system mass will assist in reducing time to completion (figure 2). Using the simulation parameters used here, an addition of only 2kg reduced time to completion by 2.14sec after 5km for the parameters space used. It is not suggested that cyclists should add mass to the system throughout the race. This will have obvious disadvantage on ascents and flat sections. However, it may be advantageous for cyclists to take additional water and food supply prior to long descents in order to gain additional reduction in descent time, provided system aerodynamics is not compromised in the process. With prior knowledge of the difference in mass between a cyclist and an opponent it is possible to use the mathematical model presented here to estimate the changes in time gap at the completion of a descent that are attributable to mass difference alone. Changes in surface gradient have large effect on time to completion. However, the authors assume that the gradient is identical to all com-

petitors and therefore does not play a major role in race strategy. It is obvious that where possible, cyclists should select the steepest descent line in order to maximise velocity. A negligible effect on time to completion was also demonstrated for simulated changes to rolling resistance (table 2). This was simulated for rolling resistance values that correspond with reduction of tyre pressure from 130psi to 100psi. The small effect suggests that little may be gained by increasing tyre pressure beyond current functional range in order to minimise rolling resistance. The applicability of the modelling technique presented here is limited by the lack of methodology to estimate Aerodynamic force in field conditions, and the inability to link laboratory measured CdA for a cyclist with the field performance. Future research should focus on developing methodology to enable estimation of cyclists' projected frontal surface area and pose characteristics from lab and field observations. The general mathematical model presented here can be applied to other sporting events where time to completion of a descent is a determinant of overall performance, such as downhill skiing or ski jumping. As is the case for road cycling, the challenge in those cases is the field quantification of CdA and friction.

6.4.1 Assumptions and limitations

A number of assumptions were used in the development of the mathematical model described here for the purpose of simplification; The sums of the forces in the x direction (parallel to the slope) as used in equation 2 represent a private case where the cyclist assumes a static position and produces no force to aid system propulsion. The general case for the forces along the slope should be

$$\sum F_x : mg\sin\theta - GRF_x - D = ma$$

, where the ground reaction force is comprised of both rolling resistance drag and the couple of propulsive forces applied by the cyclist through the transmission. The

$$GRF_x = F_{RR} + F_{prop}$$

, where F_{prop} is the couple of propulsive forces applied by the cyclist. Here the authors assume that $F_{prop} = 0$. Inspection of power meter data made available by riders reveal that mean power for steep descent sections are common to be below 10% of mean propulsive power for a race, and extensive periods of zero cadence. For this reason, the assumption is reasonable. Furthermore, when cyclists apply propulsive force through limb and crank motion, additional factors may affect aerodynamic drag associated with the cycling cadence. Controlling the limb and crank motion in a static pose here avoids the impact of cadence effects, which are poorly understood. The distribution of GRF between Front and rear wheels' contact with the ground, the effect of forces in the Z and Y directions were not considered. The reasons for this are lack of methodology to enable field empirical measurement of those parameters, lack of literature reference value and the individual race course and cyclist specificity. Consequently, moments around the contact points with the ground, lateral lean, steering

and lift forces were not considered. However, their relative contribution can be considered negligible and therefore their omission justified. The rolling resistance was assumed to be constant and independent of velocity. It was also assumed that the traction force is near its peak value such that the maximum theoretical speed may be obtained. This is consistent with Lim et al. [152]. Gravitational acceleration was assumed to be constant. In reality it is a function of location, elevation and system mass. The surface gradient was assumed to be constant. In reality it is a function of position and direction of motion. Air density was assumed to be constant and independent of altitude. In reality it is a function of elevation (pressure), humidity and temperature. Considering all cyclists in a race are likely to experience the same environmental conditions, this should not affect the time difference between the cyclists. For this reason this assumption is reasonable.

6.5 Conclusions

Benchmark empirical wind tunnel evaluation of aerodynamic drag of elite cyclists in two riding positions for road racing descents was conducted. A general mathematical model was developed to predict performance as measured by time to completion in road cycling descents. Based on the non-linear equations derived for the cyclist-bicycle system's equations of motion, closed-form solutions for determination of cycling speed and time to completion of road descents are obtained. With the closed-form solutions, the estimation of velocity and time to completion becomes straightforward. The evaluation of system performance under the influence of system parameters (percent reduction in drag area, mass), initial conditions (initial velocity) and environmental conditions (gradient, descent length, and rolling resistance) is possible. Simulation results for a range of realistic system parameters and initial conditions were presented. The magnitude of time savings simulated here is indicative of the potential that exists for reducing time to completion on descents. For this reason, the ability to apply the general mathematical model presented here to any racing situation may provide a powerful capability that enables the planning of racing strategy for a particular racing course and the evaluation of the likelihood of a strategy to succeed. In order to extend the model's ecological validity, future research is required to develop methodology to enable estimation of cyclists' projected frontal surface area and pose characteristics from lab and field observations.

Part III

Approaches to Human Detection and Tracking

Whereas previous parts proposed techniques for pose estimation and shape recovery in a common movement modality including the interaction with objects, this part looks at techniques for understanding human motion in a less common movement modality (slalom kayaking), where the task is detection and spatial tracking of the human-object complex in a very challenging environment and therefore rarely studied.

chapter 7 develops a detection and tracking framework for the challenging problem of automatic annotation of sport races from broadcast image sequences. The framework uses a discriminative cascade of regressors based detector for its prior and periodic regularisation that is trained offline, and a discriminative correlation filter based technique for its online tracking. The chapter also introduces a simple and effective technique for shot transition detection in broadcast image sequences and annotation of a course's obstacles. The ideas in this chapter have been submitted for publication.

Chapter 7

Automated Detection and Tracking of Slalom Paddlers from Broadcast Image Sequences using Cascade Classifiers and Discriminative Correlation Filters

abstract

This chapter addresses the problem of automatic detection and tracking of slalom paddlers through a long sequence of sports broadcast images comprised of persistent view changes. In this context, the task of visual object tracking is particularly challenging due to frequent shot transitions (i.e. camera switches), which violate the fundamental spatial continuity assumption used by most of the state-of-the-art object tracking algorithms. The problem is further compounded by significant variations in object location, shape and appearance in typical sports scenarios where the athletes often move rapidly. To overcome these challenges, we propose a *Periodically Prior Regularised Discriminative Correlation Filters* (PPRDCF) framework, which exploits recent successful Discriminative Correlation Filters (DCF) with a periodic regularisation by a prior that constitutes a rich discriminative cascade classifier. The PPRDCF framework reduces the corruption of positive samples during online learning of the correlation filters by negative training samples. Our framework detects rapid shot transitions to reinitialise the tracker. It successfully recovers the tracker when the location, view or scale of the object changes or the tracker drifts from the object. The PPRDCF also provides the race context by detection of the ordered course obstacles and their spatial relations to the paddler. Our framework robustly outputs the evidence base pre-requisite to derived race kinematics for analysis of performance. Experiments are performed on task-specific dataset containing Canoe/Kayak Slalom race image sequences with successful results obtained.

7.1 Introduction

Figure 7.1: Our PPRDCF framework outputs location and scale of the slalom paddler, and the location and order of the gates.

In competitive Canoe/Kayak Slalom (CK Slalom), negotiation of obstacles through gates is the fundamental skill and key determinant of overall performance. In race context where the winner is commonly decided by fractions of a second, minimising task time-to-completion is paramount. Thus, developing an optimal strategy and techniques for negotiation of gates that minimises overall course time-to-completion is critical. However, there is currently little quantitative data that characterises the trajectory of gate negotiation in Slalom.

Through extensive literature survey, we have found but only one paper that attempted to characterize the strategy employed by slalom paddlers in negotiation of upstream gates [125]. It analyzed upstream gate negotiation strategies of 17 elite Slalom paddlers using manual extraction of spatial kinematic data of the boat and athletes' head from image sequences obtained by overhead camera. The utility of the methodology used by Hunter [125] is, however, limited by the use of a custom calibration rig when there is no water on the course, obtrusive attachment of markers to the boat and athlete, and laborious object labelling for extraction of trajectory kinematic information. In order to be relevant in elite sport training environment or competition and improve the likelihood of feedback driven technical or tactical amendments, an analysis method must provide near real-time results.

In this work, we investigate the challenging problem of simultaneous human detection and long-term tracking from readily available image sequences comprised of persistent view changes obtained from multiple uncalibrated cameras typical of broadcast image sequences. This task serves as a crucial evidence base, a prerequisite to kinematic motion analysis of athletes aimed to optimise technique and performance in sport (see figure 7.2). We aim to tackle the limitations of existing visual object detection and object tracking algorithms especially for long term sequence with frequent view changes. We develop a new and unified framework for object detection and tracking from disparate multi-view image sequences that cou-

ples the advantages of each approach to overcome the limitations of the other. The method is applied to detection and tracking of CK Slalom paddlers through gate negotiation of a race course, which enables near real-time performance analysis.

Contributions A Periodically Prior Regularised Discriminative Correlation Filters (PPRDCF) framework is proposed for tracking fast moving objects in sport event using broadcast image sequences with possibly frequent shot transitions. Our framework exploits recent successfully applied Spatially Regularised Discriminative Correlation Filters (SRDCF) [58] with a periodic regularisation by a prior discriminative cascade classifier that is learnt offline. To overcome tracking failure associated with rapid shot transition, we introduce a robust adaptive shot transition detection algorithm that allows soft initialisation of the tracker. Finally, our framework provides race context through the detection of course obstacles and their spatial relations to the paddler. We perform experiments on task-specific dataset containing CK Slalom race image sequences and compare our results to state-of-the-art trackers. Our framework robustly outputs the evidence base pre-requisite to derived race kinematics for analysis of performance.

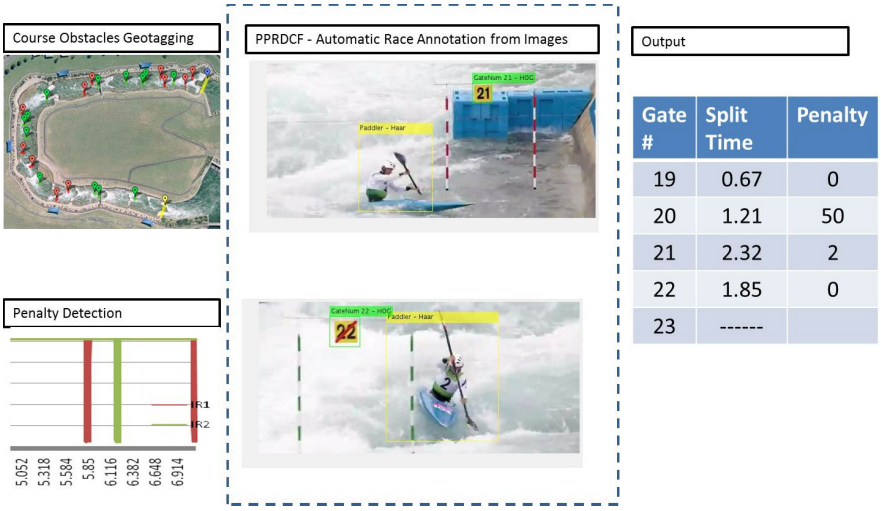


Figure 7.2: An illustrative schematic overview of a CK Slalom annotation system for the daily training environment and competition. The system includes global race course and obstacles geotagging, penalty detection and race annotations from image sequences, and outputs a detailed comprehensive race annotations including split times and penalties. This chapter focuses on the race annotations from image sequences (encompassed by the dashed line).

7.2 Related Work

A comprehensive survey of visual object tracking is outside of the scope of this chapter. Instead, this section presents a brief survey of recent techniques relevant to our

task, to provide the context for our new method.

Visual Tracking

Visual tracking is an important computer vision problem of estimating an object's kinematics from an image sequence. State-of-the-art tracking algorithms generalise the object's appearance from a small set of training samples. The tracker then performs temporal search for probabilistically matching candidates in the spatial vicinity of the object's previous image location under the assumption of trajectory smoothness [20, 119]. Online learning trackers update the model with the selected candidate [9, 59, 113].

Human movement tracking is challenging due to the varied pose and appearance caused by severe occlusions induced by the articulated body motions. The challenge is compounded in typical broadcast of sporting races, where an athlete rapidly changing pose, occlusion and appearance. In our CK Slalom task, the pose can rapidly change form front to rear view, and from top view of paddler and boat to bottom view of the boat and no visibility of the paddler. Further, the paddler is often partially or fully submerged or severely occluded from view by obstacles or water. These present critical challenges to tracking algorithms that assume small changes in the object's pose or appearance.

Many tracking algorithms model the image background [218] or extract a temporal flow field [27, 73] to aid the object tracking under the strong assumption that the object differs from the background in either appearance or motion. For example, most tracking benchmark datasets constitute image sequences of a stationary background (e.g. roads, streets or buildings) and a moving object (e.g. cars, bicycles or pedestrians). In our CK Slalom task, however, the background water often flows in the same direction as the paddler and rapidly changes in appearance due to illumination and reflections, essentially eliminating the realistic option of background modelling. Moreover, image sequences with rapid shot transition from multiple moving cameras remain the majority of available race and training data. This severely violates the global smoothness and brightness assumptions for dense correspondences requirements of tracking algorithms [27, 122]. Hence, shot transition detection and regularisation or re-initialisation of the tracker model is paramount to enhance the chance of recovery from occlusion or loss of track, contamination by negative samples, or rapid change in pose or appearance due to fast motion, or sudden change in view.

Recent best performing tracking algorithms use a Fast Fourier Transform (FFT) based Discriminative Correlation Filter (DCF) approach (see section 7.3.1). The approach, however, accumulates errors during online learning and typically drifts from the object, as detection recovery after occlusion is poor [58]. Consequently, even the current state-of-the-art tracker is not robust in long-term tracking of rapidly moving and deforming objects (see fig. 7.3).

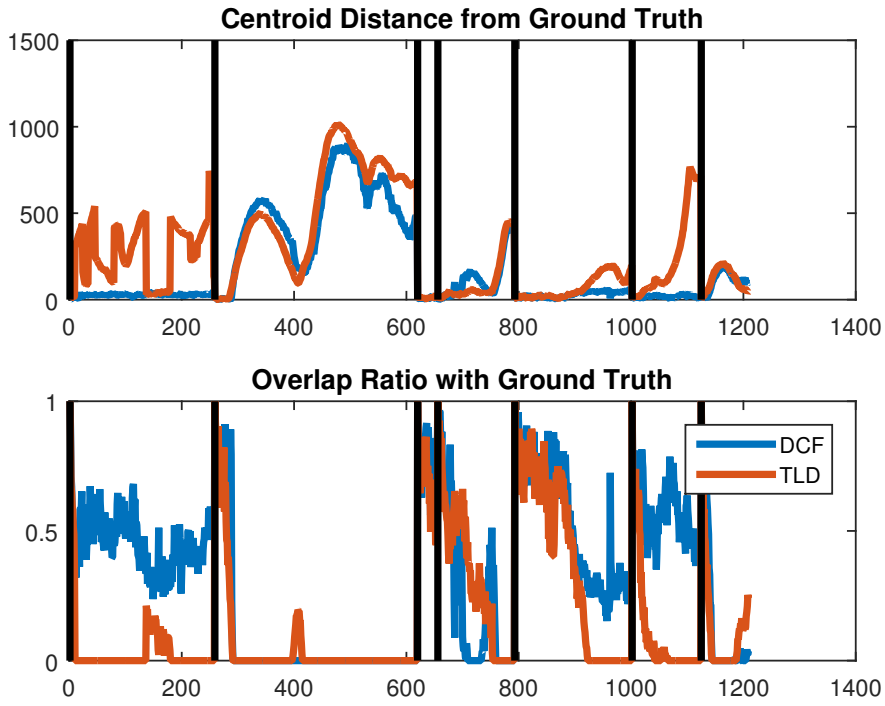


Figure 7.3: Results of state of the art trackers on our test data demonstrating rapid tracking deterioration, manifested by rapid loss of overlap with ground truth (bottom figure; value approaching 1 is indicative of good performance) and increase in precision score (top figure; value approaching 0 is indicative of good performance) within camera view (shot). Vertical black lines indicate shot transition (a sudden change of camera view). This violates the continuity assumption of the tracker and requires reinitialisation (here, by ground truth).

Object Detection

Similar to visual tracking, object detection, often a prerequisite to tracking algorithms, is a challenging computer vision problem due to the variable appearance and pose that may be present. Robust discriminative methods extract an extensive set of features at multiple scales from positive and negative image samples to train a classifier. Due to the very large feature set, learning is computationally expensive and typically performed offline. The resultant classifier is comprised of rich feature descriptors that capture the object of interest. In inference, a multi-scale sliding window search scheme scores candidate patches and the best scored patches are selected as detections. While this approach has been very successful [57, 71, 92, 100, 260], the computational cost of feature extraction and exhaustive search restricts the scheme from being used for tracking at every time step of an image sequence. Moreover, a detection score does not guarantee temporal continuity of an object detection. Post-hoc regimes are required to enforce spatial continuity [121, 240]. Furthermore, the scheme is susceptible to candidate proposal multiplicity. A heuristic, adaptive [153] or learnt score threshold influences the number of detections selected and subse-

quently the incidence of false positive and false negative detections.

7.3 Periodically Prior Regularised DCF (PPRDCF)

In this section, we introduce our Periodically Prior Regularised Discriminative Correlation Filters (PPRDCF) unified framework that couples the advantages of each technique to overcome the limitations of the other. Our framework exploits recent successfully applied Spatially Regularised Discriminative Correlation Filters (SRDCF) [58] with a periodic regularisation by a prior discriminative cascade classifier that is learnt offline. We introduce the tracking framework in section 7.3.1 followed by the object detection using a cascade of rejectors classifiers in section 7.3.2. Section 7.3.3 introduces our shot transition detection algorithm that enables soft re-initialisation of the tracking framework. In addition, our framework provides race context by detecting the ordered course obstacles and their spatial relations to the paddler (further details are provided in section 7.4.4).

7.3.1 Discriminative Correlation Filters

Discriminative approaches cast tracking as an online learning and classification problem that differentiates the tracked object from the background. Given an image patch containing the object, a classifier is learnt that discriminates the object from the environment, a process akin to tracking-by-detection. Robust object detector classifiers, however, richly characterise not only the object of interest, but importantly its environment through a very large number of negative samples. This computationally expensive and time consuming typically offline process is not feasible for online tracking algorithms, for which speed is a critical performance criterion. For this reason, discriminative tracking approaches use a compromise approach that severely under-samples negatives [9, 114], consequently, acutely hindering the tracker's performance [120].

Current state-of-the-art tracking algorithms use DCF approach, in which a correlation filter is trained from a set of samples and a periodic extension of these samples [20, 120, 151]. Significant improvements in trackers' performance has risen from recent work that formulates the convolution of two patches as an element-wise product in the Fourier domain [20]. Henriques et al. [119] demonstrated that translations of an image patch containing the object can be modelled as cyclically shifted signals using circulant matrices. Thus, the classifier training and detection computational cost is significantly reduced when the computations are efficiently performed using FFT.

The periodic extension reduces the contamination of the filter by negative samples. Consequently, a better representation of the object is learned. The approach, however, suffers from boundary contamination effects that result in inferior representation of the object and introduce inaccuracies to the learned object model. Thus, reducing its discriminative power [58, 96]. This problem was partially addressed

in Danelljan et al. [58] by utilising spatial regularisation component in the objective function.

7.3.2 Cascade of Rejectors Classification

The object detection problem involves recognition of the desired object, its location and scale in an image. We note that in recent years deformable parts based methods (DPM) outperform cascade classifiers on standard object detection tasks [45, 92, 249]. DPM, however, strongly relies on a spatial relations of parts model, which cannot handle severe occlusions that are commonly present in our task.

More recently, deep learning of convolutional neural network (CNN) produced the state-of-the-art performance on standard detection tasks [100, 192]. In CNN, high level features replace low and middle level features with improved discriminative power. Notwithstanding, these features are very expensive to compute. Selective search strategies to reduce the computational cost of using CNN [100] resulted in object localisation errors [121]. Importantly, both low computational cost and accurate object localisation are critical to our framework. For these reasons, we opt to use cascade classification for object detection in our framework tasked with tracking initialisation and periodic tracking regularisation.

A cascade detector uses a sequence of node classifiers to distinguish objects from non-objects and simultaneously select weak features to form strong ensemble classifiers using adaptive boosting (AdaBoost). The work of Viola and Jones [230] leverages the scarcity of the object of interest relative to the background to achieve efficient detection by early rejection of most easily classified negative features. They also introduced integral images for fast feature computation and utilising AdaBoost for automatic feature selection. These ideas remain a foundation for modern detectors.

Viola and Jones [231] used low level Haar features due to low computational cost achieved with the aid of integral images. However, the cascade classification approach can be used with other feature descriptors. Significant improvements were obtained by using mid-level features, such as Histogram of Oriented Gradients (HOG) in detection [57] and in speed [260].

7.3.3 Shot transition

Typical broadcast sport image sequences are characterised by frequent shot transitions. We introduce a robust adaptive shot transition detection algorithm that allows soft initialisation of the tracker. Further details are provided in 7.4.3

7.3.4 Our PPRDCF framework

Conceptually, our framework is most similar to Kalal et al. [129] in adopting a unified framework that distinguishes between the detection and tracking tasks. Kalal et al. [129] described a framework of three sub-tasks of Tracking, Learning and Detection (TLD). The TLD uses a naive geometric shape template matching method

with median flow for tracking and a cascade classifier with online learning. We argue that while operating independently, the aggregation of the TLD's three sub tasks are equivalent to a modern DCF tracker with an online learning of the tracked patch. Therefore, due to its online learning component it suffers from the error accumulation problem of DCF trackers. Instead, we opt for using a true independent offline learnt detector to complement an online DCF tracker. Furthermore, to enhance the overall framework performance, our detector uses a different feature descriptor than the tracker. This enhances the cumulative discriminative power, as the two models hold complementary characteristics of the object.

In agreement with Kalal et al. [129], we accept the view that neither tracking nor detection can solve the task independently. We support the view that the two approaches can be complementary. A detector can initialise the tracker, provide tracking validation and failure recovery to a tracker. A tracker accumulates temporal object localisation and can reduce the computational cost and running time of the detection.

We define a spatial discrepancy signal between the prior classifier and the tracker. To overcome the inherent limitation of tracking algorithms, the object evidence accumulated by the tracker and the discrepancy signal are used to prune false positive and overcome false negative detections. Within each shot sequence, our proposed PPRDCF formulation introduces a penalty term on the correlation filter coefficients during online learning. The prior regularisation reduces the corruption of positive samples during online learning of the correlation filters by negative training samples. Consequently the PPRDCF successfully recovers the tracker when the location, view or scale of the object changes or the tracker loses the object.

Alternatively, DCF tracking can be characterised as tracking-by-detection. However, unlike robust object detector classifiers, rich characterisation of the object is not possible due to the high computational cost. Hence, DCF uses a compromise approach that restricts the object region and severely under-samples negatives acutely hindering the trackers performance [120]. Our framework can then be viewed as tracking by weak detection classifier with periodic update by a rich detection classifier. Our observation is that the weak online learnt detector is likely to become contaminated, whilst the rich offline learnt detector will remain immune to contamination and will retain its strong discriminative power.

The basic structure of our framework is as follows; For initialisation and regularisation of the tracker we construct a rejection cascade classifier similar to Viola and Jones [230] and described in section 7.4.1. For tracking, we construct a DCF following Danelljan et al. [58] as detailed in section 7.4.2. For shot transition detection we use an adaptive outlier detection method described in section 7.4.3. The race annotation component is detailed in section 7.4.4. Figure 7.4 depicts the outline of our framework.

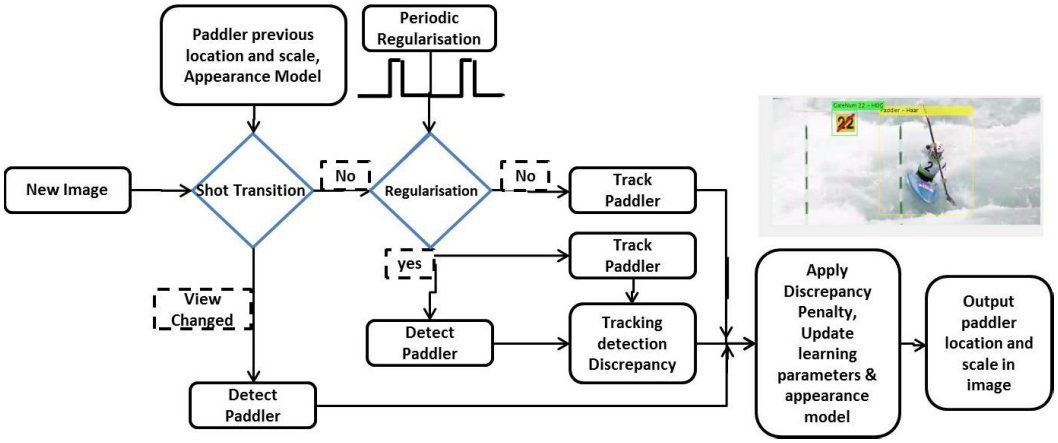


Figure 7.4: Overview of our PPRDCF algorithm overview (see details in text)

7.4 System Implementation

7.4.1 Paddler Detection

Our object detection framework overview is depicted in figure 7.5 for learning and in figure 7.6 for inference. We construct a rejection cascade similar to Viola and Jones [230]. Essentially, a cascade classifier forms a degenerate decision tree, where a negative classification of an image patch results in rejection of the patch. A positive classification is passed on for evaluation at the subsequent classifier. In this manner easily classified patches are rejected early with improved overall efficiency due to the observation that an image consists of mostly negative samples. Only positive samples will be evaluated by a classifier at every stage.

At each stage a classifier is trained on the examples that were evaluated as positives in all preceding stages. Consequently, the classifier's complexity and discriminative power increases as stages increase due to the escalating task difficulty. We use 20 stages in our implementation, predicated on our experiments described in section 7.5.3.

The computational cost of training a cascade classifier is significant. Inference, however, is fast due to the cascade of rejectors and boosting. This makes the approach suitable for complementary detection in our tracking framework.

In inference, a sliding window approach is employed to evaluate the classifier score function f over rectangular sub-regions of the image I at multiple scales. We select the object's region $\tilde{R}$ to be its maximum as

$$\tilde{R} = \underset{R \subseteq I}{\operatorname{argmax}} f(R|\tilde{\mathbf{x}}), \tag{7.1}$$

where $\tilde{\mathbf{x}}$ is the learnt object's appearance model and R ranges over all rectangular sub-regions of the image I. We incrementally scale the search region as defined by $\lceil (T_m \cdot F_s^n) \rceil$, where T_m indicates the median patch size used to train the classification model (117×124), F_s a scale factor determined by the ratio between the size of the

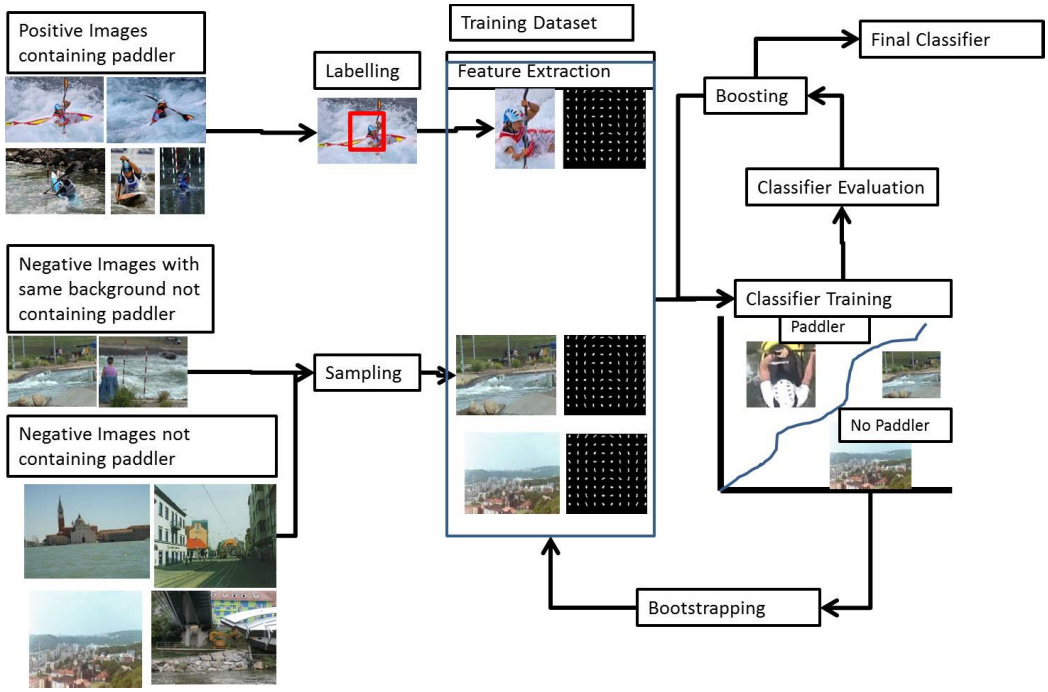


Figure 7.5: Training a cascade classifier of a Slalom paddler. See text for details.

input image I and T_m and the number of increments N(7 in our implementation), and $n \in \{1, \ldots, N\}$ is an indicator function of the current increment.

In our experiments (see section 7.5.3), using rich mid-level feature descriptors (HOG and LBP) in our detection classifier results in rare false positives, but high levels of false negatives. In contrast, using Haar features produced fewer false negatives, but significantly more false positives. In our framework, detection of false positives corresponds to a high discrepancy signal between detector and tracker. Hence, they are naturally handled by a severe penalty imposed by a penalty function (eq. (7.5)) that changes the tracker's online learning parameters.

Furthermore, Wojek and Schiele [240] showed that a combination of several feature descriptors outperforms any single feature descriptor. Considering our tracker already uses HOG features, a complementary detector using a different feature descriptor is preferred, as it is likely to capture ancillary aspects of the object. For these reasons, we opt for using Haar features in our detector model.

7.4.2 Paddler Tracking

The standard DCF tracker [119, 120] is essentially a regressor $g(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle$, where ϕ represents the mapping to the Hilbert space induced by a kernel function and $\mathbf{w}$ is the discriminative model. Considering all the previous image patches $\{\mathbf{x}^k, k = 1, ..., t-1\}$ of size $M \times N$ centred on the object, this regressor can be effectively

trained by optimizing

$$\min_{\mathbf{w}} \sum_{k=1}^{t-1} \alpha_k \sum_i (\langle \mathbf{w}, \phi(\mathbf{x}_i^k) \rangle - y_i)^2 + \lambda \|\mathbf{w}\|^2, \tag{7.2}$$

where k denotes the frame index and α_k is the frame weight. The matrix $\mathbf{x}_i^k$, $i \in \{0, \ldots, M-1\} \times \{0, \ldots, N-1\}$ is a cyclic shift version of the image patch $\mathbf{x}^k$. The scalar y_i is the Gaussian-shaped regression target based on the periodic shift of patch $\mathbf{x}^k$.

The power of the DCF tracker lays in the fact that all possible cyclic shifts of the object image patches are taken into account to train the model while the solution to the optimisation problem (eq. (7.2)) can be efficiently computed using Discrete Fourier Transform (DFT). To track an object at frame t, the responses of all cyclic shifts of a test image sample can be obtained efficiently in the same way. The location corresponding to the cyclic shift with the maximal response is treated as the final result.

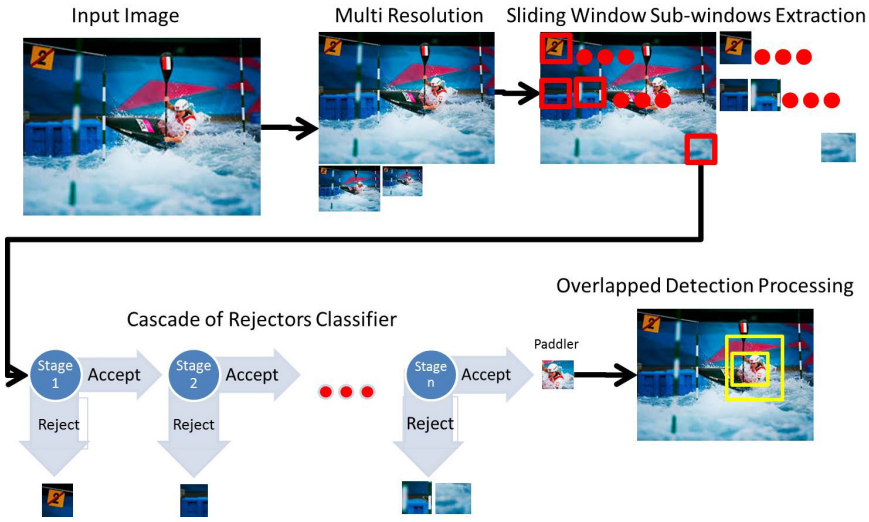


Figure 7.6: Inference detection of a slalom paddler in a new sample using a cascade classifier. See text for details

SRDCF [58] extended the standard formulation of DCF to address issues caused by the underlying periodic assumption. It also adapts to object scale change by applying the regressor at multiple resolutions similar to Yang et al. [247]. We therefore employ SRDCF as the tracking component and denote its output as the object's region $\hat{R}^t$.

The discriminative model $\mathbf{w}$ can be updated via incremental learning as the new training sample $\mathbf{x}^t$ becomes available. The weight α_k is a key factor associated with each training sample from frame k. In the original DCF formulation, it is updated by

$$\alpha_k^t = (1-\gamma)\alpha_k^{t-1} \tag{7.3}$$

where $\gamma = \gamma_0 = 0.01$ is a fixed learning rate used to control the speed of adapting to the new object appearance.

To properly incorporate the regularization information from the detector, we choose to adapt the learning rate γ according to the discrepancy signal between the detector and the tracker. Specifically, we enforce a strong impact upon the tracker's update when the discrepancy signal is large by setting γ to a higher value. This implies a detection bounding box that is far from the estimated tracking result, and indicates that the tracker has most likely experienced significant failure or drift. Thus,

$$\gamma^t = c \exp(-\frac{1}{\sigma_l^2} \| d(\hat{R}^t, \tilde{R}^t) - 1 \|^2), \tag{7.4}$$

where c and σ_l are constants and $d(\hat{R}^t, \tilde{R}^t) = 1 - (\hat{R}^t \cap \tilde{R}^t)/(\hat{R}^t \cup \tilde{R}^t)$ is the discrepancy metric based on overlap between two bounding boxes $\hat{R}^t$ and $\tilde{R}^t$, estimated from the paddler tracker and detector respectively.

We update the new region of the object from $\hat{R}^t$ and $\tilde{R}^t$ by linear interpolation using the discrepancy signal $d(\hat{R}^t, \tilde{R}^t)$, such that

$$R^t = (1 - d(\hat{R}^t, \tilde{R}^t))\hat{R}^t + d(\hat{R}^t, \tilde{R}^t)\tilde{R}^t. \tag{7.5}$$

This is to say, when the discrepancy signal is large, we enhance the impact of the detection result when updating the new position and scale of the object.

Algorithm 1 PPRDCF Paddler Tracking

Input: Image I^t, Image I^{t-1}
Previous target region R^{t-1}
The paddler tracker model $\mathbf{w}^{t-1}$, a constant learning rate $\gamma_0 = 0.01$ and the global static detector's object appearance model $\tilde{\mathbf{x}}$
Output: target region R^t and updated tracker model $\mathbf{w}^t$

if detectShotTransition$(I^t, I^{t-1}) = 0$ **then**
  if *period* mod $frameNum \neq 0$ **then** ▷ Standard SPRDCF
    $R^t \leftarrow trackPaddler(I^t, R^{t-1}, \mathbf{w}^{t-1})$
    $\mathbf{w}^t \leftarrow updateTrackerModel(R^t, \gamma_0)$
  else ▷ Periodic regularisation - detector's model
    $\tilde{R}^t \leftarrow detectPaddler(I^t, \tilde{\mathbf{x}})$ ▷ eq (7.1)
    $\hat{R}^t \leftarrow trackPaddler(I^t, R^{t-1}, \mathbf{w}^{t-1})$
    update γ^t using $d(\hat{R}^t, \tilde{R}^t)$ ▷ eq (7.4)
    update R^t using $d(\hat{R}^t, \tilde{R}^t), \hat{R}^t, \tilde{R}^t$ ▷ eq (7.5)
    $\mathbf{w}^t \leftarrow updateTrackerModel(R^t, \gamma^t)$
else ▷ Shot transition detected
  $R^t \leftarrow detectPaddler(I^t, \tilde{\mathbf{x}})$ ▷ eq (7.1)
  $\gamma^t = 1$
  $\mathbf{w}^t \leftarrow updateTrackerModel(R^t, \gamma^t)$

7.4.3 Shot Transition Detection

Frequent shot transitions are characteristic of broadcast image sequences. This severely violates the spatial continuity assumption of tracking algorithms. It is therefore necessary to re-initialise the tracker when a change of view takes place.

To detect shot transition we employ a simple yet effective method. We let the distance metric $d_{t-1,t}$ between two consecutive frames be the Mean Square Error (MSE) of pixel intensities between the frames. We fit a Gaussian distribution $\{\mu^{t-1}, \sigma^{t-1}\}$ to the accumulated distance metric of all preceding images. A new frame is considered to be a shot transition, if the likelihood of its distance metric from its predecessor is outside 3 standard deviations from the expectation, i.e., $|d_{t-1,t} - \mu^{t-1}| > 3\sigma^{t-1}$, where $d_{t-1,t}$ is the distance metric between two consecutive frames. The Gaussian model is then incrementally updated with the new data if no shot transition is detected.

Upon detection of shot transition, we employ a 'soft' re-initialisation scheme. The scheme involves re-localisation of the object's position using the detector, and short-term enhanced learning parameters through boosted γ in eq. (7.5). This reflects a higher level of trust in the rich detector's model.

7.4.4 Race Annotation

For kinematic analysis, in contrast to enhancing spectator experience and entertainment [132], athlete tracking is only useful, when the context of the motion is understood. Unlike sports like Football [124] or Athletics [90], where the field of play is known, constrained and can be modelled, in CK Slalom no two venues are the same, the water flow is rapidly changing as are the obstacles and navigation gates' positions. Hence, in order for athlete tracking to be relevant for further kinematic analysis, the context of the motion in relation to the environment is necessary.

For CK Slalom, our framework detects the location of the gates, through which the athletes need to navigate, and their assigned order (see figure 7.7). This is performed via a discriminative cascade classifier similar to our paddler detector in 7.4.1 that is trained offline, with the exception of the feature descriptor used (HOG). The classifier outputs the location of the gate poles and number. This enables further association between the athlete tracking and the race context. This analysis is outside the scope of this chapter.

Gate Number Identification Once a gate number object is detected using a cascade classifier, the gate number is identified using a trained multi-class linear Support Vector Machine (SVM) classification with a 'one-vs-one' scheme over HOG features (see figure 7.9). Since a finite maximal number of 24 gates may be used in slalom competition and to avoid aggregation of single digit models for two-digit numbers, our framework learns 24 distinct number classes. The model is learnt from 4×4 cell HOG features extracted for each training image in our gate number dataset. This dataset contains 201 32×32 image patches per number class, which were extracted from the SlalomImRV dataset (described below) and scaled. We note that a diagonal

Figure 7.7: Gate poles and number detection using discriminative cascade classifier

red line may be present on gate numbers in slalom to indicate illegal direction of gate negotiation. The gate number dataset contains both appearance types. The number identification training framework is depicted in figure 7.8.

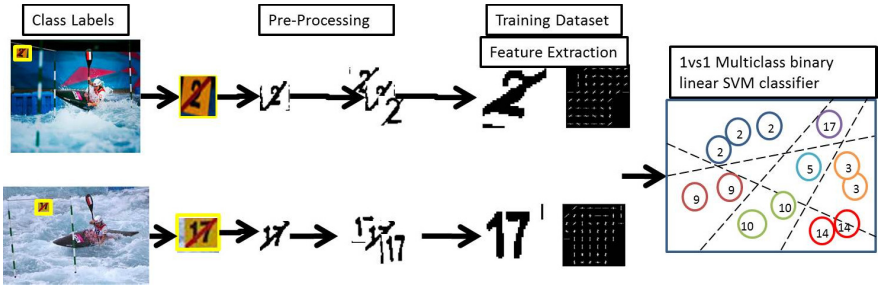


Figure 7.8: Gate number identification using a multi-class linear SVM classification

Multi-class linear SVM classification is a mature technique. Its goal is to construct a function that will correctly predict the class of a new sample to one of K different classes. In its basic form the problem can be trivially decomposed into mutually exclusive binary classification problems. The one-vs-one method constructs $k(k-1)/2$ classifiers, where each classifier is trained as a binary classifier on two classes. Thus, given a set of m training samples $(x_1, y_1), \ldots, (x_m, y_m)$, where $x_i \in R^K, i = 1, \ldots, m$ and $y_i \in \{1, \ldots, k\}$ is the class of x_i, the i, j SVM classifier solves the binary classification problem

$$D_{ij}(\mathbf{x}) = \mathbf{w}_{ij}^t x + b_{ij}, \tag{7.6}$$

where each $\mathbf{w}_{ij}$ is a m-dimensional vector, b_{ij} is a scalar and $D_{ij}(\mathbf{x}) = -D_{ji}(\mathbf{x})$. For the input vector $\mathbf{x} \in R^K$ we calculate

$$D_i(\mathbf{x}) = \sum_{j \neq i, j=1}^{k} \text{sign}(D_{ij}(\mathbf{x})) \tag{7.7}$$

and classify $\mathbf{x}$ into the class

$$\arg \max_{i=1,\ldots,k} D_i(\mathbf{x}). \tag{7.8}$$

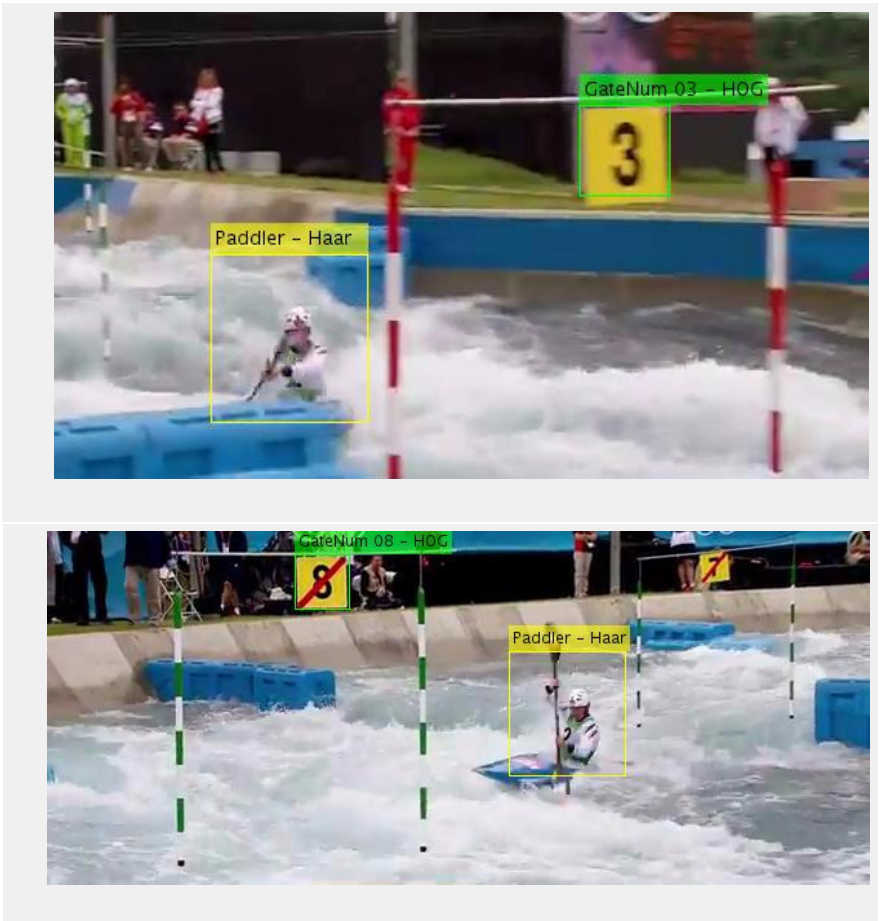


Figure 7.9: Gate Number identification results using multiclass linear SVM classifier.

7.5 Quantitative Evaluation

This section reports on a set of quantitative experiments to evaluate the performance of PPRDCF. The experiments are conducted on a new challenging task-specific dataset containing image sequences of slalom paddling with rapid shot transition, and frequent changes in object appearance and pose in unconstrained environment.

7.5.1 Datasets

The SlalomImRV dataset contains 404 images of slalom paddlers, typically in training or competition *in natura*, whose location in each image was manually annotated by a bounding box. The dataset contains images that have been either captured by the authors, or obtained from publicly available online repositories with a licence search criteria set to creative commons (Flickr, Vimeo), or labelled for reuse (Google Images). The dataset contains images of paddlers with a wide variety of appearance, pose, garments, and in varied lighting and illumination conditions. Moreover, this dataset contains images with many instances of occlusions and self-occlusion conditions including severe cases caused by partial submersion in the water environment, which present a significant challenge to the classifier's model.

To train our *detection* models, we have split the SlalomImRV dataset into a standard 80% and 20% for model learning and cross-validation sets respectively. Our negative set is comprised of 1694 images from the positive and negative training images of the INRIA Person [57] and Parse [191] datasets. In both datasets, the positive training images contain images of people, whilst the negative sets contain mostly background scenery images. Using images that contain people in our negative set ensures that our paddler detection model discriminates well between people displaying a variety of activities and paddlers for our specific task.

The SlalomVidRV dataset contains 30 broadcast image sequences of slalom competition races of 70 to 110 seconds in duration captured at 25 to 30Hz. These sequences have been either captured by the authors, or obtained from publicly available online repositories with a licence search criteria set to creative commons (Flickr, Vimeo) from three competition venue locations that distinctively differ from the images in the SlalomImRV dataset. Specifically, the datasets differ in venue locations where the images and sequences were captured. This is reflected in distinct appearance of the scene, obstacles, slalom gates, and gate numbering, as well as the environmental and lighting conditions in the SlalomVidRV dataset compared to the SlalomImRV dataset. Importantly, the majority of camera views of the capture differed. In generating our detector model from a dataset distinct from the dataset used for testing the tracking algorithm, we ensure that our framework generalises well within the activity-specific target application. Using this dataset, section 7.5.2 reports on the performance of our shot transition detector. We report on the performance of our paddler detection module and the empirical selection of feature type and number of cascade stages, which were performed on a subset of this dataset as described in section 7.5.3. Section 7.5.4 evaluates the performance of the unified tracking framework.

7.5.2 Shot Transition Results

For all image sequences of the SlalomVidRV dataset, we manually annotated all frames with a latent random binary variable indicating the ground truth for shot transition. To evaluate the performance of our shot transition detection we calculate precision and recall using $p_t/(p_t + p_f)$ and $p_t/(p_t + n_f)$ respectively, where p_t is

true positive, p_f is false positive and n_f is false negative with respect to the ground truth. The high precision and recall and low p_f achieved by our results (see table 7.1) indicate that this algorithm is very effective in detecting shot transition.

Table 7.1: Shot Transition Detection Results

#sequences	#frames evaluated	$\#p_t$	$\#p_f$	Precision	Recall
30	60,669	233	14	0.94	1

7.5.3 Paddler Detection Results

For selection of detector feature type and number of cascade stages, and to separately evaluate the performance of our paddler detection module, we empirically tested the detector on 3 levels of features (Haar, LBP and HOG) and 3 levels of number of cascade stages (15, 20 and 30) experimental conditions. The image test set that was used for these experiments consisted of 1500 images randomly extracted from 3 image sequences in our SlalomVidRV dataset (500 random images per sequence) that were manually annotated for the paddler's ground truth location with a bounding box. The images in this set distinctively differ from the images in the SlalomImRV dataset that was used to train the detector. The difference in venue location is reflected in distinct appearance of the scene and the environmental and lighting conditions.

In addition to precision and recall metrics, we considered the *precision score* defined as the centre of the paddler's bounding box relative to the ground truth, and the *success ratio* defined in 7.5.4. The scores on each comparison metric were then averaged across the image sequences and are presented for all experimental conditions in table 7.2 (best score for each metric is indicated in bold).

These results indicate that using Haar features resulted in slightly lower precision, and significantly inferior p_f, precision and success ratio scores compared to using rich mid-level feature descriptors (HOG and LBP). However, since the role of the detector in our framework is to initialise and recover the tracker, we consider the Haar feature's superiority in n_f, number of detections and recall more critical to the overall performance of the framework. Hence, the paddler model trained on Haar feature descriptor was selected for our detection module. In addition, since p_f corresponds to a high discrepancy signal between detector and tracker, they are naturally handled by the penalty imposed by the penalty function (eq. (7.5)).

Intuitively, a feature descriptor that aggregates a number of existing descriptors may result in superior detection performance. However, due to the computational efficiency of using Haar features, we decided against it.

Likewise, the empirical results in table 7.2 show the superior performance of using 20 cascade stages over the alternative experimental conditions in the number of detections, n_f, and recall. Hence, the corresponding paddler model was selected for use in our framework.

Table 7.2: Paddler Detection Results (mean per 500 images)

	Feature Type			# Cascade Stages		
	Haar	**LBP**	**HOG**	**15**	**20**	**30**
# Detections	**367.67**	203.50	268.00	257.50	**297.33**	284.33
# p_t	**298.67**	180.33	231.67	214.33	**259.67**	236.67
# p_f	69.00	**23.17**	36.33	43.17	**37.67**	47.67
# n_f	**132.33**	296.50	232.00	242.50	**202.67**	215.67
Precision	0.81	**0.89**	0.86	0.83	**0.89**	0.84
Recall	**0.69**	0.38	0.50	0.47	**0.57**	0.53
Precision Score	103.24	22.67	**16.74**	**31.60**	53.24	57.81
Success Ratio	0.25	**0.44**	0.36	**0.36**	0.34	0.35

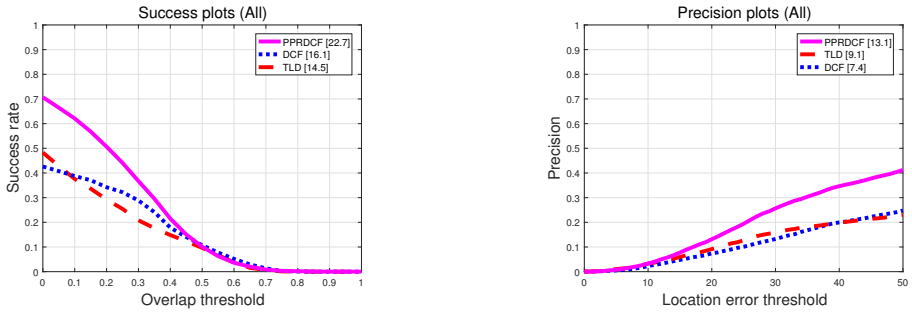


Figure 7.10: *Success* and *precision* plots comparing the performance of our PPRDCF with state of the art trackers initialised by detection. Algorithms are ranked by the area under the curve and the precision score (20 pixels threshold [244]). Our method (magenta) consistently achieves superior performance.

7.5.4 Paddler Tracking Results

To evaluate the PPRDCF, we compare its performance with two tracking algorithms; SPRDCF [58] and TLD [129]. The SPRDCF is currently the state-of-the-art tracker, and TLD is a tracker that conceptually is comparable to our framework in using detection to initialise the tracker and assist in tracker failure recovery.

The experiments in this section adopt the following evaluation protocol; We employ the one-pass evaluation that takes the ground truth at the first frame of a sequence as the initialization bounding box then run each tracker until the last frame. The produced trajectory is then compared to manually labelled ground truth using the standard *precision score* and *success ratio* metrics [244]. For each tracker, we calculate a discrepancy signal for the detected objects' location error and overlap ratio with respect to the ground truth. The *precision score* calculates the rate of frames whose centre location is within a certain threshold distance with the ground truth. Here, we use a commonly used threshold of 20 pixels following Wu et al. [244]. This metric emphasizes how well a tracker is able to clasp the target. The *success ratio* calculates the same ratio based on bounding box overlap threshold $(B^* \cap B_{gt})/(B^* \cup B_{gt})$, where B^* and B_{gt} are the estimated and ground truth bounding boxes' areas, respectively. This metric indicates how well a tracker adapts and covers the

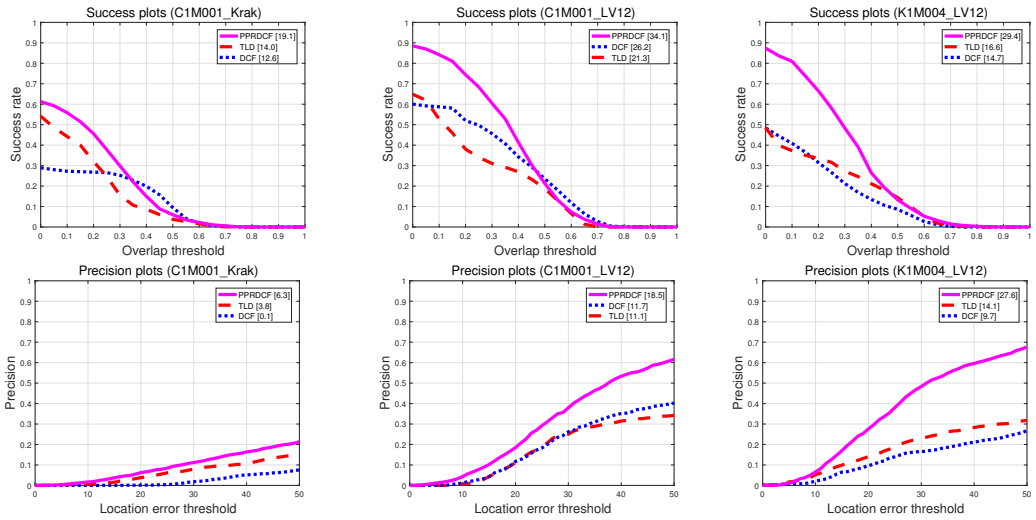


Figure 7.11: *Success* and *precision* plots comparing the performance of our PPRDCF (magenta) with state of the art trackers on three specific image sequences from two distinct venues (_krak and _LV) and two slalom paddling types (C1 - single blade canoe; and K1 - double blade kayak) reflecting varied appearance and environmental conditions. Our method consistently achieves superior performance.

target. A typical value is 0.5 as used in object detection evaluation [88]. Thus, for both metrics, higher performance is represented by a greater area under the graph. The results are summarised in figure 7.10. We also present results on 3 sample image sequences of two different slalom disciplines (C1 - single blade canoe, and K1 - double blade kayak) that were captured in different venues and present distinct variation in the appearance of the scene and environmental conditions in figure 7.11. Qualitative results from these image sequences are presented in figure 7.12. We provide the Area Under Curve (AUC) in the figures, which represents the average of all success ratios at different thresholds when the thresholds are evenly distributed.

Further, we performed initialisation experiments with two additional experimental conditions; For the TLD and SPRDCF trackers, we performed separate experiments with initialisation by our paddler detector result and with ground truth bounding box input *at each shot transition*. The former represents an equal opportunity for the three trackers tested, having an identical initialisation. The latter is a standard tracker testing procedure where initialisation is provided by the ground truth, but is only applied to the TLD and SPRDCF trackers. This places our PPRDCF at a disadvantaged starting point. Nevertheless in both experimental conditions the PPRDCF outperforms the TLD and SPRDCF trackers for both precision score and success ratio. We note that the first experimental condition represents a more realistic scenario for automatic systems, where tracker initialisation requires detection.

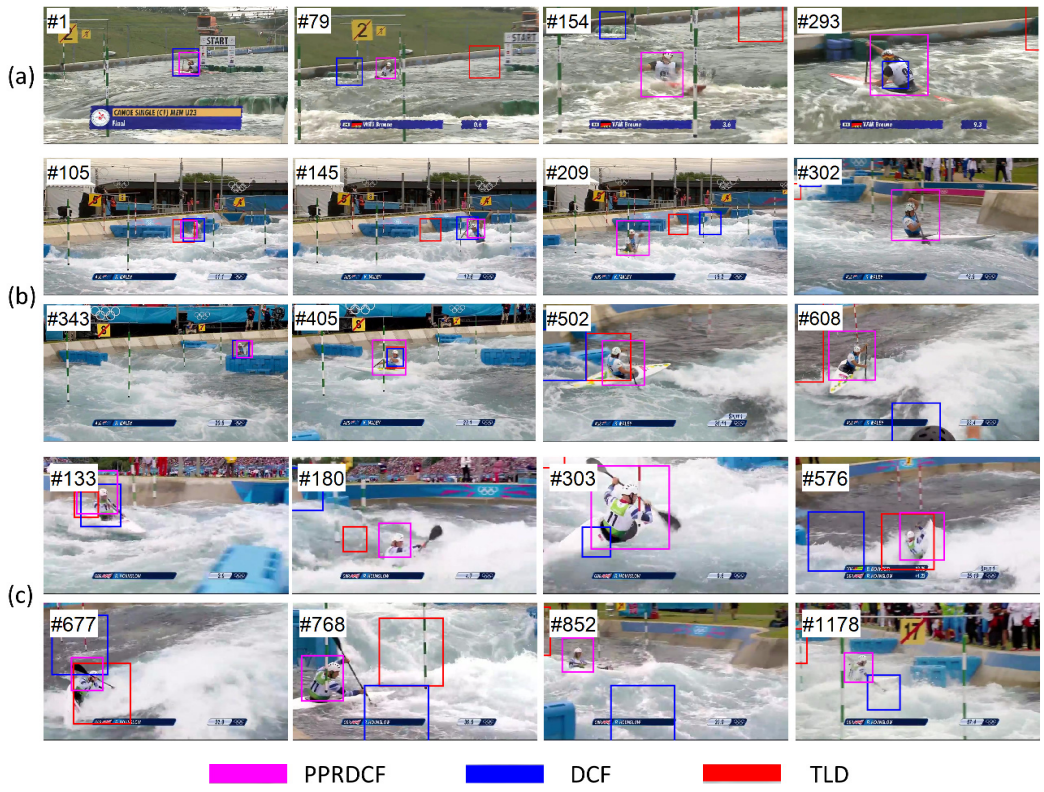


Figure 7.12: Qualitative comparisons of our PPRDCF (magenta) with state of the art trackers on sample image sequences; (a) C1M001_Krak depicts a sequence of a C1 (single blade canoe) at venue #1. (b) C1M001_LV12 depicts a sequence of a C1 at venue #2. (c) K1M004_LV12 depicts a sequence of a K1 (double blade kayak) at venue #2. Our method attains robust tracking performance and detection recovery in challenging scenarios including fast motion, motion blur, deformation, and severe occlusion resulting from paddler submersion (#180 and #852).

7.6 Discussion

In this chapter, we investigated the challenging problem of simultaneous human detection and long-term tracking from image sequences comprised of persistent transitioned shots obtained from multiple moving cameras typical of broadcast image sequences, where the object changes appearance frequently as it moves in and out of the camera view.

We introduced Periodically Prior Regularised DCF framework, which uses complementary detection and tracking models. We introduced a robust shot transition detection algorithm for tracking re-initialisation. For tracking our framework uses spatially regularised discriminative correlation filters. For detection, we use offline trained cascade of rejectors classifier. We demonstrated that exploiting the periodic regularisation and camera shot transition detection results in tracker failure and drift recovery.

Our experiments demonstrated that our framework outperforms the state-of-the-art trackers on a new task-specific dataset. The output of our framework forms a critical component and a crucial evidence base pre-requisite to kinematic motion analysis of athletes aimed to optimise technique and performance in sport.

One interesting consequence of our approach is the case where no paddlers are present in an image sequence. As afore stated, initialisation is a pre-requisite to our tracking module. This means that no tracking takes place unless the detector finds a likely object candidate. It is possible that in the absence of a paddler, a false detection initialises the tracker, but then the tracker's performance deteriorates rapidly. However, when a true detection becomes available, the framework naturally overrides the tracker and recovers the position to the paddler's location in the image. In fact, the periodic regularisation of the online tracker's object model by a fixed rich object descriptor highlights a major advantage of our approach. This is further enhanced by our use of different feature descriptors in the detector module (Haar) and the tracker (HOG), as it allows a capture of ancillary characteristics of the object's appearance.

Limitations and Future Work A number of challenges remain that need addressing for enhanced robustness of the PPRDCF. For instance, despite the periodic update by the detector, which reduces the contamination of the tracker model, the DCF tracker still suffers from drift. This strengthens the argument for enhanced negative sampling of the background used in the online training of the tracker. In the current implementation, the detection method was selected for its low computational cost in inference with disregard to its cost in offline training. It would be interesting to test the performance of recent state-of-the-art CNN detection techniques as an alternative in the framework. Furthermore, inherent to the common use of rectangular image patch to represent the object, is a degree of background contamination in the object's model. The extent to which this degree of contamination should be controlled by non-rectangular patches or advantageous is still debated by the community.

For our task, we achieved robust results with a task-specific trained classifier. It is, however, impossible to design a discriminative classifier for the general case because of the high variability in the appearance of human ambulation. The performance of the approach critically relies on time consuming and costly process and the availability of a large quantity of training samples. Further, the generalisation of the approach can only be achieved through the addition of adequate training samples, as its adaptability to unseen body postures is low and typically manifested by poor performance were occlusions exist.

An interesting extension to our framework will exploit recent advances in simultaneous detection and human pose estimation. These methods exploit in addition to the appearance of the object's parts, the spatial [249] and temporal [45] relations of the parts. This, however, requires pose estimation algorithms to better handle occlusions and self-occlusions than has so far been achieved.

Detailed performance and skill execution analysis of a paddler negotiating a slalom course requires the construction of a 3D model of the scene (slalom course)

that is outside the scope of this chapter. Nevertheless, whilst 3D reconstruction of *dynamic* scenes from images remains an open problem, the information extracted by our race annotation framework naturally provides additional scene depth information and thus can serve as a strong cue for computation of partial sparse reconstruction using known multi-view relations. In particular, the slalom gate design is standardised to include known sized poles with a known within-gate gap, as well as height-fixed alternating pole colour segments (green/white for downstream gates and red/white for upstream gates, see fig. 7.7). Intensity and feature-based tracking of these structures should simplify the correspondence problem for stereo matching techniques from which globally consistent slalom course mosaics can be generated and is intended for future work.

Appendices

A.1 Part Spatial Relation and Deformation Cost

In section 2.2.3, we represent the spatial relations between adjacent body parts (joints) by a quadratic deformation vector $\psi(\mathbf{l}_i, \mathbf{l}_j) = [dx, dy, dx^2, dy^2]^T$ from the relative position of the connected joints v_i at $\mathbf{l}_i$ and v_j at $\mathbf{l}_j$, such that $dx = x_i - x_j$ and $dy = y_i - y_j$, consistent with Yang and Ramanan [249] and others. We mention that this term can be interpreted as a negative spring energy resulting from pulling joint j from a relative position with respect to joint i. For clarity, in this section we provide context and alternate interpretations of this force. In order to avoid notational clutter, we omit the subscript ij for the remainder of this section, where the interpretation is obvious.

Consider a multivariate Gaussian distribution over the set of *relative* positions of the adjacent joints (v_i, v_j) in our training data, where $\bar{} = (\mu_x, \mu_y)$ is the mean of the distribution. We wish to impose a penalty over a relative joint location hypothesis that deviates from the mean of the distribution. Consequently, a body part location hypothesis would favour a relative location proposal that is in agreement with the learnt prior spatial relations model. Thus, we define a spring force that represents the distance of the test point from the distribution's centre of mass, which we define by

$$d_{ij}(\mathbf{l}_i, \mathbf{l}_j) = A(dx - \mu_x)^2 + B(dy - \mu_y)^2 \tag{A.1}$$

where, A and B are arbitrary real-valued constants. This can be expanded to

$$d_{ij}(\mathbf{l}_i, \mathbf{l}_j) = Adx^2 + Bdy^2 - 2A\mu_x dx - 2B\mu_y dy + (A\mu_x^2 + B\mu_y^2). \tag{A.2}$$

The last term does not depend on $\mathbf{l}_i$ or $\mathbf{l}_j$ and can be replaced with a constant F. After rearrangement the force can be written as the inner product of two vectors plus a constant F

$$\begin{aligned} d_{ij}(\mathbf{l}_i, \mathbf{l}_j) &= \langle (dx, dy, dx^2, dy^2), (-2A\mu_x, -2B\mu_y, A, B) \rangle + F \\ &= \langle (dx, dy, dx^2, dy^2), w_{ij}) \rangle, \end{aligned} \tag{A.3}$$

where only the first vector depends on the proposed body part locations $\mathbf{l}_i$ or $\mathbf{l}_j$ and therefore corresponds to our deformation vector $\psi(\mathbf{l}_i, \mathbf{l}_j)$, and w_{ij} is the vector of

weight coefficients to be learnt. Equation A.2 can also be viewed as a special case of the general canonical form of an arbitrarily oriented ellipse

$$Adx^2 + Bdy^2 + Cdxdy + Ddx + Edy + F = 0,$$

where $C = 0$. Therefore, a generalisation of our approach may represent the deformation cost by $\psi^\star(\mathbf{l}_i, \mathbf{l}_j) = [dx, dy, dxy, dx^2, dy^2]^T \in \mathbb{R}^5$ with the addition of a cross dimension term dxy.

More generally, an arbitrarily oriented ellipsoid in $\mathbb{R}^n$, centred at $\bar{}$ satisfies

$$(\mathbf{l} - \bar{})^T \boldsymbol{S} (\mathbf{l} - \bar{}) = 1,$$

where $\boldsymbol{S}$ is a positive definite matrix and $\mathbf{l}$ and $\bar{}$ are vectors. After substitution we get

$$\begin{aligned} \begin{bmatrix} dx & dy & 1 \end{bmatrix} \begin{bmatrix} S & -S\mu \\ -\mu^T S & \mu^T S \mu \end{bmatrix} \begin{bmatrix} dx \\ dy \\ 1 \end{bmatrix} &= 0 \\ \begin{bmatrix} dx & dy & 1 \end{bmatrix} \begin{bmatrix} A & C/2 & D/2 \\ C/2 & B & E/2 \\ D/2 & E/2 & 1 \end{bmatrix} \begin{bmatrix} dx \\ dy \\ 1 \end{bmatrix} &= 0, \end{aligned} \tag{A.4}$$

where the five upper triangular elements of the symmetric matrix here correspond to the five elements of (A, B, C, D, E).

For a probability distribution we can define a dissimilarity measure between two random vectors $\mathbf{l}$ and $\bar{}$ of the same distribution as

$$\boldsymbol{D}(\mathbf{l}, \bar{}) = \sqrt{(\mathbf{l} - \bar{}) \boldsymbol{S}^{-1} ((\mathbf{l} - \bar{})},$$

where $\boldsymbol{S}$ is the covariance matrix. In our application, this distance, the Mahalanobis distance, represents the distance of the test point $\mathbf{l}$ from the distribution's centre of mass $\bar{}$. The eigenvectors of $\boldsymbol{S}$ define the principal axes of the ellipsoid and the eigenvalues of $\boldsymbol{S}$ are the reciprocals of the squares of the semi-axes. This representation also provides an elegant link to Principal Component Analysis (PCA). Indeed, this is the form of the deformation cost that was used in early pictorial structures work (e.g. Felzenszwalb and Huttenlocher [92]).

Bibliography

1. Agarwal, A., Triggs, B., Jan. 2006. Recovering 3d human pose from monocular images. Pattern Analysis and Machine Intelligence (T-PAMI). IEEE Computer Society Transactions on 28 (1), 44–58. (cited on page 17)

2. Ahmad, R., Rambely, A., Lim, L., 2008. Solving biomechanical model using third-order runge-kutta methods. In: Computer Mathematics. Springer, pp. 163–168. (cited on pages 93 and 95)

3. Alam, F., Subic, A., Akbarzadeh, A., 2009. Aerodynamics of bicycle helmets (p68). In: Estivalet, M., Brisson, P. (Eds.), The Engineering of Sport 7. Springer, pp. 337–344. (cited on page 92)

4. Amamoto, M. Y., Agishita, K. Y., 2000. Scene constraints-aided tracking of human body. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. Vol. 1. pp. 151–156. (cited on page 36)

5. Andersen, M. S., Benoit, D. L., Damsgaard, M., Ramsey, D. K., Rasmussen, J., 2010. Do kinematic models reduce the effects of soft tissue artefacts in skin marker-based motion analysis? an in vivo study of knee kinematics. Journal of Biomechanics 43 (2), 268–273. (cited on pages 3 and 16)

6. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014. 2d human pose estimation: New benchmark and state of the art analysis. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 3686–3693. (cited on page 30)

7. Andriluka, M., Roth, S., Schiele, B., 2009. Pictorial structures revisited: People detection and articulated pose estimation. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. (cited on pages 17, 30, and 38)

8. Baak, A., Müller, M., Bharaj, G., Seidel, H.-P., Theobalt, C., 2013. A data-driven approach for real-time full body pose reconstruction from a depth camera. In: Consumer Depth Cameras for Computer Vision. Springer, pp. 71–98. (cited on page 17)

9. Babenko, B., Yang, M.-H., Belongie, S., Aug 2011. Robust object tracking with online multiple instance learning. Pattern Analysis and Machine Intelligence (T-PAMI). IEEE Computer Society Transactions on 33 (8), 1619–1632. (cited on pages 112 and 114)

10. Bangpeng, Y., Li, F.-F., 2012. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. Pattern Analysis and Machine Intelligence (T-PAMI). IEEE Computer Society Transactions on 34 (9), 1691–1703. (cited on page 36)

11. Barelle, C., Chabroux, V., Favier, D., 2010. Modeling of the time trial cyclist projected frontal area incorporating anthropometric, postural and helmet characteristics. Sports engineering 12 (4), 199–206. (cited on pages 73 and 92)

12. Bauer, J. J., Pavol, M. J., Snow, C. M., Hayes, W. C., 2007. Mri-derived body segment parameters of children differ from age-based estimates derived using photogrammetry. Journal of biomechanics 40 (13), 2904–2910. (cited on page 6)

13. Bini, R. R., Tamborindeguy, A. C., Mota, C. B., 2010. Effects of saddle height, pedaling cadence, and workload on joint kinetics and kinematics during cycling. Journal of Sport Rehabilitation 19 (3), 301–314. (cited on page 93)

14. Blair, K. B., Sidelko, S., 2009. Aerodynamic performance of cycling time trial helmets (p76). In: Estivalet, M., Brisson, P. (Eds.), The Engineering of Sport 7. Springer, pp. 371–377. (cited on page 92)

15. Blocken, B., Defraeye, T., Koninckx, E., Carmeliet, J., Hespel, P., July 2011. Numerical study of the interference drag of two cyclists. In: Wind Engineering (ICWE). Proceedings of the International Conference on. pp. 1–8. (cited on page 72)

16. Blocken, B., Defraeye, T., Koninckx, E., Carmeliet, J., Hespel, P., September 2012. Rans and les simulations of the interference drag of two cyclists. In: Xiang, H.-F., Ge, Y.-J., Cao, S.-Y. (Eds.), Bluff Body Aerodynamics and Applications (BBAA). Proceedings of the International Colloquium on. China Communication Press, pp. 1–10.

17. Blocken, B., Defraeye, T., Koninckx, E., Carmeliet, J., Hespel, P., 2013. Cfd simulations of the aerodynamic drag of two drafting cyclists. Computers & Fluids 71 (0), 435–445. (cited on pages 8, 70, and 72)

18. Bloom, V., Makris, D., Argyriou, V., 2012. G3d: A gaming action dataset and real time action recognition evaluation framework. In: Computer Vision and Pattern Recognition Workshops (CVPRW). Proceedings of the IEEE Computer Society Conference on. pp. 7–12. (cited on page 54)

19. Bobbert, M. F., Schamhardt, H. C., Nigg, B. M., 1991. Calculation of vertical ground reaction force estimates during running from positional data. Journal of Biomechanics 24 (12), 1095–1105. (cited on page 2)

20. Bolme, D., Beveridge, J., Draper, B., Lui, Y. M., June 2010. Visual object tracking using adaptive correlation filters. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 2544–2550. (cited on pages 112 and 114)

21. Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. Pattern Analysis and Machine Intelligence (T-PAMI). IEEE Computer Society Transactions on 23 (11), 1222–1239. (cited on page 31)

22. Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32. (cited on page 59)

23. Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A., 1984. Classification and regression trees. CRC press. (cited on page 59)

24. Bresson, X., Vandergheynst, P., Thiran, J., 2006. A variational model for object segmentation using boundary information and statistical shape prior driven by the mumford-shah functional. International Journal of Computer Vision 68 (2), 145–162. (cited on page 75)

25. Brownlie, L., Kyle, C., Carbo, J., Demarest, N., Harber, E., MacDonald, R., Nordstrom, M., 2009. Streamlining the time trial apparel of cyclists: the nike swift spin project. Sports Technology 2 (1-2), 53–60. (cited on page 92)

26. Brownlie, L., Ostafichuk, P., Tews, E., Muller, H., Briggs, E., Franks, K., 2010. The wind-averaged aerodynamic drag of competitive time trial cycling helmets. Procedia Engineering 2 (2), 2419–2424. (cited on page 70)

27. Brox, T., Bruhn, A., Papenberg, N., Weickert, J., May 2004. High accuracy optical flow estimation based on a theory for warping. In: Computer Vision (ECCV). Proceedings of the European Conference on. Vol. 3024 of Lecture Notes in Computer Science. Springer, pp. 25–36. (cited on page 112)

28. Burgess, S., 1998. Improving cycling performance with large sprockets. Sports Engineering 1 (2), 107–113. (cited on page 92)

29. Cámara, J., Maldonado-Martín, S., Artetxe-Gezuraga, X., 2012. Influence of the position on the bicycle on the frontal area in road cyclists. Coll. Antropology 36 (2), 529–534. (cited on page 73)

30. Camomilla, V., Bonci, T., Dumas, R., Cheze, L., Cappozzo, A., 2015. A model of the soft tissue artefact rigid component. Journal of Biomechanics 48 (10), 1752–1759. (cited on pages 3 and 16)

31. Camomilla, V., Donati, M., Stagni, R., Cappozzo, A., 2009. Non-invasive assessment of superficial soft tissue local displacements during movement: A feasibility study. Journal of Biomechanics 42 (7), 931–937. (cited on pages 3 and 16)

32. Candau, R. B., Grappe, F., Menard, M., Barbier, B., Millet, G. Y., Hoffman, M. D., Belli, A. R., Rouillon, J. D., 1999. Simplified deceleration method for assessment of resistive forces in cycling. Medicine and science in sports and exercise 31, 1441–1447. (cited on page 98)

33. Capelli, C., Schena, F., Zamparo, P., Dal Monte, A., Faina, M., Di Prampero, P. E., April 1998. Energetics of best performances in track cycling. Medicine & Science in Sports & Exercise 30 (4), 614–624. (cited on page 73)

34. Cappozzo, A., Della Croce, U., Leardini, A., Chiari, L., 2005. Human movement analysis using stereophotogrammetry: Part 1: theoretical background. Gait & posture 21 (2), 186–196. (cited on pages 3 and 16)

35. Caselles, V., Kimmel, R., Sapiro, G., Feb. 1997. Geodesic active contours. International Journal of Computer Vision 22 (1), 61–79. (cited on page 74)

36. Cerveri, P., Pedotti, A., Ferrigno, G., 2005. Kinematical models to reduce the effect of skin artifacts on marker-based human motion estimation. Journal of Biomechanics 38, 2228–2236. (cited on page 6)

37. Ceseracciu, E., Sawacha, Z., Fantozzi, S., Cortesi, M., Gatta, G., Corazza, S., Cobelli, C., 2011. Markerless analysis of front crawl swimming. Journal of biomechanics 44 (12), 2236–2242. (cited on page 17)

38. Chabroux, V., Barelle, C., Favier, D., 2008. Aerodynamics of time trial bicycle helmets (p226). In: Estivalet, M., Brisson, P. (Eds.), The Engineering of Sport 7. Springer, pp. 401–410. (cited on page 92)

39. Chabroux, V., Mba, M. N., Sainton, P., Favier, D., ???? Wake characteristics of time trial helmets using piv-3c technique. In: Applications of Laser Techniques to Fluid Mechanics. Proceedings of the International Symposium on. (cited on page 92)

40. Challis, J. H., 1995. A procedure for determining rigid body transformation parameters. Journal of biomechanics 28 (6), 733–737. (cited on pages 3 and 16)

41. Chan, T., Zhu, W., June 2005. Level set based shape prior segmentation. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. Vol. 2. pp. 1164–1170. (cited on page 74)

42. Chen, X., Yuille, A., 2014. Articulated pose estimation by a graphical model with image dependent pairwise relations. In: Advances in Neural Information Processing Systems (NIPS). (cited on page 18)

43. Chen, Y., Tagare, H. D., Thiruvenkadam, S., Huang, F., Wilson, D., Gopinath, K. S., Richard, Briggs, W., Geiser, E. A., 2002. Using prior shapes in geometric active contours in a variational framework. International Journal of Computer Vision 50, 315–328. (cited on page 74)

44. Cheng, C.-K., Chen, H.-H., Chen, C.-S., Lee, C.-L., Chen, C.-Y., 2000. Segment inertial properties of chinese adults determined from magnetic resonance imaging. Clinical biomechanics 15 (8), 559–566. (cited on page 6)

45. Cherian, A., Mairal, J., Alahari, K., Schmid, C., 2014. Mixing body-part sequences for human pose estimation. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. (cited on pages 17, 30, 34, 44, 115, and 129)

46. Chowdhury, H., Alam, F., Mainwaring, D., 2011. A full scale bicycle aerodynamics testing methodology. Procedia Engineering 13, 94–99. (cited on page 93)

47. Chu, X., Ouyang, W., Li, H., Wang, X., 2016. Structured feature learning for pose estimation. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. (cited on page 18)

48. Cooper, K., 1993. Bluff-body aerodynamics as applied to vehicles. Journal of Wind Engineering and Industrial Aerodynamics 49 (1), 1–21. (cited on page 92)

49. Cootes, T., Taylor, C., 1999. A mixture model for representing shape variation. Image and Vision Computing 17 (8), 567–573. (cited on page 78)

50. Cootes, T., Taylor, C.J.and Cooper, D., Graham, J., 1995. Active shape models-their training and application. Computer Vision and Image Understanding (CVIU) 61 (1), 38–59. (cited on page 74)

51. Corazza, S., Andriacchi, T. P., 2009. Posturographic analysis through markerless motion capture without ground reaction forces measurement. Journal of biomechanics 42 (3), 370–374. (cited on pages 8, 17, and 53)

52. Corazza, S., Mündermann, L., Andriacchi, T., 2007. A framework for the functional identification of joint centers using markerless motion capture, validation for the hip joint. Journal of biomechanics 40 (15), 3510–3515. (cited on page 17)

53. Corazza, S., Mündermann, L., Chaudhari, A. M., Demattio, T., Cobelli, C., Andriacchi, T. P., 2006. A markerless motion capture system to study musculoskeletal biomechanics: Visual hull and simulated annealing approach. Annals of Biomedical Engineering 34 (6), 1019–1029. (cited on pages 8, 17, and 53)

54. Cremers, D., Tischhäuser, F., Weickert, J., Schnörr, C., 2002. Diffusion Snakes: Introducing statistical shape knowledge into the Mumford–Shah functional. International Journal of Computer Vision 50 (3), 295–313. (cited on pages 74 and 75)

55. Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E., 2010. Regression forests for efficient anatomy detection and localization in ct studies. In: Medical Computer Vision. Proceedings of the International MICCAI Workshop on. Springer, pp. 106–117. (cited on page 64)

56. Cutti, A. G., Paolini, G., Troncossi, M., Cappello, A., Davalli, A., 2005. Soft tissue artefact assessment in humeral axial rotation. Gait & Posture 21 (3), 341–349. (cited on pages 3 and 16)

57. Dalal, N., Triggs, B., June 2005. Histograms of oriented gradients for human detection. In: Schmid, C., Soatto, S., Tomasi, C. (Eds.), Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. Vol. 2. pp. 886–893. (cited on pages 17, 19, 21, 27, 40, 75, 76, 113, 115, and 124)

58. Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M., 2015. Learning spatially regularized correlation filters for visual tracking. In: Computer Vision Workshops (ICCV). Proceedings of the IEEE Computer Society International Conference on. pp. 4310–4318. (cited on pages 111, 112, 114, 115, 116, 119, and 126)

59. Danelljan, M., Khan, F. S., Felsberg, M., van de Weijer, J., 2014. Adaptive color attributes for real-time visual tracking. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 1090–1097. (cited on page 112)

60. De Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.-P., Thrun, S., 2008. Performance capture from sparse multi-view video. In: ACM Transactions on Graphics (TOG). Vol. 27. p. 98. (cited on page 53)

61. De Leva, P., 1996. Adjustments to zatsiorsky-seluyanov's segment inertia parameters. Journal of biomechanics 29 (9), 1223–1230. (cited on page 6)

62. Debraux, P., Bertucci, W., Manolova, A., Rogier, S., Lodini, A., 2009. New method to estimate the cycling frontal area. International journal of sports medicine 30 (04), 266–272. (cited on pages 73 and 92)

63. Defraeye, T., Blocken, B., Koninckx, E., Hespel, P., Carmeliet, J., 2010. Aerodynamic study of different cyclist positions: Cfd analysis and full-scale wind-tunnel tests. Journal of Biomechanics 43 (7), 1262–1268. (cited on pages 51, 52, 70, 71, and 72)

64. Defraeye, T., Blocken, B., Koninckx, E., Hespel, P., Carmeliet, J., 2010. Computational fluid dynamics analysis of cyclist aerodynamics: Performance of different turbulence-modelling and boundary-layer modelling approaches. Journal of biomechanics 43 (12), 2281–2287. (cited on pages 92 and 100)

65. Defraeye, T., Blocken, B., Koninckx, E., Hespel, P., Carmeliet, J., 2011. Computational fluid dynamics analysis of drag and convective heat transfer of individual body segments for different cyclist positions. Journal of Biomechanics 44 (9), 1695–1701. (cited on pages 70 and 72)

66. Defraeye, T., Blocken, B., Koninckx, E., Hespel, P., Verboven, P., Nicolai, B., Carmeliet, J., 2014. Cyclist drag in team pursuit: influence of cyclist sequence, stature, and arm spacing. Journal of biomechanical engineering 136 (1), 011005. (cited on pages 52, 70, 72, and 73)

67. Delaitre, V., Sivic, J., Laptev, I., 2011. Learning person-object interactions for action recognition in still images. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K. (Eds.), Advances in Neural Information Processing Systems (NIPS). Curran Associates, Inc., pp. 1503–1511. (cited on page 36)

68. Della Croce, U., Leardini, A., Chiari, L., Cappozzo, A., 2005. Human movement analysis using stereophotogrammetry: Part 4: assessment of anatomical landmark misplacement and its effects on joint kinematics. Gait & posture 21 (2), 226–237. (cited on pages 3 and 16)

69. Desai, C., Ramanan, D., 2012. Detecting actions, poses, and objects with relational phraselets. In: Computer Vision (ECCV). Proceedings of the European Conference on. pp. 158–172. (cited on pages 18, 22, and 36)

70. Di Prampero, P., Cortili, G., Mognoni, P., Saibene, F., 1979. Equation of motion of a cyclist. Journal of Applied Physiology 47 (1), 201–206. (cited on pages 73, 93, and 95)

71. Dollár, P., Belongie, S., Perona, P., 2010. The fastest pedestrian detector in the west. In: British Machine Vision Conference (BMVC). (cited on pages 75 and 113)

72. Doval, P. N., 2012. Aerodynamic analysis and drag coefficient evaluation of time-trial bicycle riders. Master's thesis, Aerospace Engineering, University of Milwakee. (cited on page 72)

73. Doyle, D. D., Jennings, A. L., Black, J. T., 2014. Optical flow background estimation for real-time pan/tilt camera object tracking. Measurement 48, 195–207. (cited on page 112)

74. Drillis, R., Contini, R., et al., 1966. Body segment parameters. New York University, School of Engineering and Science. (cited on page 6)

75. Drory, A., 2012. Bike Fit and Aerodynamics. Bloomsbury, Ch. 8, pp. 104–122. (cited on pages 10 and 99)

76. Drory, A., Brown, N., Crouch, T., June 10th 2010. Brick or blade - is projected frontal surface area a good predictor of aerodynamic drag in elite cyclists? In: Cycling Science. Proceedings of the World Congress on. (cited on pages 52, 72, and 73)

77. Drory, A., Cherian, A., Li, H., Hartley, R., Submitted for Publication. Modelling human-object interactions for improved human pose estimation. (cited on page 66)

78. Drory, A., Hartley, R., Li, H., October 2014. Cyclist detection in images for pose estimation using cascade deformable part based model over histogram of oriented gradients. In: Computer Methods in Biomechanics and Biomedical engi-

neering (CMBBE). Proceedings of the international symposium on. (cited on pages 10, 75, 76, and 78)

79. Drory, A., Hartley, R., Li, H., October 2014. Statistical shape model of cyclists in level set formulation for pose estimation towards quantification of aerodynamic drag area. In: Computer Methods in Biomechanics and Biomedical engineering (CMBBE). Proceedings of the international symposium on. (cited on pages 10, 65, 75, 78, 82, 83, and 84)

80. Drory, A., Li, H., Hartley, R., December 2016. Markerless sagittal skeletal kinematics estimation from uncalibrated images using mixture of parts classification. In: Proceedings of the 10th Australasian Biomechanics Conference (ABC10). (cited on page 64)

81. Drory, A., Li, H., Hartley, R., [Article in press]. Estimating the projected frontal surface area of cyclists from images using a variational framework and statistical shape and appearance models. Journal of Sports Engineering and Technology. (cited on page 65)

82. Drory, A., Yanagisawa, M., 29 November 2011 2011. A mathematical model for estimation of time saved in road cycling descents through reduction in aerodynamic drag area achieved via modifications to static rider position. In: 8th Australasian Biomechanics Conferrence (ABC8). (cited on page 10)

83. Drory, A., Yanagisawa, M., June 1, 2012 2012. Predictive mathematical model of time saved on descents in road cycling achieved through reduction in aerodynamic drag area. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology 226 (2), 152–160. (cited on pages 10, 51, 70, and 99)

84. Durkin, J. L., Dowling, J. J., Andrews, D. M., 2002. The measurement of body segment inertial parameters using dual energy x-ray absorptiometry. Journal of biomechanics 35 (12), 1575–1580. (cited on page 6)

85. Ehrig, R. M., Taylor, W. R., Duda, G. N., Heller, M. O., 2006. A survey of formal methods for determining the centre of rotation of ball joints. Journal of Biomechanics 39 (15), 2798–2809. (cited on pages 2, 3, and 16)

86. Eichner, M., Ferrari, V., Zurich, S., 2009. Better appearance models for pictorial structures. In: British Machine Vision Conference (BMVC). Vol. 2. p. 5. (cited on page 30)

87. Eichner, M., Marin-Jimenez, M., Zisserman, A., Ferrari, V., 2012. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. International Journal of Computer Vision 99, 190–214. (cited on pages 17 and 30)

88. Everingham, M., Eslami, S. M. A., Gool, L. V., Williams, C. K. I., Winn, J. M., Zisserman, A., 2015. The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision 111 (1), 98–136. (cited on page 127)

89. Fanelli, G., Gall, J., Van Gool, L., 2011. Real time head pose estimation with random regression forests. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 617–624. (cited on page 64)

90. Fastovets, M., Guillemaut, J.-Y., Hilton, A., June 2013. Athlete pose estimation from monocular tv sports footage. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 1048–1054. (cited on pages 17 and 121)

91. Felzenszwalb, P., McAllester, D., Ramanan, D., 2008. A discriminatively trained, multiscale, deformable part model. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 1–8. (cited on pages 75, 76, 77, and 78)

92. Felzenszwalb, P. F., Huttenlocher, D. P., Jan. 2005. Pictorial structures for object recognition. International Journal of Computer Vision 61 (1), 55–79. (cited on pages 17, 21, 27, 34, 113, 115, and 132)

93. Fintelman, D., Sterling, M., Hemida, H., Li, F.-X., 2014. Optimal cycling time trial position models: Aerodynamics versus power output and metabolic energy. Journal of Biomechanics 47 (8), 1894–1898. (cited on page 70)

94. Freund, Y., Schapire, R. E., 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In: computational learning theory. Proceedings of the European conference on. Springer, pp. 23–37. (cited on page 59)

95. Gall, J., Stoll, C., De Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H.-P., 2009. Motion capture using joint skeleton tracking and surface estimation. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 1746–1753. (cited on pages 8, 53, and 64)

96. Galoogahi, H., Sim, T., Lucey, S., June 2015. Correlation filters with limited boundaries. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 4630–4638. (cited on page 114)

97. García-López, J., Rodríguez-Marroyo, J. A., Juneau, C.-E., Peleteiro, J., Martínez, A. C., Villa, J. G., 2008. Reference values and improvement of aerodynamic drag in professional cyclists. Journal of sports sciences 26 (3), 277–286. (cited on pages 51, 70, 71, 72, 73, 92, and 100)

98. Garg, R., Roussos, A., Agapito, L., 2013. Dense variational reconstruction of non-rigid surfaces from monocular video. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 1272–1279. (cited on pages 8 and 53)

99. Gerus, P., Sartori, M., Besier, T. F., Fregly, B. J., Delp, S. L., Banks, S. A., Pandy, M. G., D'Lima, D. D., Lloyd, D. G., 2013. Subject-specific knee joint geometry improves predictions of medial tibiofemoral contact forces. Journal of biomechanics 46 (16), 2778–2786. (cited on page 50)

100. Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. (cited on pages 113 and 115)

101. Girshick, R., Iandola, F., Darrell, T., Malik, J., 2015. Deformable part models are convolutional neural networks. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 437–446. (cited on page 5)

102. Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A., 2011. Efficient regression of general-activity human poses from depth images. In: Computer Vision (ICCV). Proceedings of the IEEE Computer Society International Conference on. pp. 415–422. (cited on page 64)

103. Girshick, R. B., Felzenszwalb, P. F., McAllester, D., 2012. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/. (cited on page 78)

104. Gorton, G. E., Hebert, D. A., Gannotti, M. E., 2009. Assessment of the kinematic variability among 12 motion analysis laboratories. Gait & Posture 29 (3), 398–402. (cited on page 6)

105. Gower, J. C., 1975. Generalised procrustes analysis. Psychometrika 40, 33–51. (cited on page 78)

106. Grappe, F., Candau, R., Belli, A., Rouillon, J. D., 1997. Aerodynamic drag in field cycling with special reference to the obree's position. Ergonomics 40 (12), 1299–1311. (cited on pages 51, 70, and 92)

107. Grimpampi, E., Camomilla, V., Cereatti, A., de Leva, P., Cappozzo, A., Feb 2014. Metrics for describing soft-tissue artefact and its effect on pose, size, and shape of marker clusters. IEEE Transactions on Biomedical Engineering 61 (2), 362–367. (cited on pages 3 and 16)

108. Gross, A., Kyle, C. R., Malewicki, D., 1983. The aerodynamics of human-powered land vehicles. Scientific American 249, 126–134. (cited on page 70)

109. Gupta, A., Chen, T. P., Chen, F., Kimber, D., Davis, L. S., 24-26 June 2008. Context and observation driven latent variable model for human pose estimation. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. (cited on pages 16, 36, and 37)

110. Gupta, A., Kembhavi, A., Davis, L. S., 2009. Observing human-object interactions: Using spatial and functional compatibility for recognition. Pattern Analysis and Machine Intelligence (T-PAMI). IEEE Computer Society Transactions on 31 (10), 1775–1789. (cited on page 36)

111. Hamacher, D., Hamacher, D., Taylor, W. R., Singh, N. B., Schega, L., 2014. Towards clinical application: Repetitive sensor position re-calibration for improved reliability of gait parameters. Gait & posture 39 (4), 1146–1148. (cited on page 4)

112. Hamer, H., Gall, J., Weise, T., Van Gool, L., 2010. An object-dependent hand pose prior from sparse training data. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 671–678. (cited on page 36)

113. Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M.-M., Hicks, S., Torr, P., 2016. Struck: Structured output tracking with kernels. Pattern Analysis and Machine Intelligence (T-PAMI). IEEE Computer Society Transactions on 38 (10), 2096–2109. (cited on page 112)

114. Hare, S., Saffari, A., Torr, P., Nov 2011. Struck: Structured output tracking with kernels. In: Computer Vision (ICCV). Proceedings of the IEEE Computer society International Conference on. pp. 263–270. (cited on page 114)

115. Hartley, R., Zisserman, A., 2000. Multiple view geometry in computer vision. Cambridge University Press. (cited on page 8)

116. Hatze, H., 2002. The fundamental problem of myoskeletal inverse dynamics and its implications. Journal of Biomechanics 35 (1), 109–115. (cited on pages 3 and 16)

117. Heil, D. P., 08 2001. Body mass scaling of projected frontal area in competitive cyclists. European Journal of Applied Physiology 85 (3-4), 358–366. (cited on page 73)

118. Henchoz, Y., Crivelli, G., Borrani, F., Millet, G. P., 2010. A new method to measure rolling resistance in treadmill cycling. Journal of sports sciences 28 (10), 1043–1046. (cited on page 98)

119. Henriques, J. F., Caseiro, R., Martins, P., Batista, J., 2012. Exploiting the circulant structure of tracking-by-detection with kernels. In: Computer Vision (ECCV). Proceedings of the European Conference on. Springer, pp. 702–715. (cited on pages 112, 114, and 118)

120. Henriques, J. F., Caseiro, R., Martins, P., Batista, J., 2015. High-speed tracking with kernelized correlation filters. Pattern Analysis and Machine Intelligence (T-PAMI). IEEE Computer Society Transactions on 37 (3), 583–596. (cited on pages 114, 116, and 118)

121. Hoiem, D., Chodpathumwan, Y., Dai, Q., 2012. Diagnosing error in object detectors. In: Computer Vision (ECCV). Proceedings of the European Conference on. Springer-Verlag, pp. 340–353. (cited on pages 113 and 115)

122. Horn, B. K. P., Schunck, B. G., 1981. Determining optical flow. Artificial Intelligence 17, 185–203. (cited on page 112)

123. Hornacek, M., Fitzgibbon, A., Rother, C., 2014. Sphereflow: 6 dof scene flow from rgb-d pairs. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 3526–3533. (cited on pages 8 and 53)

124. Huang, P., Hilton, A., 2006. Football player tracking for video annotation. IET European Conference on Visual Media Production, 175–175. (cited on page 121)

125. Hunter, A., 2009. Canoe slalom boat trajectory while negotiating an upstream gate. Sports Biomechanics 8 (2), 105–113. (cited on page 110)

126. Ignatz, R. N., Lim, A. C., Edwards, A. G., Birken, B. E., Samek, A., Byrnes, W. C., 2005. Frontal surface area versus drag area in predicting level cycling time trial performance. Medicine & Science in Sports & Exercise 37 (5), S86. (cited on page 73)

127. Jensen, Randall L.and Balasubramani, S., Brennan, G., Burley, K. C., Kaukola, D. R., LaChapelle, J. A., Shafat, A., 2007. Power output, muscle activity, and frontal area of a cyclist in different cycling positions. In: Menzel, H., Chagas, M. (Eds.), Biomechanics in Sports. Proceedings of the International Symposium on. (cited on page 73)

128. Jeukendrup, A. E., Martin, J., 2001. Improving cycling performance. Sports Medicine 31 (7), 559–569. (cited on page 92)

129. Kalal, Z., Mikolajczyk, K., Matas, J., July 2012. Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. 34 (7), 1409–1422. (cited on pages 115, 116, and 126)

130. Kass, M., Witkin, A., Terzopoulos, D., 1988. Snakes: Active contour models. International Journal Of Computer Vision 1 (4), 321–331. (cited on pages 74 and 84)

131. Kiefel, M., Gehler, P., 2014. Human pose estimation with fields of parts. In: Computer Vision (ECCV). Proceedings of the European Conference on. Vol. LNCS 8693 of Lecture Notes in Computer Science. Springer International Publishing, pp. 331–346. (cited on pages 30 and 44)

132. Kilner, J., Starck, J., Hilton, A., 2006. A comparative study of free viewpoint video techniques for sports events. IET European Conference on Visual Media Production, 87–96. (cited on page 121)

133. Kim, H. C., Park, H. J., Nam, K. W., Kim, S. M., Choi, E. J., Jin, S., Lee, J.-J., Park, S. W., Choi, H., Kim, M. G., 2010. Fully automatic initialization method for quantitative assessment of chest-wall deformity in funnel chest patients. Med. Biol. Engineering and Computing 48 (6), 589–595. (cited on pages 74 and 75)

134. Kjellstrom, H., Kragic, D., Black, M. J., 2010. Tracking people interacting with objects. (cited on pages 36 and 37)

135. Koppula, H. S., Gupta, R., Saxena, A., 2013. Learning human activities and object affordances from rgb-d videos. The International Journal of Robotics Research 32 (8), 951–970. (cited on page 66)

136. Krosshaug, T., Bahr, R., 2005. A model-based image-matching technique for three-dimensional reconstruction of human motion from uncalibrated video sequences. Journal of biomechanics 38 (4), 919–929. (cited on pages 16 and 29)

137. Kumar, M. P., Zisserman, A., Torr, P. H., 2009. Efficient discriminative learning of parts-based models. In: Computer Vision Workshops (ICCV). Proceedings of the IEEE Computer Society International Conference on. pp. 552–559. (cited on pages 24 and 26)

138. Kyle, C., 1989. The aerodynamics of handlebars and helmets. Cycling Science 1 (1), 22–25. (cited on page 92)

139. Kyle, C., Weaver, M., 2004. Aerodynamics of human-powered vehicles. Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy 218 (3), 141–154. (cited on page 92)

140. Kyle, C. R., 1979. Reduction of wind resistance and power output of racing cyclists and runners travelling in groups. Ergonomics 22 (4), 387–397. (cited on pages 92 and 93)

141. Kyle, C. R., Burke, E., 1984. Improving the racing bicycle. Mechanical Engineering 106 (9), 34–45. (cited on pages 70 and 92)

142. Kyle, C. R., Caiozzo, V. J., 1986. The effect of athletic clothing aerodynamics upon running speed. Medicine and science in sports and exercise 18 (5), 509–515. (cited on page 92)

143. Ladin, Z., Wu, G., 1991. Combining position and acceleration measurements for joint force estimation. Journal of Biomechanics 24 (12), 1173–1187. (cited on page 1)

144. Leardini, A., Chiari, L., Della Croce, U., Cappozzo, A., 2005. Human movement analysis using stereophotogrammetry: Part 3. soft tissue artifact assessment and compensation. Gait & posture 21 (2), 212–225. (cited on pages 3 and 16)

145. Lei, J., Ren, X., Fox, D., 2012. Fine-grained kitchen activity recognition using rgb-d. In: Ubiquitous Computing. Proceedings of the ACM Conference on. pp. 208–211. (cited on pages 64 and 66)

146. Lerasle, F., Rives, G., Dhome, M., Garcier, J., Van Praagh, E., 1997. Leg cycling tracking by dynamic vision. Journal of biomechanics 30 (8), 837–840. (cited on page 16)

147. Leventon, M., Grimson, W. E. L., Faugeras, O., 2000. Statistical shape influence in geodesic active contours. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. Vol. 1. pp. 316–323. (cited on pages 74 and 81)

148. Li, H., 2010. Multi-view structure computation without explicitly estimating motion. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 2777–2784. (cited on pages 8 and 53)

149. Li, K., Zheng, L., Tashman, S., Zhang, X., 2012. The inaccuracy of surface-measured model-derived tibiofemoral kinematics. Journal of biomechanics 45 (15), 2719–2723. (cited on pages 3 and 16)

150. Li, L., Prakash, B. A., 2011. Time series clustering: Complex is simpler! In: Machine Learning (ICML). Proceedings of the International Conference on. pp. 185–192. (cited on pages 58 and 64)

151. Li, Y., Zhu, J., 2014. A scale adaptive kernel correlation filter tracker with feature integration. In: Computer Vision Workshops (ECCVW). Proceedings of the European Conference on. Springer, pp. 254–265. (cited on page 114)

152. Lim, A. C., Homestead, E. P., Edwards, A. G., Carver, T. C., Kram, R., Byrnes, W. C., 2011. Measuring changes in aerodynamic/rolling resistances by cycle-mounted power meters. Medicine and science in sports and exercise 43 (5), 853–860. (cited on pages 98 and 103)

153. Lisani, J., Nov 2014. Adaptive thresholds for robust face detection with a short cascade of classifiers. In: Signal-Image Technology and Internet-Based Systems (SITIS). Proceedings of the International Conference on. pp. 27–31. (cited on page 113)

154. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M. J., 2015. Smpl: A skinned multi-person linear model. ACM Transactions on Graphics (TOG) 34 (6), 248. (cited on page 4)

155. Lu, T.-W., O'connor, J., 1999. Bone position estimation from skin marker coordinates using global optimisation with joint constraints. Journal of biomechanics 32 (2), 129–134. (cited on page 3)

156. Lu, T.-W., O'connor, J., 1999. Bone position estimation from skin marker coordinates using global optimisation with joint constraints. Journal of biomechanics 32 (2), 129–134. (cited on page 16)

157. Lu, Z., Pearlman, W. A., 2000. Semi-automatic semantic video object extraction by active contour model. In: IEEE International Conference on Multimedia and Expo (II). pp. 645–648. (cited on page 74)

158. Luinge, H. J., Veltink, P. H., 2005. Measuring orientation of human body segments using miniature gyroscopes and accelerometers. Medical and Biological Engineering and computing 43 (2), 273–282. (cited on page 4)

159. Lukes, R., Chin, S., Haake, S., 2005. The understanding and development of cycling aerodynamics. Sports Engineering 8 (2), 59–74. (cited on page 70)

160. Lukes, R. A., Chin, S. B., Hart, J. H., Haake, S. J., 2004. The aerodynamics of mountain bicycles - the role of computational fluid dynamics. (abstract). Sports Engineering 7 (4), 173–173. (cited on pages 70 and 72)

161. Magalhaes, F. A., Sawacha, Z., Di Michele, R., Cortesi, M., Gatta, G., Fantozzi, S., 2013. Effectiveness of an automatic tracking software in underwater motion analysis. Journal of sports science & medicine 12 (4), 660. (cited on page 16)

162. Malleson, C., Klaudiny, M., Hilton, A., Guillemaut, J.-Y., 2013. Single-view rgbd-based reconstruction of dynamic human geometry. In: Computer Vision Workshops (ICCV). Proceedings of the IEEE Computer Society International Conference on. pp. 307–314. (cited on page 8)

163. Mariani, B., Hoskovec, C., Rochat, S., Büla, C., Penders, J., Aminian, K., 2010. 3d gait assessment in young and elderly subjects using foot-worn inertial sensors. Journal of biomechanics 43 (15), 2999–3006. (cited on page 4)

164. Marqués-Bruna, P., Grimshaw, P., 2008. Aerodynamic effects of road topography and meteorological conditions on time-trialling cycling performance. International journal of Sports Science & Coaching 3 (2), 155–167. (cited on page 94)

165. Martin, J. C., Milliken, D. L., Cobb, J. E., McFadden, K. L., Coggan, A. R., 1998. Validation of a mathematical model for road cycling power. Journal of applied biomechanics 14, 276–291. (cited on page 94)

166. Martin, P. E., Mungiole, M., Marzke, M. W., Longhill, J. M., 1989. The use of magnetic resonance imaging for measuring segment inertial properties. Journal of Biomechanics 22 (4), 367–376. (cited on page 6)

167. Masood, S. Z., Ellis, C., Nagaraja, A., Tappen, M. F., LaViola, J. J., Sukthankar, R., 2011. Measuring and reducing observational latency when recognizing actions. pp. 422–429. (cited on page 54)

168. Mclean, B., Ellis, L., July 1994. Frontal surface area as a predictor of cycling performance. In: Barabás, A. Fábián, G. (Ed.), Biomechanics in Sports. Proceedings of the International Symposium on. (cited on page 73)

169. Middleton, J., Sinclair, P., Patton, R., 1999. Accuracy of centre of pressure measurement using a piezoelectric force platform. Clinical Biomechanics 14 (5), 357–360. (cited on page 2)

170. Miranda, D. L., Rainbow, M. J., Crisco, J. J., Fleming, B. C., 2013. Kinematic differences between optical motion capture and biplanar videoradiography during a jump–cut maneuver. Journal of biomechanics 46 (3), 567–573. (cited on pages 3, 16, 54, and 58)

171. Müller, M., Röder, T., 2006. Motion templates for automatic classification and retrieval of motion capture data. In: Computer animation. Proceedings of the ACM SIGGRAPH/Eurographics symposium on. pp. 137–146. (cited on pages 58 and 64)

172. Mumford, D., Shah, J., 1989. Optimal approximations by piecewise smooth functions and associated variational problems. Communications on Pure and Applied Mathematics, 577–685. (cited on page 74)

173. Nagase, Y., Yasunaga, H., Horiguchi, H., Hashimoto, H., Shoda, N., Kadono, Y., Matsuda, S., Nakamura, K., Tanaka, S., 2011. Risk factors for pulmonary embolism and the effects of fondaparinux after total hip and knee arthroplasty: a retrospective observational study with use of a national database in japan. J Bone Joint Surg Am 93 (24), e146. (cited on page 6)

174. Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation. In: Computer Vision (ECCV). Proceedings of the European Conference on. pp. 483–499. (cited on page 5)

175. Norberg, C., 1993. Flow around rectangular cylinders: Pressure forces and wake frequencies. Journal of Wind Engineering and Industrial Aerodynamics 49 (1), 187–196. (cited on page 71)

176. Norberg, C., 2001. Flow around a circular cylinder: aspects of fluctuating lift. Journal of Fluids and Structures 15 (3), 459–469. (cited on page 71)

177. Nott, C. R., Zajac, F. E., Neptune, R. R., Kautz, S. A., 2010. All joint moments significantly contribute to trunk angular acceleration. Journal of Biomechanics 43 (13), 2648–2652. (cited on page 2)

178. Nowozin, S., Shotton, J., 2012. Action points: A representation for low-latency online human action recognition. Microsoft Research Cambridge, Tech. Rep. MSR-TR-2012-68. (cited on page 58)

179. Oggiano, L., Leirdal, S., Saetran, L., Ettema, G., 2008. Aerodynamics optimization and energy saving of cycling postures for international elite level cyclists. In: Estivalet, M., Brisson, P. (Eds.), The engineering of sport 7. (cited on pages 92 and 95)

180. Oikonomidis, I., Kyriazis, N., Argyros, A. A., 2011. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In: Computer Vision Workshops (ICCV). Proceedings of the IEEE Computer Society International Conference on. pp. 2088–2095. (cited on page 36)

181. Olds, T. S., Norton, K., Craig, N., 1993. Mathematical model of cycling performance. Journal of Applied Physiology 75 (2), 730–737. (cited on pages 73 and 93)

182. Olds, T. S., Norton, K., Lowe, E., Olive, S., Reay, F., Ly, S., 1995. Modeling road-cycling performance. Journal of Applied Physiology 78 (4), 1596–1611. (cited on pages 73 and 93)

183. Oreifej, O., Liu, Z., 2013. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 716–723. (cited on pages 53 and 54)

184. Ormoneit, D., Black, M. J., Hastie, T., Kjellström, H., 2005. Representing cyclic human motion using functional analysis. Image and Vision Computing 23 (14), 1264–1276. (cited on page 65)

185. Pearsall, D. J., Reid, J. G., Livingston, L. A., 1996. Segmental inertial parameters of the human trunk as determined from computed tomography. Annals of biomedical engineering 24 (2), 198–210. (cited on page 6)

186. Peters, A., Galna, B., Sangeux, M., Morris, M., Baker, R., 2010. Quantification of soft tissue artifact in lower limb human motion analysis: A systematic review. Gait & Posture 31 (1), 1–8. (cited on pages 3 and 16)

187. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B., 2013. Poselet conditioned pictorial structures. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. (cited on page 17)

188. Pons-Moll, G., Baak, A., Gall, J., Leal-Taixe, L., Mueller, M., Seidel, H.-P., Rosenhahn, B., 2011. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In: Computer Vision Workshops (ICCV). Proceedings of the IEEE Computer Society International Conference on. pp. 1243–1250. (cited on page 53)

189. Prest, A., Schmid, C., Ferrari, V., Mar. 2012. Weakly supervised learning of interactions between humans and objects. Pattern Analysis and Machine Intelligence (T-PAMI). IEEE Computer Society Transactions on 34 (3), 601–614. (cited on page 36)

190. Ramakrishna, V., Munoz, D., Hebert, M., Bagnell, J. A., Sheikh, Y., 2014. Pose machines: Articulated pose estimation via inference machines. In: Computer Vision (ECCV). Proceedings of the European Conference on. pp. 33–47. (cited on page 5)

191. Ramanan, D., 2007. Learning to parse images of articulated bodies. Advances in Neural Information Processing Systems (NIPS) 19, 1129. (cited on pages 27, 40, and 124)

192. Ranzato, M. A., lan Boureau, Y., Cun, Y. L., 2008. Sparse feature learning for deep belief networks. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (Eds.), Advances in Neural Information Processing Systems (NIPS). Curran Associates, Inc., pp. 1185–1192. (cited on page 115)

193. Rao, G., Amarantini, D., Berton, E., Favier, D., 2006. Influence of body segmentsâĂŹ parameters estimation models on inverse dynamics solutions during gait. Journal of Biomechanics 39 (8), 1531–1536. (cited on page 6)

194. Raptis, M., Sigal, L., 2013. Poselet key-framing: A model for human activity recognition. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 2650–2657. (cited on page 54)

195. Riemer, R., Hsiao-Wecksler, E. T., Zhang, X., 2008. Uncertainties in inverse dynamics solutions: A comprehensive analysis and an application to gait. Gait & Posture 27 (4), 578–588. (cited on pages 3 and 16)

196. Rosario, H., Page, A., Besa, A., Mata, V., Conejero, E., 2012. Kinematic description of soft tissue artifacts: quantifying rigid versus deformation components and their relation with bone motion. Medical & Biological Engineering & Computing 50 (11), 1173–1181. (cited on pages 3 and 16)

197. Rosenhahn, B., Schmaltz, C., Brox, T., Weickert, J., Seidel, H.-P., 2008. Staying well grounded in markerless motion capture. In: Pattern Recognition. Springer, pp. 385–395. (cited on page 36)

198. Rossi, M., Lyttle, A., Amar, E.-S., Benjanuvatra, N., Blanksby, B., 2013. Body segment inertial parameters of elite swimmers using dxa and indirect methods. Journal of sports science & medicine 12 (4), 761. (cited on page 6)

199. Sadeghi, M. A., Farhadi, A., 2011. Recognition using visual phrases. (cited on page 36)

200. Salzmann, M., Pilet, J., Ilic, S., Fua, P., 2007. Surface deformation models for non-rigid 3d shape recovery. Pattern Analysis and Machine Intelligence (T-PAMI). IEEE Computer Society Transactions on 29 (8), 1481–1487. (cited on page 17)

201. Sandau, M., Koblauch, H., Moeslund, T. B., Aanæs, H., Alkjær, T., Simonsen, E. B., 2014. Markerless motion capture can provide reliable 3d gait kinematics

in the sagittal and frontal plane. Medical engineering & physics 36 (9), 1168–1175. (cited on page 16)

202. Sanders, R. H., Gonjo, T., McCabe, C. B., 2016. Reliability of three-dimensional angular kinematics and kinetics of swimming derived from digitized video. Journal of sports science & medicine 15 (1), 158. (cited on pages 16 and 29)

203. Savelberg, H., Van de Port, I. G., Willems, P. J., 2003. Body configuration in cycling affects muscle recruitment and movement pattern. Journal of Applied Biomechanics 19 (4), 310–324. (cited on page 93)

204. Schache, A. G., Baker, R., 2007. On the expression of joint moments during gait. Gait & Posture 25 (3), 440–452. (cited on page 2)

205. Schache, A. G., Baker, R., Lamoreux, L. W., 2006. Defining the knee joint flexion-extension axis for purposes of quantitative gait analysis: an evaluation of methods. Gait & Posture 24 (1), 100–109. (cited on page 2)

206. Schache, A. G., Baker, R., Lamoreux, L. W., 2008. Influence of thigh cluster configuration on the estimation of hip axial rotation. Gait & Posture 27 (1), 60–69. (cited on page 6)

207. Schache, A. G., Baker, R., Vaughan, C. L., 2007. Differences in lower limb transverse plane joint moments during gait when expressed in two alternative reference frames. Journal Of Biomechanics 40 (1), 9–19. (cited on page 2)

208. Schepers, H. M., Van Asseldonk, E. H., Baten, C. T., Veltink, P. H., 2010. Ambulatory estimation of foot placement during walking using inertial sensors. Journal of biomechanics 43 (16), 3138–3143. (cited on page 4)

209. Schindler, K., Van Gool, L., 2008. Action snippets: How many frames does human action recognition require? In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 1–8. (cited on page 54)

210. Schmiedmayer, H.-B., Kastner, J., 1999. Parameters influencing the accuracy of the point of force application determined with piezoelectric force plates. Journal of Biomechanics 32 (11), 1237–1242. (cited on page 2)

211. Schmiedmayer, H.-B., Kastner, J., 2000. Enhancements in the accuracy of the center of pressure (cop) determined with piezoelectric force plates are dependent on the load distribution. Journal of Biomechanical Engineering 122 (5), 523–527. (cited on page 2)

212. Schmitz, A., Ye, M., Shapiro, R., Yang, R., Noehren, B., 2014. Accuracy and repeatability of joint angles measured using a single camera markerless motion capture system. Journal of biomechanics 47 (2), 587–591. (cited on page 3)

213. Sempena, S., Maulidevi, N. U., Aryan, P. R., 2011. Human action recognition using dynamic time warping. In: Electrical Engineering and Informatics (ICEEI)j. Proceedings of the International Conference on. pp. 1–5. (cited on pages 54, 55, and 58)

214. Shim, V. B., Fernandez, J. W., Gamage, P. B., Regnery, C., Smith, D. W., Gardiner, B. S., Lloyd, D. G., Besier, T. F., 2014. Subject-specific finite element analysis to characterize the influence of geometry and material properties in achilles tendon rupture. Journal of biomechanics 47 (15), 3598–3604. (cited on page 50)

215. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R., 2013. Real-time human pose recognition in parts from single depth images. Communications of the ACM 56 (1), 116–124. (cited on pages 53, 54, 55, 64, and 65)

216. Singh, V. K., Khan, F. M., Nevatia, R., 2010. Multiple pose context trees for estimating human pose in object context. In: Computer Vision and Pattern Recognition Workshops (CVPRW). Proceedings of the IEEE Computer Society Conference on. pp. 17–24. (cited on pages 36 and 37)

217. Sreenivasa, M., Chamorro, C. J. G., Gonzalez-Alvarado, D., Rettig, O., Wolf, S. I., 2016. Patient-specific bone geometry and segment inertia from {MRI} images for model-based analysis of pathological gait. Journal of Biomechanics 49 (9), 1918–1925. (cited on page 6)

218. Stauffer, C., Grimson, W. E. L., 1999. Adaptive background mixture models for real-time tracking. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. Vol. 2. pp. 246–252. (cited on page 112)

219. Straka, M., Hauswiesner, S., Rüther, M., Bischof, H., 2012. Rapid skin: Estimating the 3d human pose and shape in real-time. In: 3D Imaging, Modeling, Processing, Visualization & Transmission. Proceedings of the International Conference on. pp. 41–48. (cited on page 7)

220. Sun, M., Telaprolu, M., Lee, H., Savarese, S., June 2012. An efficient branch-and-bound algorithm for optimal human pose estimation. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. (cited on page 17)

221. Swain, D., 1994. The influence of body mass in endurance bicycling. Medicine and Science in Sports and Exercise 26 (1), 58–63. (cited on page 73)

222. Swain, D., Coast, J., Clifford, P., Milliken, M., Stray-Gundersen, J., Feb 1987. Influence of body size on oxygen consumption during bicycling. Journal of Applied Physiology 62 (2), 668–72. (cited on page 73)

223. Tautges, J., Zinke, A., Krüger, B., Baumann, J., Weber, A., Helten, T., Müller, M., Seidel, H.-P., Eberhardt, B., 2011. Motion reconstruction using sparse accelerometer data. ACM Transactions on Graphics (TOG) 30 (3), 18:1–18:12. (cited on page 4)

224. Taylor, W., Kornaropoulos, E., Duda, G., Kratzenstein, S., Ehrig, R., Arampatzis, A., Heller, M., 2010. Repeatability and reproducibility of ossca, a functional approach for assessing the kinematics of the lower limb. Gait & posture 32 (2), 231–236. (cited on pages 2, 3, and 16)

225. Taylor, W. R., Ehrig, R. M., Duda, G. N., Schell, H., Seebeck, P., Heller, M. O., 2005. On the influence of soft tissue coverage in the determination of bone kinematics using skin markers. Journal of Orthopaedic Research 23 (4), 726–734. (cited on pages 2, 3, and 16)

226. Underwood, L., Jermy, M., 2010. Optimal hand position for individual pursuit athletes. Procedia Engineering 2 (2), 2425–2429. (cited on pages 92 and 95)

227. Underwood, L., Schumacher, J., Burette-Pommay, J., Jermy, M., 2011. Aerodynamic drag and biomechanical power of a track cyclist as a function of shoulder and torso angles. Sports Engineering 14 (2). (cited on pages 70 and 71)

228. Urtasun, R., Fleet, D. J., Fua, P., 2005. Monocular 3d tracking of the golf swing. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. Vol. 2. pp. 932–938. (cited on page 37)

229. Veksler, O., 2008. Star shape prior for graph-cut image segmentation. In: Computer Vision (ECCV). Proceedings of the European Conference on. Springer, pp. 454–467. (cited on page 31)

230. Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. Vol. 1. pp. I511–518. (cited on pages 115, 116, and 117)

231. Viola, P., Jones, M., 2001. Robust real-time face detection. In: Computer Vision (ICCV). Proceedings of the IEEE Computer Society International Conference on. Vol. 2. pp. 747–747. (cited on page 115)

232. von Marcard, T., Pons-Moll, G., Rosenhahn, B., 2016. Human pose estimation from video and imus. Pattern Analysis and Machine Intelligence (T-PAMI). IEEE Computer Society Transactions on 38 (8), 1533–1547. (cited on pages 4 and 53)

233. von Marcard, T., Rosenhahn, B., Black, M. J., Pons-Moll, G., 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In: Computer Graphics Forum. Vol. 36. Wiley Online Library, pp. 349–360. (cited on page 4)

234. Wang, J., Liu, Z., Wu, Y., Yuan, J., 2012. Mining actionlet ensemble for action recognition with depth cameras. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 1290–1297. (cited on pages 54, 64, and 66)

235. Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y., 2016. Convolutional pose machines. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 4724–4732. (cited on page 5)

236. Weinhandl, J. T., Armstrong, B. S., Kusik, T. P., Barrows, R. T., O'Connor, K. M., 2010. Validation of a single camera three-dimensional motion tracking system. Journal of biomechanics 43 (7), 1437–1440. (cited on page 3)

237. Weiss, A., Hirshberg, D., Black, M. J., 2011. Home 3d body scans from noisy image and range data. In: Computer Vision Workshops (ICCV). Proceedings of the IEEE Computer Society International Conference on. pp. 1951–1958. (cited on page 7)

238. Wicke, J., Dumas, G. A., 2010. Influence of the volume and density functions within geometric models for estimating trunk inertial parameters. Journal of applied biomechanics 26 (1). (cited on page 6)

239. Windolf, M., Götzen, N., Morlock, M., 2008. Systematic accuracy and precision analysis of video motion capturing systems exemplified on the vicon-460 system. Journal of biomechanics 41 (12), 2776–2780. (cited on page 3)

240. Wojek, C., Schiele, B., 2008. A performance evaluation of single and multi-feature people detection. In: Pattern Recognition. Proceedings of the DAGM Symposium on. Springer-Verlag, pp. 82–91. (cited on pages 113 and 118)

241. Wu, C., Varanasi, K., Theobalt, C., 2012. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In: Computer Vision (ECCV). Proceedings of the European Conference on. Springer, pp. 757–770. (cited on page 53)

242. Wu, G., Siegler, S., Allard, P., Kirtley, C., Leardini, A., Rosenbaum, D., Whittle, M., D'Lima, D. D., Cristofolini, L., Witte, H., Schmid, O., Stokes, I., 2002. {ISB} recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion-part i: ankle, hip, and spine. Journal of Biomechanics 35 (4), 543–548. (cited on page 19)

243. Wu, G., van der Helm, F. C., Veeger, H. D., Makhsous, M., Roy, P. V., Anglin, C., Nagels, J., Karduna, A. R., McQuade, K., Wang, X., Werner, F. W., Buchholz, B., 2005. {ISB} recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion-part ii: shoulder, elbow, wrist and hand. Journal of Biomechanics 38 (5), 981–992. (cited on page 19)

244. Wu, Y., Lim, J., Yang, M.-H., 2013. Online object tracking: A benchmark. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 2411–2418. (cited on page 126)

245. Xia, L., Chen, C.-C., Aggarwal, J., 2012. View invariant human action recognition using histograms of 3d joints. In: Computer Vision and Pattern Recognition Workshops (CVPRW). Proceedings of the IEEE Computer Society Conference on. pp. 20–27. (cited on page 54)

246. Yang, J., Li, H., Jia, Y., 2013. Go-icp: Solving 3d registration efficiently and globally optimally. In: Computer Vision (ICCV). Proceedings of the IEEE Computer Society International Conference on. pp. 1457–1464. (cited on page 8)

247. Yang, S. X., Christiansen, M. S., Larsen, P. K., Alkjær, T., Moeslund, T. B., Simonsen, E. B., Lynnerup, N., 2014. Markerless motion capture systems for tracking of persons in forensic biomechanics: an overview. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 2 (1), 46–65. (cited on pages 17 and 119)

248. Yang, X., Tian, Y. L., 2012. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: Computer Vision and Pattern Recognition Workshops (CVPRW). Proceedings of the IEEE Computer Society Conference on. pp. 14–19. (cited on page 54)

249. Yang, Y., Ramanan, D., Dec 2013. Articulated human detection with flexible mixtures of parts. Pattern Analysis and Machine Intelligence (T-PAMI). IEEE Computer Society Transactions on 35 (12), 2878–2890. (cited on pages 17, 18, 26, 27, 29, 34, 39, 40, 41, 89, 115, 129, and 131)

250. Yanover, C., Weiss, Y., 2004. Finding the ai most probable configurations using loopy belief propagation. Advances in neural information processing systems (NIPS) 16, 289. (cited on page 37)

251. Yao, A., Gall, J., Van Gool, L., 2012. Coupled action recognition and pose estimation from multiple views. International journal of computer vision 100 (1), 16–37. (cited on page 57)

252. Yao, B., Fei-Fei, L., 2010. Modeling mutual context of object and human pose in human-object interaction activities. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 17–24. (cited on page 36)

253. Yeo, D., Jones, N. P., 2008. Investigation on 3-d characteristics of flow around a yawed and inclined circular cylinder. Journal of Wind Engineering and Industrial Aerodynamics 96 (10), 1947–1960. (cited on page 71)

254. Yun, K., Honorio, J., Chattopadhyay, D., Berg, T. L., Samaras, D., 2012. Two-person interaction detection using body-pose features and multiple instance

learning. In: Computer Vision and Pattern Recognition Workshops (CVPRW). Proceedings of the IEEE Computer Society Conference on. pp. 28–35. (cited on pages 54 and 66)

255. Yushkevich, P. A., Piven, J., Hazlett, C. H. C., Smith, R. G., Ho, S., Gee, J. C., Gerig, G., 2006. User-guided 3d active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. NeuroImage 31, 1116–1128. (cited on page 74)

256. Zanfir, M., Leordeanu, M., Sminchisescu, C., 2013. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In: Computer Vision (ICCV). Proceedings of the IEEE Computer Society International Conference on. pp. 2752–2759. (cited on pages 54 and 56)

257. Zatsiorsky, V., Seluyanov, V., 1983. The mass and inertia characteristics of the main segments of the human body. Biomechanics viii-b 56 (2), 1152–1159. (cited on page 6)

258. Zhang, C., Tian, Y., 2012. Rgb-d camera-based daily living activity recognition. Journal of Computer Vision and Image Processing 2 (4), 12. (cited on page 54)

259. Zhang, J., Fernandez, J., Besier, T., 2015. Rapid lower limb geometry and muscle insertion estimation from motion-capture landmarks. Australia New Zealand Orthopaedic Research Society Annual Conference. (cited on page 50)

260. Zhu, Q., Yeh, M.-C., Cheng, K.-T., Avidan, S., 2006. Fast human detection using a cascade of histograms of oriented gradients. In: Computer Vision and Pattern Recognition (CVPR). Proceedings of the IEEE Computer Society Conference on. pp. 1491–1498. (cited on pages 75, 113, and 115)

261. Zollhöfer, M., Nießner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., et al., 2014. Real-time non-rigid reconstruction using an rgb-d camera. ACM Transactions on Graphics (TOG) 33 (4), 156. (cited on page 8)