

ABSTRACT

Title of dissertation: SYMMETRY IN HUMAN MOTION ANALYSIS:
THEORY AND EXPERIMENT

Yang Ran, Doctor of Philosophy, 2006

Dissertation directed by: Professor Rama Chellappa
Department of Electrical Computer Engineering

Video based human motion analysis has been actively studied over the past decades. We propose novel approaches that are able to analyze human motion under such challenges and apply them to surveillance and security applications.

Part I analyses the cyclic property of human motion and presents algorithms to classify humans in videos by their gait patterns. Two approaches are proposed. The first employs the computationally efficient periodogram, to characterize periodicity. In order to integrate shape and motion, we convert the cyclic pattern into a binary sequence using the angle between two legs when the toe-to-toe distance is maximized during walking. Part II further extends the previous approaches to analyze the symmetry in articulation within a stride. A feature that has been shown in our work to be a particularly strong indicator of the presence of pedestrians is the X-junction generated by bipedal swing of body limbs. The proposed algorithm extracts the patterns in spatio-temporal surfaces. In Part III, we present a compact characterization of human gait and activities. Our approach is based on decomposing an image sequence into x-t slices, which generate twisted patterns defined as the

Double Helical Signature (DHS). It is shown that the patterns sufficiently characterize human gait and a class of activities. The features of DHS are: (1) it naturally codes appearance and kinematic parameters of human motion; (2) it reveals an inherent geometric symmetry (Frieze Group); and (3) it is effective and efficient for recovering gait and activity parameters. Finally, we use the DHS to classify activities such as carrying a backpack, briefcase etc. The advantage of using DHS is that we only need a small portion of 3D data to recognize various symmetries.

HUMAN MOTION ANALYSIS IN SPACE AND TIME: THEORY
AND EXPERIMENT

by

Yang Ran

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2006

Advisory Committee:
Professor Rama Chellappa, Chair/Advisor
Dr. Qinfen Zheng, Co-Advisor
Professor David Jacobs
Professor Ankur Srivastava
Professor Min Wu

© Copyright by
Yang Ran
2006

DEDICATION

To my wife and my parents.

ACKNOWLEDGMENTS

I owe my gratitude to all who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost I'd like to thank my advisor, Professor Rama Chellappa for giving me an invaluable opportunity to work on challenging and extremely interesting projects over the past four years. He has always made himself available for help and advice and there has never been an occasion when I've knocked on his door and he hasn't given me time. It has been a pleasure to work with and learn from such an extraordinary individual.

I would also like to thank my co-advisor, Dr. Qinfen Zheng. Without his extraordinary theoretical ideas and computational expertise, this thesis would have been a distant dream. I also want to express my sincere thanks to Professor Larry Davis for his expertise and his tremendous support. Thanks are due to Professor Min Wu, Professor Ankur Srivastava and Professor David Jacobs for agreeing to serve on my thesis committee and for sparing their invaluable time reviewing the manuscript.

My colleagues at the CfAR laboratory have enriched my graduate life in many ways and deserve a special mention.

I owe my deepest thanks to my family - my wife, my mother and father who have always stood by me and guided me through my career, and have pulled me

through against impossible odds at times. Words cannot express the gratitude I owe them.

It is impossible to remember all, and I apologize to those I've inadvertently left out.

Lastly, thank you all!

TABLE OF CONTENTS

| | |
|---|------|
| List of Tables | viii |
| List of Figures | viii |
| 1 Introduction | 1 |
| 1.1 Biomechanics | 1 |
| 1.2 Kinematics | 3 |
| 1.3 Video Surveillance | 4 |
| 1.4 Contributions | 5 |
| 1.5 Thesis Organization | 6 |
| 1.6 Acknowledgement | 7 |
| 2 Pixel Based Periodicity Analysis | 8 |
| 2.1 Introduction | 9 |
| 2.1.1 Research motivation | 9 |
| 2.1.2 Algorithm overview | 9 |
| 2.2 Related Work | 11 |
| 2.3 Pedestrian Detection by Periodogram | 14 |
| 2.3.1 Pixel periodicity extraction | 15 |
| 2.3.2 Periodicity verification | 18 |
| 2.4 Experimental Results | 21 |
| 2.4.1 Infrared sensor | 21 |
| 2.4.2 Color/gray sensor | 23 |
| 2.4.3 Alignment | 26 |
| 3 Model based Periodicity Analysis | 28 |
| 3.1 Cyclic Motion | 29 |
| 3.2 Extraction of Motion Pattern | 30 |
| 3.3 Estimation of Period | 33 |
| 3.3.1 Phase-locked loop | 34 |
| 3.3.2 Recursive period estimation | 35 |
| 3.4 Experimental Results | 37 |
| 3.5 Sensitivity Analysis | 39 |
| 3.5.1 Object size | 39 |
| 3.5.2 Number of frames | 40 |
| 3.5.3 Frame rate | 43 |
| 3.5.4 Walking direction | 43 |
| 3.6 Conclusion | 45 |
| 4 Gait Pattern in Space and Time | 47 |
| 4.1 Introduction | 47 |
| 4.2 Related Work | 49 |
| 4.2.1 Pedestrian detection | 50 |

| | | |
|-------|--|-----|
| 4.2.2 | Tracking | 53 |
| 4.2.3 | Temporal Video Analysis | 54 |
| 4.2.4 | Contributions | 55 |
| 4.3 | Methodology | 56 |
| 4.3.1 | Kinematic body model | 56 |
| 4.3.2 | Signature surface | 59 |
| 4.3.3 | X Junctions in space and time | 60 |
| 4.4 | Extraction | 64 |
| 4.4.1 | Support region | 64 |
| 4.4.2 | Learning X junctions | 67 |
| 4.5 | System Design | 68 |
| 4.5.1 | Graph construction | 68 |
| 4.5.2 | Initialization | 70 |
| 4.6 | Experimental Results | 72 |
| 4.6.1 | Video sensor | 72 |
| 4.6.2 | Range sensor | 75 |
| 4.6.3 | Activities in sport videos | 75 |
| 4.7 | Conclusion | 78 |
| 5 | A Compact Characterization of Human Motion | 79 |
| 5.1 | Introduction | 79 |
| 5.2 | Double Helical Signature | 82 |
| 5.2.1 | Spatio-temporal gait and activity volume | 82 |
| 5.2.2 | Geometric symmetries | 84 |
| 5.2.3 | Kinematic chain model | 86 |
| 5.2.4 | Animated human activity | 90 |
| 5.2.5 | DHS in images | 94 |
| 5.3 | DHS Extraction | 98 |
| 5.3.1 | 1D curve approximation | 98 |
| 5.3.2 | Extraction of the helical pattern | 100 |
| 5.3.3 | Degenerate DHS | 103 |
| 6 | Applications of Double Helical Signatures | 104 |
| 6.1 | Overview | 104 |
| 6.2 | Pedestrian Segmentation | 104 |
| 6.2.1 | Simultaneous segmentation and body part labeling | 106 |
| 6.2.2 | Robustness analysis | 110 |
| 6.3 | Severe Occlusion Handling | 113 |
| 6.4 | Matching | 116 |
| 6.4.1 | Across cameras | 116 |
| 6.4.2 | Across time | 117 |
| 6.5 | Load Carrying Event Detection | 120 |
| 6.6 | Summary and Discussion | 124 |

| | | |
|-----|-----------------------------|-----|
| 7 | Conclusions | 127 |
| 7.1 | Summary | 127 |
| 7.2 | Future Directions | 129 |
| | Bibliography | 131 |

LIST OF TABLES

| | | |
|-----|---|-----|
| 2.1 | Comparison of sensitivity to background clutter | 25 |
| 2.2 | Comparison of classification results at different alignment error levels | 26 |
| 6.1 | USF data: DHS Matching across cameras. | 117 |
| 6.2 | USF data: DHS Matching across Time. | 120 |
| 6.3 | Outdoor Event Classification. First row: 3 categories; Second row: total number for each category; Each cell: the number classified as the index in the left. | 124 |

LIST OF FIGURES

| | | |
|-----|---|----|
| 2.1 | Left: illustration of periodicity extraction for an infrared surveillance video. Middle: filtered periodic mask for the pixels in the (red) bounding box. Right: a distinct peaks shows frequency corresponding to human gait. | 20 |
| 2.2 | Detection of a pedestrian in low contrast infrared sequence. Top row: original object sequence; bottom row (from left to right): original image; mask generated by hypothesis testing; A peak in histogram showing gait rate; similarity matrix using [44]. | 22 |
| 2.3 | ROC analysis for infrared dataset from static/moving sensors. | 23 |
| 2.4 | Period detection in Color/gray sensor dataset. top: first frames of original sequences; bottom: corresponding histogram | 24 |
| 2.5 | ROC analysis for color/gray dataset from static and moving sensors. | 25 |
| 2.6 | Pedestrian detection with alignment errors in color/gray sensor data. | 27 |
| 3.1 | Examples of cyclic motion (a) a rotating fan, (b) a running dog, (c) a walking pedestrian. | 29 |
| 3.2 | Motion signature in synthesized sequences [45]. The one with the (red) boundary represents the best fit using a twin-pendulum model | 30 |
| 3.3 | Principle Gait Angle in original and gradient images. Only those with MPGA exhibit a salient angular pattern. | 31 |

| | | |
|------|---|----|
| 3.4 | Illustration of twin-pendulum model fitting. White pixels indicate edges; Green pixels show the detected lines by Hough Transform; Red pixels (line segment pairs along legs) show the fitted lines forming Principle Gait Angle. | 33 |
| 3.5 | Diagram of a Digital PLL. | 34 |
| 3.6 | PLL VCO output voltage vs. locking time for the infrared data (a) Representative frames. (b) VCO output. | 37 |
| 3.7 | PLL output voltage vs. locking time for color/gray sensor. (a) Representative frames. (b) VCO output. | 38 |
| 3.8 | ROC comparison between PLL+shape and shape matching for both datasets. | 39 |
| 3.9 | ROC analysis for the two methods at different object sizes. | 40 |
| 3.10 | ROC analysis for the pixel-based method with various lengths. | 41 |
| 3.11 | Locking time (in terms of frame number) vs. number of lockings for the MPGA based method. Left: infrared sensor; right: color/gray sensor. | 42 |
| 3.12 | ROC analysis for the two methods at different frame rate. | 43 |
| 3.13 | Detection with various walking directions. | 45 |
| 3.14 | ROC analysis at different walking directions. Left for the pixel based method; Right for the MPGA method. | 45 |
| 4.1 | Gait pattern in spade and time. Left: the spatio-temporal slice slightly higher than ground plane contains the twisted pattern with red dots showing the X shape crossing. Right: one frame from original sequence. | 49 |
| 4.2 | Horizontal slices in different camera motion: static, pan, tilt and zoom and their 2D gradient histograms (discussed in section 4.4.1). | 49 |
| 4.3 | Kinematic body model. | 57 |
| 4.4 | An example to illustrate X junctions generated by kinematic chain model. | 60 |

| | | |
|------|--|----|
| 4.5 | Signature surface in videos. Top: Red curve is the trajectory; blue lines form the horizontal strips in S for each frame along the trajectory. Middle: signature surface. Bottom: corresponding poses. Notice that the X junction appears for most part of the signature surface except for the radial viewing directions. | 62 |
| 4.6 | Extracting the support region in a slice. | 66 |
| 4.7 | Learning the X junctions. Left: the eigenvalues of the PCA covariance matrix drop sharply after 20. Right: positive examples in the training set. | 68 |
| 4.8 | Flow chart for the proposed system. | 70 |
| 4.9 | Continuously detecting pedestrian in PETS. | 71 |
| 4.10 | Continuously detecting pedestrians in video sequences acquired by a moving platform. | 73 |
| 4.11 | Continuously detecting pedestrians in IR sequences. We list 3 initialized regions (with different color) and only one contains the X junction. | 73 |
| 4.12 | ROC analysis for UMD color/gray dataset and IR dataset. | 74 |
| 4.13 | Simulated range sensor slice data. | 75 |
| 4.14 | ROC analysis for range sensor data. | 76 |
| 4.15 | Soccer sequence and the classified events with overlaid motion signature. Top: left image is one frame with detected players. The right image has the color dots coding the trajectories for all players and the associated motion signature representing activities. Bottom: activity classification results for two players. | 77 |
| 5.1 | Top: Original frames and silhouette; Middle: Activity Sequence; Bottom: Selected 2D slices containing helical structure. | 83 |
| 5.2 | Gait parameters coded in DHS. | 84 |
| 5.3 | An example to illustrate 7 patterns in Frieze Groups. T: translation, V: vertical reflection, H: horizontal reflection, G: glide reflection [140]. | 85 |

| | | |
|-----|---|-----|
| 5.4 | DHS generated by a twin-pendulum model. First row: A twin-pendulum moving across an image; second row: X-t slices containing periodic helical structure | 89 |
| 5.5 | Hip-to-toe kinematic chain model | 91 |
| 5.6 | Unsuccessful learning of the DHS by directly applying principal curve analysis. Left: Original DHS; Middle: Manually labeled DHS; Right: Fitted principal curves. | 99 |
| 5.7 | Learning DHS by using Frieze Group symmetry. Left: A complete DHS. Middle: Dividing into strides. Right: Dividing into quadrants. . | 100 |
| 5.8 | Learning DHS in quadrant. Left: helical pattern in quadrants, Middle: line segment approximation; Right: Extracted DHS curves | 101 |
| 5.9 | Example of extracting DHS. (a),(b) connected curves for two legs in one DHS; (c) superimposed DHS; (d) degenerate DHS for torso. . . . | 102 |
| 6.1 | Pedestrian segmentation for a video sequence captured by a moving camera. (a) X-t slices for $y = 50, 60, 70, 80$; (b) silhouettes; (c) comparison of segmentation results between not using DHS (left binary images) and using the DHS as a feature (right binary images). | 107 |
| 6.2 | USF sequence 03507C0AL segmentation: (a) extracted DHS; (b) silhouettes generated for a complete stride; (c) comparison of segmentation results between background subtraction [106] (left binary images) and the proposed method (right binary images). | 109 |
| 6.3 | Comparison of segmentation accuracy between our method (red) and background subtraction (blue) for selected USF Sequences. (02463G2AR, 02539G1AR, 03500G0AR, 03507C0AL, 03509C0AL, 03516G0AR, 03521G0AR, 03526G0AR, 03529G0AR, 03532G0AR) | 111 |
| 6.4 | Comparison of body parts labeling accuracy for USF Data. The color boxes on the body parts show the labeling results at the slice at various heights. We successfully label legs and one arm. The other arm is occluded during the walking and cannot be labeled. | 111 |
| 6.5 | Robustness to view points. Each row contains original and segmented DHS for torso and limbs under different views for the USF sequence 03507C0AR. | 112 |
| 6.6 | Robustness to size. Each group contains original and segmented image with extracted DHS under different sizes and video rates for the USF sequence 02463C0AR. | 113 |

| | | |
|------|--|-----|
| 6.7 | Severe occlusion for pedestrians. First row: original sequences; Second row: extracted mask. | 114 |
| 6.8 | Segmentation under severe occlusion. First column: the original and extracted foreground at two different heights; Second column: separated individual DHS; Right two images: two frames of restored silhouettes. | 115 |
| 6.9 | More results for segmentation under severe occlusion. Notice that even the two pedestrians in the right are wearing same color pants, the proposed method still successfully segments them from occlusion by using DHS. | 116 |
| 6.10 | DHS matching using DTW for gait activities with different lengths. | 119 |
| 6.11 | Comparison of DHS in the hand regions for different activities. | 121 |
| 6.12 | Examples of different activities: four slices (0.25,0.35,0.75,0.85 of object height) are chosen for analysis and are illustrated in the left corner with activity name superimposed. The two leftmost insets show DHS of elbow and wrist; the two rightmost insets show DHS of the knee and ankle. | 123 |
| 6.13 | Examples of recognizing activities as leaving and picking up objects. Each row shows the DHS change before and after picking an object. | 125 |

Chapter 1

Introduction

Human motion analysis is receiving increasing attention from several communities of researchers. Computer vision researchers are developing many new theories and mathematical models to manipulate the human structure inside the computer world [5], [41, 52, 129]. This interest is motivated by applications over a wide spectrum of topics. In computer vision, segmenting the parts of the human body in a image, tracking the movement of joints over an image sequence, and recovering the 3D body structure are useful for precise analysis of athletic performance or medical diagnostics. Law enforcement officials are interested to the capability to automatically monitor human activities using computers in airport, borders, and other secured sites. Another application domain is the computer animation and video game industry, where the human avatars are very common. So, the potential number of applications is very high and in the next few sections we introduce several main areas in computer based human motion research. Detailed introductions could be found in the first section of following chapters.

1.1 Biomechanics

Human motion contains biological information about the identity of an actor as well as about his or her actions, intentions, and emotions. The human visual

system is highly sensitive to biological motion and capable of extracting socially relevant information from it. Researchers have investigated the question of how such information is encoded in biological motion patterns and how such information can be retrieved. Decades ago, experimental psychology researchers introduced a visual stimulus display designed to separate biological motion information from other sources of information that are normally intermingled with motion information. They attached small point lights to the main joints of a persons body and filmed the scene so that only the lights were visible in front of an otherwise homogeneously dark background. Using these displays, he demonstrated the compelling power of perceptual organization from biological motion using just a few light points. After that, marker based approaches have been a dominant tool for human motion analysis in biomechanics for quite a long time.

The principles of classical mechanics have been applied to the study of human motion to provide an understanding of the internal and external forces acting on the body during movement. The role of muscles in generating force and controlling movement is emphasized. Researchers compare the biomechanics of various motions by collecting and analyzing motion data. Many algorithms have been proposed to describe motions of the body during typical activities, predict which muscles are responsible for controlling movement, quantify the forces acting on the body during movement, understand the limitations of different experimental and analytical techniques used to quantify human movement, interpret motion data accurately, and evaluate studies of human movement.

1.2 Kinematics

Kinematic-based human motion has been researched and used commercially for a number of years with applications found in animation and biometrics [135]. The use of markers however is intrusive, necessitates the use of expensive specialized hardware and can only be used on footage taken especially for that purpose. A markerless system of human motion capture could be run using conventional cameras and without the use of special apparel or other equipments. Combined with today's powerful PC, cost-effective and real-time markerless human motion capture has for the first time become a possibility. Such a system has a greater number of applications than its marker based predecessor ranging from intelligent surveillance to character animation and computer interfacing. For this reason the field of human motion capture has recently seen somewhat of a renaissance.

The problem with using articulated models is the high dimensionality of the configuration space and the exponentially increasing computational cost that results. A realistic articulated model of the human body usually has several tens of Degree of Freedom (DOF). There are several possible strategies for reducing the dimensionality of the configuration space. First it is possible to restrict the range of movement of the subject. Such an approach greatly restricts the resulting trackers generality. Another way to constrain the configuration space is to perform a hierarchical search. Without the assistance of a kinematic model it is very hard to independently localize specific body parts in realistic scenarios. This is mainly due to the problem of self occlusion and pose change.

1.3 Video Surveillance

The surge in the global need for automated and reliable security and surveillance systems has elicited a significant response from both industry and academia in the domain of video analysis based sensing, processing and decision support. Computer vision research and development has advanced the state-of-the-art in video surveillance related algorithms in conjunction with the exploitation of increasing processing power of standard computing platforms for deployed and experimental systems.

Visual surveillance in dynamic scenes, especially for humans and their activities, is currently one of the most active research topics in computer vision. It has a wide spectrum of promising applications, including access control in special areas such as human identification at a distance, crowd counting statistics and congestion analysis, detection of anomalous behaviors, and interactive surveillance using multiple cameras, etc. In general, the processing framework of pedestrian surveillance in dynamic scenes includes the following stages: modeling of environments, detection of motion, classification of moving objects, tracking, understanding and description of activities, human identification, and fusion of data from multiple cameras. Some other possible research directions includes occlusion handling, motion analysis for biometrics, content-based retrieval.

Visual surveillance has been investigated worldwide under several large research projects. For example, the Defense Advanced Research Projection Agency (DARPA) supported the Visual Surveillance and Monitoring (VSAM) project [42]

since 1997, whose purpose was to develop automatic video understanding technologies that enable a single human operator to monitor behaviors over complex areas such as battlefields and civilian scenes. Furthermore, to enhance protection from terrorist attacks, the Human Identification at a Distance (HID) program sponsored by DARPA in 2000 aimed to develop a full range of multimodal surveillance technologies for successfully detecting, classifying, and identifying humans at great distances [106]. The European Unions Framework V Programme sponsored Advisor, a core project on visual surveillance in metro-stations. The Army's Collaborative Technology Alliance (CTA) program was formed in 2001 to establish partnerships among research communities in the Army Laboratories and Centers, private industry and academia. Under the CTA program, Advanced Sensors and Robotics have a direct orientation towards the cutting edge of surveillance and security applications.

1.4 Contributions

To solve the problem described in the previous section, this thesis introduces an architecture that provides algorithms to reason about the symmetry of human motion in space and time. Our contributions have been four fold.

First it has achieved a framework that supports the cascading extraction of periodicity, i.e. temporal symmetry [123, 115]. This process is manipulated by hypothesis testing that is able to detect a small amount of cyclic pixels out of background clutters and adapts to dynamic environments such as IR sensors, moving platforms or object sizes as small as 20×20 pixels.

Then the thesis provide a method from the observation of an X type junction in slices representing the bipedal motion of limbs [118]. By training the detector by a gradient based histogram feature, it is robust to pose, size and appearance change etc.

Third, the thesis then builds up a novel representation for gait and activities using geometric group theory and kinematics [116, ?]. The approach differs from previous work in this area by reasoning from temporal slices. In addition, Frieze Group Theory from architecture and cystography defines a Double Helical Signature (DHS) for human body articulation. Our framework is prepared to represent the dynamic shape deformation by a compact set of DHS.

Finally, we extend our study from symmetry to asymmetry in human motion due to load carrying events as well as sport video [117]. To our best knowledge, the proposed algorithm is the first one capable of real time recognition of gait activities without segmentation of silhouette.

1.5 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 discusses how hypothesis testing can be cascaded to detect periodicity and use it for pedestrian classification. Another model based method based on MPGA fitting with detailed experiments can be found in Chapter 3.

Chapter 4 gives a view of a novel architecture for pedestrian detection by spatio-temporal X junctions.

Chapter 5 describes the DHS based representation of gait and activities. Chapter 6 presents the results of experiments with the DHS. It discusses the performance and robustness in different applications such as segmentation, occlusion handling and activity recognition.

Finally Chapter 7 discusses several possibilities for extending this work in the future and conclusions, emphasizing this thesis's original contributions.

1.6 Acknowledgement

This research effort was supported by General Dynamics Robotic Systems (GDRS) and the Collaborative Technology Alliance (CTA) under contract DAAD19-012-0012 ARL-CTA-DJH. We would like to thank Dr. Larry Davis, Dr Kevin S. Zhou and Dr Isaac Weiss for providing discussions. We would also like to thank Dr. Wael Abd-Almageed, Mr. Feng Guo, Mr. Aswin Sankaranarayanan, Mrs. Jie Shao, Mr. Haibing Ling, Mr. Jian Li, Mr Seong-Wook Joo and all the other CfAR members for their helpful comments.

Chapter 2

Pixel Based Periodicity Analysis

2.1 Introduction

2.1.1 Research motivation

Pedestrian monitoring is critical in many surveillance systems in order to detect intrusions [73, 74]. Recently there has been a growing interest in using infrared cameras for human detection using robot vision techniques because of the sharply decreasing prices of such cameras and their ability to work in low light condition. Sensor noise and target pose variations present challenges. Existing systems for automatic detection of humans by shape and/or motion using thermal infrared cameras are sometimes confused by moving foliage, passing headlights, some building windows, traffic signs and more.

For close range pedestrian detection such as driver assistance systems, shape information can be reliably extracted. But for targets in mid or far range, it is no longer dependable and motion cue has to be integrated. The goal of this work is to develop a general motion-based pedestrian detector in low or regular lighting situations, where previous shape based methods exhibit high false alarm rate.

2.1.2 Algorithm overview

Object motions that repeat themselves are common in both nature and man-made environments. Many real-life motions are periodic such as the wings of flying birds, a puma's chasing and so on. Most human locomotory motions (e.g., walking, running, skipping, shuffling) are periodic in a frame of reference that moves with the person. Knowing that an object's motion is periodic is a strong cue for object

and action classification at a distance. Natural repeating motions tend not to be perfectly regular, i.e., the period varies slightly from one cycle to the next, or from one body part to another. For human gait, different parts of the human body share approximately the same period.

The two methods proposed here are based on detecting periodic motion. One is a bottom-up approach based on hypothesis testing over periodograms and the other relies on global gait fitting. The first one is designed to be computationally efficient by identifying periodicity in pixels. The second method is based on a gait based feature called Maximal Principal Gait Angle (MPGA). It is insensitive to alignment error and does not require segmentation but it is more computationally expensive. The two methods can be used individually or can be combined.

We initialize the target candidates using independent motion detection [112] or by a method that uses a hierarchal shape structure reported in [99], and track them with a method reported in [78]. The target locations are defined by bounding boxes in each frame. We focus on periodic motion because gait characterizes pedestrians and is more reliable when the targets are observed at a distance. Whether using only periodic property of motion or both motion and shape, the core task is how to efficiently use them. Our methods for detecting humans:

1. Is effective in different poses and from various distances.
2. Exhibits stable statistical performance.
3. Has efficient implementation.

The outline of this chapter is as follows. Section 2.2 discusses related work

on pedestrian detection. In Section 2.3 we discuss the pixel-based method with preliminary experiments. Section 2.4 provides a detailed analysis of sensitivity to several key factors.

2.2 Related Work

In recent years, automatic pedestrian detection in video has become an active research area in computer vision [41, 91]. This task is especially difficult for video sequences acquired by moving platforms and low quality sensors in situational awareness applications. Some of the difficulties in these applications are 1) non-rigid motion of pedestrians; 2) target pose and range change; 3) cluttered backgrounds and low video quality and 4) arbitrary camera motion. Reviews of some of the prior research on this topic can be found in [52, 81]. Useful criteria for classifying pedestrian detectors are the cues they use such as shape or motion.

Examples of algorithms in the first category can be found in [52] with learning tools such as wavelets [100], neural networks [150] and others. Nanda [99] builds a probabilistic shape hierarchy to achieve efficient detection at different scales. The method in [60, 127], uses handcrafted human models for pedestrian detection, but requires segmentation into body parts which is very difficult. A system [101] proposed by Pai *et al* recognizes pedestrians by measuring the distance between leg silhouette after background subtraction, which is not effective for moving platforms. Lipton [87] uses a skeleton based 'star' model to identify humans, which also depends on the extraction of a foreground mask. Another approach involves extracting low-

level features such as edges or responses to filter banks, and using standard pattern classification techniques to determine the presence of a pedestrian as in [107], where the authors extract wavelet features and then use an SVM to classify. Fang [49] compares the multi dimensional features between visible and infrared images and uses vertical projections of bright pixels specifically for infrared sensors. Objects in the background clutter such as windows, traffic signs and moving foliage often confuse shape based methods leading to high false alarm rates at acceptable detection rates. Besides, shape based detectors works better for close range targets than those far away.

In the motion based category, the gait feature has been analyzed based on pixel-wise or region-based oscillations. Statistical periodic behavior provides classification. For example, Little [84] used the Discrete Fourier Transform (DFT) to measure pixel oscillations. Tsai et al also described a similar method using DFT to extract pixel period in [130]. Efros and Berg [90] identified the cyclic motion in the optical flow domain. Liu and Picard [85] examined the pixel oscillations over the XT plane to extract the fundamental frequency of gait. Seitz and Dyer in [128] presented a novel concept, referred to as period trace to detect motion trends. Boyd [21] uses vPLLs (video Phase-Locked Loops) to measure the period contained in every pixel due to gait. Allmen and Dyer, in [?], proposed an approach to measure periodicity using a curvature scale space at each pixel. Polana and Nelson, in [108], showed that the recognition of human or animal locomotion can be done using low-level, non-parametric representations and matching against a spatio-temporal template of motion features. The main limitations of prior approaches in this sec-

ond category are the sensitivity to alignment as well as to changing background. For videos acquired from moving platforms, accurate alignment is hard to achieve and hence pixel-wise periodicity can be corrupted. A method that is closely related to this paper and motivates our work can be found in Cutler and Davis [44]. The authors look for the gait period by calculating a similarity matrix for every image pair in a sequence. The approach is computationally expensive and sensitive to background clutter. Furthermore, video sensors in infra-red band contain higher noise levels than in the visible band, which makes the similarity calculation easier to corrupt and good alignment harder to achieve.

A significant feature in some other methods is to combine shape and motion. Some of them directly trained the detector over shape and motion information simultaneously. For example, Viola's Adaboost detector cascade in [132] is a real-time pedestrian detection algorithm for a static camera. It was trained using patterns of frame difference as well as static shape features. Because of the static camera, those regions which have human-like shapes such as windows, stop signs and trees etc., are filtered out as non-moving background by preprocessing and do not enter the classifier cascade.

There are many multi-stage systems for detecting pedestrians by using different cues at different steps. One cue (shape or contour) is used for initial detection and others (motion, gait) are used as verification. For instance, Curio [45] proposed a method for the detection, tracking, and final recognition of pedestrians crossing a moving observer's path. The initial detection process was based on texture analysis and geometric features. The classification was obtained by a temporal analysis

of the walking process. However their algorithm "is restricted to the detection of pedestrians that cross the road" [45] and hence is not general enough for robot's situational awareness such as intrusion detection.

Another class of methods tried to fit a 3D human model to a 2D image to determine the articulation. For example, A. Broggi *et al* [28] compare several approaches relying on the matching between image features and model features stored in a predefined or dynamically updated database. A challenge to using model based fitting is that the complicated nature of human gait and variations of pose requires a large number of Degree Of Freedom. Hence it is very difficult to map the non-rigid dynamics. The authors in [28] also concluded that "it is difficult to obtain an exhaustive model set that gives good results on very different scenes".

2.3 Pedestrian Detection by Periodogram

Although gait period is used widely to analyze walking motion, few of the proposed methods are suitable for detecting pedestrians. The reason lies in two factors. One is the high complexity and the other is the sensitivity to pose change.

Different known forms of frequency detection are studied and applied in this work for pedestrian detection. Phase-Locked Loop and autoregressive moving average models (ARMA) have been traditionally used for estimating frequencies of sinusoidal time series data. In [111], the frequency is estimated using a second order ARMA model in an efficient fashion. Other frequency estimation techniques are parametric minimum entropy and subspace methods such as multiple signal classi-

fication (MUSIC). Our challenge here is how to use them efficiently to address the very specific pedestrian detection problem.

2.3.1 Pixel periodicity extraction

Objects with periodic motion are similar in many aspects, including appearance, motion flow, and shape [44]. However, environmental conditions (such as lighting, shadows, cluttered backgrounds) and internal conditions (pose, shape) variation contribute to wide signal variation and adversely affect detection. Besides, periodic behavior may only exist in some portion of an object. We carefully studied and implemented this in [44] and found that focusing on the overall similarity between images may fail due to a large number of non-periodic pixels. In this section, we describe a new algorithm which first tests the periodicity on a pixel-wise level and then analyzes the overall distribution of periods.

We start from a sequence of bounding boxes. Pre-processing is carried out to adjust for small alignment errors and to normalize for size. Assuming that the intensity at a periodic pixel (i, j) is a sum of a periodic signal $M(i, j)(t)$ and additive noise $n(t)$ while a non-periodic pixel contains only noise, we expand the signal in a frequency domain.

$$\begin{aligned} x_t(i, j) &= M_t(i, j) + n(t) \\ &= \mu(i, j) + \sum_{k=1}^{\infty} [\alpha_k \cos(k\omega t) + \beta_k \sin(k\omega t)] + n(t) \end{aligned} \quad (2.1)$$

where $n(t)$ is noise, $M_{(i,j)}(t)$ is the oscillatory signal and t is time.

To simplify the equations we linearize them and only use the first 3 coefficients

to approximate the original signal, yielding:

$$x_t(i, j) \approx \mu(i, j) + \alpha(i, j) \cos(\omega t) + \beta(i, j) \sin(\omega t) + n(t), \quad (2.2)$$

This approximation enables efficient estimation with low computation cost. With N observations at $t = 0, 1, \dots, N - 1$, we have N linear equations and 3 unknown parameters to estimate. We rewrite the N equations in a matrix form as:

$$A(\omega) \begin{bmatrix} \mu(i, j) \\ \alpha(i, j) \\ \beta(i, j) \end{bmatrix} = b \quad (2.3)$$

where

$$A(\omega) = \begin{bmatrix} 1 & 1 & 0 \\ 1 & \cos(\omega) & \sin(\omega) \\ \vdots & & \vdots \\ 1 & \cos((N-1)\omega) & \sin((N-1)\omega) \end{bmatrix}$$

and

$$b = [x_0(i, j) - n(0), x_1(i, j) - n(1), \dots, x_{N-1}(i, j) - n(N-1)]^T.$$

For a given period, or frequency ω , the least square estimator (LSE) for the parameters is given by:

$$\begin{bmatrix} \hat{\alpha}(i, j) \\ \hat{\beta}(i, j) \\ \hat{\mu}(i, j) \end{bmatrix} = (A(\omega)^T A(\omega))^{-1} A(\omega)^T b. \quad (2.4)$$

The residual sum of squares, for a given ω , is calculated as:

$$RSS(\omega) = \sum_{t=0}^{N-1} (x_t(i, j) - \hat{x}_t(i, j))^2 \quad (2.5)$$

$$= \sum_{t=0}^{N-1} (x_t(i, j) - \hat{x}_t(i, j))x_t(i, j) \quad (2.6)$$

$$= \sum_{t=0}^{N-1} x_t^2(i, j) - b^T A(\omega) (A(\omega)^T A(\omega))^{-1} A(\omega)^T b \quad (2.7)$$

$$= \sum_{t=0}^{N-1} [x_t(i, j) - \bar{x}]^2 - \{[A(\omega)b]^T \begin{bmatrix} \hat{\alpha}(i, j) \\ \hat{\beta}(i, j) \\ \hat{\mu}(i, j) \end{bmatrix} - \bar{x}^2\}.$$

The second derivation given above results from the orthogonality between the original signal and the estimation error under the Gaussian noise assumption [111]. Thus, the estimate of the period for the object sequence should be that ω which generates the smallest $RSS(\omega)$ over all possible frequencies (periods). Notice that minimizing $RSS(\omega)$ is equivalent to maximizing the second term in (2.7) with respect to ω . This enables us to estimate the period directly. Moreover, $(A(\omega)^T A(\omega))^{-1}$ could be approximated as [111]:

$$(A(\omega)^T A(\omega))^{-1} \approx \frac{1}{N} \begin{bmatrix} 1 & o(1) & o(1) \\ o(1) & 2 & o(1) \\ o(1) & o(1) & 2 + o(1) \end{bmatrix}, \quad (2.8)$$

where $o(1)$ denotes terms tending to zero, so the cost function is simplified to:

$$\begin{aligned}
I_{i,j}(\omega) &= \sum_{t=0}^{N-1} \{ [A(\omega)b]^T \begin{bmatrix} \hat{\alpha}(i,j) \\ \hat{\beta}(i,j) \\ \hat{\mu}(i,j) \end{bmatrix} - \bar{x}^2 \} \\
&\approx b^T \frac{1}{N} \begin{bmatrix} 1 + 2 \cos 0\omega & \dots & 1 + 2 \cos(N-1)\omega \\ 1 + 2 \cos \omega + 2 \sin \omega & \dots & 1 + 2 \cos(N-1)\omega \cos \omega + 2 \sin(N-1)\omega \sin \omega \\ \vdots & \vdots & \vdots \\ 1 + 2 \cos(N-1)\omega + 2 \sin(N-1)\omega & \vdots & 1 + 2 \cos^2(N-1)\omega + 2 \sin^2(N-1)\omega \end{bmatrix} b - N \frac{1}{N^2} \bar{x}^2
\end{aligned} \tag{2.9}$$

Substituting b and expanding the result we obtain:

$$I_{i,j}(\omega) = \frac{2}{N} \left\{ \sum_{t=0}^{N-1} x_t(i,j) \cos(\omega t) \right\}^2 + \frac{2}{N} \left\{ \sum_{t=0}^{N-1} x_t(i,j) \sin(\omega t) \right\}^2 = \frac{2}{N} \left\| \sum_{t=0}^{N-1} x_t(i,j) e^{i\omega t} \right\|^2 \tag{2.11}$$

The quantity $I_{i,j}(\omega)$ is the well-known periodogram [111] of the pixel. It has been showed in [111] that the maximizer of the periodogram over all frequencies cannot be improved on, in terms of asymptotic variance, by any other technique without extensive knowledge of the distribution of the noise $n(t)$.

2.3.2 Periodicity verification

Periodograms can be regarded as the signal response of the system at different frequencies. We verify the existence of a period via hypothesis testing to confirm the existence of a well-pronounced peak in the periodogram for each pixel. By filtering out non-periodic or stationary pixels, we are able to focus on periodic pixels only. Given the signal model as in Equation (2.2), where the noise is Gaussian, we perform the following statistical hypothesis test for every pixel:

$$\mathcal{H}_0 : \lambda(i,j) = 0 \quad \text{vs.} \quad \mathcal{H}_\lambda : \lambda(i,j) > 0. \tag{2.12}$$

where λ is the oscillatory amplitude for function $M_{(i,j)}(t)$ at the pixel (i,j) . It could be approximated from Equation 2.2 as $\lambda \approx (\alpha^2 + \beta^2)^{1/2}$. H_0 stands for non-periodic pixels (amplitude is zero) and H_λ for the periodic pixels. We use Bayesian decision rule based on the posteriori probability

$$P(H_\lambda|X) \ll P(H_0|X) \quad (2.13)$$

Using Bayes theorem, the decision rule can be transformed as:

$$P(X|H_\lambda)P(H_\lambda) \ll P(X|H_0)P(H_0) \quad (2.14)$$

Under a Gaussian noise assumption, the test rejects the null hypothesis \mathcal{H}_0 for large values of the ratio of the maximized likelihood under \mathcal{H}_λ to the maximized likelihood under \mathcal{H}_0 , i.e.:

$$-\frac{N}{2} \log(2\pi\sigma_\lambda^2) - \frac{N}{2} \log(2\pi\sigma_0^2) - \frac{N}{2}, \quad (2.15)$$

where

$$\sigma_\lambda^2 = \frac{1}{N} \sum_{t=0}^{N-1} [x_t(i, j) - \bar{x}]^2 - \max_{\omega_k \in \Theta} I_{i,j}(\omega_k)$$

and

$$\sigma_0^2 = \frac{1}{N} \sum_{t=0}^{N-1} [x_t(i, j) - \bar{x}]^2.$$

We reject \mathcal{H}_0 if $\sigma_\lambda^2/\sigma_0^2$ is too small, or equivalently,

$$r = \frac{\max_{\omega_k \in \Theta} I_{i,j}(\omega_k)}{\frac{1}{N} \sum_{t=0}^{N-1} [x_t(i, j) - \bar{x}]^2}$$

is too large [44].

After performing hypothesis testing, we are left with only periodic pixels and the most likely periods for them. We then compute the histogram of these periods

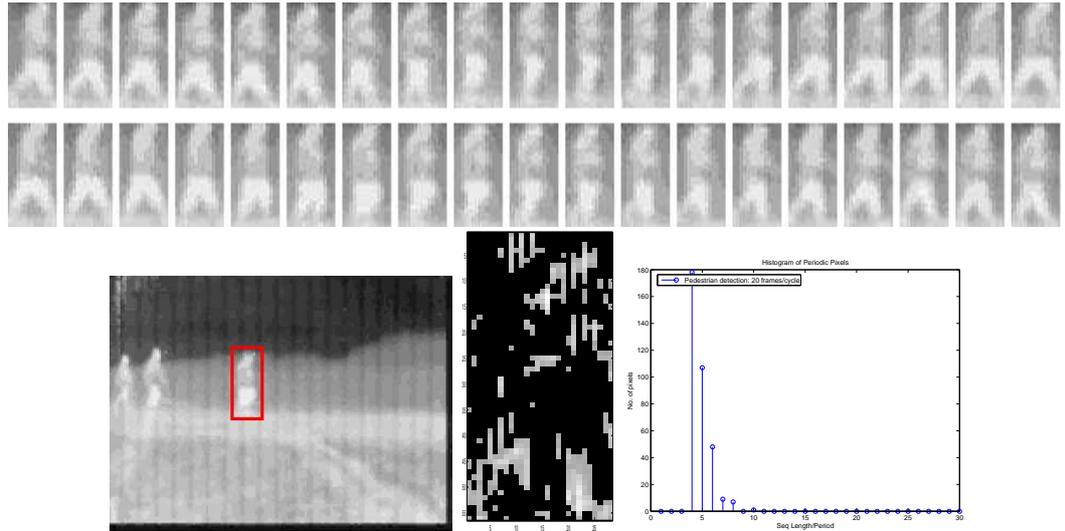


Figure 2.1: Left: illustration of periodicity extraction for an infrared surveillance video. Middle: filtered periodic mask for the pixels in the (red) bounding box. Right: a distinct peaks shows frequency corresponding to human gait.

and look for maximum period in this histogram. This is done using the same hypothesis testing method given above for the pixel based periodograms. This gives us the period of the global object.

An example of periodicity extraction is shown in Figure 2.1. After the testing, 'good pixels' are filtered out as the non-black pixels in the middle mask image for the human on the left image. The higher intensity stands for stronger periodicity. The histogram is shown on the right, with a well-pronounced peak representing gait rate. We also display a complete cycle of the walking sequence as a reference at the top of the figure. Because of the low resolution of the infrared camera, only a small portion of the image shows periodicity. Most of the periodicity comes from the region around the lower part (legs) of the human body, which captures most of

the motion for human gait. In spite of the difficulties, the algorithm still correctly detects a distinct peak at the periodic frequency in Figure 2.1.

As an extension of such a two-stage testing method, we apply shape constraints to lower the false positives rate. For example, the symmetry and relatively fixed location of periodic pixels for human gait could help us discriminate pedestrian walking from other periodic motion.

2.4 Experimental Results

Two datasets were tested. One was obtained using infrared cameras and the other employed color ground-based sensors. The infrared data (HONDA dataset, UMD dataset I) consists of 40 sequences ranging from 3 minutes to 7 minutes containing more than 80 objects (55 pedestrians and 35 vehicles) from both static and moving sensors. The other dataset (UMD dataset II) contains 55 color/gray sequences with 90 pedestrians and is also acquired by moving and static platforms. They include typical scenes such as parking lots, roads and other urban settings containing pedestrians varying in terms of size, speed, clothes and poses. We get a successful detection for an object only if the bounding boxes cover major portion of a human body and the motion based classification is correct.

2.4.1 Infrared sensor

In Figure 2.2, we illustrate the detection process for a more challenging sequence captured by a very low quality interlaced thermal sensor. The sensor blurs

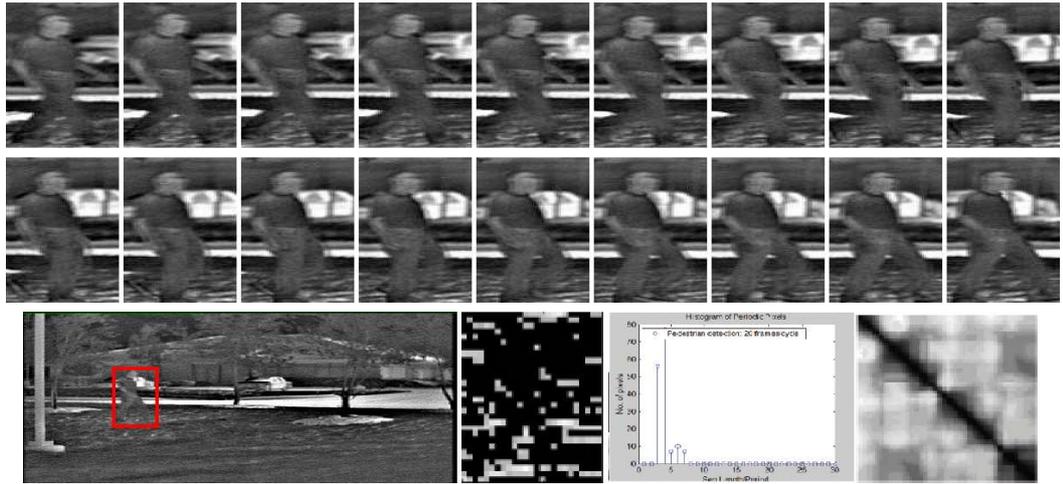


Figure 2.2: Detection of a pedestrian in low contrast infrared sequence. Top row: original object sequence; bottom row (from left to right): original image; mask generated by hypothesis testing; A peak in histogram showing gait rate; similarity matrix using [44].

foreground target’s contour and appearance with background, which makes most shape and motion based methods fail. For example, the similarity based method such as [44] yields a weak correlation matrix as shown in the left bottom image in Figure 2.2. Each pixel in that matrix represents correlation between two frames. If the contrast is high enough, we will observe darker lines parallel to the diagonal, which is caused by the similarity between two images in the same gait phase. Although no periodicity is observed in this matrix, our method successfully filters out the ‘good’ periodic pixels (even a very small portion of the whole image sequence) for estimating the correct gait rate.

To evaluate the accuracy of our method, we compute the ROC (Receiver Operating Characteristics) curve. The ROC curves plots the false positive rate against

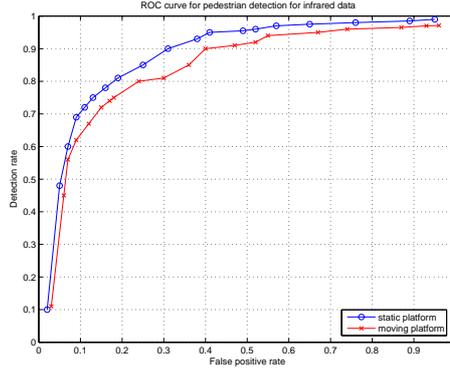


Figure 2.3: ROC analysis for infrared dataset from static/moving sensors.

the detection rate, when the classification criterion is varied. The false positive rate is defined as the total number of false positive detections divided by the total number of objects in all sequences; the true detection rate is the ratio between total number of correct detections and total number of detections in all sequences. Since we use a whole sequence for each classification, we do not divide the above rates by the frame number. Our classification criterion is the likelihood ratio, Equation. 2.15. This criterion depends on other parameters, namely the frame length N and the noise variance σ_A^2, σ_0^2 . We adjust only one of these parameters at a time to get different pairs of detection and false positives rate.

For infrared videos, our method maintains the detection rate above 80% with a false positives rate lower than 10% for static and the moving platforms.

2.4.2 Color/gray sensor

Figure 2.4 shows the pixel-wise classification results for three representative pedestrians from the data set. Two are from the same sequence (static camera)

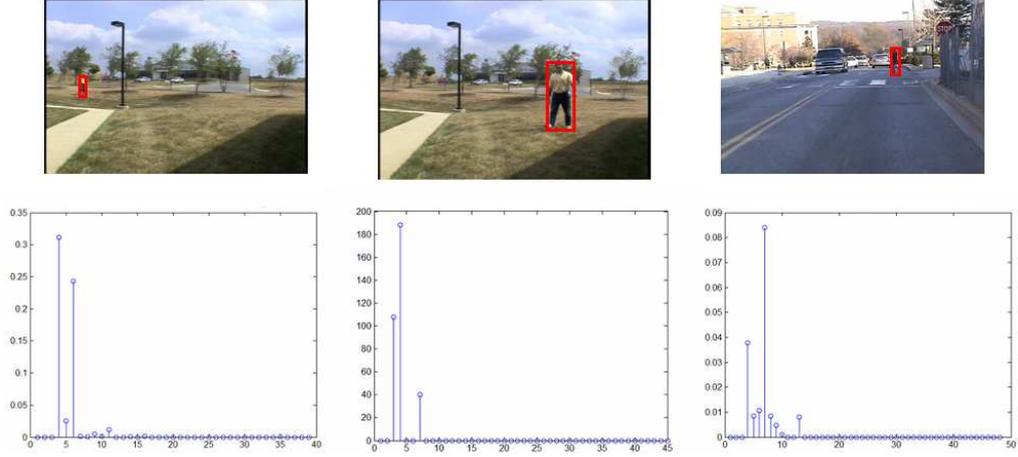


Figure 2.4: Period detection in Color/gray sensor dataset. top: first frames of original sequences; bottom: corresponding histogram

with different views: lateral (across the camera) and radial (towards or away from the camera), and the other is from a moving sensor in the lateral view. In all cases, the method detects the correct period and classifies the objects as pedestrians.

We plot the ROC curves for the static and moving platform for the UMD dataset II in Figure 4.12. Compared to the results for the infrared data, we obtain higher performance in terms of detection rate at the same false negative rate. This is due to the higher contrast and lower sensor noise level.

To compare to the method in [44] where we started our research, we implemented a version of the latter. Before presenting results of comparison, we make the following observations. Ours is more robust to the background clutter. Correlation over all pixels for an image pair computed in [44] will inevitably include background. When the background intensity is not homogenous, the correlation score for image pairs in the same gait phases will decrease, which is demonstrated in Table 2.1. We

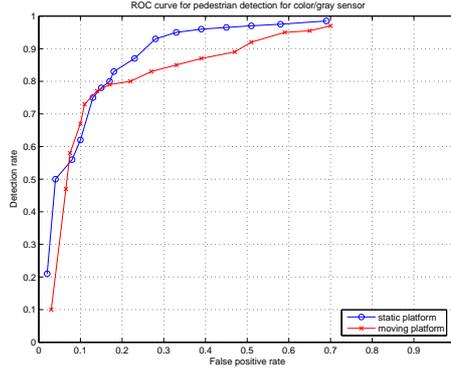


Figure 2.5: ROC analysis for color/gray dataset from static and moving sensors.

Table 2.1: Comparison of sensitivity to background clutter

| Object size increasing ratio (%) | 0 | 2.5 | 5 | 10 | 15 |
|--------------------------------------|------|------|------|------|------|
| Correlation score ratio for [44] | 8.63 | 4.32 | 3.50 | 1.93 | 2.02 |
| Maximum-to-mean ratio for our method | 12.9 | 11.3 | 14.5 | 10.0 | 9.8 |

calculate the maximum-to-variance ratio of one row in the correlation matrix for a sequence in Fig. 2.4. By systematically increasing the bounding box size, we include more and more background clutter into correlation calculation. The same ratio in the histogram from our new method is also calculated. In Table. 2.1, with the increase of the box size and hence the portion of background clutter, the correlation score ratio decreases sharply for a method like [44], while our method successfully filters out the periodic pixels and extracts the gait rate correctly.

The second advantage is that our method needs less computational power and is amenable for parallel hardware implementation. Assuming an object sequence with normalized size X by Y and frame number is N , the major operation in periodogram based estimation is FFT whose complexity is $N \log(N)$. The total number

Table 2.2: Comparison of classification results at different alignment error levels

| σ | 0 | 2.0 | 4.0 | 6.0 | 8.0 | 10.0 |
|----------------|-------|-------|-------|-------|-----------|-----------|
| Period | 34 | 32 | 35 | 30 | - | - |
| Classification | Human | Human | Human | Human | Non-Human | Non-Human |

of operations will be $X \cdot Y \cdot N \log(N)$. In order to have a robust detection, the method in [44] calculates the full correlation matrix to detect the lattice pattern. Each correlation between two images adds up to $X \cdot Y$ operations. The overall correlation matrix requires $X \cdot Y \cdot C_2^N = X \cdot Y \cdot 2/N(N-1)$. The computational saving is the ratio of the above two: $N \log(N)/C_2^N$, or $(N-1)/(\log N \cdot 2)$ times faster.

2.4.3 Alignment

This method works better when we have accurate alignment of the frames since it uses pixel-wise temporal information. Current detection and tracking algorithms cannot provide error-free alignment. We selected a subset sequences (more than 150) with a length of 64 frames and a box size around $40 * 80$. To obtain a quantitative estimate of the error in periodicity estimation resulting from misalignment, every box of the probing sequence is shifted in both directions by a quantity (dx, dy) , which obeys a zero-mean uniform distribution $U(0, \sigma)$. The periodicity is re-calculated with various σ . We list in Table 2.2 the detected period for one sequence as a function of the shift parameter σ .

With the increase in alignment error range, the performance of hypothesis

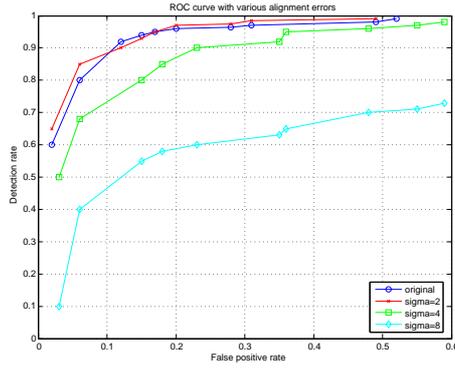


Figure 2.6: Pedestrian detection with alignment errors in color/gray sensor data.

testing approach deteriorates. It classifies the pedestrian as non-human when the alignment error is more than 6 pixels. Furthermore, we present the ROC curves in Fig. 2.6 for a subset of our data.

The result shows that the method is sensitive to large alignment errors and works reasonably well when alignment error is within a reasonable range. In order to reduce the sensitivity to alignment, we present another method based on cyclic property but less sensitive to alignment errors.

Chapter 3

Model based Periodicity Analysis

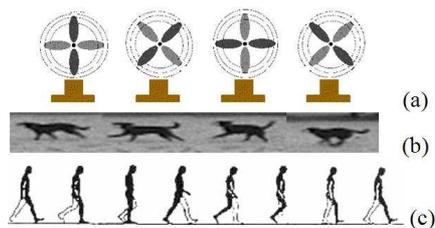


Figure 3.1: Examples of cyclic motion (a) a rotating fan, (b) a running dog, (c) a walking pedestrian.

3.1 Cyclic Motion

Beside the speed of motion (such as walking or running), the concept of gait also captures the style or manner in which a human moves. Periodicity differentiates a pedestrian from other non-periodic motions such as moving vehicles, while gait differentiates humans from other cyclic moving objects such as machines or animals. Gait also differentiate individuals. By comparing different kinds of periodic motion extracted from various objects illustrated in Fig. 3.1, we can identify a distinctive pattern that applies only to pedestrians. In particular, the swing of the two legs characterizes this pedestrian-specific oscillation.

We start by investigating the kinematics of human gait from a synthesized sequence as in Figure 3.2. The figure displays a complete cycle of a pedestrian’s legs. We develop a computationally efficient human motion analysis algorithm based on the twin-pendulum model introduced in [5, 45]. The twin-pendulum model has a very simple form that captures the inherent nature of gait. It focuses on the motions of the legs. Each leg is represented by two jointed cylinders. The diameters of the

cylinders are constant but the lengths of the cylinders are changing over time.

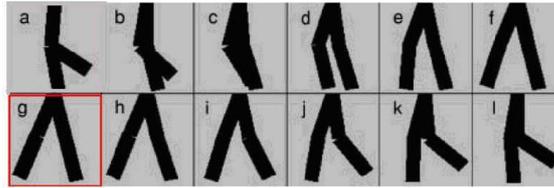


Figure 3.2: Motion signature in synthesized sequences [45]. The one with the (red) boundary represents the best fit using a twin-pendulum model

3.2 Extraction of Motion Pattern

The discussion above suggests that we classify a moving object as a human by features related to cyclic motion pattern. But changes in appearance, non-rigid deformation of human body and arbitrary motion of camera present challenges. What is a good feature to analyze the cyclic motion pattern unique to a walking human? The answer involves two issues. First, good features should be global and shape-based rather than pixel-based to reduce sensitivity to temporal alignment. Second, since precise shape extraction (segmentation) is very difficult, we prefer features derived from the human contour. A closer look at Figures 3.2 and 3.3 reveals that the relative angle between the two thighs can be used as such a feature. The Principle Gait Angle is defined as the angle between two legs during walking. But the non-rigid deformation and self-occlusion of two legs as well as the arbitrary pose makes it difficult to continually observe this angle in a complete stride. Instead we focus on a special case in which the two legs are maximally separated as in (g) in Figure 3.2, enclosed in a box. We refer to it as the Maximal Principal Gait

Angle (MPGA). This corresponds to a unique phase in the cycle in which the toe-to-toe distance approaches a maximum. The periodicity of the angle is a strong cue for detection. An example is given in Figure 3.3. We apply an edge detector to pedestrians at different principle gait angles of walking. Only those with MPGA exhibit a salient angular pattern formed by two line segments.

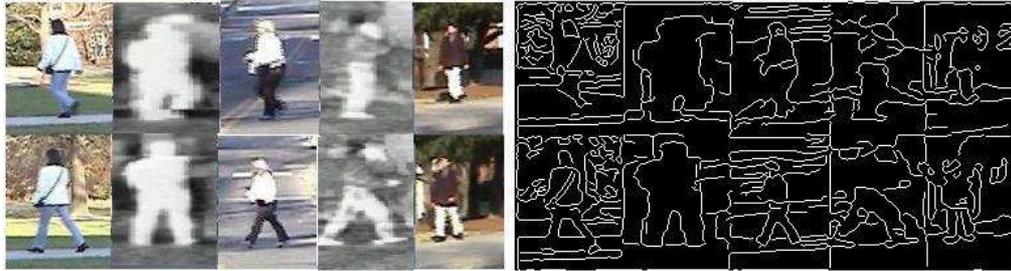


Figure 3.3: Principle Gait Angle in original and gradient images. Only those with MPGA exhibit a salient angular pattern.

We next describe how to extract the critical Principle Gait Angle from a cluttered background. We first apply an edge operator (i.e. Canny operator) to the image. Then a Hough line detector scans the edge map and generates a list of the candidate lines in the image with length above a pre-specified threshold. We pair these lines by checking symmetry and slope and choose the pair with the longest average length to be our candidate in that frame. Finally, we use a Bayes classifier to identify the occurrence of the MPGA. Intuitively, the MPGA should arise from line segments with sufficient length forming the gait angle and the angle should be related to the model and the pose. The distance between segments should fall into a narrow range. We formulate the detection of MPGA in a Bayes linear classifier

framework. An observation vector X is defined as

$$X = \{l, d, \alpha\} \quad (3.1)$$

where l is the average length of the twin-line, d is the center distance and α is the angle formed by that pair. The first two are normalized by the height of the object bounding box. However, in order to construct the likelihood ratio, the conditional probability must be in closed form for each class. In applications like ours, we have to estimate this distribution using samples from a training set. Since this is impractical, we use an approximation, namely a linear classifier. We are looking for a vector V and a scalar v_0 such that $y = V^T X <> -v_0$ is the discriminant function for this two-class problem. When X is normally distributed, y is also normal and we outline the process to generate a linear Bayesian classifier from the training set as follows (adapted from [50]).

1. Compute the sample mean \hat{M}_i and the covariance matrix $\hat{\Sigma}_i$ of vector $\{l, d, \alpha\}$ from a manually labeled training set;
2. Calculate V for a given weight s by $V = [s\hat{\Sigma}_1 + (1 - s)\hat{\Sigma}_2]^{-1}(\hat{M}_2 - \hat{M}_1)$;
3. Using the V computed as above, obtain $y_j^{(i)} = V^T X_j^{(i)}, i = 1, 2, 3 \dots N. X_j^{(i)}$ is the j th i -class sample
4. $y_j^{(1)}$ and $y_j^{(2)}$, which do not satisfy $y_j^{(1)} < -v_0$ and $y_j^{(2)} > -v_0$, are counted as errors.
5. Vary v_0 to find the v_0 which gives the smallest error

6. Vary s from 0 to 1 and repeat steps 2-5; choose the s giving the smallest error as well as the corresponding V and v_0 to form the discriminant function

We give some sample fitting results in a sequence of the lower part of a pedestrian.

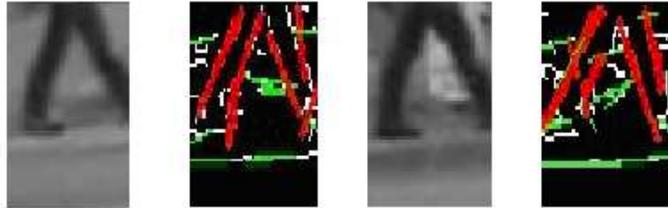


Figure 3.4: Illustration of twin-pendulum model fitting. White pixels indicate edges; Green pixels show the detected lines by Hough Transform; Red pixels (line segment pairs along legs) show the fitted lines forming Principle Gait Angle.

Such a Bayesian classifier gives us a binary sequence representing the classification decision for each frame in the video sequence. That is, for an image with a positive detection we have a 1 in the binary sequence and 0 otherwise. Intuitively, the sequence should be quasi-periodic and its instantaneous frequency should be the gait rate. In fact, even with false alarms, we still can observe a strong periodic oscillation in such a sequence and a more accurate solution will be provided in the next section.

3.3 Estimation of Period

The motivation for this section is to integrate shape and appearance with motion, which is expected to give better detection rate and fewer false alarms. Having

detected the presence or absence of the MPGA for each frame in a sequence, we can test for periodicity by the hypothesis testing methods described in the previous chapter.

3.3.1 Phase-locked loop

A PLL, or Phase-Locked Loop, is basically a close-loop feedback control system, whose operation is based on the detection of the phase difference between the input and output signals of a voltage controlled oscillator (VCO). Phase-locked loops are widely used in communications. An introduction to PLL can be found in [18].

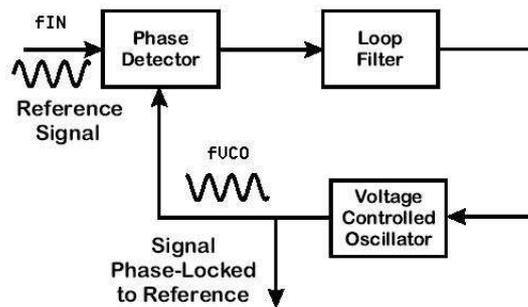


Figure 3.5: Diagram of a Digital PLL.

We use a software version of PLL [13]. Figure 3.5 shows the classic configuration. The phase detector is a device that compares two input frequencies, generating an output approximately proportional to their phase difference (if, for example, they differ in frequency, it gives a periodic output whose frequency is the difference frequency). Let's denote the reference signal frequency and the output of VCO frequency as f_{IN} and f_{VCO} . If f_{IN} doesn't equal f_{VCO} , the phase-error signal

causes the VCO frequency to deviate in the direction of f_{IN} . If conditions are right, the VCO will quickly "lock" to f_{IN} , maintaining a fixed relationship with the input signal.

3.3.2 Recursive period estimation

We use the output of the previous stage, namely the binary sequence provided by fitting the Principle Gait Angles, as the input to the digital PLL module for estimating both the frequency and phase of gait.

The input is a 0-1 sequences representing the critical phases corresponding to presence or absence of maximum toe-to-toe distances. This signal is passed through a low pass filter to remove high frequency components and obtain a smoothed signal:

$$V_i(t) = Lowpass(V_1(t)) \quad (3.2)$$

Without loss of generality, we write the input signal and the output signal from the VCO as

$$V_i(t) = A \cdot \sin(\omega_i(t) + \theta_i), V_o(t) = \cos(\omega_o(t) + \theta_o) \quad (3.3)$$

If we use a multiplier as the phase detector, the signal after multiplication will be

$$V_{PD}(t) = K \cdot A \cdot \sin(\omega_i(t) + \theta_i) \cdot \cos(\omega_o(t) + \theta_o) \quad (3.4)$$

where K is the gain of the phase detector (multiplier in our case). Furthermore,

we could write it as

$$V_{PD}(t) = 1/2 \cdot A \cdot K \cdot \{\sin[(\omega_i(t) + \omega_o(t)) + \theta_i + \theta_o] + \cos[(\omega_o(t) - \omega_o(t)) + \theta_i - \theta_o]\} \quad (3.5)$$

When $\omega_o \approx \omega_i$, the first item in the above representation is attenuated by the low pass filter (inside the loop filter) in Figure 3.5. The input of the VCO after low pass filtering can be approximated as

$$V_{VCO,IN}(t) = 1/2 \cdot A \cdot K \cdot \sin(\theta_i - \theta_o) \quad (3.6)$$

When the phase difference is small enough, this equation can be simplified to

$$V_{VCO,IN}(t) = 1/2 \cdot A \cdot K \cdot (\theta_i - \theta_o) \quad (3.7)$$

$V_{VCO,IN}$ is proportional to $\theta_i - \theta_o$. We can now explain how the dPLL locks the gait period. Suppose at first that the object's period is unknown. The initial frequency of the VCO output is set to an approximate value of gait frequency, ω_0 (20 frames/cycle). When the gait period (frequency ω_t of V_i) changes, the difference between V_o and V_i is detected by the phase detector which controls $V_{VCO,IN}$, causing the VCO frequency to deviate in the direction of ω_t . Hence, the period is estimated. Only when the rate falls into an interval representing a normal gait range, will the object be classified as a pedestrian.

In practice, the dPLL loop is activated by the result of initial hypothesis testing. As soon as a period is detected for the first time, the dPLL module works on the following frames with the initial VCO frequency being set to the detected period.

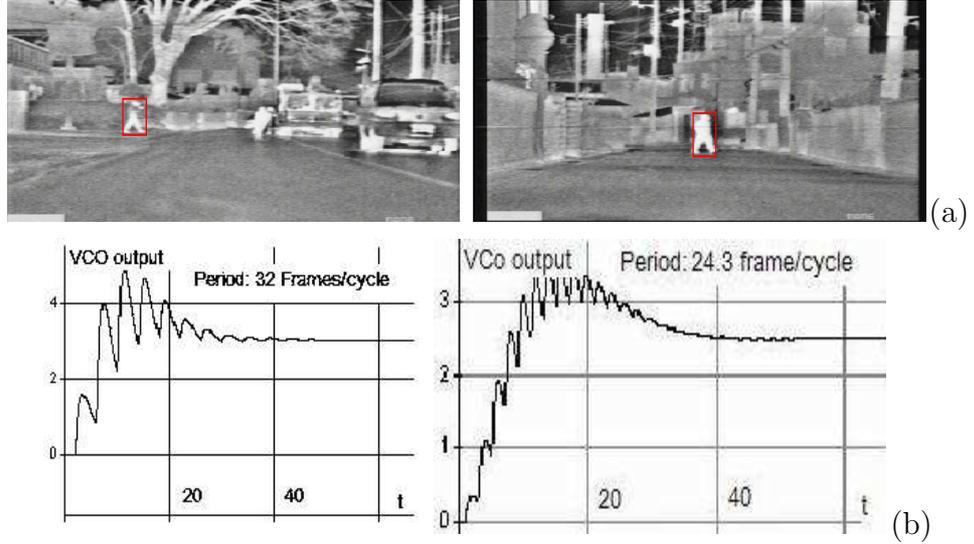


Figure 3.6: PLL VCO output voltage vs. locking time for the infrared data (a) Representative frames. (b) VCO output.

3.4 Experimental Results

We use a module reported in [99] to initialize detection and the tracking method reported in [78, 65] to track bounding boxes surrounding the targets. We test the system using the Infrared and Color/gray datasets as in the pixel-based method.

In Figure 3.6, we show the results based on tracking two pedestrians for 200 frames. The period for the first object is locked around a frequency of 32 frames/cycle, which corresponds to the gait rate. The second is locked at around 24 frames/cycle. We plot the PLL VCO voltage output vs. locking time in the second row to illustrate how fast the method adapts to the real signal. In Figure 3.7, we present a sequence from the color/gray set. We track pedestrians for 150 frames; the PLL locks to the period after 40 frames.

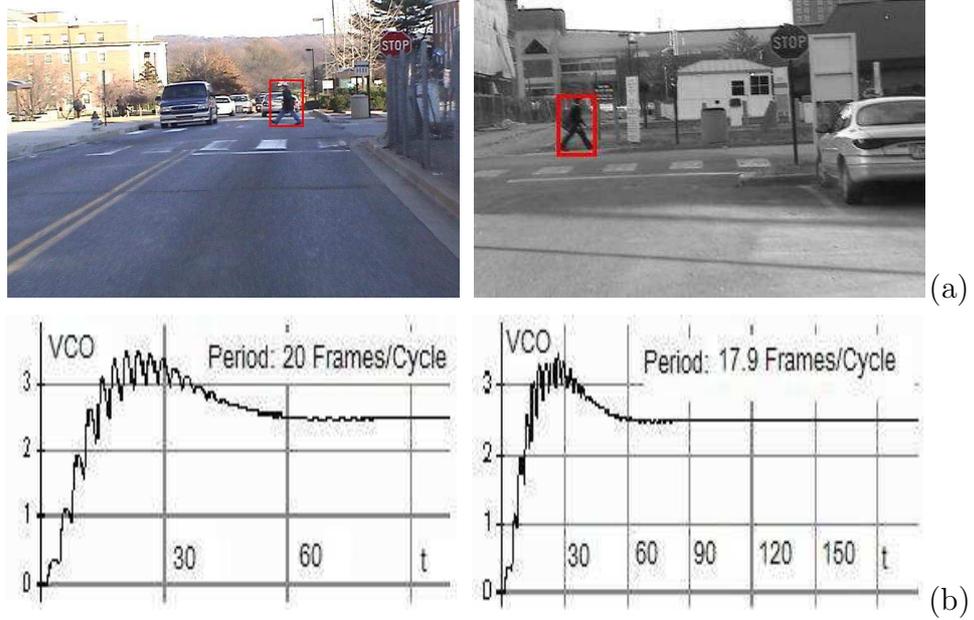


Figure 3.7: PLL output voltage vs. locking time for color/gray sensor. (a) Representative frames. (b) VCO output.

In order to evaluate the performance of the detector in finding the MPGA, we plot the ROC curves in Figure 3.8. We present the curves for infrared and color/gray data set respectively with the detection results obtained using the shape matching method reported in [99] and compare the performance improvement when cascaded with the MPGA fitting method. In this experiment, the training set is composed of 827 positive samples (boxes containing a pedestrian with maximum toe-to-toe distance) and 3270 negative samples (images containing pedestrians in other gait phases, other objects or background). For each data set, two cases are compared. One is the direct results obtained purely by using a shape hierarchy [99] in every frame and the other cascades matching and cyclic motion verification stages. As we can observe, the cascaded detectors successfully use the gait angle to identify true

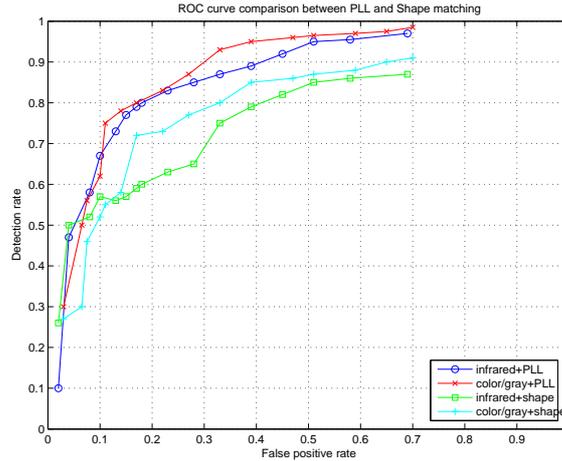


Figure 3.8: ROC comparison between PLL+shape and shape matching for both datasets.

pedestrians with a higher detection rate under the same false positive rate with the help of MPGA fitting algorithm and PLL-based gait rate estimation.

3.5 Sensitivity Analysis

In this section we study the sensitivity of detection accuracy of the two algorithms presented above to several important independent variables. The variables considered here are object size (determined by the distance to the camera), signal length, frame rate and movement directions.

3.5.1 Object size

Sensitivity to object size is important for judging a system’s ability to detect targets at various distances. We present the result for a subset of the two data sets with different down-sample ratios in Figure. 3.9. Typical object sizes from

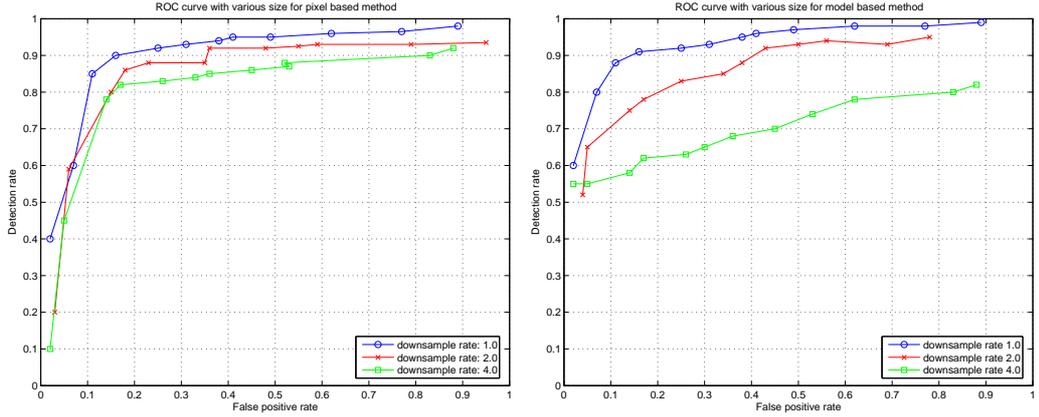


Figure 3.9: ROC analysis for the two methods at different object sizes.

this subset of sequences are greater than 3200 pixels (based on a bounding box of 40×80). We obtain consistently correct results for the first method even when the target size is gradually reduced to 10×20 . Notice that during the down-sampling, the detected period does not change. This demonstrates that the pixel based method exhibits only a weak dependence on object size. But the second method works only on relatively large objects, which is not surprising since the MPGA is extracted from edges.

This results also gives us a promising way to reduce the computational cost when using the pixel-based method. By reducing the object size by 2 or even 4, it reduces computations while maintaining the performance.

3.5.2 Number of frames

An interesting issue is the minimal sequence length needed to reliably extract the gait rate. We would like the detector to make a decision with minimal delay and to yield a reasonable detection rate. For shorter lengths, the signal may not be

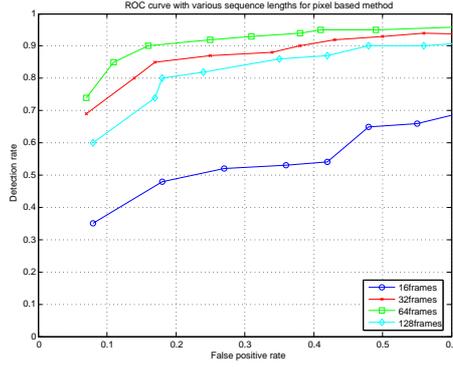


Figure 3.10: ROC analysis for the pixel-based method with various lengths.

long enough to exhibit periodicity. When the length increases, tracking is harder and the change of external variables such as pose, size, lighting etc will corrupt the cyclic signal.

Suppose we estimate the frequency directly from the periodogram output without any further processing [111]. If the true period for a pixel is ω , and it falls into two adjacent bins: k and $k+1$,

$$\omega \in [k \times 2\pi F_{sample}/N, (k+1) \times 2\pi F_{sample}/N] \quad (3.8)$$

where F_{sample} is the sample frequency and N is the signal length, we will have a bias up to the width of the bin. Hence this method requires longer sequences for higher resolution. But it is not always easy to maintain tracks of small objects in long low quality video acquired by a moving platform. We test the first method for object sequences with various lengths and the resulting ROC curves are shown in Fig. 3.10. Using the first method, for a typical human, we need only about two to three stride cycles (30-40 frames for a 30 fps video) to estimate the correct period.

For the MPGA method, a more meaningful measure will be the PLL locking

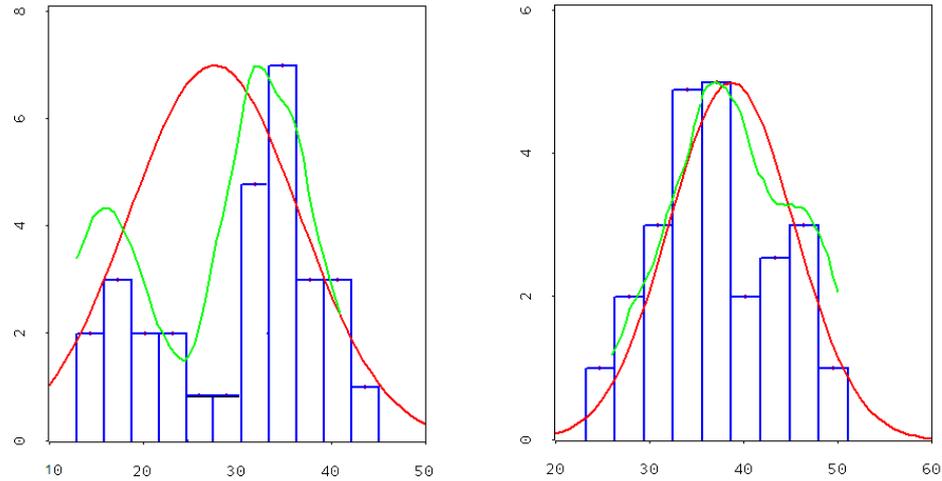


Figure 3.11: Locking time (in terms of frame number) vs. number of lockings for the MPGA based method. Left: infrared sensor; right: color/gray sensor.

time. Given the initial guess of the internal oscillator to be a regular gait rate (for example, 1Hz or 30 frames/cycle in color sensor and 2Hz or 15 frames/cycle in IR video for full frame rate), we plot the locking time (in terms of frames) vs. number of lockings in Fig 3.11. The left image shows the results for the infrared sensor and the right is for the visible spectrum sensor, together with the approximate Gaussian distributions. The histogram for the color sensor has a clear peak at 36 frames with a narrow bandwidth, showing the quick locking time. We observe two peaks for infrared data due to the fact that some low quality thermal sensor has the same response for left and right legs and so the real 'period' is half of the gait rate because of the symmetry in walking motion. After locking to the correct frequency, the module adapts to it without any re-initialization at a very high speed.

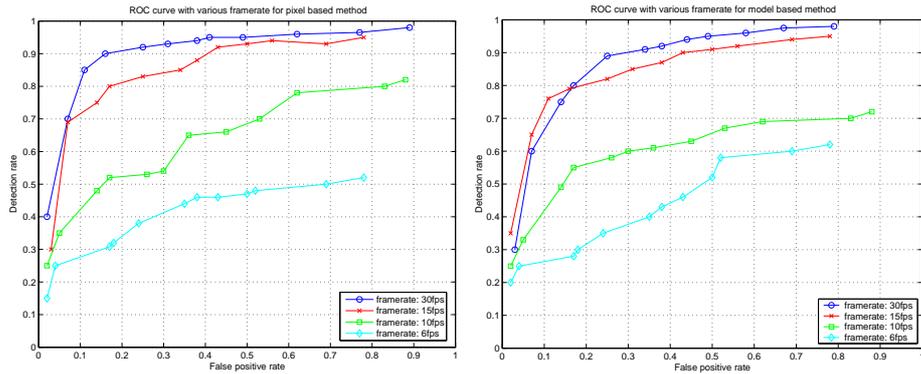


Figure 3.12: ROC analysis for the two methods at different frame rate.

3.5.3 Frame rate

Due to sensor limitation, we may be unable to always obtain the full frame rate ($> 25fps$). In addition, robustness to frame rate reduction could be useful for reducing overall computational cost. We present results for the two data sets with different frame rates in Fig. 3.12. The original rate is 30 frames/second and we reduce it systematically to 6 frames/second. We still obtain good results for both of the methods even when the frame rate is about 10 frames/second.

Comparing it with the ROC curves from sensitivity analysis against size, the results show that the pixel and model based methods are more sensitive to frame rate than to object size. At lower frame rates, longer sequence length could be used to compensate for the loss in periodic signal strength.

3.5.4 Walking direction

The observed oscillation amplitude of walking in images varies with different walking directions. It will approach a minimum when the pedestrian is walking

in a radial direction and will increase gradually to a maximum when the walker is moving sideways. The change in the amplitude of periodic signal will directly affect the detector performance.

We divide parts of the two data sets into subgroups according to walking directions and compare the results for the pixel based detector in Fig. 3.13. The results show that the first method correctly classifies a human under different poses ranging from radial to lateral. This could be explained by the 2-step hypothesis testing. When a pedestrian is walking towards the camera, many locations no longer exhibit strong cyclic pattern. Yet subtle oscillations still exist around body parts such as arms, legs and shoulder. By filtering out the non-periodic pixels, a small number of 'good' pixels with reasonable periodicity amplitude can be extracted to support correct estimation. In Fig. 3.13, with a box size of 40*80, the number of 'good' pixels decreased from thousands to hundreds and even to several tens as the movement direction changes from lateral to radial. A portion of good samples less than $1/16$ ($= 200/(40 * 80)$) is used in the final case when the target is moving towards the sensor.

A comparison of the ROC curves of sensitivity to walking directions is shown in Fig. 3.14 for both detectors. The angle of walking direction is the angle between it and the image plane. The results demonstrate that the pixel based method is more stable to walking directions due to the selectivity of periodic pixels. The MPGA method, as expected, drops performance sharply when the walking direction is over $\pi/3$.

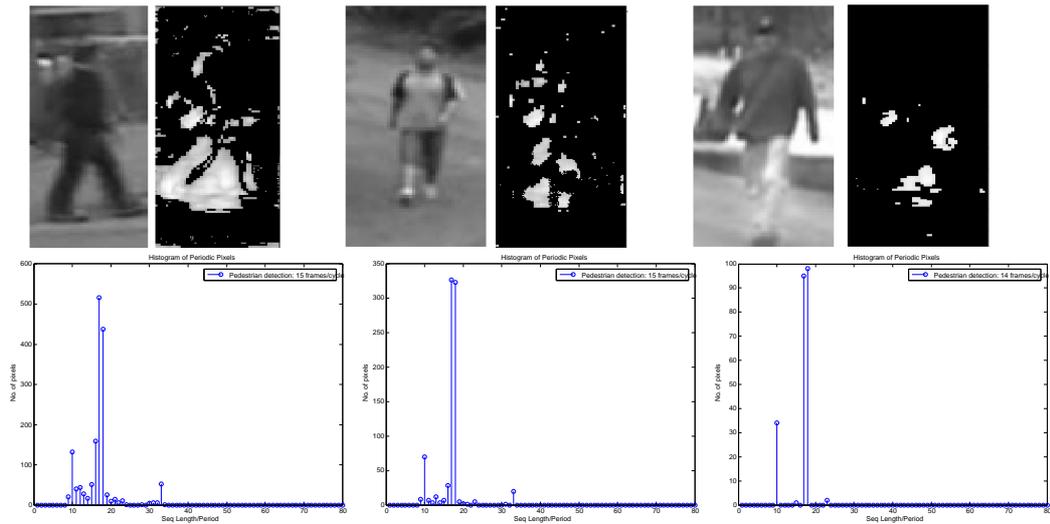


Figure 3.13: Detection with various walking directions.

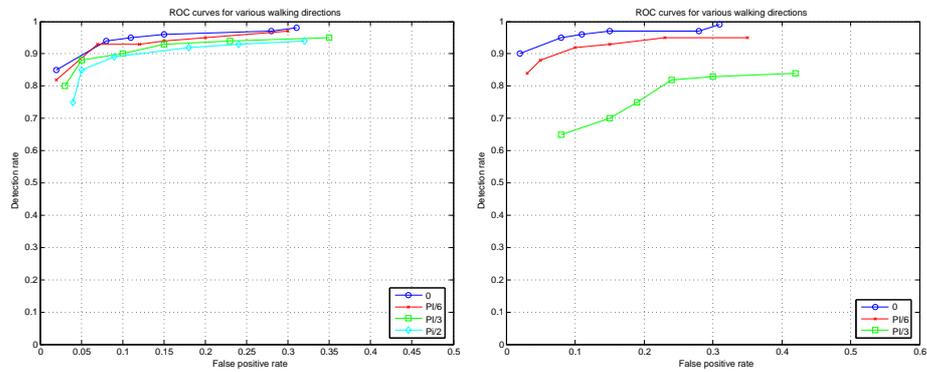


Figure 3.14: ROC analysis at different walking directions. Left for the pixel based method; Right for the MPGA method.

3.6 Conclusion

We presented two algorithms for pedestrian classification based on periodic motion. The first method is simple, efficient and robust to camera motion, sensor noise and walking directions but depends on good alignment accuracy. The MPGA method uses a global descriptor combining shape and motion, which is robust to

alignment and recursively estimates the gait rate. It integrates appearance and motion cues to classify objects. Both methods can detect pedestrians within a short time period (less than 2 seconds). Sensitivity analysis shows the robust behavior of the proposed methods with respect to a number of important factors such as frame rate, walking directions and object size.

The pixel based method monitors the oscillation at each pixel site and statistically extracts the overall frequency. It works better when the alignment is reliable. The method works well for both lateral and radial views and is computationally efficient.

The model-fitting method obtains classification cues from global shape and appearance and then examines gait dynamics. It extracts the MPGA in special gait phases and uses a phase lock to continuously classify targets. It does not require accurate alignment between frames.

A promising direction is to use a shape detector such as Viola's Adaboosting method [132] or Nanda's shape hierarchy [99] followed by the cyclic motion detector or pixel periodic detector as a verification module to obtain higher performance. By doing so we do not need to search the whole image in every frame and hence it is more computationally efficient. As part of the results shown in Fig. 3.8, this will form an automatic pedestrian detection system with lower false alarm rate and faster speed.

Chapter 4

Gait Pattern in Space and Time

4.1 Introduction

Vision based pedestrian detection is a natural choice based on a human's own experience. The human visual perception system is perhaps the best example of what is achieved possibly with these vision sensors. Although video cameras can obtain much richer information about the pedestrians and their surrounding environments compared to radar or a laser scanner, the image sequences can not be used for anything directly without further interpretation. How to extract useful information effectively from available image sequences is not a trivia task due to several reasons. First, video surveillance involves a complex uncontrolled indoor or outdoor environments. The illumination conditions may change due to weather and lighting. Pedestrians are found in surveillance conditions where the background texture (e.g. nearby buildings, vehicles, poles and trees) form a highly cluttered environments. Second, a wide range of variations exist in pedestrian appearance because of clothing, pose, occlusion, shadow, motion and size. Third, a moving platform will increase the difficulty in differentiating between background objects such as trees, windows, traffic signs and pedestrians.

Researchers have done a lot work to address such challenges. There are still a number of serious questions to be answered as to how to efficiently combine various

cues and how to maintain a balance in the system between computational cost and robustness. We realize that some proposed systems fall into a dilemma. On the one hand, to continuously locate targets, one must incorporate a tracking module. No matter what cue it uses, searching and matching is computationally expensive and subject to the non-rigid shape change. Tracking does not always provide reliable results for multiple non-rigid objects in cluttered scenes. On the other hand, detecting targets in each frame requires the detector to scan in every possible location. Such detectors are also easily influenced by background clutter, target appearance change and shape variations and hence cannot provide reliable results.

Although a large portion of work for localization is based on human shape, it is recognized that one of the most routine actions that humans perform is walking. Because of the upright pose and a piecewise translational trajectory, analyzing pedestrian motion in spatio-temporal domain is more reasonable and efficient than in a single image. Intuitively, if we could verify the presence of the gait pattern in the spatio-temporal volume occupied by the human, we can efficiently integrate shape and gait without tracking. We assume that the human does not change movement directions dramatically within a short period, such as a half second. This is equivalent to assume that the trajectory is locally linear.

The motivation for our approach arises from an analysis of pedestrian walking motion pattern in footsteps shown in Fig. 4.1. Instead of considering the cyclic property, we focus on the shape lying on a spatio-temporal surface within a stride: the crossing due to bipedal swing of the legs. In our notion, a signature surface cuts through the X-Y-t volume along a pedestrians trajectory and contains the pattern



Figure 4.1: Gait pattern in space and time. Left: the spatio-temporal slice slightly higher than ground plane contains the twisted pattern with red dots showing the X shape crossing. Right: one frame from original sequence.

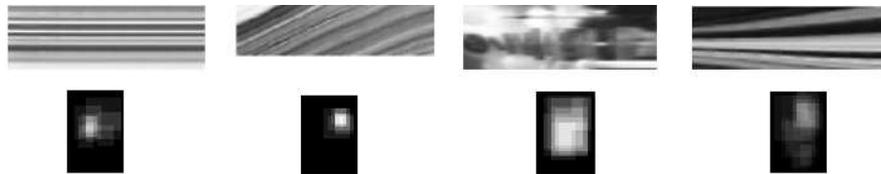


Figure 4.2: Horizontal slices in different camera motion: static, pan, tilt and zoom and their 2D gradient histograms (discussed in section 4.4.1).

that interests us. Another motivation lies in the observation of the orientation in temporal slices due to motion parallax and depth discontinuity between foreground and background. Figure 4.2 shows four types of orientations in horizontal surfaces with corresponding 2D gradient histogram (described in Section. 4.4.1). The orientation of the background pixels in the surface encodes camera motion as well as the scene depth structure.

4.2 Related Work

To develop human motion analysis algorithms, we need to integrate biomechanics, graphics and computer vision. In this section, we survey several related

areas.

4.2.1 Pedestrian detection

Different algorithms have been proposed to detect pedestrians in the image sequences acquired from video cameras. Some of the methods have been reviewed in [52, 81]. Recent research shows two main trends: 1) two-step approaches as screening candidate regions followed by recognizing the rhythmic gait after tracking; 2) one-step approaches as recognizing in every frame. Methods for the initialization in two-step approaches are either independent motion detection or shape matching. Rhythmic motion based approaches take into account temporal information and try to detect periodic features of human gait. On the other hand, shape based approaches rely on shape features and scans the image to locate humans.

Shape based methods recognizes both moving and stationary pedestrians from static or moving platforms. The primary difficulty associated with these approaches are how to accommodate the wide range of variations in pedestrian appearances due to pose, articulations of body parts ([17, 82, 83]), lighting, clothing, occlusion etc. The key issues are: i) to find a concise yet sufficient human shape feature representation that could achieve high inter-class variability with low intra-class variability; ii), to maintain a balance between accuracy of detection and processing time. In [107], an over-complete dictionary of Harr wavelets is used as representation of human shape characteristics followed by a support vector machine for classification. In order to detect partial occluded pedestrian, the same system is modified to de-

tect components of human body (e.g. head, torso or limbs) first, then the detected body parts are assembled together in [107]. In [151], the distance measurement from stereo vision is used for the segmentation step. A neural network trained by example pedestrian images is then used to classify segmented foreground objects. The stereo vision system developed for the ARGO vehicle is introduced in [28]. The vertical symmetry of a pedestrian is used in the segmentation step. Then pedestrian candidates are filtered with head shape, distance, size and aspect ratio. In [51], the Chamfer system developed by Daimler-Chrysler is introduced. The segmentation step is implemented by matching distance transformed images with different pedestrian shape templates. To reduce the processing time, the templates are organized in a certain hierarchy. A radial basis function based fine analysis is then implemented to reduce the false positive rate. A single feature is not enough to hold the human body shape change and multiple features requires a lot more time to compute. Although the shape-based method is general, the major drawbacks associated with them are: 1) high false positive rates due to variation of human shape and changing lighting conditions, 2) heavy computational cost when performing feature matching over every frame, 3) tracking is still needed for continuously localizing targets. Hence they are often used as an initialization step.

On the other hand, motion based approaches use rhythmic features or motion patterns unique to human beings. Periodicity of the human gait is a strong cue that can be used for the recognition of walking pedestrians [71, 72]. In [145], the maximum entropy method is applied to observe the periodic change of image intensity caused by walking. In [44], Fourier Transformation with Hanning window is

used to find periodicity in the correlation matrix of the acquired image sequences. Little *et al* analyzed the shape variations due to motion for classification [84]. Boyd [21] introduced the video phase locked loop for perceiving the oscillations at a pixel level. Seitz *et al* presented [128] a novel concept called period trace for detecting motion trends. Allmen *et al* [10] proposed an approach for measuring periodicity using a curvature scale space at each pixel. The work was extended into surface flow in [11, 12, 13]. Polana *et al* [109] showed that the recognition of periodic locomotion can be matched against a temporal template. Tsai *et al* [130] described a method using DFT to extract the pixel period. However, most of the above pixel-based methods do not exploit gait kinematics except for periodicity. Hence they are unable to capture the articulations of body parts present within a stride. The major drawback associated with rhythmic motion based methods is the delay. The detector has to accumulate a sufficient number of (basically 2-4 period) frames to observe the rhythmic features.

Another trend is to combine motion and shape. Viola's Adaboost detector cascade [132] is a fast pedestrian detection algorithm for a static camera. It was trained using patterns of frame difference as well as the static shape features. Because of the static camera, those regions which have human-like shapes such as windows, stop signs and trees etc., are filtered out as non-moving background by preprocessing and do not enter the classifier cascade. An interesting idea mixing shape and motion analysis has been developed by Curio *et al*. [45]. The torso is first tracked so that the lower part of the region can be located to reveal the relative motion of legs. A rough model of two legs consisting of two rod-like pieces each,

jointed at the knees, is juxtaposed on the image area below the tracked torso. The detected periodic movement is correlated with an experimental curve derived from the statistical average of human gait periods. High peaks of the correlation function indicate the presence of a person. The approach pursued by Niyogi and Adelson [98] analyzes the pattern in a slice of a XYT cube by a persons ankles. The tracks assume a characteristic plait shape due to the relative periodic motion, and may be easily identified by shape recognition methods, like snakes. Unfortunately, these ideas encounter difficulties in applications involving a moving camera.

4.2.2 Tracking

To continuously locate pedestrians in a video, human body tracking has been incorporated in many existing methods [136, 137, 141, 142, 143, 64]. There have been many approaches to solve such a problem. They differ mainly by whether the recovered motion description is 2D or 3D, and whether there is an explicit model of the human body. Bobick and Davis [48] developed a template based action recognition system, the Motion History Image, which records the recency of activity at each pixel, with each action having its own MHI template. Morris and Rehg [93] built a 2D stick figure kinematic model of the body for tracking the motion of the body through single source video. Bregler [26, 27] employed a mathematical result from robotics, the twist/screw motion of kinematic chain, and related it to image gradients to solve for differential motion of the body joints. Leventon et al [79] collected statistical data of the likelihood of body configurations and used them

to recover body joint angles in each frame. Particle filtering or mean shift based algorithm have been a major trend for adaptive trackers [38, 80, 43, 14, 29, 30]. Zhou *et al* [154] designed a particle filter framework to estimate the deformation between frames, which works fairly well for pedestrians at a distance. But because of the complicated articulating motion, tracking tends to fail due to non-rigidity and occlusion of body parts . It is hard to predict the body shape variations. Besides, prediction and matching need significant computation.

4.2.3 Temporal Video Analysis

To fuse motion and appearance for achieving a more reliable algorithm, we need to consider in space and time simultaneously, which directly brings up the temporal video analysis [22, 23, 24]. In the overwhelming majority of studies to date, image sequences are primarily analyzed and processed in groups of two frames, as by differentiating one frame from the other, one is able to infer the dynamics occurring in an image sequence. Although the single or two-frame approach has been very successful in some applications, such as the shape or shape context [17] based methods and the Adaboost based human classifier, it faces considerable difficulties if used, for example, to reason about non-rigid human motion. This subsection reviews the developments made in processing an alternative image sequence structure; the spatio-temporal surface (or slice), which has been proposed to alleviate the shortcomings of the traditional pair-wise approach.

One way to analyze the spatio-temporal volume is to consider it as being

formed by a stack of two-dimensional temporal slices or surfaces [104, 105, 138, 139]. For example, if the cube were to be sliced horizontally, one slice per scan line, then each slice exhibit structures related to the image features which pass over that scan line over time. Unlike the spatial features such as points and corners, the spatio-temporal paths contain some special "strip" pattern since they consist of the temporal dimension. If we compare such strips to the regions in 2D images, strips requires less effort to extract because of the continuity in scene depth change. In the case of locally linear camera motion, the orientations of such strips are usually aligned along one direction. A special type of surfaces, slices of the spatio-temporal volume, was first investigated by Bolles et al. [20], which focused on the geometric recovery of static scene structure. The particular class of slices analyzed were termed epipolar plane images (EPIs), and by restricting camera motion to linear paths, with a fixed orientation orthogonal to the direction of motion, depth information could be extracted from the relative angles of paths formed by features in the EPI. Following that Generalized EPI was proposed by Bolles [19] to handle non-linear camera motion. Ngo *et al* [95, 94, 97, 96] used spatio-temporal slices for the detection of cuts and wipes, where the task of detecting scene breaks was reformulated as the detection of boundaries in spatio-temporal slices.

4.2.4 Contributions

Our work can be categorized as a two-step algorithm combing both shape and motion. The distinct feature of our method lies in the fact that we find a

strong indicator within a stride instead of the periodicity. The main contributions of this work are threefold. First, we fuse motion and appearance within a very short interval (a half second). Second, there is no tracking module employed. Third, by embedding kinematic analysis and graph theory into the framework, it results in a real time system capable of simultaneously localizing pedestrians and monitoring some of their activities.

The rest of this chapter is organized as follows. Section 4.3 presents the analysis of body kinematics and propose a strong spatio-temporal indicator for pedestrians: X junction. Section 4.4 provides a solution to extract the gait pattern in spatio-temporal surfaces. Section 4.5 discusses a system for detecting pedestrians. Results are presented in Section 4.6 and section 6.6 summarizes the approach.

4.3 Methodology

4.3.1 Kinematic body model

To measure the articulation of a human body in video, we need to relate the points on the body to the images. Extensive discussion of this topic can be found in [110, 135]. For our purposes it is sufficient to model the motion of a human body as an 3D articulated motion of rigid body parts. We model the geometry of body parts by 3D volumetric primitives. Articulated motion can be represented by kinematic chains. There are many parameter systems that can be used to model kinematic chains [135]. We expect that the choice of a small set of kinematic parameter would be sufficient to understand and estimate temporal motion patterns in our framework.

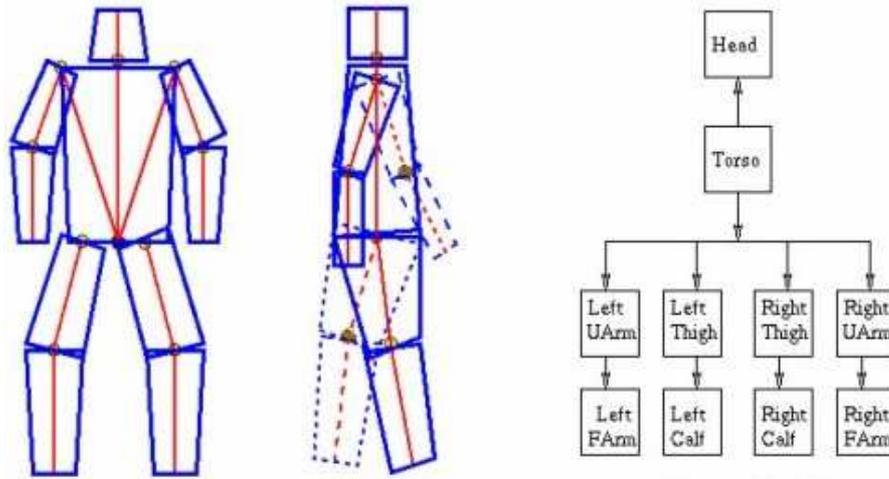


Figure 4.3: Kinematic body model.

Regardless of the parameter number, we can characterize the body motion by a mapping describing the forward kinematics of the underlying mechanical structure. The model used here is the well-established kinematic tree model where the torso and head form the base part and each limb as a separate chain connected to the base as shown in Fig. 5.1. A given human motion can be described as a global motion in the base and the articulations in limbs.

The forward kinematics mapping specifies the position and orientation of any body point from body parts and joints articulation. For any kinematical limb, its articulation is given by a series of matrix multiplications, parameterized by joint angles $\theta_1, \dots, \theta_n$ (θ_i is generally a time-varying multi-dimensional signal). In homogeneous coordinates, we can calculate the 3D position x of a point attached to the k -th body segment after motion as follows:

$$x = g \cdot x^0 = g_k(\theta_1, \dots, \theta_k) \cdot x^0 \quad (4.1)$$

where x^0 is the original position of the point x given in the local body coordinate system and g_k is the mapping describing the forward kinematics of the first k segments in the kinematic chain. For each fixed parameter set at time t , $\theta_1, \dots, \theta_k$, $g(\theta_1, \dots, \theta_k)$ is a 4 by 4 matrix specifying a rigid body transformation in homogeneous coordinates. Unlike many problems in robotics where the position and the orientation of the base is fixed, we allow the base frame to move. This motion can be caused either by human body motion or by camera motion. Hence x is given by:

$$x = D \cdot g_k(\theta_1, \dots, \theta_k) \cdot x^0 \quad (4.2)$$

where $D = [R, T]$ is the homogeneous matrix corresponding to rotation R and translation T of the base coordinate frame with respect to the camera frame.

Given the body and the camera model as well as the current body posture, the relationship between body points and the corresponding image pixels can be expressed by a combination of kinematic and camera transformations. We use the perspective projection to model the geometry of the camera transformation. This results in a mapping $u = f(F, x)$ between the image point u and the corresponding world point x .

Let the coordinates of a body point in a local body coordinate frame be denoted by x^0 . Assuming that this point belongs to the k -th body segment and that the configuration of the body at time t is given by $R(t), T(t), \theta_1(t), \dots, \theta_k(t)$. Its image position $u(t)$ after motion and projection can be calculated by a series of mappings as:

$$u(t) = f[F(t), D(t) \cdot g_k(\theta_1(t) \dots \theta_k(t)) \cdot x^0] \quad (4.3)$$

4.3.2 Signature surface

Our approach addresses the problem from a new angle to describe the foreground objects and background. It seeks to find a surface passing through foreground object that cuts the space-time volume. This reliance on appearance in time allows us to deal foreground objects with both static and dynamic scenes. Instead of focusing on the scene structure with fixing the coordinate system at the camera center, we study the case when the coordinate system is fixed on the object. Because of the vertical pose of the human body in many of the surveillance videos, we propose a spatio-temporal surface S as follows.

Let F be a video sequence, which is an ordered set of frames $\{F_t\}_1^N$. Each frame in turn contains an ordered set of horizontal strips F_t^j . The space-time volume is the set of strips $\{F_t^j\}$ (t is coordinate on the temporal axis, j is coordinate on the y axis). Using these notations, the desired surface can be denoted by a list of strips:

$$S = \{F_{t_k}^{j_k}\}_{k=1}^K \quad (4.4)$$

The height of the strips is set to one pixel as wider strips may form naturally when needed. We seek to find the surface which cuts the object with an equal distance to its top in every frame. An example of such a surface is shown in Fig. 4.5.

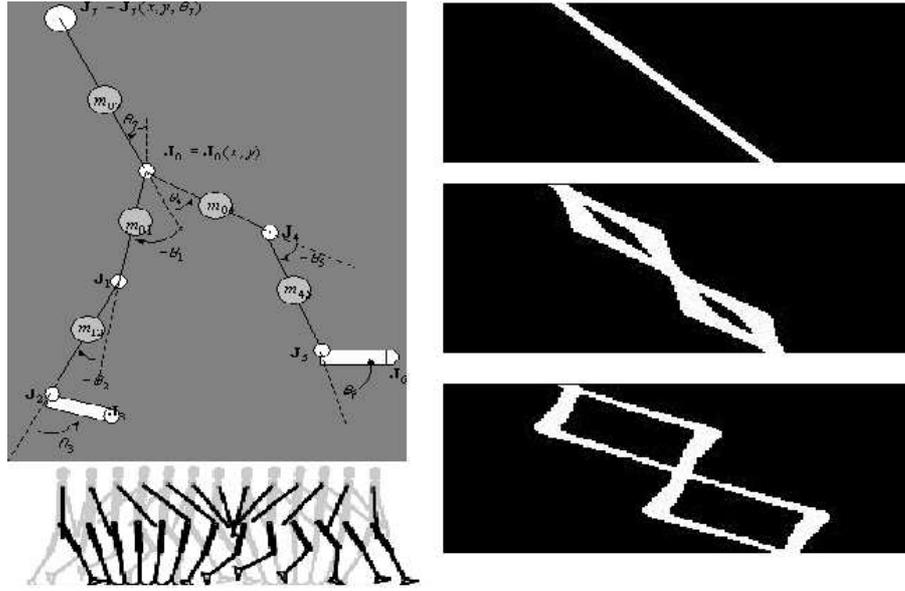


Figure 4.4: An example to illustrate X junctions generated by kinematic chain model.

4.3.3 X Junctions in space and time

In the described human body model, each limb is modeled as a kinematic chain centered at torso with a 3D motion $g(R, T)$. As in [110], a nine DOF chain model is used to describe the body structure (head, arms, torso and legs). The articulations are due to the mass of rigid body parts and motion (rotation and translation) of joints as shown in the left row in Fig. 4.4. Since the volumetric body parts are of some width, the pattern generated by them at time t in one surface could be represented by a point set as the intersection of the body and the tangent plane to the signature surface. By connecting these points we obtain the trajectory shown in the right row in Fig. 4.4. Assuming that the object is at a distance far enough from the camera, $F(t)$ can be set as a constant F over short time period.

The surface S is formed by combining various horizontal strips from frames throughout a sequence. For a rigid motion, S will contain a continuous path as shown in top-right image in Fig. 4.4. For non-rigid motion, the path is more complicated such as being twisted. We illustrate this in the middle and bottom images in Fig. 4.4.

The motion of the limbs divides gait articulation into approximately two levels: intra- and inter- gait motion. Intra gait motion represents the articulation of body parts relative to the body mass center and inter gait motion describes the overall body motion by the rotation and translation of the mass center. We make each row of S at a given vertical distance to the mass center as in Fig. 4.5 so that the inter-gait motion such as body translation and mass center variation is canceled. The remaining temporal pattern is only due to the intra-gait motion.

When considering the swing of limbs, different individual will have different styles and articulations and hence trajectories. But one feature is shared among almost all pedestrians while walking: two legs swing approximately out-of-phase, making an X-junction in space and time when they occlude each other. We naturally define X junction as the volume occupied by the two legs during their occlusion in walking motion. To prove the existence of such special shape we continue to use the kinematic chain model introduced above.

We set the world coordinate system at the pelvis center of a human body with directions towards sagittal plane, coronal plane and axial plane. Although the articulation of legs is subtle during the whole gait, their motion before and after both toes contacting the ground could be approximated by a twin-pendulum model.

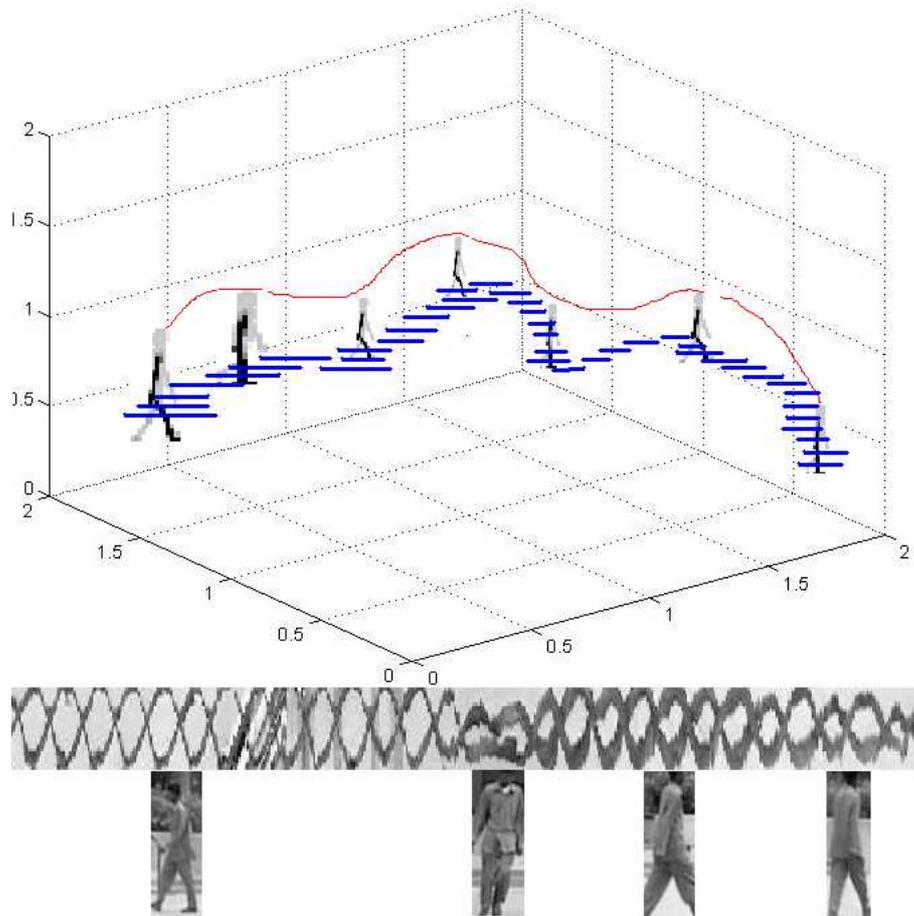


Figure 4.5: Signature surface in videos. Top: Red curve is the trajectory; blue lines form the horizontal strips in S for each frame along the trajectory. Middle: signature surface. Bottom: corresponding poses. Notice that the X junction appears for most part of the signature surface except for the radial viewing directions.

If we set the time as t_0 when both toes contact the ground plane, it also corresponds to the time when the twin-pendulum's arms are in the lowest location. Following Eqn. 4.3, we consider the same point at time $t_0 - \delta t$ and $t_0 + \delta t$:

$$\begin{aligned}
 u(t_0 + \delta t) &= f[F(t_0 + \delta t), D(t_0 + \delta t) \cdot g_k(\theta_1(t_0 + \delta t) \dots \theta_k(t_0 + \delta t)) \cdot x^0] \\
 u(t_0 - \delta t) &= f[F(t_0 - \delta t), D(t_0 - \delta t) \cdot g_k(\theta_1(t_0 - \delta t) \dots \theta_k(t_0 - \delta t)) \cdot x^0] \quad (4.5)
 \end{aligned}$$

It is shown in biomechanics that when δt is small enough compared to the gait period, $\theta(t)_1^k$ is symmetric. They could be approximated by Taylor series. Hence we have:

$$u(t_0 + \delta t) = u(t_0) + \Delta u \quad (4.6)$$

$$u(t_0 - \delta t) = u(t_0) - \Delta u \quad (4.7)$$

The above equation shows that the volume carved by one leg is symmetric to the position $u(t_0)$. Because of the bipedal motion, both legs generate a 3D X-shape junction in space and time. If we cut such a volume by a horizontal surface S defined as in Eqn. 4.4, we obtain a 2D X junction.

The nice features of the X-junctions are: 1) Inherent to the bi-pedal motion of pedestrians. 2) Robust to environmental conditions such as viewing angles, anthropometric parameters and platform motion. 3) Distinct from background clutter. We fuse human shape and motion by considering X junctions in space and time. Because of the vertical pose of human body while walking, S contains the intersection by slicing the 3D volume after compensating for global body motion.

4.4 Extraction

4.4.1 Support region

The surface S contains foreground patterns and background camera motion patterns. The motion of the pixels in the surface is due to either camera or foreground objects. When the camera is not static, the orientation of pixels provides information to estimate and compensate ego motion just like optical flow. Instead of calculating the flow as in [61, 15], we design an approach based on the gradient computation introduced in [149] to estimate the overall orientation of the background pixels in a surface for static as well as moving cameras. The distribution of local orientations across time inherently reflects the camera motion in a surface. In most of the videos, it is sufficient to assume that the camera motion is locally linear. If the surface is taken with a length of half a second from a video, there will be a dominant motion for that surface which is featured by the global orientation in that surface. To measure such a orientation, let S' denote the collapsed slice surface (fit each row of S in a plane) with an orientation angle θ . To estimate θ we pose the problem as finding the minimum gray level axis within a local neighborhood. More specifically, we find the local directional derivative that vanishes,

$$u^T \nabla S' = 0 \tag{4.8}$$

The above equation is often solved by minimizing a cost function within a local area R as:

$$C(u) = \min_{\|u\|=1} \int_R \|u^T \nabla S'\|^2 dR \quad (4.9)$$

Expanding the above equation we get:

$$C(u) = \min_{\|u\|=1} u^T \int_R (\nabla S' \nabla S'^T dR) u \quad (4.10)$$

Denote the integral part as

$$T = \int_R \nabla S' \nabla S'^T dR = \int_R \begin{bmatrix} S_x^2 & S_x S_t \\ S_t S_x & S_t^2 \end{bmatrix} dR$$

Hence the optimization becomes the following with T denoted as the spatio-temporal structure tensor.

$$C(u) = \min_{\|u\|=1} u^T T u \quad (4.11)$$

It is shown that the above optimization process corresponds to finding eigenvectors and eigenvalues. The eigenvalues can be used to classify regions as: homogeneous (i.e., no dominant orientation $\gamma_1 = \gamma_2 = 0$), single orientation ($\gamma_1 \gg 0, \gamma_2 = 0$) and multiple orientations ($\gamma_1, \gamma_2 \gg 0$). Interestingly, though motivated differently the structure tensor and its subsequent eigen-analysis corresponds to the SIFT corner detector analysis [89]. We seek to classify the dominant regions and thus avoid the expense involved in explicitly calculating the eigenvalues and instead analyze the determinant D and trace Tr of T ,

$$D = \det(T) = \gamma_1 \gamma_2, \quad Tr = \text{tr}(T) = \gamma_1 + \gamma_2 \quad (4.12)$$



Figure 4.6: Extracting the support region in a slice.

By calculating the 2D structure tensor for each pixel within a small neighborhood, we obtain the distribution of the gradient. We divide the 2D motion direction in a surface into a 2D histogram with $M * N$ bins. Each pixel in the surface votes for the bin $((\alpha_x, \alpha_t))$ with its gradient confidence G (a DOG operator) at direction along x and t as:

$$H(\alpha_{x0}, \alpha_{t0}) = \sum_{\alpha_x=\alpha_{x0}, \alpha_t=\alpha_{t0}} G(\alpha_x, \alpha_t) \quad (4.13)$$

The orientation caused by the camera motion can be detected by finding the peak in the histogram as shown in the bottom row of Fig. 4.2 because the peak corresponds to the dominant motion among the bins. It indicates the ego motion direction and amplitude of the platform. By subtracting the pixels belonging to the bins close to the global peak, we roughly eliminate the background and obtain a support region for the foreground. Examples of extracted support region are given in Fig. 4.6.

4.4.2 Learning X junctions

Since the foreground objects are at different depths than the background, they could be segmented using the method described above. We obtained a database of exemplar gray scale image patches of size 20×20 that covered a broad range of X-junctions in the spatiotemporal domain. In total, 400 X-junctions were labeled with an accompanying set of 500 anti X-junctions for classification.

The spatial domain analysis are presented in [47] using oriented histogram of gradients. Here the X-junction patches were normalized to have zero mean and unit variance and the principal components [50] were calculated to represent the dominant features of X junctions. Not surprisingly, the principal components capture certain characteristics of X-junctions that are physically plausible C namely, the occluding edge and the intensity profile of that being occluded. Projecting the data back onto the top three principal components, we can see a distinct structure in the data. To learn a classification model for X-junctions we tested the PCA descriptor using a Bayes classifier. Consider the a large vector composed of each image's intensities in the training set, PCA reduces the model dimensionality. In our system, the descriptor projects the data matrix onto the top 20 principal components and the energy distribution due to PCA is shown in Fig. 4.7.

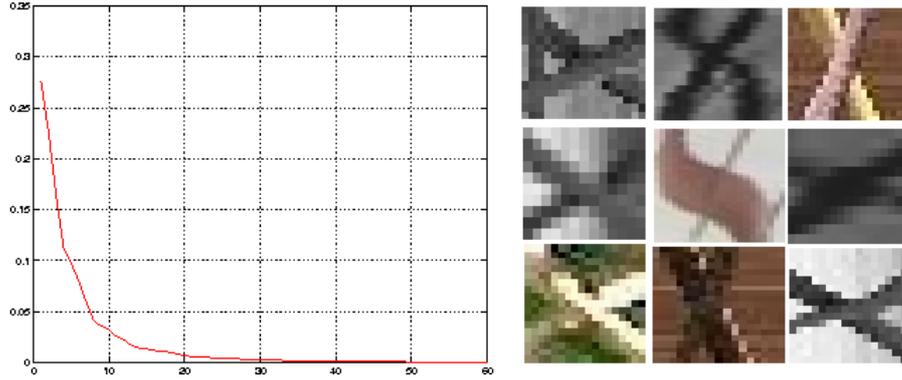


Figure 4.7: Learning the X junctions. Left: the eigenvalues of the PCA covariance matrix drop sharply after 20. Right: positive examples in the training set.

4.5 System Design

4.5.1 Graph construction

In this section we construct a complete system for reliable pedestrian detection regardless of camera motion. The flowchart in Fig. 4.8 shows the modules in the proposed system. First an initialization module is applied to a sequence at time $0, T_0, 2 * T_0, \dots$. This step gives a bunch of potential areas (boxes) $B_i(t)$ in those frames. Then we connect two boxes $B_i(t_1)$ and $B_j(t_2)$ to form a candidate trajectory only if they satisfy the following conditions:

1. The time-stamping difference between t_1 and t_2 are close enough (T_0 in our case).
2. The spatial distance between $B_i(t_1)$ and $B_j(t_2)$ are close enough for a regular walking speed.
3. The normalized appearance difference between $B_i(t_1)$ and $B_j(t_2)$ is less than

a preset threshold.

Or equivalently;

$$|t_1 - t_2| \leq \gamma_1 T_0 \quad (4.14)$$

$$\|O_i - O_j\| \leq \gamma_2 (w_i + w_j)(h_i + h_j) \quad (4.15)$$

$$\text{mean} \left[\sum_{P \in B_i(t_1), Q \in B_j(t_2)} (I_P^2 - I_Q^2) \right] \leq \gamma_3 \quad (4.16)$$

After that, we construct an infinite graph G to organize the candidates. The graph is defined as:

$$G(V, E) \doteq G(B_i(t), S(z)) \quad (4.17)$$

where $B_i(t)$ is the vertex. An edge E is defined as the spatial temporal surface formed by cutting the data volume connecting two vertices. It is valid if and only if the two vertices belong to the same pedestrian. We generate spatio-temporal surface(s) for each pair of vertices (boxes) satisfying the above criterion by cutting the X-Y-t volume by planes parallel to the trajectory. Only a small set of surfaces are selected for extracting the possible pattern. In our experiment, only the surface passing at 1/4 height of the body is chosen. After estimating the global camera motion at the surfaces, we use the obtained support regions to locate the foreground. Finally the support region is partitioned and verified for the presence of the gait pattern. Such a procedure simultaneously affirms initial detections and generates piecewise linear trajectories as the *paths* in Graph G .

It should be emphasized that by connecting candidate regions, this system no longer needs to track the objects, which is unreliable and computationally expensive in some cases.

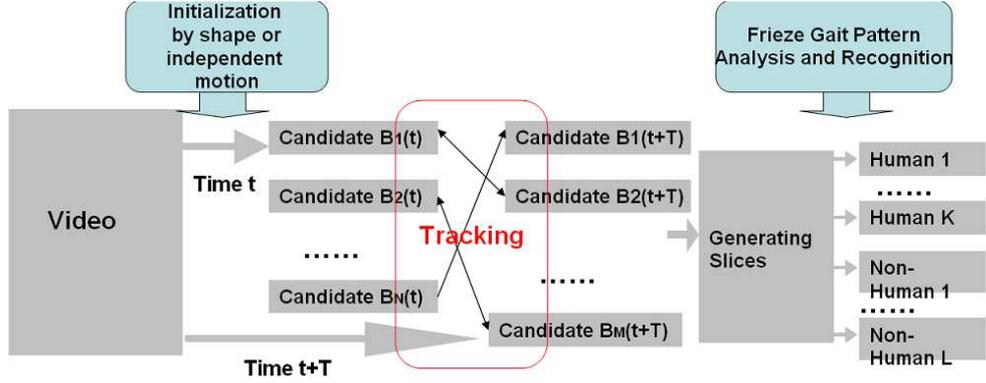


Figure 4.8: Flow chart for the proposed system.

4.5.2 Initialization

This system can take any algorithms which are capable to initialize pedestrians in a frame. We use a similar method developed in [99] or the OpenCV version of Viola’s Adaboosting method [132] to find the candidate areas using probabilistic template matching. For the first method, the training data for developing the probabilistic template consists of more than 100 rectangular images containing human upper body segments in different poses and orientations. The reasons that we only consider the upper body lie in the following facts: (i) head and shoulder articulations are relatively small compared to limbs. (ii) upper body shares a similar ω contour shape for different individuals, which is easier to describe than a complicated model for the whole body. Since pose is a challenging factor to estimate [2, ?, 4], we have a few templates to take into account different poses. In our case three poses are used, left, right and upfront at different scales. For each pixel of the templates, the probability of it being pedestrian at that pose is calculated based on how frequently it appears as 1 in the training data. This upper body template in effect gives the

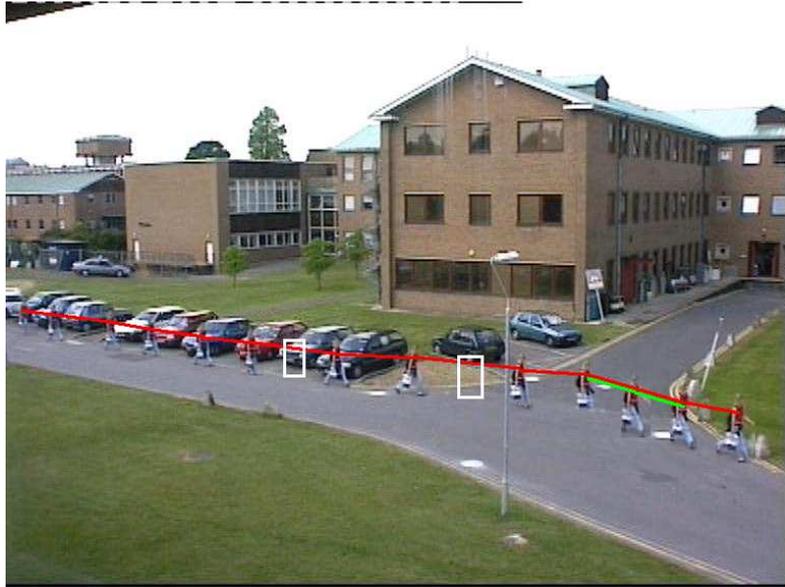


Figure 4.9: Continuously detecting pedestrian in PETS.

probability of seeing a foreground at different pixel locations for pedestrians. During initialization, we apply a gradient operator to the smoothed image and scan the gradient map to calculate the sum of these probabilities for all pixels. This gives us the combined probability of the given window containing a human. The value of the threshold is decided using a Bayesian classifier reported in [99].

In order not to miss a true target, we relax the parameters in both methods to include as many real pedestrians as possible and a reasonable amount of false alarms to be rejected later.

4.6 Experimental Results

4.6.1 Video sensor

We first consider the detection of pedestrians in videos acquired from a single video camera. By connecting the verified edges containing the X junctions, we continuously find the pedestrian trajectories . Figure 4.9 shows the typical detection results for a video from the PETS data where a static camera is deployed in a surveillance area. The initial detector works every T_0 frames (15 or 30 in our case) and gives a bunch of false alarms as well as true detections (red lines). For example, the windows in the background building were initialized as candidates in some frames. But the verification successfully rejects them. It also misses two detections which are illustrated by the white bounding boxes. Even with possible false negatives, our algorithm picks it up by allowing the linking of two initial detections $2 * T_0$ frames apart in the graph. Finally we superimpose the detections and trajectories in red color in one frame. The path in G reveals the human's movement direction. In general, it correctly classifies and locates the target consistently. In the right part of the trajectory, one segment of the trajectories is verified as non-human because the lower half of the human body is occluded by a car passing in front of the human during that time.

Results for videos acquired by a moving camera are shown in Fig. 4.10 when the platform is mounted on a mobile sentry. Only those corresponding to the same pedestrian give positive results and affirm the initial detections (in red) by rejecting false alarms (in green and blue).

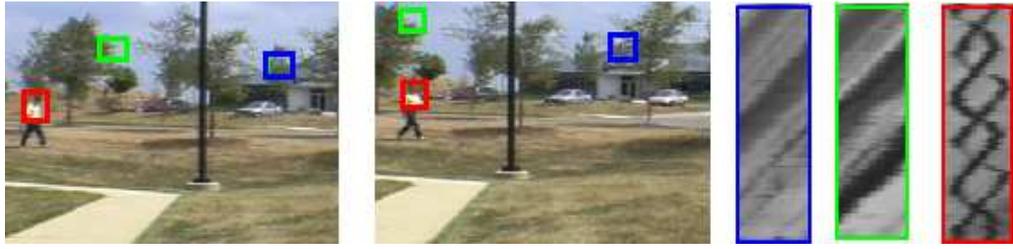


Figure 4.10: Continuously detecting pedestrians in video sequences acquired by a moving platform.

We then apply the detector to an IR sensor and the result is shown in Fig. 4.11. Because of the sensor noise, IR video tends to have more false alarms in the initialization stage. But for infrared videos, our method maintains the detection rate above 80% with a false positives rate lower than 10%.

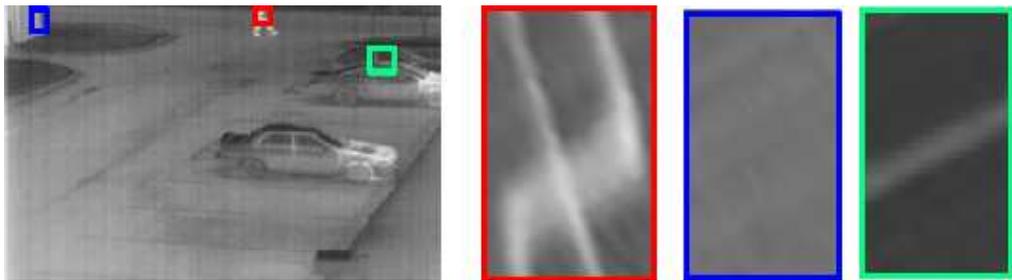


Figure 4.11: Continuously detecting pedestrians in IR sequences. We list 3 initialized regions (with different color) and only one contains the X junction.

We plot the ROC curves for the UMD and PETS dataset in Figure 4.12 with 369 pedestrian clips. The ROC curves plot the false positive rate against the detection rate, when the classification criterion is varied. The false positive rate is defined as the total number of false positive detections divided by the total number of objects in all sequences; the true detection rate is the ratio between the total

number of correct detections and the total number of detections in all sequences. Since we use a sequence for one classification, we do not divide the above rates by frame number. We compare the proposed method to the results from the OpenCV version of Adaboosting classifier and our implementation of a shape based pedestrian detector [99]. We obtain better performance in terms of higher detection rate at the same false negative rate for both color and IR sensors. This is due to the efficient verification of temporal symmetrical motion coherence. During our experiments we also notice a better performance of the proposed method for the static platform case than for the moving platform. This is because the accuracy of 'supporting region' is better in the static case than in the moving case. Since the camera motion is more complicated (may including rotation) for moving sensors, the quality of support region and hence the performance of verification stage drops.

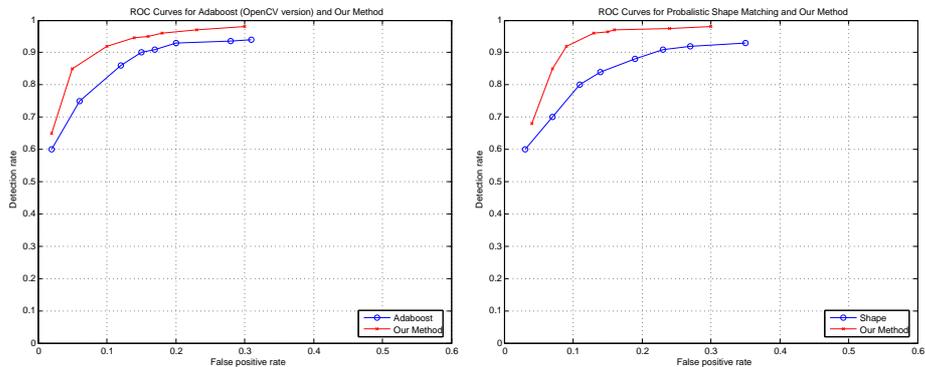


Figure 4.12: ROC analysis for UMD color/gray dataset and IR dataset.

An important factor is the viewing angle, or walking direction. The temporal X junction will degenerate into ribbon(s) in radial views. Under such conditions, the classifier cannot differentiate it from other rigid moving objects. Another situation when this method may fail is when the pedestrians are stationary. But our algorithm

is able to quickly pick it up as soon as the target resumes walking.

4.6.2 Range sensor

Range sensor models for ultrasonic sensors and laser range finders have been around for many years. These sensors are frequently deployed as components on autonomous systems for navigation or surveillance. We are interested in the Horizon Infrared Surveillance Sensor (HISS) system, which is installed in a static or a moving platform, scanning horizontal lines at a very high speed. The output is the range data at a given height, which naturally forms a slice (a horizontal surface) containing the desired motion signature as shown in Fig. 4.13.



Figure 4.13: Simulated range sensor slice data.

We simply apply the same method to the output signal from range sensors for the purpose of detecting pedestrians and the result as well as the ROC curve are shown in Fig. 4.14. During our experiments, we are using an SRI stereo head (base line 10 inches) to simulate the range sensor response. The sensor is mounted at one foot from the ground plane to best capture the X-junctions during the leg swing.

4.6.3 Activities in sport videos

Given the detected trajectories and the verified X-junctions associated with continuous time instants, we can parse each trajectory into segments where each

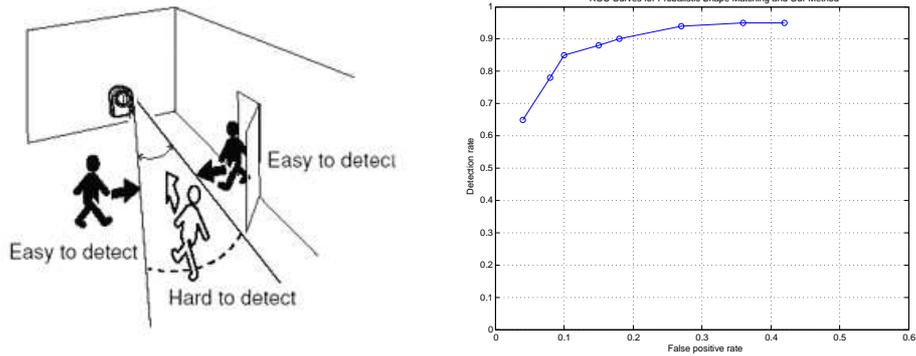


Figure 4.14: ROC analysis for range sensor data.

segment contains a single atomic motion such as walking, standing, running etc. The goal in this section is to bridge the gap between low-level visual features and mid-level semantics. Here we are focusing on automatic analysis of sports videos, or more specifically, soccer videos. This is especially useful in applications of digital video game understanding such as summarizing and retrieving.

Traditional sports video analysis toolbox focus on the silhouette, which is hard to extract in many cases and may not be reliable. Since the players's trajectory is almost piecewise linear and the patterns from various activities are distinct from each other, we can apply some simple classification for activity recognition based on the features of X junctions. For example, a standing player will degenerate the X-junctions into straight ribbons. A running player makes the distance between adjacent X-junctions much shorter than walking (faster in time). When he/she is passing or receiving the ball, the X-junction will be distorted into curves of other form. Such observations suggest to analyzing the activity by the deformation of X-junctions in real time. We add a post-processing module after the extraction of

X-junctions to measure its closeness (in time) and induce the repeating speed of the X junctions for each player. We also train a similar ribbon detector (PCA feature and Bayesian classifier) to handle the degenerate X-junctions for standing/stopping players.

Our approach keeps updating and classifying the motion signature for each player at every half second. This interval is fast enough to analyze the activity on-the-fly. By doing so, we can not only know the positions but also understand the strategy of both teams. For each player, we assign an activity tag to him/her such as *running*, *walking*, *stopping*, *others*. An example is shown in Fig. 4.15.

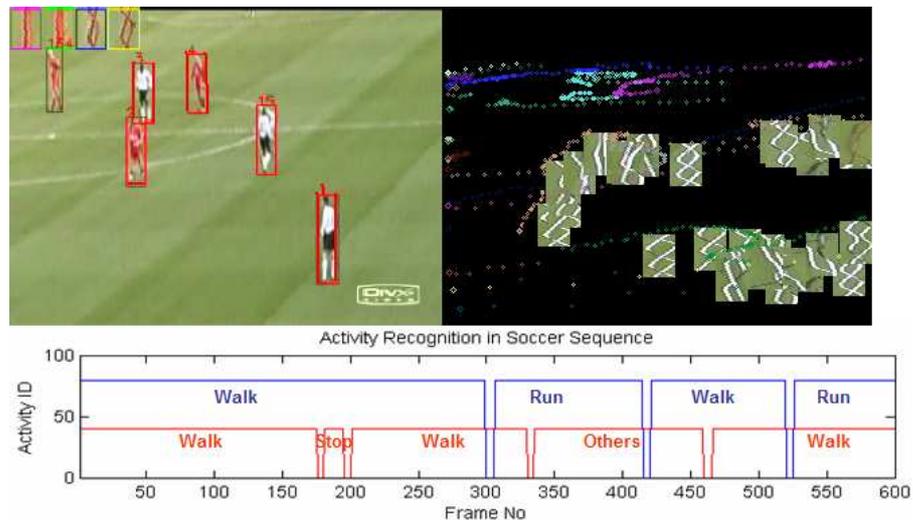


Figure 4.15: Soccer sequence and the classified events with overlaid motion signature. Top: left image is one frame with detected players. The right image has the color dots coding the trajectories for all players and the associated motion signature representing activities. Bottom: activity classification results for two players.

The advantage of our method is that so far no feature point tracking or background subtraction [37] is required and hence it is computationally efficient and

robust to camera motion.

4.7 Conclusion

We presented a method for analyzing pedestrians in space and time and implemented a real time monitoring system. This two-step system first initializes potential candidates and then verifies the gait patterns in temporal surfaces formed by connecting close candidates in an infinite graph structure. The feature used as a particularly strong indicator of pedestrians is the X-junction in space and time. Unlike traditional rhythmic motion based detectors, the X junction lies within a stride and the detector does not need long period tracking. Our approach fuses appearance and motion in a very efficient manner to reject the false alarms in the initialization stage. The proposed algorithm is independent of tracking modules. It can be applied to surveillance applications such as mobile sentry or intrusion detection. The experimental results demonstrate the effectiveness of the proposed method using visible and IR videos with arbitrary camera motion. It is also compatible with scanning data from range sensors. The ROC curve analysis supports the high statistical performance. We also show its effectiveness in sport video analysis by paring the players' activities in a soccer game. It successfully recognizes the temporal pattern change due to walking, standing or running without tracking any feature point or extracting the silhouette.

Chapter 5

A Compact Characterization of Human Motion

5.1 Introduction

Human motion analysis has been an active research area over the last decade with applications in surveillance, clinical and sports video analysis [33, 34, 35, 36, 131, 148, 67, 7, 8, 9, 32, 31]. Although the traditional approaches are based on markers, silhouette or feature points extracted from the human body [63, 68, 69, 70, 77, 76], we are interested in the landmark-free methods. Comprehensive reviews on human motion and activity analysis may be found in [52, 81]. In [52], methods are classified into 2D approaches that do not use explicit models, 2D approaches that use explicit models and 3D approaches. To capture the cyclic human gait patterns, many solutions have been proposed for characterizing the periodic motion at pixel level or in 3D space.

In the first category, Little *et al* analyzed the shape variations due to motion for real time classification [84]. Boyd [21] introduced the video phase locked loop for perceiving the oscillations at a pixel level. Seitz *et al* presented [128] a novel concept called period trace for detecting motion trends. Allmen *et al* [10] proposed an approach for measuring periodicity using a curvature scale space at each pixel. Polana *et al* [108] showed that the recognition of periodic locomotion can be matched against a temporal template. Tsai *et al* [130] described a method using

DFT to extract pixel period. However, most of the above pixel-based methods do not exploit gait kinematics except for periodicity. Hence they are unable to capture the articulations of body parts present in a stride.

Methods in $x - y - t$ space analyzed the topology of human motion in space and time [6]. Niyogi *et al* in [98, ?] analyzed the periodic patterns and used them to estimate gait parameters by Active Contours proposed by Cootes *et al* [46] and Active Snakes/Surfaces by Kass *et al* in [75]. Liu *et al* in [85] used the power spectrum for extracting the periodic motion. Again, they neither studied the gait animation/kinematic constraints nor considered multiple view geometric constraints.

To understand the spatio-temporal pattern in activity sequences, a straightforward direction is to measure the similarity between each instances. The work closely related to this paper is by Liu *et al* [?] who used the Frieze Group Theory for modeling gait silhouette profile in X and Y axes. They proposed a method to estimate the viewing angles from different symmetries in 1D patterns. This approach requires segmented silhouettes and does not study other activities than walking. Cutler *et al* [44] measured the motion similarity between image pairs of a gait sequence and extracted a lattice representing human walking motion. However the body articulation is lost during the correlation stage. There are several major differences between our method and those discussed above: (1) We study the topology of activities guided by a kinematic model, while they discard such information either by focusing on correlation or silhouette histogram. (2) These methods assume that objects have been segmented and/or aligned. We focus on simultaneous detection and segmentation with and without occlusion. (3) We analyze the geometric con-

straints between DHS from multiple views or individuals. (4) Finally, we investigate its usefulness in activity recognition.

A systematic analysis of human activities can be done using state-of-the-art techniques from computer graphics, biomechanics and computer vision. Inspired by studies in human motion animation and kinematics, we introduce the DHS to characterize the gait or activity topology in spatio-temporal domain. Humans walk at a stable frequency. The body and limbs maintain the center of gravity above the point of contact and minimize the muscular effort needed for balancing the whole body for various events such as natural walking or carrying a brief case. If we stack all the instances for an activity in the X-Y-Z-t space, we observe a 4D point set with special topology containing both global body movement and local subtle variations for that specific event. Although they vary for different individuals, the topology for one class of activity retains strong similarity. Following Felix Klein's work [?] we look for a compact representation of human walking motion.

The main feature of the proposed approach with respect to the techniques mentioned above is that we do not match image features such as regions, points and markers from frame to frame. Our work makes two major contributions. First we show that the sliced spatio-temporal pattern generated by a 3D kinematic model belongs to a geometric symmetry group and forms a compact signature. Then we propose a robust pedestrian monitoring system that segments and labels targets and recognizes a class of events by applying DHS.

The paper is structured as follows. Section 5.2 discusses the DHS pattern and then establishes its properties such as symmetry and compactness in representing

human gait and activities. Section 5.3 proposes a method for simultaneously identifying and extracting such signatures. Section 6.1 discusses a pedestrian monitoring system covering applications from segmentation with and without occlusion to event detection supported by experiments. Section 6.6 summarizes the work.

5.2 Double Helical Signature

5.2.1 Spatio-temporal gait and activity volume

In subsequent discussions, we interchangeably use gait and activity. Human gait can be represented by a set of 3D body points. Each body point generates its own trajectory in a four dimensional space (X, Y, Z, t) . The sequence captured by cameras is formed by stacking the 2D frames at every time instant and hence is in a 3D space (x, y, t) . Two corresponding points X_P , x_p from the two spaces are related by a projection as:

$$x_p = \begin{bmatrix} a_1 & a_2 & a_3 & 0 & a_4 \\ b_1 & b_2 & b_3 & 0 & b_4 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot X_P = P \cdot X_P \quad (5.1)$$

where X_P and x_p are in homogeneous format. By doing so, temporal information is naturally coded as one of the dimensions. We define the Activity Volume as a subset of points in the 4D space as follows:

Definition: Activity Volume G is the 4D spatio-temporal (X-Y-Z-t) volume occu-

ped by the human performing the activity. **Activity Sequence g** is the projection of the Activity Volume to the 3D spatio-temporal (x-y-t) frame domain.

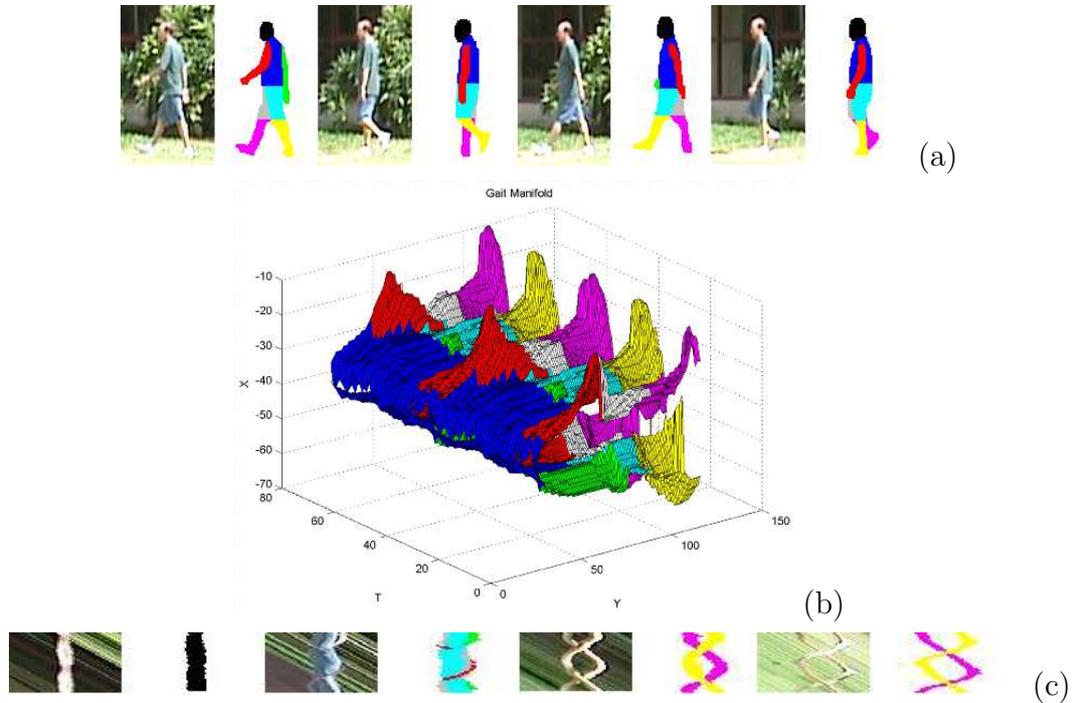


Figure 5.1: Top: Original frames and silhouette; Middle: Activity Sequence; Bottom: Selected 2D slices containing helical structure.

What are preserved in many gait related activities are the global periodic (or symmetrical) body translational motion, bipedal limb articulation and the vertical pose as in Fig. 5.1. It is our goal to find an efficient representation derived from the raw video frames. The proposed spatio-temporal slices along the body movement direction generates a geometric repetitive pattern. More interestingly, a close look at the horizontal slices reveals the embedding of some body articulation parameters as Fig. 5.2.

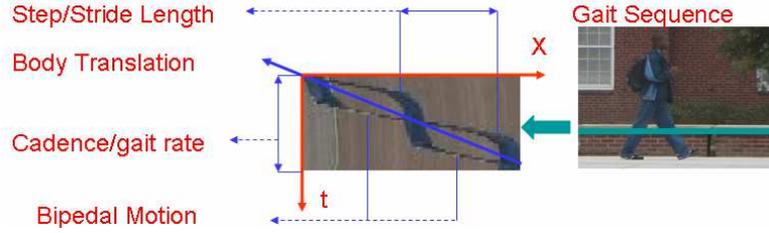


Figure 5.2: Gait parameters coded in DHS.

5.2.2 Geometric symmetries

Symmetry is a fundamental concept for understanding repetitive patterns in art decoration, crystallography etc. This has been a primary motivation for developing the branch of mathematics known as Geometric Group Theory [140]. A geometric figure is said to be symmetric if there exist isometries that permute its parts while leaving the object as a whole unchanged. An isometry of this kind is called a symmetry. The symmetries of an object form a group called the symmetry group of the object. A symmetrical group spanning in 1D is defined as a Frieze Group and is defined as a Wallpaper Group in 2D space.

Because human walking motion generates translation along planes parallel to the direction of global body translation, we are more interested in planar symmetries such as reflections and half turn. There are seven distinct subgroups (up to scaling) in the discrete Frieze group generated by translation, reflection (along the same axis or a vertical line) and a half turn (180 rotation). Sample patterns are shown in the Fig. 5.3. The seven different groups can be described as follows:

1. Translations only.

| | Group generated by: | Pattern left invariant |
|-------|---------------------|------------------------|
| (i) | T | ...P P P P P P P... |
| (ii) | G | ...P Ъ P Ъ P Ъ P... |
| (iii) | T, R | ...B B B B B B B... |
| (iv) | T, V | ...A A A A A A A/... |
| (v) | T, H | ...S S S S S S S;... |
| (vi) | G, V | ...A V A V A V A/... |
| (vii) | everything | ...H H H H H H H... |

Figure 5.3: An example to illustrate 7 patterns in Frieze Groups. T: translation, V: vertical reflection, H: horizontal reflection, G: glide reflection [140].

2. Glide-reflections and translations.
3. Translations, reflections in the horizontal axis and glide reflections.
4. Translations and reflections across certain vertical lines.
5. Translations and 180 rotations.
6. Glide-reflections, translations and rotations.
7. Translations, glide reflections, reflections in both axes and 180 rotations.

In our work, we consider the Frieze Group Theory of 1D pattern to repeat along arbitrary directions in R^2 . Such a direction represents the instantaneous translation

of the human body. We do not have to limit ourselves to rely on the cropped and aligned images as in [44, ?].

5.2.3 Kinematic chain model

It can be very difficult to analyze the complicated nature of human shape and motion without using a shape model. Several possibilities for human shape modeling exist ranging from stick figures and ellipsoids to more sophisticated deformable models [?]. Modeling the human body as rigid parts linked in a kinematic structure is an effectively approximate model for many purposes. In biometric research, body locomotion is specified at two levels: inter-gait (global) parameters such as speed v , step/stride length s , and cadence f (or period T) resulting in global translation as well as intra-gait (local) articulation capturing subtle deviations. The kinematic chain offers a useful approximation of such locomotion. A humanoid, or a walking robot, consists of a number of serial sub chains: legs, arms, and head, all connected to the same trunk. Consider a human model H , where H has articulated parts $L_1, L_2, L_3 \dots L_N$ and joints $J_1, J_2, J_3 \dots J_M$. The parts L_i are rigid and the joints J_i between parts are small enough compared to the parts L_i they connect. Without loss of generality, we assume that body parts are of the unit width, forming a stick model. The number of independent body parts and joints can vary for different accuracy requirements.

We consider an open kinematic chain composed of parts $\{L_1, L_2, \dots L_K\}$ and joints $J_0, J_1, \dots J_K$ connected to the base body in the order of $J_0, L_1, J_1, L_2, \dots J_K, L_K$.

The body frame coordinates system $\overline{X} = (X, Y, Z)$ is fixed in the torso region. The position of any point P in the (X, Y, Z, t) 4D space of the kinematic chain at the k^{th} part at time $t + 1$ can be represented as a product of a series of matrices and the previous position at time t as:

$$X_P(t + 1) = T(t) \cdot X_P(t) = \prod_{i=0}^{k_0} T_i(t) \cdot X_P(t) \quad (5.2)$$

where $T_i(t)$ is the 4×4 transformation matrix describing the articulation of joint J_{i+1} from J_i . The forward kinematic equation (5.2) gives the position in the 4D space using the homogeneous transformation. Moreover, the matrix $T_i(t)$ defines both the link transformations (determined by the link geometry) and the joint transformations (determined by the joint position). The 4×4 matrix $T_i(t)$ can be decomposed as a product of a matrix D_i describing a position vector and a 3×3 rotation matrix $R_{i-1,i}$ for the i^{th} relative to the $(i - 1)^{th}$ part as:

$$X_P(t + 1) = R_{0,1}(t) \cdot D_1(t) \cdot R_{1,2}(t) \cdot D_2(t) \dots R_{k-1,k}(t) \cdot X_P(t) \quad (5.3)$$

On one hand, from the bio-mechanical point of view, the human body keeps the mass center in the middle during the swing of the limbs. On the other hand, from the computer animation point of view, each joint in the body chain model generates its own trajectory in space and time. With many physically possible combinations, the normal articulation parameters for a pair of limbs $\Theta = \{\theta_l, \theta_r\}$ are identical except for a phase difference of π because of the bipedal property:

$$\theta_l(t) = \theta_r(t - T/2) \quad (5.4)$$

$$\theta_l(t) = \{R_{0,1,l}(t), D_{1,l}(t), R_{1,2,l}(t), \dots, R_{k-1,k,l}(t)\} \quad (5.5)$$

$$\theta_r(t) = \{R_{0,1,r}(t), D_{1,r}(t), R_{1,2,r}(t), \dots, R_{k-1,k,r}(t)\} \quad (5.6)$$

Also, assuming T to be the gait rate, we have the following constraint due to periodic motion:

$$\theta_l(t) = \theta_l(t + T), \quad \theta_r(t) = \theta_r(t + T) \quad (5.7)$$

If we slice G horizontally along the direction of body movement (assumed to be piecewise linear) into planes, we observe a view-dependent twisted pattern. Each row (t^{th}) of such pattern at height Y_0 is the Y_0^{th} row of t^{th} instant pose for that activity. Intuition tells us that the pattern is also periodic. The *stride* and *step length* as well as *cadence* can also be directly obtained from such a pattern. These patterns uniquely characterize an individual's articulation in gait and activities.

Definition: An **Activity Signature** is the set of shapes $S = \{S_1, S_2 \dots S_Y \dots\}$ formed by slicing the Activity volume G at all heights covering the whole human body during a complete stride. Each S_Y corresponding to limbs is defined as a **Double Helical Signature**.

Combining G with the articulation parameters Θ , S_Y is given as $S_{Y_0} = G_{\Theta}(X, Y, Z, t)|_{Y=Y_0}$. If we define $G_{\Theta}(X, Y, Z, t)|_{Y=Y_0}$ as $G(Y_0)$, decomposing it from $t = 1 \dots T$ results in: $S_Y = [G(Y, 1)|G(Y, 2)|\dots G(Y, T)]$, where $G(Y, t)$ denotes the subset of G at height Y and time t . It is a $3 \times n$ matrix that contains the coordinates of n points. S_Y

is a $3 \times N$ matrix where N is the sum of all n_s . In particular, we can divide $S_Y(t)$ into two halves when it corresponds to the DHS generated by a pair of limbs as:

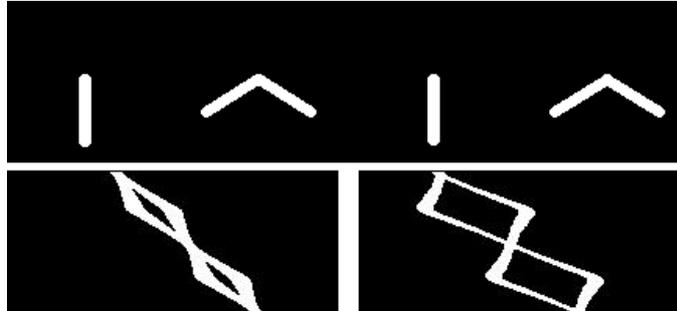
$$S_Y(t) = [G(Y, t)_l | G(Y, t)_r]$$


Figure 5.4: DHS generated by a twin-pendulum model. First row: A twin-pendulum moving across an image; second row: X-t slices containing periodic helical structure

Before providing a detailed analysis, we first give a simple example in Fig. 5.4. In a symmetrical twin pendulum model approximating the hip-to-toe motion, each leg is modeled as a line segment with out-of-phase oscillating angle $\alpha(t)$, period T and a translation speed v . The generated slices are shown in the second row of Fig 5.4 where the twin pendulum is translating across the image from left to right. They contain symmetries such as reflection symmetry along horizontal and/or vertical axis and even 180° rotation for any $\alpha(t)$. Since the 'legs' are of unit width, the signature at time t (t^{th} row) consists of up to two points, which are the intersections of two chains with the plane $Y = Y_0$. Assuming that the center O is located at pelvis with

Y axis pointing downwards, we have

$$S_Y(t) = [G(Y, t)_l | G(Y, t)_r] = [P_1, P_2] = \begin{bmatrix} Y \tan \alpha(t) + v * t & -Y \tan \alpha(t) + v * t \\ 0 & 0 \\ Y & Y \end{bmatrix} \quad (5.8)$$

The pattern S_Y at height Y has a translational speed v along the X axis with period $T/2$. This immediately suggests the applicability of the repetitive pattern analysis using Frieze Group Theory [140]. When v equals to zero, it degenerates into the case studied in [44, ?] where the objects are already aligned. Another interesting fact is that the DHS patterns at different heights only differ by the oscillating amplitude, which is decided by the distance from that plane to the center O (pelvis complex). From a DHS at $Y = Y_1$ one can derive the signature at another plane $Y = Y_2$ by simply scaling in the direction perpendicular to the translational displacement (line $(v * t, t)$). We can compress the volume by using only one DHS because it carries enough information to estimate the articulation (the speed v and style $\Theta(t)$).

5.2.4 Animated human activity

Actual human activities are more complicated but they do share some simple kinematic properties. We propose the following theorem relating the bipedal and periodic nature of a real gait with Frieze Group Theory.

Theorem 1: The DHS generated by limb articulation belongs to a Frieze Group.

Proof: It is obvious that the DHS has a translational period T because of the cyclic gait property. Given the relationship of intra-gait model configuration $\Theta(t) =$

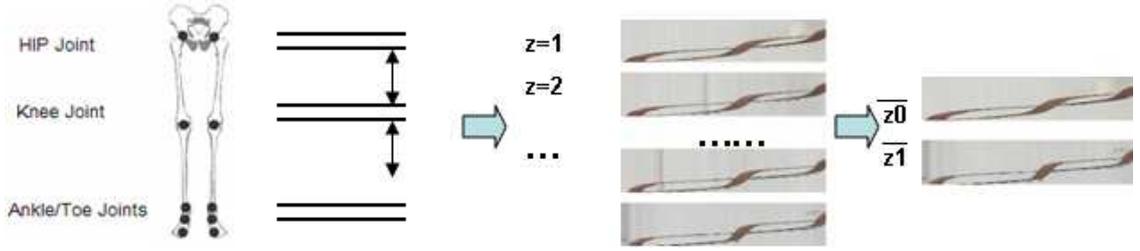


Figure 5.5: Hip-to-toe kinematic chain model

$\{\theta_r(t), \theta_l(t)\}$ for limbs as specified in Eqns. (5.6) and (6.4), the lower body's DHS at height Y and time t has two points at the t^{th} row in that slice. For a given t we have the following for the matrix of S_Y :

$$\begin{aligned}
 S_Y(t) &= [G(Y, t)_l | G(Y, t)_r] = [G(Y, t - T/2)_r | G(Y, t + T/2)_l] \\
 &= [G(Y, t + T/2)_r | G(Y, t + T/2)_l] = [G(Y, t + T/2)_l | G(Y, t + T/2)_r]
 \end{aligned}
 \tag{5.9}$$

which directly suggests a translational symmetry. The inter-gait parameters such as speed will shift the t^{th} row by vt along the X axis and cadence will scale the length along t as shown in Fig. 5.5. But the DHS still preserves the symmetry because the shear transformation does not affect the symmetry along the shifting direction ($X = vt$). One can observe from Eqn. 5.9 that the signature at t^{th} row repeats itself at $(t + T/2)^{\text{th}}$ row. If we consider the tiles of the DHS as a *set* and define the operation f as the translation (assuming that the walking direction does not change) of one tile by multiples of $(T/2, vT/2)$ and I as the identity transformation, we obtain a *group* since it satisfies axioms such as the existence of inverse transfor-

mation (backward translation), *associativity* and *commutativity*. In particular, the pattern belongs to the Frieze Group because it has a period of $T/2$ characterized by 1D translation $(v * t, t)$.

It is apparent that different walking styles will result in various point or line symmetries in DHS. For example, approximating each limb by a pendulum generates the half-turn (180 *rotation*) symmetry in DHS. Asymmetry between the left and right limbs changes the translational vector to (vT, T) . These observations can be proved using arguments given above. In summary, the possible symmetries in gait and activities are listed below.

1. Asymmetrical limbs: Translations with a vector (vT, T) only.
2. Symmetrical limbs: Translations with a vector $(vT/2, T/2)$.
3. Symmetrical limbs approximated by pendulums : Translations with a vector (vT, T) and half turns.
4. Symmetrical limbs approximated by identical pendulums : Translations with a vector $(vT/2, T/2)$, horizontal reflections and half turns.
5. Symmetrical limbs approximated by constant-speed pendulums : Translations with a vector $(vT/2, T/2)$, horizontal , vertical reflections and half turns.

The spatio-temporal planes are formed by slicing an image sequence. The transformation of the original pattern in 4D activity volume to 3D shape sequence with both projection (Eqn. (5.1)) and articulation (Eqn. (5.2)) are represented as follows under orthographic projection:

$$x_p(t+1) = P \cdot X_P(t+1) = P \cdot T \cdot X_P = P \cdot \prod_{i=0}^k T_i(t) \cdot X_P \quad (5.10)$$

where P is the projection matrix. Besides the symmetry preserved in the slices, there is also redundancy in gait. The following theorem establishes the efficiency of DHS in representing the gait volume G :

Theorem 2: There exists a finite set of DHS as a compact representation for the hip-to-toe Activity Volume G .

Proof: Given an articulation parameter set $G(\Theta)$, we consider the animated gait by the limb part L_i at different slices along the Y axis. Because of the vertical body configuration, axis Y can be divided into intervals $\{Y_0, Y_1\}, \{Y_1, Y_2\}, \dots, \{Y_{N-1}, Y_N\}$ (shown in Fig. 5.5). Each interval contains the DHS generated by the n -th pair of limb parts $L_{n,l}, L_{n,r}$.

The first limb pair generates DHS in horizontal planes from $Y = Y_0$ to $Y = Y_1$. From Eqn. (5.2), any two points P_1, P_2 lying on the same part L_n share the same articulation parameters T_0, T_1, \dots, T_{n-1} and differ only in their location vectors D . Hence all the DHS in this interval are identical except for a scale factor in the X direction. Given the DHS at $Y = Y_1$, we could derive others at $Y = 0, 1 \dots Y_1 - 1$. Actually this holds for any $\overline{Y_0} \in [Y_0, Y_1)$. Thus we conclude that the DHS at $Y = \overline{Y_0}$ has enough information to reconstruct or estimate the topology of G at interval $Y = 0, 1 \dots Y_1 - 1$. Similar conclusion can be drawn for the volume at following intervals $[Y_1, Y_2), [Y_2, Y_3), \dots, [Y_{N-1}, Y_N)$. Hence the DHS at $\{\overline{Y_0}, \overline{Y_1}, \overline{Y_2}, \dots, \overline{Y_{N-1}}\}$ form a complete set to reconstruct G and estimate Θ . The number of DHS in the above

set is less than the total number of slices. In real applications the number of required DHS is decided by the complexity of underlying kinematic motion and the accuracy required to approximate G .

In conclusion, to capture every subtle variations in gait variations, we need to include all the slices. But as an approximation to the topology of activity volume and designing an efficient pedestrian monitoring system, a few slices may be adequate.

5.2.5 DHS in images

When a human performs an activity, we only observe the images at a specific camera location. Various positions and poses result in different activity sequences. This section studies the variation of features such as symmetry of the extracted DHS when the subject is captured by multiple cameras. Most existing methods deal with geometric constraints for repeated patterns between images. Schaffalitzky and Mittal's approaches in [?, 92] automatically detected and grouped image elements repeating on a planar scene. Baker *et al* [?] proposed the epipolar-plane image (EPI) to analysis geometric structure constraint in slices.

Because it is difficult to establish accurate correspondence between landmarks extracted from a human body, we take an alternative approach to find the geometric constraint. During the period of an activity, the y^{th} row of each image contains points generated by the intersection of the body parts (unit width) and the plane at y . Denote the compact set of slices used for an activity as $s = \{\overline{s_0}, \overline{s_1}, \overline{s_2}, \dots, \overline{s_{N-1}}\}$ at $y = \{\overline{y_0}, \overline{y_1}, \overline{y_2}, \dots, \overline{y_{N-1}}\}$. For one row in a slice, denoted by $R(y_0, t_0) = g(x, y, t) |_{t=t_0, y=y_0}$,

the spatio-temporal point set (one for each kinematic chain in our case) is arranged in a matrix as: $s_{y_0} = (x_1^T, x_2^T, \dots, x_n^T) = (x_1, x_2, \dots, x_n)^T$, where x_i^T are the coordinates of the intersecting points and n is the number of points in row R . Generally speaking, if the human was sufficiently far from the camera and could be regarded as a planar object, there exists a homography matrix H between corresponding points x, x' in the images from two cameras as $x' = Hx$. For two calibrated cameras, H is of the form $H = R + \frac{1}{d}T\hat{n}^T \in R^{3 \times 3}$ [?]. R and T are the rotation matrix and translation vector between the two cameras in the world coordinates system and \hat{n} is the surface norm. In most surveillance applications we can assume R to be I , i.e. there is no rotation between two cameras. The above equation reduces to:

$$H = I + \frac{1}{d}[T_x \ T_y \ T_z]^T [n_1 \ n_2 \ n_3] = \begin{bmatrix} 1 + t_x n_1 & t_x n_2 & t_x n_3 \\ t_y n_1 & 1 + t_y n_2 & t_y n_3 \\ t_z n_1 & t_z n_2 & 1 + t_z n_3 \end{bmatrix} \quad (5.11)$$

where t_x, t_y, t_z are T_x, T_y, T_z multiplied by $1/d$. In our case, we are considering horizontal slices cut from the image volume. We want horizontal lines in one image "mapped" to horizontal lines in the other image. Hence H satisfies $[\gamma, 0, 0]^T = H[1, 0, 0]^T$. This directly gives the constraint as $t_y n_1 = 0; t_z n_1 = 0$. Or equivalently $n_1 = 0$ or $t_y = t_z = 0$. The answers correspond to two camera settings. One is when the object plane normal is in horizontal plane ($n_1 = 0$, i.e. vertical pose). The other is when the translation between the two cameras is along the x direction only (i.e., a stereo rig). Actually, these two cases are common in many surveillance applications. In our case, each row R in one slice corresponds to

one row in an image at height y . There is no geometric equation in a general sense. But given the vertical pose of the gait and the activities we are interested in, we still can derive an epipolar constraint between the DHS patterns. From now on we assume that the following facts hold:

1. A pedestrian is sufficiently far from the cameras so that he/she can be regarded as a planar shape.
2. A pedestrian keeps a vertical pose when walking or performing a class of activities.

Under such conditions, the surface normal n satisfies the constraint above as $n = [0, n_2, n_3]^T$. The points from one row in one image will be mapped to the other row, which makes H in turn change into:

$$H = \begin{bmatrix} a_1 & 0 & a_2 \\ 0 & a_3 & a_4 \\ 0 & 0 & 1 \end{bmatrix} \quad (5.12)$$

If we normalize the two imaged object heights, we obtain the point correspondence for the pair of rows at the same height y_0 and time t_0 . Thus the transform between them is as:

$$R'(t_0, y_0) = H(t_0)R(t_0, y_0), \quad F(t_0) = \begin{bmatrix} a_1(t_0) & 0 & a_2(t_0) \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (5.13)$$

Because of the body translation, $H(t)$ is a time varying transformation. Since the distance between the object and the camera is sufficiently larger than the object

size, H becomes almost constant in a short time period. If we cascade two corresponding slices, we obtain the geometry constraint for two slices $s'(y_0), s(y_0)$ under multiple views as:

$$s'(y_0) = H s(y_0) \quad (5.14)$$

Let the points in the slice set for an activity sequence be represented in a matrix form as $g = [s_{\bar{y}_0} | s_{\bar{y}_1} | \dots | s_{\bar{y}_{N-1}}]$. Each $s_{\bar{y}_0}^T$ contains all the homogeneous point coordinates in the slice at height \bar{y}_0 . For two sequences g, g' captured by two cameras for the same G , we directly obtain the epipolar constraint as:

$$g' = Hg, \quad H = \begin{bmatrix} a_1 & 0 & a_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (5.15)$$

which states that there is a special homography connecting DHS images in different views. Since we already scale the sequence in vertical direction and pair the corresponding rows, H only contains the horizontal translation and scaling. Such a relationship in Eqn. (5.15) leads to several conclusions. First, since 1D scaling and translation does not change the isomorphism in a plane [140], the symmetries are preserved in the projection from the 4D activity volume G to the 3D sequence g . We can directly apply *Theorems 1, and 2* in image sequences. The parameter a_2 becomes zero when the object sequence is aligned, which corresponds to the case studied in [44] and [?]. When the view angle changes from lateral to radial, the

scaling may degenerate the DHS into a line. Second, it provides a solution to match DHS patterns in images under different views. Actually most of the applications aim at analyzing gait and activity at a distance and hence the assumptions are almost always satisfied. However, it should be noticed that the planar assumption is an approximation. The actual human body is not of unit width and occlusion can violate the assumption. Some of the cases are studied in Sec. 5.3.3.

5.3 DHS Extraction

The activity volume results from both intra- and inter- gait motion. Since analysis in the original space or in framed-up sequence can be represented by various global body motion vector v , we no longer differentiate between the two cases. To focus on the extraction method, we first apply a preprocessing step as in [85] to estimate the global trajectory of heads and align the center of every row of DHS to compensate for inter-gait motion.

5.3.1 1D curve approximation

Inspired by the two Theorems in Sec 5.2.4, we try to embed 1-D curves for DHS and stack these curves to generate an approximation to the topology of the activity volume G . Definitions of several nonlinear curves have been evaluated in [?]. We use a definition based on self-consistency, i.e., the DHS curve should coincide at every position with the projected expected value of the pixels from an articulating human body. Verbeek et al proposed a k-segment algorithm [?] as an efficient local

linear model fitting method. However, a disadvantage with their approach is that for self-intersecting data, it often fails to capture the complete structure. An example is shown in Fig. 5.6. We directly applied Verbeek’s algorithm but failed to generate meaningful curves. Unfortunately, self-intersection is inevitable in DHS under most view points.

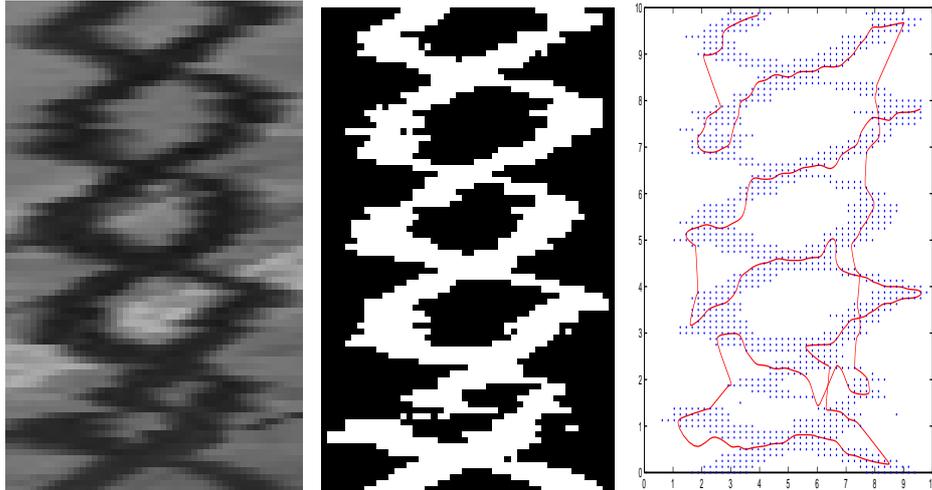


Figure 5.6: Unsuccessful learning of the DHS by directly applying principal curve analysis. Left: Original DHS; Middle: Manually labeled DHS; Right: Fitted principal curves.

To solve this problem, we adopt a divide-and-conquer strategy. We divide a DHS within a stride into four quadrants so that we can separate it into non-self-intersecting curves. Because it belongs to a Frieze Group and has a translational periodicity ($T/2$) and an approximate horizontal reflection as illustrated in Fig. 5.7, *Theorem 1* guarantees the success of such an approach. Each quadrant image contains a single curve capturing the articulation of gait activities in a stride for one limb as in Fig. 5.7.

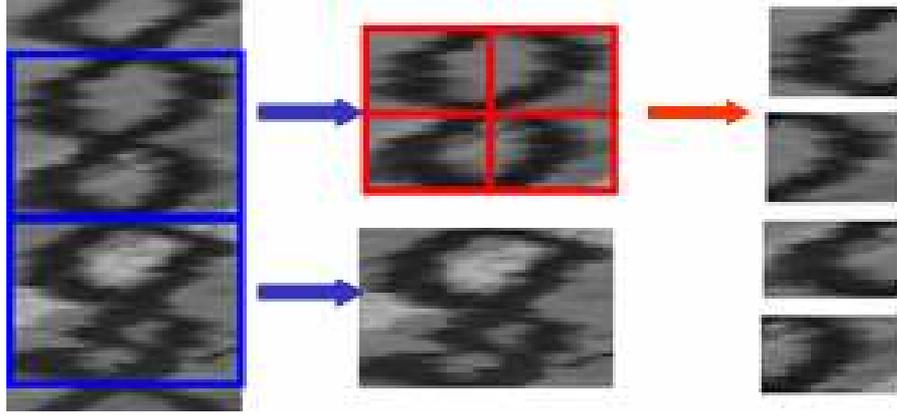


Figure 5.7: Learning DHS by using Frieze Group symmetry. Left: A complete DHS. Middle: Dividing into strides. Right: Dividing into quadrants.

5.3.2 Extraction of the helical pattern

Following *Theorem 1* and [?], the DHS (white pixels in left column of Fig. 5.8) in a quadrant is modeled as

$$p(x) = \int_0^{t_0} p(x|t)p(t)dt + \int_{t_0}^{T/2} q(x|t)q(t)dt, \quad (5.16)$$

where t is the latent variable uniformly distributed over a quadrant and $p(x|t)$ and $q(x|t)$ are the distributions of features at point t . Instead of starting with one line segment corresponding to the most dominant principal component and then increasing the number of segments gradually as in [?], *Theorem 1* suggests the use of only two segments. So if t_0 denotes the time when the DHS approaches the maximum oscillating amplitude, then only one t_0 exists in each quadrant. The segments are defined as: $s_1 = \{s(t)|t \in (0, t_0)\}$, $s_2 = \{s(t)|t \in (t_0, T/2)\}$. After defining a distance metric, the principal curve analysis algorithm will recursively look for the curve that minimizes a cost function. The distance from a point in the

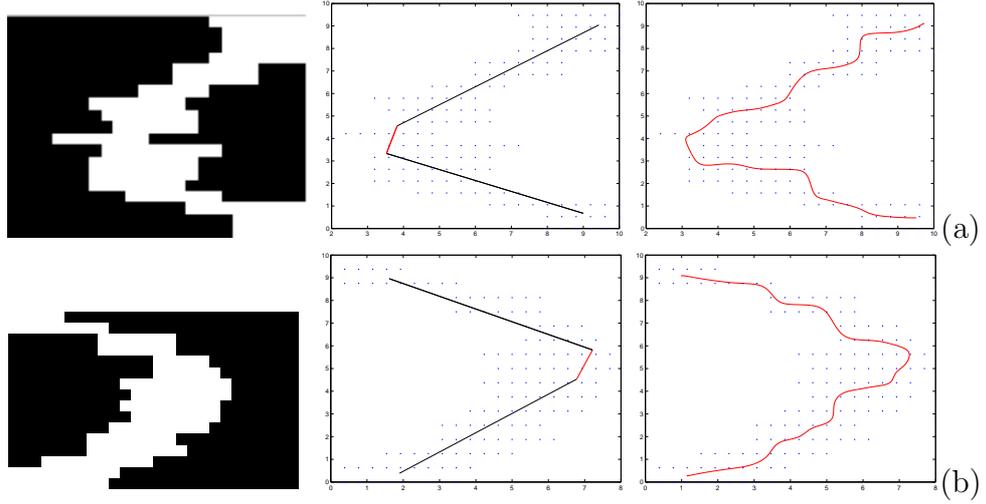


Figure 5.8: Learning DHS in quadrant. Left: helical pattern in quadrants, Middle: line segment approximation; Right: Extracted DHS curves

quarter slice to any of the line segment is defined as:

$$d(x, s_i) = \min_{s_i(t) \in s_i} \|s_i(t) - x\|. \quad (5.17)$$

Our goal is to find two segments s_1 and s_2 such that the overall distance is minimized over all DHS data points:

$$C_{s_1, s_2} = \int_0^{t_0} d(x, s_1)^2 dt + \int_{t_0}^{T/2} d(x, s_2)^2 dt. \quad (5.18)$$

We connect segments into polygonal lines and form a smoothed curve for each limb. The results of learning the structure from quadrants are shown in Fig. 5.8. The fitted curves are shown in red in the right column. We assemble these curves to restore a complete DHS as shown in Fig. 5.9. There is a significant improvement when compared to Fig. 5.6, showing the effectiveness of Theorem 1. Actually the whole process does not require the compensation of global body translation. It can work directly on slices cut from activity volume with global body translation.

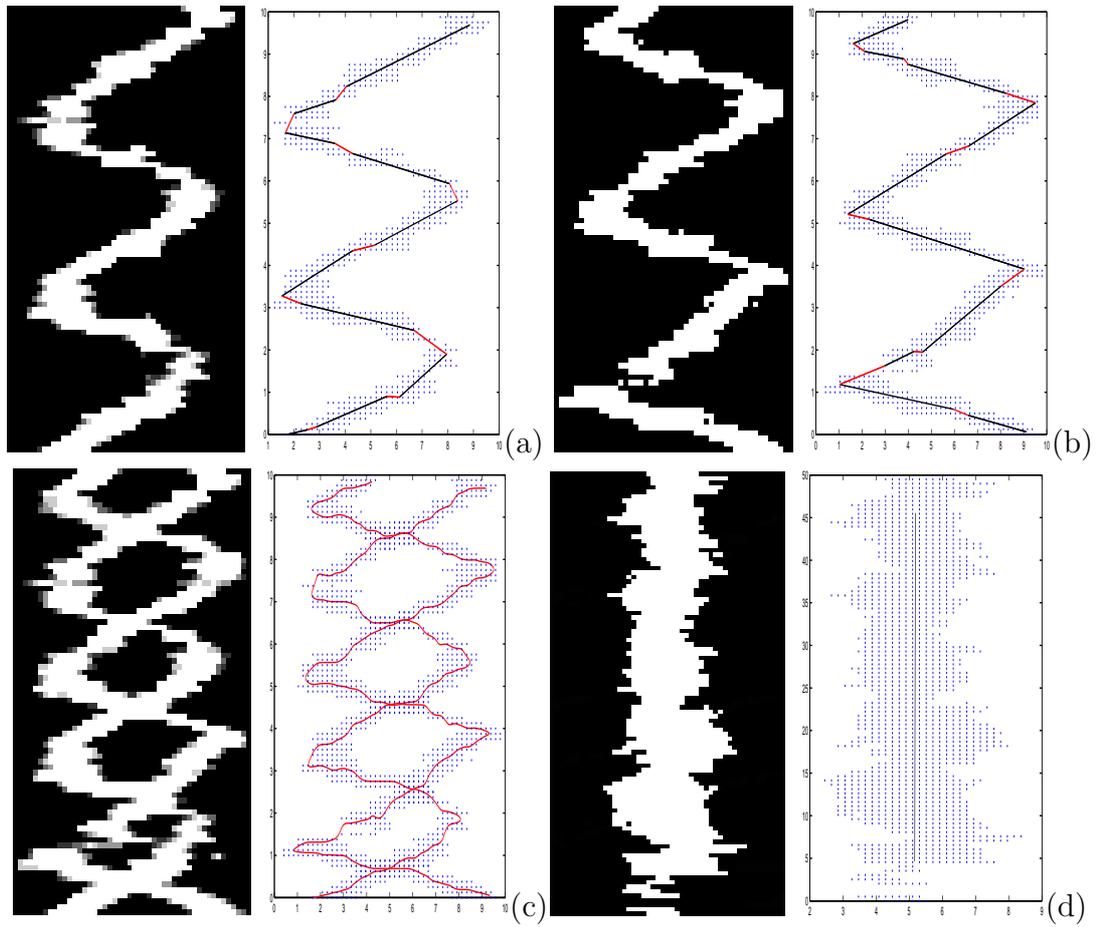


Figure 5.9: Example of extracting DHS. (a),(b) connected curves for two legs in one DHS; (c) superimposed DHS; (d) degenerate DHS for torso.

5.3.3 Degenerate DHS

The extraction algorithm works well for limbs (arms and legs). For torso and head regions we do not observe strong periodicity. We simplify the twisted helix into a single-line and apply it to the upper body slices as in Fig. 5.9(d). It should be noted that the magnitude of swing for the two limbs will decrease from the lateral view to the frontal view and approach the minimum in the latter case. Thus the motion signature will degenerate to a straight ribbon when the target is moving towards/away from the camera.

Chapter 6

Applications of Double Helical Signatures

6.1 Overview

A complete pedestrian monitoring system must be capable of detecting and segmenting humans and their body parts and recognizing events related to carrying a backpack, briefcase etc. Most of the current approaches are based on x-y domain (frames) [39, 40, 56, 57, 62, 146, 147, 148, 153, 152, 25, 88]. In this section, we present a new system based on analyzing DHS in planes parallel to the ground, which is especially useful for applications such as surveillance in parking lot or indoors.

6.2 Pedestrian Segmentation

The goal in this step is to generate the silhouettes of a pedestrian and to label the body parts. To achieve this we need to integrate segmentation and learning. We use DHS to provide a distance constraint based on the fact that pixels close to the curves are more likely to belong to the body. Hence it acts as a 'temporal skeleton' just like the real skeleton for the human body. While the skeleton of a human body is hard to extract, the DHS is easier to obtain as described in Sec. 5.3. Intuitively, describing each pixel by its appearance and closeness drives the background regions away from targets and makes segmentation easier than using appearance alone. The

input to the clustering algorithm is a feature vector that combines the pixel intensity (or color) and the distance measure calibrating the closeness to the DHS. Many major clustering methods work well. We invoke the spectral clustering algorithm [?] and iterate the extraction and clustering steps for each slice. Spectral optimization eventually leads to eigenvectors. At the core of spectral clustering is the Laplacian of the graph adjacency (pairwise similarity) matrix, represented by the similarity in appearance and distance closeness. We use *Theorem 2* to speed up the process: the extracted DHS for one slice is used as the initial condition for the one above it. The algorithm is described as below.

- Input: Slices cut from an activity volume
- Output: Silhouettes with body parts labeled
- Algorithm:
 1. Divide each slice into strides and each stride into four quadrants.
 2. Initialize the DHS and fit principal curves for every quadrant image. Connect and smooth into right and left spirals for the two limbs respectively.
 3. Use the extracted DHS for spectral clustering.
 4. Re-calculate the intersection points from the two spirals and re-partition the slice.
 5. Repeat steps 2-5 until convergence or a pre-specified number of iterations is reached for each slice.

6. Smooth and stack the signatures from each slice and output the mask with labeled body parts.

An interesting issue is the initialization of this iteration. In the current implementation, we use the detected gait period to generate a generic DHS tile made up of two connected lines for each quadrant as in Sec. 5.2 Fig. 5.4. The trajectory serves as the center line for the Frieze gait pattern. The only degrees of freedom are t_0 and the oscillating magnitude x_0 in each quadrant. In our experiment, we vary (t_0, x_0) to generate different DHS and select initial condition as the one with the minimum normalized intensity variance within it. The assumption behind such a method is justified by the fact that the DHS appearance is the temporal repetition of the target at a given height. It is more like a single Gaussian than the appearance distribution in the whole human body.

We use a shape based pedestrian detector reported in [52] to locate the boxes only at key frames (every 2 seconds in our case). The boxes in the intermediate frames are linearly interpolated, assuming that humans move at a constant speed between the key frames.

6.2.1 Simultaneous segmentation and body part labeling

We tested the algorithm on videos acquired from static and moving sensors. The frames are of size $720 * 480$ from interlaced color cameras. Most of the objects we try to segment contain $30 * 60$ to $50 * 100$ pixels (An example of the original frame is shown in Fig. 6.2(c)). Fig. 6.1 shows the result for a moving camera

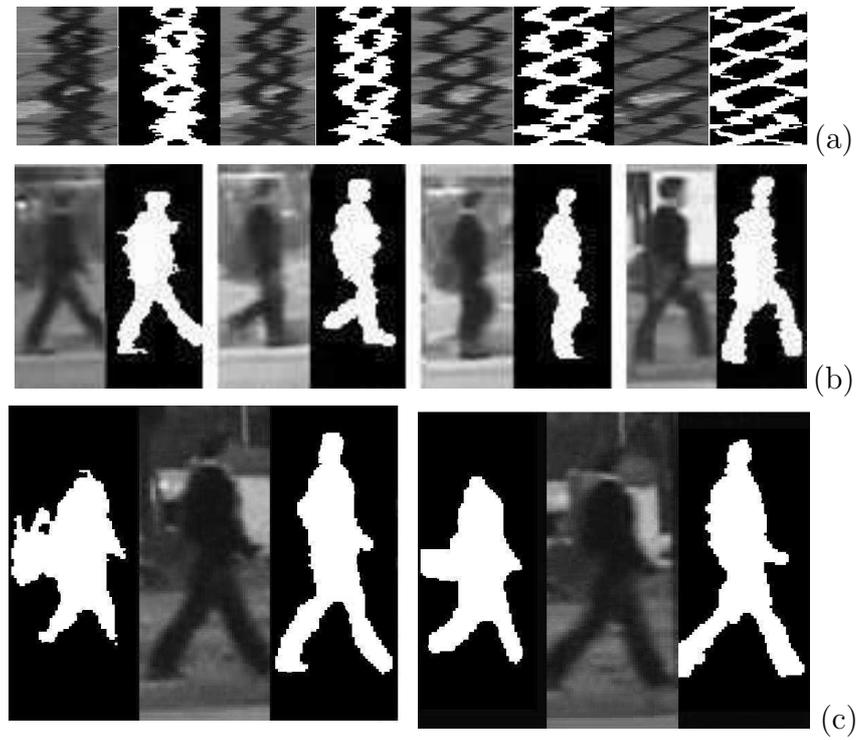


Figure 6.1: Pedestrian segmentation for a video sequence captured by a moving camera. (a) X-t slices for $y = 50, 60, 70, 80$; (b) silhouettes; (c) comparison of segmentation results between not using DHS (left binary images) and using the DHS as a feature (right binary images).

and provides a comparison between the proposed method using DHS+appearance as features and the same method using only the appearance. A problem in moving object segmentation is that similar color in background and foreground will corrupt the silhouettes as shown in the left of Fig. 6.1 (c) (the misclassified region close to the pedestrian's back). This cannot be solved by applying morphological operators. The motion signature in the X-t slice helps to reject the background pixels because they are far away from DHS skeleton.

Next we show the results for the USF data in Fig. 6.2(a), (b) and compare them to a state-of-the-art background subtraction (BGS) methods [106] widely used. The BGS methods are sensitive to object size, shadow, non-rigidity and lighting changes and may contain holes and spurs. To demonstrate the improvement, we enlarge and compare the results in Fig. 6.2(c). Over smoothing makes the BGS method merge the two legs and insufficient smoothing may leave holes in the human body. But our approach correctly segments them due to the distance constraint introduced by DHS. Also, our algorithm is robust to shadows due to the fact that shadows do not exhibit periodicity. In summary, although several pre- and post-processing techniques are available to improve the quality of BGS methods, they share some common disadvantages:

1. Pixel based background subtraction methods do not effectively model spatial and temporal neighborhoods.
2. Morphological operations do not consider the coherence of body articulation.
3. Parameter tuning may result in better performance for a given video but may

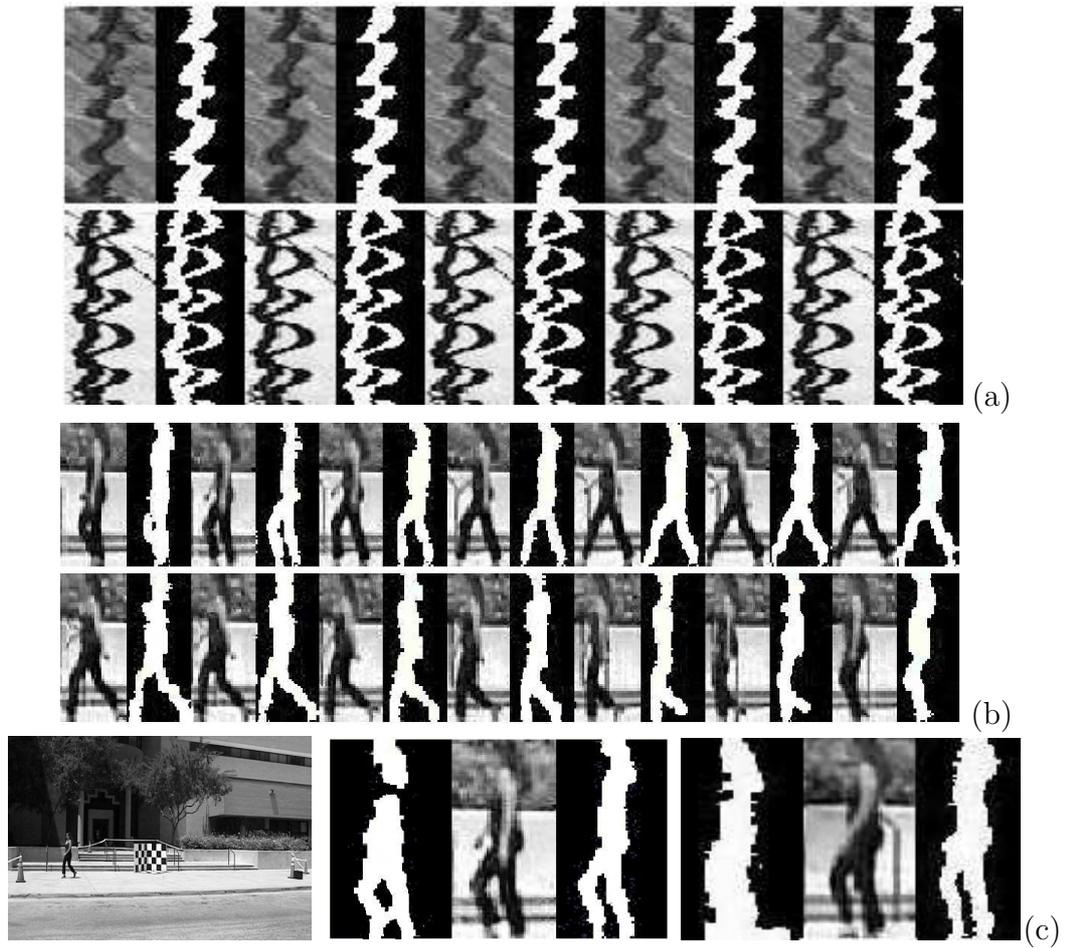


Figure 6.2: USF sequence 03507C0AL segmentation: (a) extracted DHS; (b) silhouettes generated for a complete stride; (c) comparison of segmentation results between background subtraction [106] (left binary images) and the proposed method (right binary images).

fail for others.

For example, a strong post processing step may fill holes but over-smooth the contours, while a weak one may leave too many spurious edge or holes along and inside the body. The main improvement is due to the accurate capture of the limb positions which are crucial for gait and activity analysis. Holes or spurious areas are much less likely to appear in our segmentation result due to using DHS as a skeleton. To have a quantitative comparison, we use the ground truth silhouettes set to compare the accuracy (% of misclassified pixels) of the two methods for a subset of USF data containing 10 subjects in Fig 6.3. Our method generates consistently better results.

Furthermore, we present the accuracy of the part labeling to the ground truth provided in [106] as shown in Fig. 6.4. An advantage of our method is due to exploiting the spatio-temporal coherence of human motion. But we can infer the part location when they are completely occluded by body such as the right arm of the target.

6.2.2 Robustness analysis

In the algorithm pipeline, each step may be sensitive to noise. For example, unreliable period detection or asymmetry between left and right legs will yield inaccurate quadrant partitions. Since we re-partition between each iteration from the previous step, the extracted DHS will eventually converge to the true location in our experiment. In our experiment, we found that 2-3 iterations are sufficient.

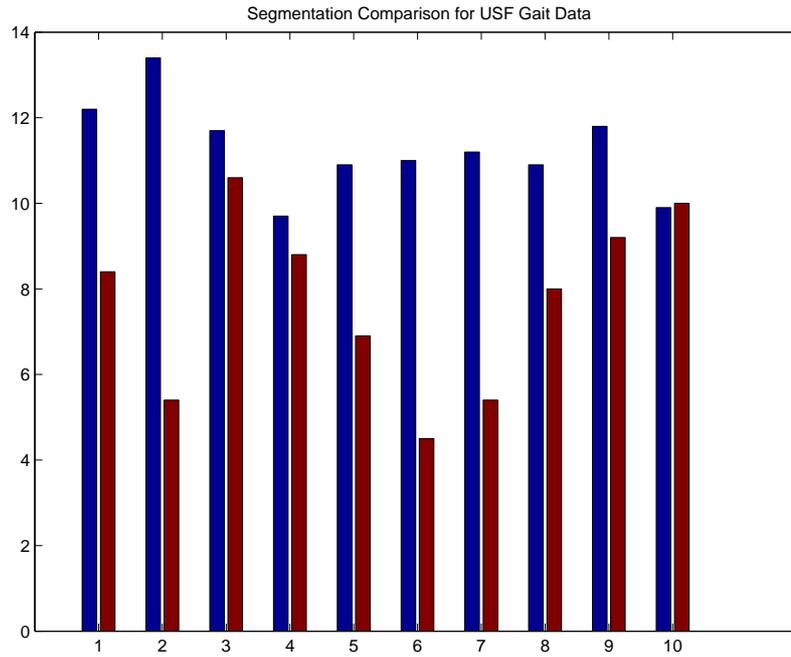


Figure 6.3: Comparison of segmentation accuracy between our method (red) and background subtraction (blue) for selected USF Sequences. (02463G2AR, 02539G1AR, 03500G0AR, 03507C0AL, 03509C0AL, 03516G0AR, 03521G0AR, 03526G0AR, 03529G0AR, 03532G0AR)

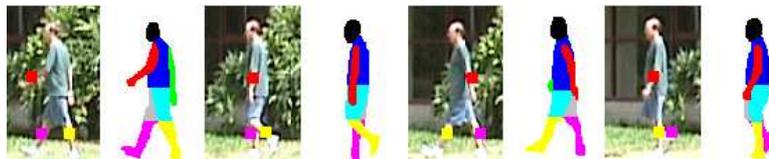


Figure 6.4: Comparison of body parts labeling accuracy for USF Data. The color boxes on the body parts show the labeling results at the slice at various heights. We successfully label legs and one arm. The other arm is occluded during the walking and cannot be labeled.

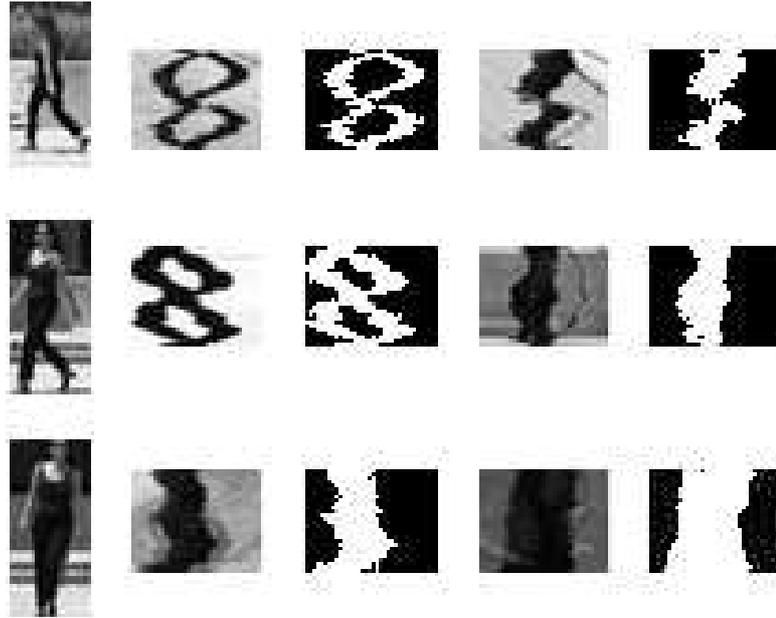


Figure 6.5: Robustness to view points. Each row contains original and segmented DHS for torso and limbs under different views for the USF sequence 03507C0AR.

In order to demonstrate robustness, we first present the results of evaluation at different viewing angles (or equivalently body movement directions) in Fig. 6.5. View dependency is of importance for human motion [114, 102, 103, 113]. The viewing angle varies from 0 (lateral) to $\pi/2$ (radial) and the proposed method automatically handles the changes in DHS magnitude by taking the degenerate case into consideration and gives satisfactory results.

Secondly, we present results for different sizes and frame rate by applying various down-sampling ratios in Figure. 6.6. The original object size is greater than 3200 pixels (based on a bounding box of 80×120) in a 30 frames per second (fps) video, which corresponding to a DHS size of 80×80 in one stride. The object size and sampling rate is reduced by 2 (40×60 , 15 fps) and by 4 (20×30 , 8 fps) respectively.

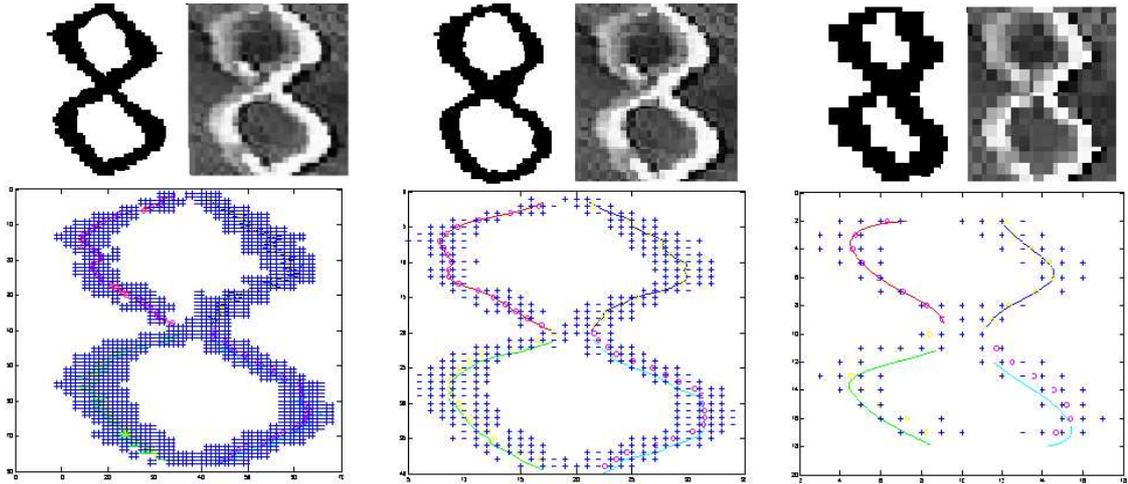


Figure 6.6: Robustness to size. Each group contains original and segmented image with extracted DHS under different sizes and video rates for the USF sequence 02463C0AR.

We obtain consistently correct segmentation and labeling results for the legs and torso even when the target size is reduced to 20×30 at a frame rate of 8fps. However, during down-sampling, the labeling for upper limbs is successful only when the size is above 40×60 because the arms can be hardly separated from the torso at lower resolution. The results demonstrate that the segmentation performs well for objects at close and middle range (less than 50 meters in our experiments) and various video frame rates.

6.3 Severe Occlusion Handling

In most segmentation/tracking algorithms the trackers are unable to accurately label walkers and locate their body parts during severe occlusion. The invisible parts make it hard to extract the body contour. An example is given in Fig. 6.7

where two targets walk towards each other and one completely occludes the other in some frames.

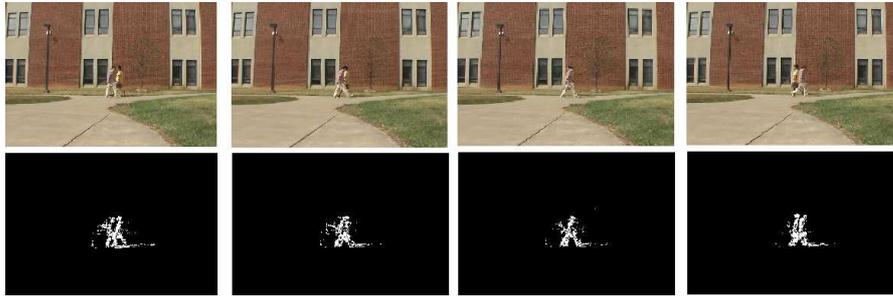


Figure 6.7: Severe occlusion for pedestrians. First row: original sequences; Second row: extracted mask.

The DHS provides a promising approach to handle occlusion. The intuition behind our approach is that helical pattern should be consistent in time because of periodicity and symmetry. To classify the foreground pixels in each $x-t$ slice to the walker they belong to during occlusion, we compare them with the different DHS using a Bayesian classifier and assign them to the most similar walker. The details of the algorithm is given below:

1. Obtain the trajectories for each individual $1, 2, \dots, n$ by detecting lines of heads in slices. Notice that if a pedestrian changes his walking direction, his trajectory will be split.
2. Occlusion $O_i(j_1, j_2, \dots)$ $i = 1, 2, \dots, m$ are detected by finding the intersections of any two trajectories. i is the occlusion segment index and j_1, j_2, \dots are the individuals' indices.
3. Extract DHS for each object using the method proposed in Sec. 5.3.

4. For each occlusion O_i , repeat the following steps:
5. Divide the x-y-t data for occlusion O_i into x-t slices
6. For each individual, translate its extracted DHS into the occlusion slice
7. For each pixel in a slice, test its probabilities in all the corresponding objects' DHS distribution and assign it to the one giving the highest probability.

The occluded pixels can be restored almost perfectly by filling with the un-occluded DHS. An example is given in Fig. 6.8, where our algorithm precisely captures the contour with one walker completely behind the other. A limitation of our algorithm is that the prediction fails when the targets change their gait during occlusion. This could be handled by allowing the DHS to deform during occlusion.

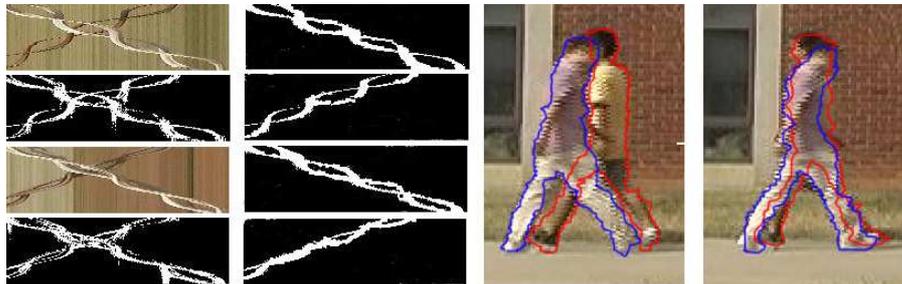


Figure 6.8: Segmentation under severe occlusion. First column: the original and extracted foreground at two different heights; Second column: separated individual DHS; Right two images: two frames of restored silhouettes.



Figure 6.9: More results for segmentation under severe occlusion. Notice that even the two pedestrians in the right are wearing same color pants, the proposed method still successfully segments them from occlusion by using DHS.

6.4 Matching

6.4.1 Across cameras

Because of the equivalency of viewing angle change and pose change in generating image sequences, we study them in a unified framework. As in Eqn. (5.14), the normalized $x-t$ DHSs in various views (poses) are related to a specific transformation. If we can estimate such an H and measure the quality of restored geometry, we obtain a similarity measure between two activity sequences. There are quite a few methods available for solving Eqn. (5.14) given point coordinates. We take the method reported in [?] to estimate H by using a set of linear equations. After calculating the homography H , we obtain the error as the overall average point wise difference for a set of DHS slices $s = \{\bar{s}_0, \bar{s}_1 \dots \bar{s}_{N-1}\}$ at $y = \{\bar{y}_0, \bar{y}_1 \dots \bar{y}_{N-1}\}$ for a pair of gait sequences g and g' assuming that they are from the same volume G :

Table 6.1: USF data: DHS Matching across cameras.

| Error% | 02463G2AR | 02539G1AR | 03500G0AR | 03507C0AR | 03509C0AR |
|-----------|------------|------------|------------|------------|------------|
| 02463G2AL | 1.3 | 5.1 | 7.2 | 4.0 | 5.0 |
| 02539G1AL | 3.7 | 1.0 | 5.2 | 7.5 | 5.0 |
| 03500G0AL | 7.2 | 2.7 | 1.1 | 10.0 | 5.0 |
| 03507C0AL | 4.1 | 5.3 | 7.9 | 1.4 | 5.4 |
| 03509C0AL | 5.5 | 7.1 | 8.4 | 5.0 | 1.5 |

$$Error(g, g') = \frac{1}{\sum N_i} \sum_{s_i \in s} \sum_{p \in s_i} |x_p - \hat{F}x'_p|^2 \quad (6.1)$$

where \hat{F} is the estimated transformation and N_i is the number of points in slice $S_{\bar{y}_i}$. To validate the proposed method, we present the results for matching targets from a subset of USF data (02463G2AR, 02539G1AR, 03500G0AR, 03507C0AR, 03509C0AR) and (02463G2AL, 02539G1AL, 03500G0AL, 03507C0AL, 03509C0AL) captured by two cameras. The confusion matrix (not symmetrical) is given in Table 6.1. In this experiment we divide each sequence into clips containing DHS for one stride. Each entry in the table presents the average *Error*. It shows that the proposed method consistently matches the DHS across cameras and does not confuse with other subjects.

6.4.2 Across time

When the two clips to be matched are from similar viewing directions but have different speeds, the DHS from the same individual will have different length. This

may due to videos captured at various time instants, under different environmental conditions etc. The matching above considers translation and scale change both in spatial and time, which works for DHS from multiple views for the same individual at the same time. However, if the two DHS are captured at different times, the matching may become nonlinear. Eqn. 5.14 is not enough to incorporate the subtle change in gait.

Temporal alignment deals with matching two sequences with different lengths and plays a key role in human identification by gait. Among many existing methods, Dynamic Time Warping (DTW) is the most used technique for non-linear alignment. It uses dynamic programming to find the optimal alignment of two sequences of different lengths. But a major challenge is accurately determining the location of start and end points in samples. In our case the complete DHS are partitioned into strides, which makes the DTW [?] an effective tool. DTW yields an accumulated distance matrix (AD matrix) representing the best possible match between the input pattern and the template. The DHS pattern giving the lowest accumulated distance is the best match for the input pattern.

A challenge in DTW for temporal multi-dimensional signal warping is the cost involved to compute the minimum-cost assignment path, which also holds for multi dimensional activity space. But since we decompose g into slices and each DHS contains two spirals for left and right limbs, we can view them as two 1D temporal signals $s(t)_l, s(t)_r$. 1D DTW is a standard and efficient process given the end points. To compare two DHS at the same height from two clips, the cost function is given as:

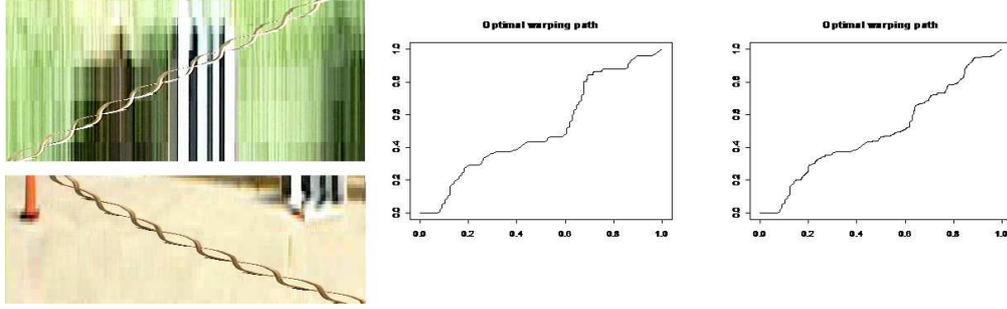


Figure 6.10: DHS matching using DTW for gait activities with different lengths.

$$D(s, s') = \min\{d(s(t)_l, s'(t)_l) + d(s(t)_r, s'(t)_r), d(s(t)_l, s'(t)_r) + d(s(t)_r, s'(t)_l)\} \quad (6.2)$$

where $d(s(t), s'(t))$ is the DTW cost function for two single spirals and $D(s, s')$ is the overall cost for a DHS containing multiple spirals. Each DHS contains two spirals for left and right limbs. We compare the two possible permutations and choose the one with minimum error. When a set of DHS is used to represent an activity g , we average the cost for each and write the cost function as:

$$D(g, g') = \frac{1}{N} \sum_{i=1, s_i \in g, s'_i \in g'}^N D(s_i, s'_i) \quad (6.3)$$

Comparing two sequences by measuring similarity in DHS is a major advantage in our framework. The shape sequence is represented by a compact set of 1D temporal signal pairs. An example of the proposed method is given as Fig 6.10. In our experiment, only 3 DHS are used. They are at heights (0.4, 0.6, 0.8) in the bounding boxes. The confusion matrix (symmetrical) are given in Table. 6.2. This experiment suggests the potential of using DHS for human identification.

Table 6.2: USF data: DHS Matching across Time.

| $D(g, g')$ | 02463G2AR | 02539G1AR | 03508G0BR | 03521G0AR | 03603G0AR |
|------------|------------|------------|------------|------------|------------|
| 02463G0AR | 2.1 | 6.2 | 6.3 | 5.7 | 7.0 |
| 02539G0AR | 6.2 | 1.7 | 5.7 | 6.7 | 5.0 |
| 03508G1AR | 6.3 | 5.7 | 3.1 | 8.0 | 6.7 |
| 03521C0AR | 5.7 | 6.7 | 8.0 | 2.3 | 6.4 |
| 03603C0AR | 7.0 | 5.0 | 6.7 | 6.4 | 1.9 |

6.5 Load Carrying Event Detection

In activity analysis, gait abnormalities are typically detected by measuring changes from silhouettes or landmark trajectories [16, 53, 57, 55]. Methods in these domains will be easily affected by segmentation and tracking errors due to non-rigidity and occlusion. In many cases, it is difficult to separate the person and the object he/she is carrying using shape alone or by tracking body points. Variations in viewing directions make it even harder. In this section, we show how the symmetries in a small set of DHS can be used for detecting load carrying events. Our approach does not require segmentation or landmark tracking. It is effective even for moving platforms, where other approaches depending on silhouettes do not work that well.

In the proposed model in Fig 5.5 in Sec. 5.2.3, the motion of the limbs is represented as a pair of kinematic chains oscillating out of phase. The hands would like to maintain the center of gravity above the point of contact and minimize the energy to balance the body during bi-pedal leg swing. We expect that the presence of a sufficiently heavy object will (at least in hand regions) distort the DHS pattern.

| Activity | Sequences | Hands slices | Periodicity | Symmetry |
|----------------------|---|---|-------------|---------------------------|
| walking |  |  | T/2 | Horizontal & Vertical |
| Load on Shoulder |  |  | T | 180 Rotational & Vertical |
| Load in Hand |  |  | T | Vertical |
| Holding load in Arms |  |  | N/A | N/A |

Figure 6.11: Comparison of DHS in the hand regions for different activities.

Theorem 2 enables us to look at only one slice to understand the arm articulation. *Theorem 1* enables us to use the presence and absence of Frieze Group symmetries to classify events. We list three activities: natural walking, carrying an object with one hand, holding an object in hands and examine the different symmetries in the DHS in Fig. 6.12 associated with activities. For example, vertical symmetry exists for all events but only natural walking has horizontal symmetry. In summary, one side of the signature disappears when one carries objects in one hand and the whole DHS disappears when holding objects in arms.

To prove the presence and change of symmetries due to different object carrying activities, we invoke the kinematic chain model in Section 5.2.3. Each limb is described by a kinematic chain L_1, \dots, L_k and corresponding articulation T_0, T_1, \dots, T_k . Carrying an object will distort the symmetrical gait and hence change the articulations of limbs. Assume that a human is carrying a briefcase in one of his/her hands,

the body movement will adapt the gait to such load change and minimize the energy consumption. The periodicity remains but we no longer observe the same symmetry as in natural walking:

$$\theta_l(t) = \theta_l(t + T), \quad \theta_r(t) = \theta_r(t + T) \quad \theta_l(t) \neq \theta_r(t - T/2) \quad (6.4)$$

Such a change results in a new translational vector T in DHS. Similarly when the target is holding an object in both arms, the two halves of the DHS degenerate into a line since the arms do not oscillate anymore. By comparing the translational vector as well as symmetries in hand DHS and leg DHS, we can detect those activities.

We first apply 1D autocorrelation along t axis for each slice with sufficiently long length (2 seconds) after compensating for global motion. Local peaks are extracted within a local neighborhood (5*5 in our case) representing the lattices. The distances between adjacent peaks in a DHS is the vector for translational symmetry. If the two vectors for hands and legs are the same, we reject it from further analysis and classify it as hand-free. If we do not observe strong horizontal symmetry in the hand DHS, we label the event as 'holding in two hands'.

We then transform the original DHS slice S by various symmetry rules (such as glide reflection or half turn) along the trajectory to obtain a transformed slice S' . Then we compute the average residual error between S and S' . If the error is less than a preset threshold, the specific symmetry is said to exist and the event is labeled as 'holding in one hand'. In our experiment we considered the slices around 1/4 of the body height from the head as hand DHS and around 3/4 as leg DHS. We have tested on 30 sequences collected by cameras attached to a building and 63 ground-

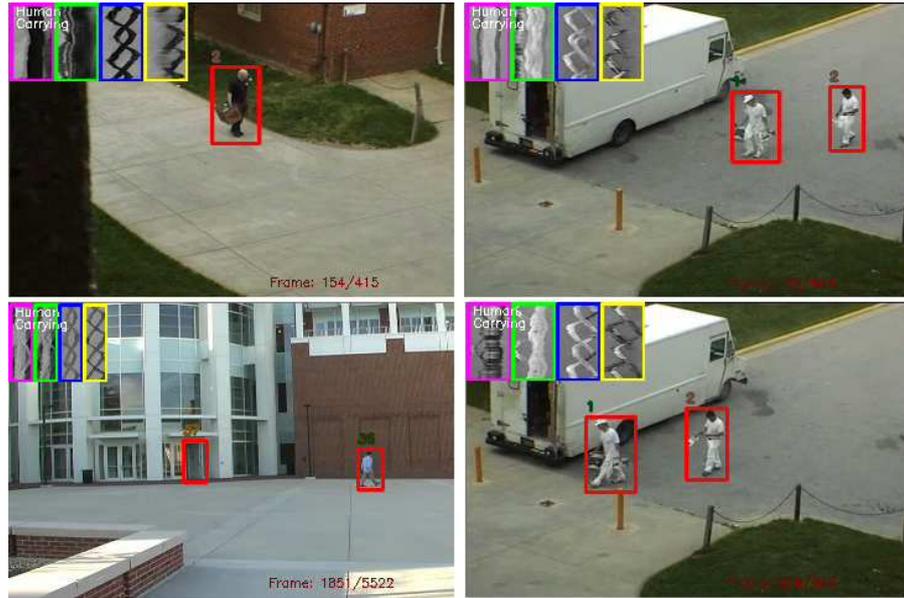


Figure 6.12: Examples of different activities: four slices (0.25,0.35,0.75,0.85 of object height) are chosen for analysis and are illustrated in the left corner with activity name superimposed. The two leftmost insets show DHS of elbow and wrist; the two rightmost insets show DHS of the knee and ankle.

based sensor sequences recorded for various activities, distances, viewing angles, backgrounds and lightings. All the video segments were collected from azimuth angles from $\pi/2$ to $\pi/6$ and elevation angles from 0 to $\pi/3$. The preliminary results are shown in Fig. 6.12.

Table 6.3 presents detection rates (the percentage of observed events that were correctly classified) and false alarm rates (the percentage of events that were incorrectly classified). We list 3 events: *None* for natural walking, *One hand* for carrying objects, *Arms* for holding in arms. The average detection rate is 88.7% and the false alarm rate is at 11.7%. Most of the false alarms are due to the self-occlusion of the human body parts, which suggests using multiple cameras in future

Table 6.3: Outdoor Event Classification. First row: 3 categories; Second row: total number for each category; Each cell: the number classified as the index in the left.

| | None | One hand | Arms | False alarm |
|---------------|-------|----------|-------|-------------|
| Total | 40 | 31 | 22 | N/A |
| None | 37 | 1 | 2 | 7.5% |
| One hand | 2 | 27 | 1 | 10.0% |
| Arms | 1 | 3 | 19 | 17.4% |
| Recognition % | 92.5% | 87.1% | 86.4% | |

work. The proposed method could be directly applied with line scan based LASER or range sensors without capturing the whole scene. By doing so, computational power and resource is dramatically reduced. The limitation of this method is when the target is walking towards the camera. The DHS degenerated into a ribbon and no strong symmetry is observed. To attack such challenges, shape cue may be integrated into our system.

Furthermore, we examine the activities of leaving or picking objects by checking the symmetry change and the results are shown in Fig. 6.13. It demonstrate the power of DHS to differentiate the symmetry change due to carrying objects, which is of importance in the security surveillance.

6.6 Summary and Discussion

We presented a method for understanding the gait and activity volume using DHS in layered slices. The proposed method naturally integrates temporal body

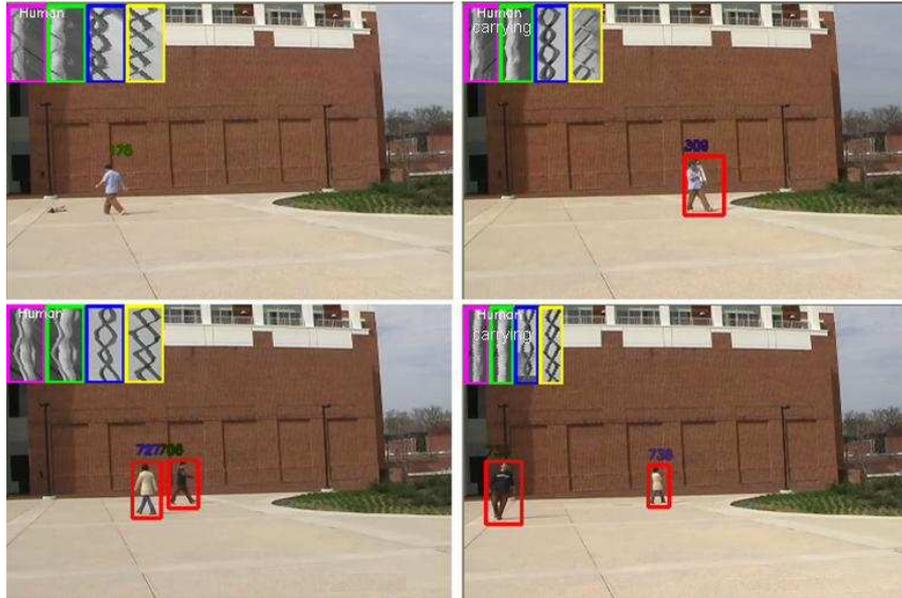


Figure 6.13: Examples of recognizing activities as leaving and picking up objects. Each row shows the DHS change before and after picking an object.

dynamics with 2D shape information. It does not require silhouettes and feature tracking. Our approach has two major features. First, the twisted pattern belongs to a Frieze Group, enabling separation of self-intersecting curves for robust and efficient learning. Second, only a finite set of DHS is needed for compact and sufficient representation for activity volume topology estimation and articulation parameters estimation such as cadence, step/stride length and style.

Moreover, we have implemented a pedestrian monitoring system capable of simultaneously segmenting and labeling body parts, matching across various cameras and time as well as recognizing load carrying events. The experimental results demonstrate the effectiveness and robustness under lighting changes, shadows, camera motion, various viewing angles as well as severe occlusions. The approach is robust against several key factors such as body movement direction, viewing angle and

target size. The work indicates that considering human motion in spatio-temporal domain is an efficient method to analyze gait and activities.

Chapter 7

Conclusions

7.1 Summary

This dissertation presented several approaches for human motion modeling and analysis with specific focus on the applications in surveillance. Our general approach can be understood in terms of three aspects. The first aspect is to model gait by periodic signals. The second aspect is to use kinematic and geometric constraints to characterize the articulation within a period. The third aspect is to present a compact representation using video sequences to analysis human motion in space and time simultaneously. In summary, here are some of the key contributions made in the thesis:

1. In Part I, we proposed a new algorithms to extract the periodic motion pattern and used it as a cue for pedestrian classification. It is very compact yet efficient encoding of the gait. In the literature, usually periodicity based detection methods are designed using shape or contour. This assumption might be hard to satisfy in moving platform or when the pixels on target is small. By using the cascaded hypothesis testing and symmetry constraints, we can achieve a robust algorithm that recovers the periodicity under more general and difficult setting, i.e., when the platform is moving under different illuminations. In particular, this algorithm can handle IR as well as visible images with object

at a distance and yields a robust performance.

2. The MPGA approach in Chapter 2 is image-based and does not require explicit 3D body model. It models the bipedal swing of limbs and is computationally efficient and is also able to deal with images of small size. In contrast, traditional 3D model-based approaches are computationally intense and need more pixels on the target. By employing a PLL module, it could track the gait rate continuously.
3. Periodic analysis of human gait in literature requires long sequences to provide sufficient temporal information. In Part II Chapter 3 we used a kinematic model which characterizes the high order statistical deformations of a human body. The bipedal movement suggests a strong attribute: the X Junction in space and time to act as an effective feature. This specific feature enables the analysis human gait within a short time of period.
4. The analysis of redundancy in gait signatures from different heights is presented in Part III, Chapter 5. The proposed method naturally integrates temporal body kinematics with 2D shape information. It does not require silhouettes and feature tracking. Our approach has two major features. First, the twisted pattern belongs to a Frieze Group, enabling separation of self-intersecting curves for robust and efficient learning. Second, only a finite set of DHS is needed for compact and sufficient representation for activity volume topology and estimation of articulation parameters such as cadence, step/stride length and style.

5. In Chapter 6, we presented a pedestrian monitoring system capable of simultaneously segmenting and labeling body parts, matching across various cameras and time as well as recognizing load carrying events. The experimental results demonstrate the effectiveness and robustness under lighting changes, shadows, camera motion, various viewing angles as well as severe occlusions. The approach is robust against several key factors such as body movement direction, viewing angle and target size. The work indicates that considering human motion in spatio-temporal domain is an efficient method to analyze gait and activities.
6. We presented, in Chapter 7, a summary of how we model periodic human motion and articulation in space and time in video sequences. This framework provides a complete description of activities related to human waling motion. Various current schemes are just instances of this generic framework.

7.2 Future Directions

Human motion analysis and recognition can be expanded in many ways. The following just lists some potential avenues to explore in the context of the proposed approaches.

The approaches taken in this dissertation by no means cover the whole spectrum of the unconstrained human motion analysis problem and address only a small portion of all available issues. Existing todays video surveillance systems while providing the basic functionality, fall short of providing the level of information need to

change the security paradigm from investigation to preemption. Automatic visual analysis technologies can move today's video surveillance systems from the investigative to preventive paradigm. Smart Surveillance Systems provide a number of advantages over traditional video surveillance systems, including:

1. the ability to preempt incidents – through real time alarms for suspicious behaviors
2. enhanced forensic capabilities – through content based video retrieval
3. situational awareness – through joint awareness of location, identity and activity of objects in the monitored space

There are still a number of technical challenges that need to be addressed. These include challenges in robust object detection, tracking objects in crowded environments, challenges in tracking articulated bodies for activity understanding, combining biometric technologies like face recognition with surveillance to achieve situational awareness. In addition, performance characterization of surveillance systems is very challenging and requires significant amounts of annotated data. Typically annotation is a very expensive and tedious process. Additionally, there can be significant errors in annotation. All of these issues make performance evaluation a significant challenge.

BIBLIOGRAPHY

- [1] E.H. Adelson and J.R. Bergen, "Spatiotemporal Energy Models for the Perception of Motion," *Journal Optical Society of America A*, Vol. 2(2), 1985.
- [2] A. Agarwal and B. Triggs. 3D Human Pose from Silhouettes by Relevance Vector Regression. In Proc.IEEE Computer Vision and Pattern Recognition, 2004.
- [3] A. Agarwal and B. Triggs. Tracking Articulated Motion with Piecewise Learned Dynamic Models. In Proc.European.Conf.of Computer Vision, LNCS, Springer-Verlag, 2004.
- [4] A. Agarwal and B. Triggs. Recovery of 3D Human Pose from Monocular Images . IEEE Trans. Pattern Analysis and Machine Intelligence, 28(1), 2006.
- [5] J.K. Aggarwal and Q. Cai, "Human Motion Analysis: a Review," *Computer Vision and Image Understanding*, Vol. 73(3), pp. 428-440, 1999.
- [6] J.K. Aggarwal, Q. Cai, W. Liao, and B. Sabata. Articulated and Elastic Non-Rigid Motion: A Review. In Workshop on Motion of Non-Rigid and Articulated Objects, pages 2C14, Austin, Texas, USA, 1994.
- [7] J. K. Aggarwal, L. S. Davis, and W. N. Martin, "Correspondence Process in Dynamic Scene Analysis", Proc. of the IEEE, 69(5):562-572, 1981.
- [8] J. K. Aggarwal and N. Nandhakumar, "On the Computation of Motion of Eequences of Images - a Review", Proc. of the IEEE, 76(8):917-934,1988.
- [9] K. Akita, "Image Sequence Analysis of Real World Human Motion," *Patten Recognition*, 17(1):73-83, 1984.
- [10] M. C. Allmen, "Image Sequence Description using Spatiotemporal Flow Curves: Toward Motion-Based Recognition", Ph.D. Dissertation, Computer Sciences Department Technical Report 1040, University of Wisconsin - Madison, August 1991.
- [11] Allmen, M., Dyer, C.R., Computing Relations for Dynamic Perceptual Organization, Computer Vision, Graphics and Image Processing: Image Understanding 58, 338-351, 1993
- [12] Allmen, M., Dyer, C.R., "Long-range Spatiotemporal Motion Understanding using Spatiotemporal Flow Curves", in *Proceeding of IEEE International Conference of Computer Vision and Pattern Recognition*, 303-309, 1991

- [13] Allmen, M., Dyer, C.R., "Computing Spatiotemporal Surface Flow", *in Proceedings of IEEE International Conference on Computer Vision*, 3, 47-50, 1990
- [14] A. Azarbayejani and A. Pentland, "Real-time Self-calibrating Stereo Person Tracking using 3-d Shape Estimation from Blob Features," *In Proc. of Intl. Conf. on Pattern Recognition*, pages 627- 632, Vienna, Austria, August 1996.
- [15] J. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques", *International Journal of Computer Vision*, 12(1):42C77, 1994.
- [16] C. BenAbdelkader, R. Cutler, and L. Davis, "Motion-based Recognition of People in EigenGait Space", *In International Conference on Automatic Face and Gesture Recognition*, Washington D.C., USA, May 20-21 2002.
- [17] S. Belongie, J. Malik, and J. Puzicha. Matching Shapes. In *International Conference on Computer Vision*, Vancouver, Canada, July 9-12 2001.
- [18] A. Blanchard, *Phase-Locked Loops*. New York, NY: John Wiley and Sons 1976
- [19] Baker, H.H., Bolles, R.C, "Generalizing epipolar-plane image analysis on the spatiotemporalsurface," *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp 2-9, 1988.
- [20] R. Bolles, H. Baker, and D. Marimont. Epipolar-plane Image Analysis: An Approach to Determining Structure from Motion, *International Journal of Computer Vision*, 1(1):7-56, 1987.
- [21] J.E. Boyd, "Synchronization of Oscillations for Machine Perception of Gaits," *Computer Vision and Image Understanding*, Vol. 96(1), pp. 35-59, Oct 2004.
- [22] A. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. *Philosophical Trans. Royal Soc. London*, 352:1257C1265, 1997.
- [23] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(3):257C267, 2001.
- [24] O. Boiman and M. Irani, "Detecting irregularities in images and in video", *In Proc. Int. Conf. on Computer Vision*, (ICCV05), 2005.

- [25] K. Bradshaw, I. Reid, and D. Murray, "The active recovery of 3d motion trajectories and their use in prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(3):219C234, March 1997.
- [26] C. Bregler and J. Malik. "Tracking People with Twists and Exponential Maps," *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp 8-16, 1998.
- [27] C. Bregler, J. Malik, and K. Pullen. Twist based acquisition and tracking of animal and human kinematics. *International Journal of Computer Vision*, 56(3), 2004.
- [28] A. Broggi, M. Bertozzi, A. Fascioli, M. Sechi, "Shape-based Pedestrian Detection," *Proc. IEEE Intell. Veh. Symp.*, pp. 215-220, 2000.
- [29] Q. Cai and J. K. Aggarwal, "Tracking human Motion using Multiple Cameras", *In Proc. of Intl. Conf. on Pattern Recognition*, pages 68-72, Vienna, Austria, August 1996.
- [30] Q. Cai, A. Mitiche, and J. K. Aggarwal, "Tracking human motion in an indoor environment", *In Proc. of 2nd Intl. Conf. on Image Processing*, volume 1, pages 215-218, Washington, D.C., October 1995.
- [31] L. W. Campbell and A. F. Bobick, "Recognition of Human Body Motion using Phase Space Constraints", *In Proc. of 5th Intl. Conf. on Computer Vision*, pages 624-630, 1995.
- [32] Z. Chen and H. J. Lee, "Knowledgeguided Visual Perception of 3D Human Gait from a Single Image Sequence", *IEEE Trans. on Systems, Man, and Cybernetics*, 22(2):336-342, 1992.
- [33] M. Chen, G. Ma, and S. Kee, "Pixels Classification for Moving Object Extraction", *In IEEE Workshop on Motion and Video Computing (MOTION05)*, Breckenridge, Colorado, Jan 2005.
- [34] F. Cheng, W.J. Christmas, and J. Kittler, "Recognising Human Running Behaviour in Sports Video Sequences", *In International Conference on Pattern Recognition*, Quebec, Canada, August 11-15 2002.
- [35] G. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette for Articulated Objects and its use for Human Body Kinematics Estimation and Motion Capture", *International Journal of Computer Vision*, 2003.

- [36] K. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette across Time Part ii: Applications to Human Modeling and Markerless Motion Tracking. *International Journal of Computer Vision*, 63(3), 2005.
- [37] S.S. Cheung and C. Kamath. Robust Techniques for Background Subtraction in Urban Traffic Video. *In Visual Communications and Image Processing. Proceedings of SPIE*, volume 5308, 2004.
- [38] K. Choo and D.J. Fleet. People Tracking Using Hybrid Monte Carlo Filtering. *In International Conference on Computer Vision*, Vancouver, Canada, July 9-12 2001.
- [39] A. Roy Chowdhury and R. Chellappa. A factorization approach for event recognition. In CVPR Event Mining Workshop, 2003.
- [40] A.R. Chowdhury and R. Chellappa. A factorization approach for activity recognition. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Washington DC, June, 2004.
- [41] R.T. Collins, A.J. Lipton, and T. Kanade, "Introduction to the Special Section on Video Surveillance," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22(8), pp. 745-746, 2000.
- [42] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L. Wixson, A system for Video Surveillance and Monitoring, Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep., CMU-RI-TR-00-12, 2000.
- [43] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-Based Object Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5), 2003.
- [44] R. Cutler, L.S. Davis, "Robust Real-time Periodic Motion Detection, Analysis, and Applications," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22(8), pp. 781-796, 2000.
- [45] Curio, C., J. Edelbrunner, T. Kalinke, C. Tzomakas and W. von Seelen, "Walking Pedestrian Recognition," *IEEE Transactions on Intelligent Transportation Systems* 1(3) September pp. 155-163, 2000
- [46] Cootes, T.F., Cooper, D., Taylor, C.J., Graham, J., Active Shape Models - Their Training and Application, *Computer Vision and Image Understanding*. Vol. 61, 38-59, 1995

- [47] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Computer Vision and Pattern Recognition*, San Diego, CA, USA, June 20-25 2005.
- [48] J. Davis and A. Bobick. "The representation and recognition of action using temporal templates", In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 928-934, 1997.
- [49] Fang, Y., Yamada, K., Ninomiya, Y., Horn, B. and Masaki, I., "Comparison between. Infrared-image-based and Visible-image-based Approaches for Pedestrian Detection," *IEEE. Intelligent Vehicles Symposium*, pp.505-510, 2003
- [50] K. Fukunaga. *Introduction to statistical pattern recognition* (2nd ed). Academic Press, Boston 1990.
- [51] D. M. Gavrilu. "Protecting pedestrians in traffic: Sensor-based approaches," *IEEE Intelligent Systems*, 2001.
- [52] D.M. Gavrilu, "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image understanding*, Vol. 73(1), pp. 82-98, 1999.
- [53] I. Hariatoglu, R. Cutler, D. Harwood, and L.S. Davis, "Backpack: Detection of People Carrying Objects using Silhouettes", *Computer Vision and Image Understanding*, 81(3), 2001.
- [54] I. Haritaoglu, M. Flickner, and D. Beymer, "Ghost3D: Detecting Body Posture and Parts Using Stereo", In *Workshop on Motion and Video Computing*, Orlando, Florida, November 7 2002.
- [55] I. Haritaoglu, D. Harwood, and L.S. Davis, "Ghost: A Human Body Part Labeling System Using Silhouettes", In *International Conference on Pattern Recognition*, 1998.
- [56] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4: Who? When? Where? What? - A Real Time System for Detecting and Tracking People," In *International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998.
- [57] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4: Real-Time Surveillance of People and Their Activities", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000.

- [58] A. Hilton and P. Fua, Foreword: modeling people toward vision-based understanding of a persons shape, appearance, and movement, *Comput. Vis. Image Understanding*, vol. 81, no. 3, pp. 227C230, 2001.
- [59] Hirokawa, S., "Normal Gait Characteristics Under Temporal and Distance Constraints," *J. Biomed. Engr.*, 11: 449 - 456. 1989
- [60] D. Hogg. "Model-based Vision: a Program to See a Walking Person," *Image and Vision computing*, Vol. 1, No. 1, pp. 5-20, 1983.
- [61] Horn, B. K. P, Schunck, B. G., Determining optical flow, *Artificial Intelligence*, vol. 17, pp 185-203, 1981
- [62] G. Hua, M-H. Yang, and Y.Wu. Learning to Estimate Human Pose with Data Driven Belief Propagation. In Proc.IEEE Computer Vision and Pattern Recognition, 2005.
- [63] E. Huber, "3D Real-time Gesture Recognition using Proximity Space", *In Proc. of Intl. Conf. on Patten Recognition*, pages 136-141, Vienna, Austria, August 1996.
- [64] S. S. Intille and A. F. Bobick, "Closed-world Tracking," *In Proc. Intl. Conf. Comp. Vis.*, pages 672- 678, 1995.
- [65] M. Isard and A. Blake, "Contour Tracking by Stochastic Propagation of Conditional Density," *In Proceedings of the 1996 European Conference on Computer Vision*, pages 343C356, 1996.
- [66] Jablan, S., Theory of Symmetry and Ornament, Mathematical Institute, Belgrade, 1995.
- [67] S. Ju, M. Black, and Y. Yacoob, "Cardboard people: A parameterized model of articulated image motion," *In Proceedings of International Conference on Face and Gesture Analysis*, 1996.
- [68] I. A. Kakadiaris and D. Metaxas, "Model based Estimation of 3D Human Motion with Occlusion Based on Active Multi-viewpoint Selection", *In Proc. of IEEE Comp. Soc. Conf. on Computer Vision and Patten Recognition*, pages 81-87, San Francisco, CA, 1996.
- [69] I. A. Kakadiaris and D. Metaxas, "3d Human Body Model Acquisition from Multiple Views," *In Proc. of 5th Intl. Conf. on Computer Vision*, pages 618-623, 1995.

- [70] I. A. Kakadiaris, D. Metaxas, and R. Bajcsy, "Active Part-decomposition, Shape and Motion estimation of Articulated Objects: A Physics-based Approach. In Proc. CVPR, pages 980-984, Seattle, WA, 1994.
- [71] A. Kale, A. Rajagopalan, N. Cuntoor, and V. Krueger, "Human identification using Gait", *In Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Washington, DC, USA, May 21-22, 2002.
- [72] A. Kale, A. Sundaresan, A.N. Rjagopalan, N. Cuntoor, A.R. Chowdhury, V. Krger, and R. Chellappa, "Identification of Humans using Gait", *IEEE Trans. Image Processing*, 9:1163C1173, 2004.
- [73] T. Kanade, R. Collins, A. Lipton, P. Anandan, and P. Burt, "Cooperative Multisensor Video Surveillance," *In Proceedings of the 1997 DARPA Image Understanding Workshop*, volume 1, pages 3C10, May 1997.
- [74] T. Kanade, R. Collins, A. Lipton, P. Burt, and L. Wixson, "Advances in cooperative multisensor video surveillance," *In Proceedings of the 1998 DARPA Image Understanding Workshop*, volume 1, pages 3C24, November 1998.
- [75] Kass, M., Witkin, A., Terzopoulon, D., Snakes: Active contour models , *International Journal of Computer Vision*, pp 321-331, 1987
- [76] P. H. Kelly, A. Katkere, D. Y. Kuramura, S. Moezzi, S. Chatterjee, and R. Jain, "An Architecture for Multiple Perspective Interactive Video," *In Pmc. of ACM Conf. on Multimedia*, pages 201-212, 1995.
- [77] W. Kinzel, "Pedestrian Recognition by Modelling Their Shapes and Movements", Intl. Conf. on Image Analysis and Processing 1999, pages 547-554, Singapore, 1999.
- [78] S. Zhou, R. Chellappa, and B. Moghaddam. "Visual Tracking and Recognition Using Appearance-adaptive Models in Particle Filters", *IEEE Transactions on Image Processing* , Vol. 11, pp. 1434-1456, November 2004.
- [79] M. Leventon and W. Freeman. "Bayesian estimation of 3-d human motion from an image sequence," *Technical Report TR-98-06, Mitsubishi Electric Research Laboratory*, Cambridge,MA, July 1998.
- [80] B. Li, R. Chellappa, and H. Moon, "Monte Carlo Simulation Techniques for Probabilistic Tracking", *In Thirty-Fifth Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, Californina, November 4-7 2001.

- [81] L. Wang, W. Hu and T. Tan, "Recent Developments in Human Motion Analysis," *Pattern Recognition* Volume 36, Issue 3, pp 585-601, March 2003
- [82] H. Ling and D.W. Jacobs, "Deformation Invariant Image Matching", *IEEE International Conference on Computer Vision (ICCV)* Vol. II, pp. 1466-1473, 2005
- [83] H. Ling and D.W. Jacobs, "Using the Inner-Distance for Classification of Articulated Shapes", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. II, pp. 719-726, 2005.
- [84] J.J. Little and J.E. Boyd, "Recognizing People by Their Gait: the Shape of Motion", *Videre: Journal of Computer Vision Research*, The MIT Press, Vol. 1(2), pp. 24-42, 1998
- [85] F. Liu, and R.W. Picard, "Finding Periodicity in Space and Time," *Proceedings of the Sixth International Conference on Computer Vision*, pp. 376-382, 1998.
- [86] W. C. Lindsey and C. M. Chie, eds., *Phase-Locked Loops*. IEEE PRESS Selected Reprint Series, New York, NY: IEEE Press, 1986.
- [87] J. Lipton, H. Fujioshi, R. S. Patil. "Moving Target Classification and Tracking from Real-Time Video," *Workshop on Applications of Computer Vision*, Princeton, NJ, Oct. pp. 8-14, 1998.
- [88] A. Lipton, "Local Application of Optic Flow to Analyse Rigid versus Non-rigid Motion," In *ICCV99 Workshop on Frame-Rate Applications*, September 1999.
- [89] D.G. Lowe. Distinctive Image Features From Scale-Invariant Keypoints. *International Journal on Computer Vision*, 60(2):91C110, 2004.
- [90] Efros, AC Berg, G. Mori, J. Malik, "Recognizing Action at A Distance," *Proceedings of IEEE International Conference on Computer Vision*, pp. 726-733, October 2003
- [91] S. Maybank and T. Tan, "Introduction to Special Section on Visual Surveillance," *International Journal of Computer Vision*, Vol. 37(2), pp. 173-173, 2000.
- [92] Anurag Mittal and Larry S. Davis, "M2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene", *International Journal of Computer Vision*. Vol. 51 (3), Feb/March 2003.

- [93] D. Morris and J Rehg. "Singularity analysis for articulated object tracking," *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp289-296, 1998.
- [94] Ngo, C.W., Pong, T.C., Zhang, H.J., Motion analysis and segmentation through spatio-temporal slice processing, *IEEE Transactions on Image Processing*, pp 341-355, 2003
- [95] Ngo, C.W., Pong, T.C., Zhang, H.J., Motion-based video representation for scene change detection, *International Journal of Computer Vision*, pp 127-142 , 2000
- [96] Ngo, C.W., Pong, T.C., Zhang, H.J., Chin, R.T., Motion characterization by temporal slice analysis, *IEEE Conference on Computer Vision and Pattern Recognition*, 2000
- [97] Ngo, C.W., Pong, T.C., Chin, R.T., Detection of gradual transitions through temporal slice analysis, *in Proceedings of International Conference of Computer Vision and Pattern Recognition*, 1999
- [98] S.A. Niyogi and E.H. Adelson, "Analyzing and Recognizing Walking Figures in XYT," *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 469-474, 1994.
- [99] Harsh Nanda, Larry Davis, "Probabilistic Template Based Pedestrian Detection in Infrared Videos", *IEEE Intelligent Vehicle Symposium*, Versailles, France, June 18-20, 2002
- [100] Oren, M., Papageorgiou, C.P., Sinha, Osuna, E., Poggio, T., "Pedestrian Detection Using Wavelet Templates," *IEEE Conference on Computer Vision and Pattern Recognition* , 193-199, 2003.
- [101] Chia-Jung Pai, Hsiao-Rong Tyan, Yu-Ming Liang, Hong-Yuan Mark Liao, Sei-Wang Chen, "Pedestrian Detection and Tracking at Crossroads," *Pattern Recognition* 37(5): 1025-1034, 2004
- [102] V. Parameswaran and R. Chellappa. "View Invariants for Human Action Recognition", *In Computer Vision and Pattern Recognition*, Madison, Wisconsin, June 16-22 2003.
- [103] V. Parameswaran and R. Chellappa, "View Independent Human Body Pose Estimation from a Single Perspective Image", *In Proc. IEEE Computer Vision and Pattern Recognition*, 2004.

- [104] Peng, S.L., Medioni, G., Interpretation of image sequences by spatio-temporal analysis, *Workshop on Visual Motion*, pp 344-351, 1989
- [105] Peng, S.-L., Temporal slice analysis of image sequences, *in Proceedings of Computer Vision and Pattern Recognition*, pp 283-288, 1991
- [106] P. J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. W. Bowyer, "The Gait Identification Challenge Problem: Data Sets and Baseline Algorithm," *International Conference on Pattern Recognition*, pp. 385-388, Aug 2002.
- [107] C. Papageorgiou, T. Evgeniou, T. Poggio. "A Trainable Pedestrian Detection System," *IEEE Int. Conf. on Intelligent Vehicles*, pp. 241-246, Germany, Oct 1998.
- [108] R. Polana, C. Nelson, "Detection and Recognition of Periodic, Nonrigid Motion," *International Journal of Computer Vision* Volume 23, Issue 3, pp. 261 - 282, 1997
- [109] R. Polana. "Temporal Texture and Activity Recognition", PhD thesis, Department of Computer Science, University of Rochester, 1994.
- [110] Qian Wang and Jasbir S. Arora, "Digital Human Modeling and Virtual Reality for FCS," Technical Report No. VSR-04.02, The University of Iowa, 2004
- [111] B.G. Quinn and E.J. Hannan, *The Estimation and Tracking of Frequency*, Cambridge University Press, ISBN 0-521-80446-9 2001.
- [112] Qinfen Zheng, Rama Chellappa "Automatic Registration of Oblique Aerial Images," *IEEE International Conference on Image Processing* pp. 218-222. 1991.
- [113] D. Ramanan, D.A. Forsyth, and A. Zisserman, "Strike a Pose: Tracking People by Finding Stylized Poses", *In Proc.IEEE Computer Vision and Pattern Recognition*, 2005.
- [114] C. Rao, A. Yilmaz, and M. Shah, "View-Invariant Representation and Recognition of Actions," *Journal of Computer Vision*, 50(2), 2002.
- [115] Yang Ran, Isaac Weiss, Qinfen Zheng, Larry Davis, "Pedestrian Detection via Periodic Motion", *the International Journal of Computer Vision*, to appear in 2006 June's volume.

- [116] Yang Ran, Rama Chellappa, Qinfen Zheng, "A Compact Representation of Human Gait and Activities", *IEEE Trans on Pattern Analysis and Machine Intelligence*, submitted.
- [117] Yang Ran, Rama Chellappa, "Real-time Pedestrian Monitoring", *The Video Demo Proceedings of The 2006 International Conference on Computer Vision and Pattern Recognition (CVPR2006)*, New York City, USA, to appear.
- [118] Yang Ran, Rama Chellappa, "Finding Gait in Space and Time", *The 2006 International Conference on Pattern Recognition (ICPR2006)*, HongKong, P.R.China, to appear.
- [119] Yang Ran, Rama Chellappa, Qinfen Zheng, "Gait DNA and its applications in surveillance systems", *University of Maryland Invention Disclosure IS-2005-108, 2005*.
- [120] Yang Ran, Qinfen Zheng, Isaac Weiss, Larry S. Davis, Wael Abd-Almageed, "Pedestrian Classification from Moving Platforms Using Cyclic Motion Pattern", *International Conference of Image Processing (ICIP) 2005*, Genova, Italy (Best 10 presentations selected by all attendees), II pp 854-857.
- [121] Hussein, M.; Abd-Almageed, W.; Yang Ran; Davis, L. "Real-Time Human Detection, Tracking, and Verification in Uncontrolled Camera Motion Environments", *ICVS '06. IEEE International Conference on Computer Vision Systems*, 2006 Page(s):41 - 41
- [122] Yang Ran, Qinfen Zheng, Isaac Weiss, Larry Davis, "Reliable Segmentation of Pedestrians in Moving Scenes", *The 2005 International Conference on Acoustics, Speech, and Signal Processing (ICASSP2005)*, Philadelphia, USA, pp229 - 232
- [123] Yang Ran, Isaac Weiss, Qinfen Zheng, Larry Davis, "An Efficient And Robust Human Classification Algorithm Using Finite Frequencies Probing", *the Joint IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum with CVPR2004*, Washington DC, USA, 2004, Page(s):132 - 140.
- [124] Yang Ran, Qinfen Zheng, "Multi Moving people detection From Binocular sequences", *Multimedia and Expo, ICME '03. Proceedings. 2003 International Conference on* Page(s):II - 297-300 vol.2
- [125] Yang Ran, Qinfen Zheng, "Multi Moving people detection From Binocular sequences", *The 2003 International Conference on Acoustics, Speech, and Signal Processing (ICASSP2003)*, HongKong, P.R.China, Page(s):II - 297-300 vol.2.

- [126] C. Regazzoni and V. Ramesh, Special issue on video communications, processing, and understanding for third generation surveillance systems, *Proc. IEEE*, vol. 89, pp. 1355C1367, Oct. 2001.
- [127] K. Rohr. "Towards Model-Based Recognition of Human Movement in Image Sequences," *CVGIP: Image Understanding*, Vol. 59, No. 1, pp. 94-115, Jan, 1994.
- [128] S. M. Seitz and C. R. Dyer, "View-Invariant Analysis of Cyclic Motion," *Int. J. Computer Vision*, 25(3), pp. 231-251, 1997
- [129] C. Tomasi and T. Kanade, "Shape and Motion from Image Streams: Factorization Method," *International Journal of Computer Vision*, 9(2), 1992.
- [130] P. Tsai, M. Shah, K. Keiter, and K. Kasparis, "Cyclic Motion Detection," *Pattern Recognition*, Vol. 27, No. 12, 1994
- [131] N. Vasvani, A. Roy Chowdhury, and R. Chellappa, "Activity Recognition using the Dynamics of the Configuration of Interacting Objects," *In Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Madison, WI, June 16-22, 2003.
- [132] Viola, P., Jones, M., Snow, D. "Detecting Pedestrians using patterns of motion and appearance," *Ninth IEEE International Conference on Computer Vision* pp 734-782, 2003.
- [133] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," *In Computer Vision and Pattern Recognition*, Kauai Marriott, Hawaii, December 9-14 2001.
- [134] P. Viola, M.J. Jones, and D. Snow, "Detecting Pedestrians Using Patterns of Motion and Appearance," *International Journal of Computer Vision*, 63(2), 2005.
- [135] Vladimir M. Zatsiorsky, Kinematics of Human Motion, Human Kinematics, 1998
- [136] S.Wachter and H.-H. Nagel, "Tracking of Persons in Monocular Image Sequences", In Workshop on Motion of Non-Rigid and Articulated Objects, Puerto Rico, USA, 1997.
- [137] S. Wachter and H.-H. Nagel, "Tracking Persons in Monocular Image Sequences. *Computer Vision and Image Understanding*, 74(3):174C192, 1999.

- [138] Wang, J., Adelson, E., Representing Moving Images with Layers, *IEEE Transactions on Image Processing*, 3(5) pp 625-638, 1994
- [139] Willis, C., Video Stack Image Analysis, Edinburgh University MSc Dissertation, 2004
- [140] Weyl, H., Symmetry, Princeton University Press, Princeton, 1952.
- [141] C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder: Real-Time Tracking of the Human Body", *Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780C785, 1997.
- [142] B. Wu and R. Nevatia, "Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detection", In *International Conference on Computer Vision*, Beijing, China, Oct 15-21 2005.
- [143] Y.Wu, G. Hua, and T. Yu. Tracking Articulated Body by Dynamic Markov Network. In *International Conference on Computer Vision*, Nice, France, 13-16 October 2003.
- [144] Y. Liu, R.T. Collins, Y. Tsin, "Gait Sequence Analysis Using Frieze Patterns," *ECCV* pp. 657-671 2002
- [145] S. Yasutomi and H. Mori, A Method for Discriminating of Pedestrian Based on Rhythm, in *Proc. of IEEE Intl. Conference on Intelligent Robots and Systems*, pp. 988-995, 1994.
- [146] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation", *In Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, CVPR05, 2005.
- [147] A. Yilmaz and M. Shah, "Recognizing human actions in videos acquired by uncalibrated moving cameras," *In Proc. Int. Conf. on Computer Vision*, (ICCV05), 2005.
- [148] J. Zhang, R. Collins, and Y. Liu, "Bayesian Body Localization Using Mixture of Nonlinear Shape Models," *In Proc. IEEE Int. Conf. of Computer Vision*, 2005.
- [149] Chong-Wah Ngo, Ting-Chuen Pong, Hong-Jiang Zhang: "Motion Analysis and Ssegmentation Through Spatio-temporal Slices Processing," *IEEE Transactions on Image Processing* 12(3): 341-355, 2003

- [150] L. Zhao and C. Thorpe, "Stereo and Neural Network-based Pedestrian Detection," *IEEE Transactions on Intelligent Transportation Systems*, Vol. 1, No. 3, pp. 148 -154, September, 2000.
- [151] L. Zhao, C. Thorpe. "Stereo- and Neural Network Based Pedestrian Detection," *Proceedings ITSC*, Tokyo, Japan, 1999.
- [152] T. Zhao and R. Nevatia, "Stochastic Human Segmentation from a Static Camera", *In Workshop on Motion and Video Computing*, Orlando, Florida, November 7 2002.
- [153] T. Zhao and R. Nevatia, "Tracking Multiple Humans in Crowded Environments," *In Computer Vision and Pattern Recognition*, Washington DC, USA, June 2004.
- [154] S. Zhou, R. Chellappa, and B. Moghaddam. "Visual Tracking and Recognition Using Appearance-adaptive Models in Particle Filters," *IEEE Transactions on Image Processing* , Vol. 11, pp. 1434-1456, November 2004.