Estimation of Pedestrian Walking Direction for Driver Assistance System

Zhao Guangzhe

Abstract

Road traffic accidents are a serious problem around the world, where the cost of human life is impossible to evaluate, and cause massive and continuous government spending. Different solutions have been proposed to reduce the effects of accidents, one of which, Advanced Driver Assistance Systems, as their name suggest, assist the driver by providing vital information on the traffic environment or by acting under specific circumstances to safeguard the occupants of the vehicle, or to facilitate driving. In case that an accident cannot be prevented, collision mitigation devices that are incorporated into vehicle design enhancement can be deployed to reduce the impact of the collision on the pedestrian.

In this thesis we present a practical approach to the problem. Pedestrian protection is a crucial component of driver assistance systems. Our aim is to develop a video-based driver assistance system for the detection of the potential dangerous situation, in order to warn the driver. We address the problems of detecting pedestrian in real-world scenes and *estimating walking direction with a single camera from a moving vehicle*. The challenge is of considerable complexity due to the *varying appearance of people* (e.g., clothes, size, pose, shape, etc.), and the *unstructured moving environments* that urban scenarios represent. In addition, the required performance is demanding both in terms of *computational time* and *detection rates*.

Considering all the available cues for predicting the possibility of collision is very important. The “direction” in which the pedestrian is facing is one of the most important cues to predict where the pedestrian may move in future. Therefore we first emphasize the core problem of *pedestrian orientation estimation* in real-world scenes. The method is designed to estimate major eight orientations of different appearances. By taking account head-part orientation into the estimation, accuracy of overall estimation is drastically improved.

Consequently, we construct and propose a three-stage method: (i) pedestrian detection, (ii) orientation estimation for single-frame and (iii) walking direction estimation for multi-frame. The first two stages employ and extend computationally

light Adaboost algorithms and Haar-like features. Therefore, the overall method meets a practical need with embedded on-board computing performance.

In order to perform the experiments, we also present a real-world pedestrian dataset to the evaluation of all stages. It achieved a very good performance, 64% accuracy as a 8-class recognition problem. Then, we estimate the *pedestrian walking direction using multi-frame* based on the result of single-frame orientation estimation. The results presented in this thesis not only end with a proposal of a pedestrian detection and pedestrian orientation estimation but also go one step beyond by estimating the pedestrian walking direction over an image sequence, introducing new techniques and evaluating their performance, which will provide new foundations for future research in the area.

Following is a brief summary of each chapter in the dissertation. Chapter 1 presents the background and motivations of driver assistance systems and pedestrian protection systems. Technique issue and current trends of research and development are addressed. Chapter 2 describes related researches in pedestrian detection, pedestrian orientation estimation and pedestrian walking trajectory estimation. Chapter 3 presents the proposed whole system architecture consisting of three pyramid stages, and introduces the conventional approach for pedestrian detection for the stage 1. Chapter 4 presents our proposed approach for the estimation of pedestrian orientation based on the same feature extraction for the stage 1 with a newly designed multi-class classifier and describes the experimental results. Chapter 5 introduces a new approach on the study of the problem of pedestrian walking direction estimation. Chapter 6 summarizes the results and identifies other possible application areas and topics for future research.

Acknowledgments

Composing a Ph.D. thesis is without doubt the hardest and most tiring project I have ever made. My Ph.D. dissertation could never have been completed without the generous support of colleagues, peers, friends, and family. I would like to take a moment to thank these individuals and acknowledge their support.

Professor Kenji Mase offered to me, in late 2007, the opportunity to pursue Ph.D. studies at Nagoya University under his guidance. I was grateful for the opportunity to improve my research abilities, but was not yet able to fully appreciate the depth of challenges that would await me. I began my Ph.D. studies in April, 2008. With Professor Mase's constant guidance, I became accustomed to the practices of research in the Pedestrian Protection Project which is performed as a joint research between the TOYOTA CENTRAL R&D LABS., INC. and Nagoya University. Professor Mase not only helped me with professional and personal guidance, but also was directly responsible for finding financial support, by means of the CREST project. This fund allowed me to continue the latter part of my studies at Nagoya University. Thank Professor Mase for enabling me to take my first steps in the world of academic research. This thesis would not have been completed without his commitment and diligent efforts which are not only influenced the content of the thesis but also the language in which it has been conveyed. He not only showed how to exhaustively explore different methodologies and analysis results, but also imbibed in me the need for perfection performing research and communicating the results. I am also grateful to him who always encouraged me to make my Ph.D research more applicable and offered me the opportunity to do a Ph.D on such an interesting and challenging topic and providing me with the platform for this thesis.

I am also indebted to the other members of my doctoral thesis committee: Professor Toyohide Watanabe, Professor Takami Yasuda and Associate Professor Jien Kato. Their insightful comments and criticisms helped me improve my research and my research style, and I am thankful for their time and their advice.

I would also like to thank the members from TOYOTA CENTRAL R&D LABS,

giving me an opportunity to perform such an interesting and challenging topic and providing with me the supplyment for this research such as the pedestrian dataset. I will never forget thanks to Takashi Naito who gave me valuable comments and encouraged me to challenge this research topic.

I would like to thank all my friends from Mase's lab where has been a fun and comfortable research place. Much respect to my lab mates Yuchi Koyama, Haijun Song, Fushi Li, Fei Yang, for putting up with me during these past years. I shared many interesting discussions and fun work time with them. The help received from them both personally and professionally is also immense. I also want to thank Ms. Yasuyo Kawamura; she helped make me to be sure that administrative matters relating to my studies and life.

Finally, I would like to thank my family, for their absolute confidence in me, which help me to have the opportunity to do a PhD in Japan.

Contents

List of Figure

Chapter 1 Introduction

Chapter 2 Background and Literature Review

Chapter 3 Advanced Driver Assistance System

Chapter 4 Pedestrian Orientation Estimation

Chapter 5 Pedestrian Walking Direction Estimation

Chapter 1

Introduction

As the modern industry develops, human seek happiness and life convenience. Humans have been designed intelligent machines since the appearance of early civilizations. For example, the first programmable machines have been traced back to the Ancient Greece in the 1st century BC [1]. Furthermore, it is said that human sized automaton was built from the Renaissance until the 20th century in the previous centuries [2]. The vehicle provide the convenience for human life, as the same time, vehicle accidents are one of the main causes of accidental death in the modern civilized world. The car seems to be one of the most deathly man-made objects used in the civilized world.

The percentage of pedestrian deaths is even higher in many countries of Asia and Europe. Intelligent vehicle systems have the capability which can reduce the pedestrian deaths and injuries effectually. However, in order to provide effective protection, such systems need to not only detect pedestrians in varying environmental conditions, but also predict the possibility of collision. They should relay the information to the driver in efficient and non-distracting manner or to the control system of the vehicle in order to take preventive actions.

1.1 Pedestrian safety

Vehicles represent one of the key advanced technologies for human development in the modernization of the industry area. About 50 million passenger cars and 20 million commercial vehicles are being produced worldwide every year [3]. At this rate, the number of automobiles in the world will reach one billion units in the later years, especially due to emerging economies of Asian countries such as India and China. **Fig.1.1** is referred from image.baidu.com.

Fig. 1.1 Large numbers of automobiles

Unfortunately, as the same time, followed with these many benefits, such the technology has also carried a bad effect: traffic accidents. The first death by a motor vehicle was registered in Ireland on 1869 [4]. Nowadays, according to the World Health Organization, road accidents represent the 6th cause of death in high-income countries and the 11th worldwide [5,6]. Every year almost 1.2 million people are killed in traffic collisions while the number of injured rises to 50 million. Furthermore, these numbers are expected to increase a 65% between 2000 and 2020, especially in low and middle-income countries.

According to the World Bank website [11], pedestrians account for 65% of the fatalities out of the 1.17 million traffic related deaths around the world. In the United States, according to the National High-way Traffic Safety Administration [12], there were 4641 pedestrian fatalities during 2004, which accounted for 10.9% of the total 42636 traffic-related fatalities. In Britain, pedestrians are twice as likely to be killed in accidents as vehicle occupants [13]. In Japan, about 2000 pedestrian fatalities every year (see **Fig.1.2**), especially many accidents owed the pedestrian make a sudden crossing (see **Fig.1.3**). The pedestrian death of number is total 18 in 2012 (12 is from right to left and 6 is from left to right), and from the 2009 to 2011, these kind of accidents is 132 in total.

In developing countries such as India and China, the problem is much worse. During 2001, there were 80000 fatalities on Indian roads, which grew in last decade at 5% per year [14]. In fact, 60%–80% of the road fatalities are the vulnerable road users [15],

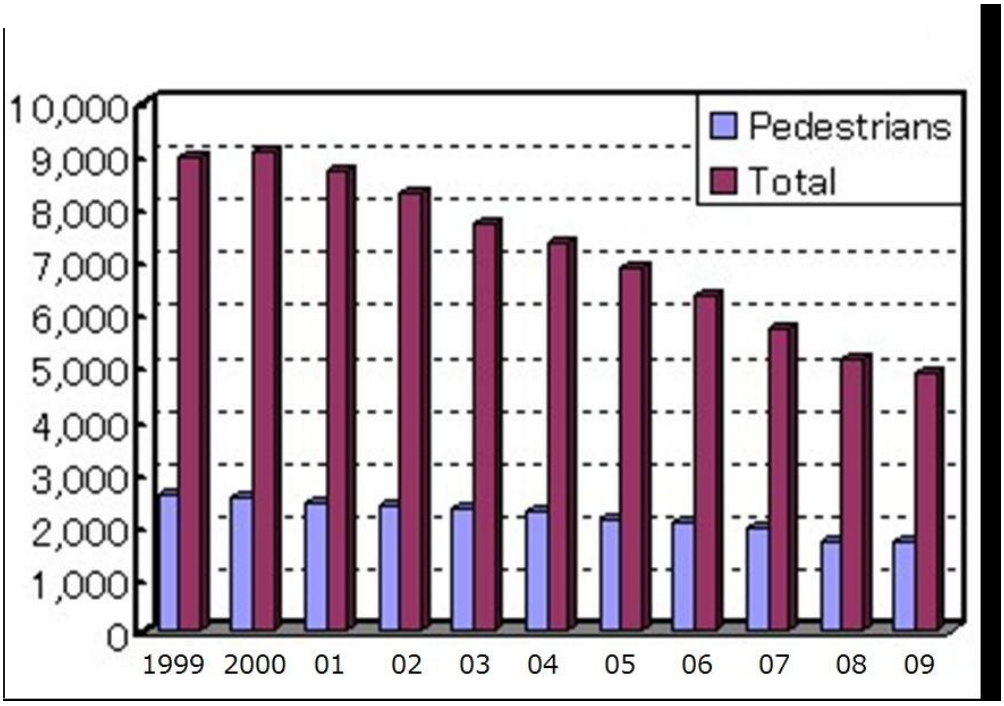


Fig. 1.2 Death toll in road traffic accidents in Japan (The blue bar represent the death number of pedestrian and the red bar represent the number of total accidents)

many of them from low-income groups. With the rapid increase in the number of vehicles in these countries, the number of accidents and fatalities is likely to increase before they can be reduced. Furthermore, the problems faced by developing countries are often different from those faced by developed countries. In developing countries, there are a large number of two wheelers, three wheelers, bicyclists, and pedestrians sharing the same road space with cars, buses, and trucks [16], [17]. Hence, the solutions for developed countries may not all be directly applicable for developing countries. In fact, the first steps for these countries lie in improving infrastructure design and developing appropriate infrastructure-based solutions.

1.2 Approach for improving pedestrian safety

Pedestrian safety can be improved at several stages. Long-term measures include infrastructure design enhancements in vehicles to reduce the fatalities. These enhancements can be complemented by systems that detect the pedestrians and prevent

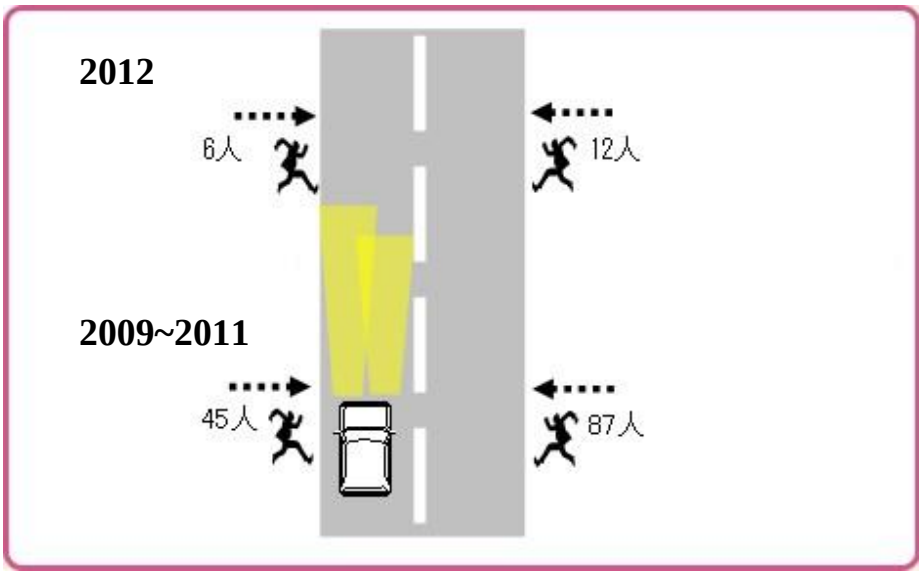


Fig. 1.3 Accidents owed the pedestrian make a sudden crossing
(The number is the death number of pedestrian)

accidents by warning the driver or triggering autonomous braking. In the cases where an accident cannot be prevented, collision mitigation devices that are incorporated into vehicle design enhancement can be deployed to reduce the impact of the collision on the pedestrian. In this thesis, we focus on developing a video-based driver assistance system for the detection of potentially dangerous situation with the pedestrian, in order to warn the driver. **Fig.1.4** is an introduction for traffic scene schematically. Infrastructure like road mark and signal lamb design enhancement and vehicle design can be reduced the fatality, autonomous driving system may be useful for reduce the accident. In this these, we focus on a driver assistance system which is able to detect the pedestrian and estimate the pedestrian walking direction.

1.2.1 Infrastructure Design Enhancements

Infrastructure enhancements to reduce pedestrian-related accidents are divided into three categories of countermeasures: speed control, pedestrian–vehicle separation, and measures to increase visibility of pedestrians.

1. Reduction of speed results in fewer injuries due to the lowering of kinetic energy as well as greater reaction time. The techniques for speed control include single-lane roundabouts, speed bumps, pedestrian refuge islands, and use of multi way stop signs.

2. Separation of pedestrians and vehicles can be performed by measures such as installing traffic signals, allocating exclusive time for pedestrian signals, in-pavement

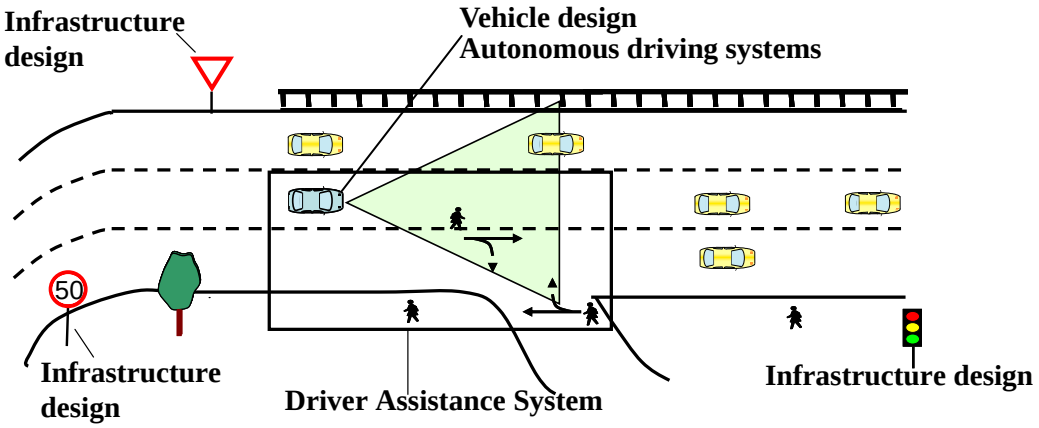


Fig. 1.4 Pedestrian safety approach

flashing lights to warn drivers, and automatic pedestrian detection at walking signals.

3. Pedestrian visibility can be increased by improving roadway lighting, since a majority of pedestrian fatalities occur at night time.

1.2.2 Vehicle design for safety systems

The design of the vehicle has great impact on the extent of injury that a pedestrian sustains in case of a collision. Most of the injuries take place in lower limbs, whereas most of the fatal injuries are head injuries. In order to minimize the effect of these injuries, various collision-absorbing components such as compliant bumpers, pop-up bonnets, and windscreen airbags are suggested. Ford and Volvo research laboratories have developed the use a finite element model of a pedestrian to simulate accidents. Such models help in predicting the effects of collision and in improving the vehicle design to minimize these effects. The EU has recently mandated the incorporation of pedestrian safety systems in cars. They designed an active hood, which automatically rises in case of collision with pedestrian, so that the surface that comes in contact with the head is deformable and flexible instead of hard and rigid. They have also developed a pair of airbags at the windshield pillar to prevent impact of windshield on the head. It is claimed that their pedestrian protection system can reduce risk of life-threatening injuries to 15% from nearly 100% for a collision, which is enough to satisfy the EU requirement. Pedestrian airbags can reduce head injuries by 90% and upper body

injuries by 50%.

1.2.3 Autonomous driving systems

The first stone in the area of advanced drive assistance systems was put by E. Dickmanns group in 1986 with an autonomous highway driving system [7, 8]. They presented a system which can able to drive through closed highways at speeds of up to 96km/h by exploiting cameras, rudimentary image processors and based on Kalman filtering algorithms. This research would later lead to the first European project on autonomous vehicles: Prometheus.

Nowadays many advanced drive assistance systems have already been commercialized which can be found in the market as practical applications. For example, the first adaptive cruise control systems were introduced in high class Lexus, Mercedes and Jaguar in the late 1990s [9]. Adaptive cruise control systems keep a constant distance to the front vehicle by slowing or accelerating the host one. Lane departure warning systems warn the driver when the vehicle moves out of its lane, unless the corresponding direction turn signal is on. This technology was first included in trucks in 2000 [10] and later extensively used in sedans. This technology is currently being improved by assisting the steering action or warning/intervening in lane changing in case of danger. One of the current-hot research topic is advanced front lighting systems, which control the headlight parameters so that the beam is optimized for different conditions like driving speed and direction.

1.2.4 Advanced driver assistance systems

Considerable research is being conducted by various groups for designing pedestrian detection systems. Such systems can employ various types of sensors and computer vision algorithms in order to detect pedestrians and to predict the possibility of collisions. The output of the systems can be used to generate appropriate warnings for the driver or to perform autonomous braking or maneuvering in the case of an imminent collision.

Pedestrian Protection Systems are a particular type of advanced driver assistance systems devoted to pedestrian safety. A pedestrian protection system is formally defined as a system that detects both static and moving people in the surroundings of the vehicle typically in the front area in order to provide information to the driver and perform evasive or braking actions on the host vehicle if needed. Pedestrian detection before the

Fig. 1.5 The variability of pedestrian is high as a result of the different possible illuminations, size, poses, view angels, clothes, etc.

impact is crucial given that the severity of injuries for the pedestrian decreases with speed of the crashing vehicle. Thus, any reduction in the speed can drastically reduce the severity of the crash. According to [25], pedestrians have a 90% chance of surviving to car crashes at 30km/h or below, but less than 50% chance of surviving to impacts at 45km/h or above.

Without assistance, the human reaction time is long and consequently the brakes are auctioned about 1 second after the dangerous situation. A pedestrian is likely to suffer severe harm if he or she is at less than 25m. With assistance, the benefits are twofold. First, they can reduce the reaction time to 100ms or less [26, 27]. Second, since they can anticipate the potential accident they can not only provide warnings to the driver in a reduced time but also control the different active measures like airbags or brakes. Hence, the distance where pedestrians can be severely damaged is significantly reduced.

1.3 The role of Computer Vision

The central problem of pedestrian protection systems corresponds to the task of detecting pedestrians so as to provide the useful information to the driver. In order to detect objects (e.g., vehicles, pedestrians, obstacles) in the distance, advanced driver assistance systems make use of sensors that provide data to a computer/controller that processes them and performs the corresponding actions. A comprehensive analysis of these sensors is made.

The most widely used sensors for pedestrian detection are cameras working either in the visible or infrared spectra. they provide the rich information, such as cues like edges, contours, texture or even relative temperature in the case of infrared cameras. Therefore, it is clear that Computer Vision plays a key role in the task of pedestrian detection, which in fact is the central problem in pedestrian protection systems.

Once the problem and sensors involved in detection have been introduces, the

challenges for pedestrian protection systems can be summarized in the following points:

1. Appearance variability is very high in pedestrians, given that they can change pose, wear different clothes, carry different objects and their range of sizes is considerable (see **Fig.1.5**).

2. Pedestrians shall be identified in outdoor urban scenarios. That is, they shall be detected in a *cluttered background* because urban areas are more complex than highways under *different illumination and weather conditions* that add variability to the quality of the sensed information (e.g., shadows and poor contrast in the visible spectrum).

3. In addition, pedestrian can be partially occluded by different urban elements such as parked vehicles or street furniture. Maybe the people of the advertisement stand at the roadside will be considered as pedestrian.

4. Pedestrian shall be identified in very dynamic scenes given that not only the pedestrians move but also the camera does, which makes tracking and movement analysis difficult. Because the background is not static means the background subtraction method is no more efficient.

5. Furthermore, pedestrians appear under different viewing angels (e.g., lateral and front/rear positions) and a big range of distances shall be reached. Most of the systems are focused on the distances from 5 to 25m to the camera, namely high risk area. However, extending the detection to 50m, that is, covering also the low risk area, represents a great aid for pedestrian protection systems in the long term accident prevention.

6. Nighttime detection with infrared cameras is affected by temperature, distance.

7. The required performance is quite demanding in terms of system reaction time and robustness (false alarms and misdetections).

In this thesis, we focused on the point of 1,2,4,5, and 7. Point 3 which is occlusion problem is a difficult problem current, and will be developed the research as future work. Point 7 will be benefit from the using vehicle sensors, and will be develop the research as future work.

Our long term aim is to develop the driver assistance system which informs the dangerous situation with predicting the collision probability to the driver. Pedestrian orientation estimation can potentially improve the predication of future trajectories that the pedestrian may take and improve collision.

1.4 Approach of the thesis

Pedestrian protection systems not only need to detect pedestrians, but also to predict the possibility of collisions between pedestrians and vehicles. The system must efficiently relay information to drivers so that they can take preventive actions.

Our goal is to detect the pedestrian and estimate the pedestrian walking direction with a single camera on a moving vehicle, and long term aim is to develop a video-based driver assistance system for the detection of potentially dangerous situations with pedestrians, in order to warn the driver.

It is certain that using some equipment such as senor, radar or stereo camera can increase the detect rate efficiently because these equipment can provide more useful information, but the cost is high and make the vehicle equipment complicate. Monocular imaging systems are less expensive and simpler to set up. So we propose to detect pedestrian and estimate the pedestrian walking direction using a single camera from a moving vehicle based learning approach.

The challenges for our system can be summarized in the following points:

1. Single camera. Single camera makes the detection difficult, but the cost is low and easier to embed on-board.
2. Dynamic Scene. Not only the pedestrians move, but also the camera does. It makes the movement analysis difficult, because the background subtraction method is no more efficient.
3. Outdoor/Various background. The urban traffic situation is more complex, various background make the problem of pedestrian detection more difficult.
4. Varying Illumination. There are different light conditions in real traffic world.
5. Arbitrary clothing. Pedestrians wear different clothes.
6. Various distances. It makes the pedestrian have different sizes
7. Daytime.
8. No occlusion.

Future research will develop in different situation, such as in nighttime and occlusion.

Considering all the available cues for predicting the possibility of collision is very important. The “direction” in which the pedestrian is facing is one of the most important cues to predict where the pedestrian may move in future. Therefore we first address the problem of single-frame pedestrian orientation estimation in real-world scenes. The method is design to estimate major eight orientations by taking account head-part orientation into the estimation. Then, we estimate the pedestrian walking direction using multi-frame based on the result of single-frame orientation estimation finally.

Consequently, we construct a three-stage method: pedestrian detection for

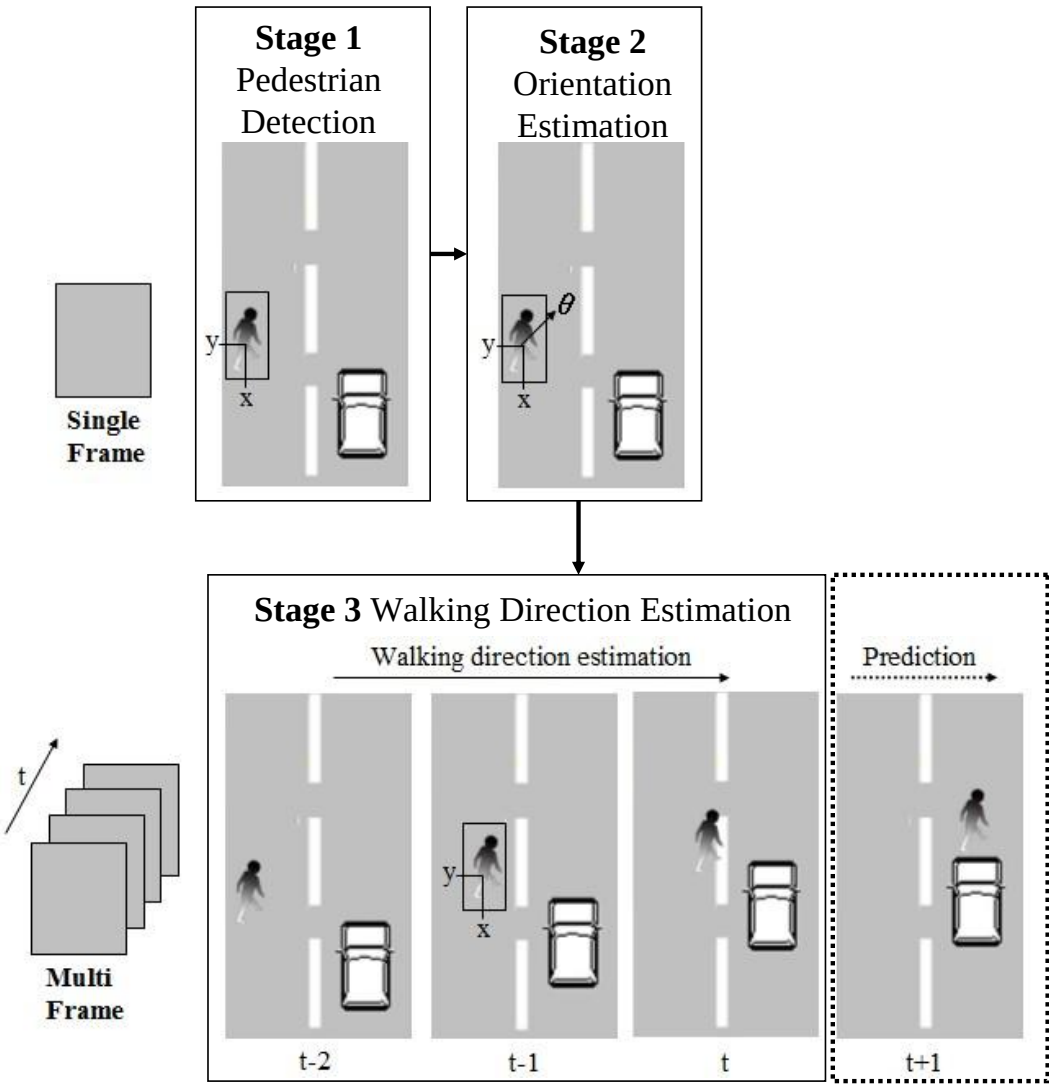


Fig. 1.6 The Schematic representation of whole system.

single-frame single-frame stage, orientation estimation for single-frame stage and walking direction estimation for multi-frame stage.

Consider the system should be realized in real-time, the first two stages employ and extend computationally light Adaboost algorithms and Haar-like feature. Therefore, the overall method meets a practical need with embedded on-board computing environment. The results presented in this thesis not only end with a proposal of a pedestrian detection and pedestrian orientation estimation but also go one step beyond by estimating the pedestrian walking direction over an image sequence, introducing new techniques and evaluating their performance, which will provide new foundations for future research in the area.

1.5 Thesis outline

In this thesis we present a practical approach to the problem. Pedestrian protection is a crucial component of driver assistance systems. Our aim is to develop a video-based driver assistance system for the detection of the potential dangerous situation, in order to warn the driver. We address the problems of detecting pedestrian in real-world scenes and *estimating walking direction with a single camera from a moving vehicle*. The challenge is of considerable complexity due to the *varying appearance of people* (e.g., clothes, size, pose, shape, etc.), and the *unstructured moving environments* that urban scenarios represent. In addition, the required performance is demanding both in terms of *computational time* and *detection rates*.

Considering all the available cues for predicting the possibility of collision is very important. The "direction" in which the pedestrian is facing is one of the most important cues to predict where the pedestrian may move in future. Therefore we first emphasize the core problem of *pedestrian orientation estimation* in real-world scenes. The method is designed to estimate major eight orientations of different appearances. Then, we estimate the *pedestrian walking direction using multi-frame* based on the result of single-frame orientation estimation.

Consequently, we construct and propose a three-stage method: (i) pedestrian detection, (ii) orientation estimation for single-frame and (iii) walking direction estimation for multi-frame (see **Fig.1.6**). The first two stages employ and extend computationally light Adaboost algorithms and Haar-like features. Therefore, the overall method meets a practical need with embedded on-board computing performance.

The results presented in this thesis not only end with a proposal of a pedestrian detection and pedestrian orientation estimation but also go one step beyond by estimating the pedestrian walking direction over an image sequence, introducing new techniques and evaluating their performance, which will provide new foundations for future research in the area.

Following is a brief summary of each chapter in the dissertation.

Chapter 1 presents the background and motivations of Intelligent Transport System (ITS), driver assistance systems and pedestrian protection systems. Technical issue and current trends of research and development are addressed.

Chapter 2 describes related researches in ITS, pedestrian detection, pedestrian

orientation estimation and pedestrian walking trajectory estimation as well as image processing and human image processing with pattern recognition. Some details are addressed in Appendix.

Chapter 3 presents the proposed whole system architecture consisting of three pyramid stages, and introduces the conventional approach for pedestrian detection for the stage 1. Pedestrian detector is trained using Adaboost algorithm and Haar-like feature based on a large number training data which were obtained from the pictures of walking pedestrian in real world with a single camera on a moving vehicle. We confirm the performance of the detector for pedestrian detection problem.

Chapter 4 presents the most important issue with our proposed approach for the estimation of pedestrian orientation based on the same feature extraction for the stage 1 with a newly designed multi-class classifier and describes the experimental results. In this chapter, we outline the problem of estimation into one of eight orientations. We propose a cascade orientation estimation that integrated the head orientation estimation by a multi-Bayesian model. We also compares the performance of our proposed approach to one vs. one and one vs. all multi-class classification approach. Also, we compared it with a state of art 4 orientation detection method and confirmed a comparable performance on a possible semi-fair condition.

Chapter 5 introduces a new approach on the study of the problem of pedestrian walking direction estimation. We propose an average orientation estimation method to estimate the pedestrian walking direction using the result of orientation estimation in a single frame, which is obtained from stage 2. The accuracy of pedestrian walking direction in a straight walking case reached to 83%-98%.

Chapter 6 summarizes the results and achievements of pedestrian walking direction estimation, and it identifies other possible application areas and topics for future research in a short range and a long range aspect.

Chapter 2

Background and Literature Review

The survey starts by an overview of the different approach for pedestrian detection in advanced driver assistance systems in Section 2.1, describing the properties of the ones used for pedestrian protection systems in detail. Then, the review is made divided in review and analysis; the first is to enumerat the existing techniques and the later is to highlighting the advantages, disadvantages and future trends for each stage. Note that not all the reviewed techniques are strictly used in pedestrian protection systems but we also include the ones we find special importance for the area. Pedestrian pose estimation is described in Section 2.2. Finally, pedestrian walking trajectory estimation is presented in Section 2.3.

2.1 Pedestrian detection

Pedestrian detection is a rapidly evolving area in computer vision with key applications in intelligent vehicles, surveillance, and advanced driver assistance systems. Finding pedestrian in images is a key ability for those important applications. It is a difficult task from a machine vision perspective. Many approaches propose the problem using a vehicle-based sensor and get significant result with high cost. Other approaches are proposed to detect the pedestrian using classification method with learning algorithms.

2.1.1 Video based monocular pedestrian detection

In the case of imaging sensors, the shape and appearance of the pedestrians can be used to separate them from the background. For this purpose, characteristic features are

extracted from images, and a trained classifier is used to separate pedestrian from the background and other objects. Some of the features used for appearance-based detection are raw sub images [35], size and aspect ratio of bounding boxes [36], Haar wavelets [37], Gabor filter outputs [38], symmetry [37], [39], intensity gradients [31] and their histograms [40], and active contours [41]. In [42], texture information is extracted using simple masks, and classification is performed based on integrating the weak classifiers obtained from these masks. In thermal IR images, pedestrians that are warmer than the background form hot spots, which are used for detection, as in [43]. In [44], features based on histogram, inertia, and contrast is used to distinguish pedestrians.

Motion is also an important cue in detecting pedestrians. In the case of stationary infrastructure-based cameras, background subtraction is used to separate moving objects from static background. However, in the case of moving platforms, the background undergoes ego-motion that depends on the camera motion as well as the scene structure. For laterally moving pedestrians, it is usually feasible to separate the pedestrian motion from ego-motion. However, for longitudinally moving pedestrians, the image motion is parallel to the ego-motion and, therefore, difficult to separate. The vehicle ego-motion can be split into rotation and translation. Rotational motion in video does not depend on the distance of the scene feature and is sometimes neglected [45], [46] or compensated for using gyro sensors [47]. The translational motion is inversely proportional to the distance to the scene and, hence, can be used in determining the scene structure. In the absence of rotational motion, the image motion vectors converge at a single point in the image called the focus of expansion. In [48], ego-motion estimation is performed using sparse optical flow at corner-like features. Motion of outliers corresponding to independently moving objects does not pass through focus of expansion and are clustered using region-growing segmentation on the residual image. In [47], a two-stage stereo correspondence and motion-detection procedure is developed to distinguish an object motion that is inconsistent with the background. This procedure does not need explicit ego-motion computation. Motion information can also be combined with texture information, as in [42]. An extremely efficient representation of image motion is developed based on five types of shifted image differences.

Features characteristic to periodicity of human body motion which is within a frequency range are useful in detecting pedestrians and separating them from other moving and stationary objects. Spatial motion distribution represented by moment features [48], power spectral distribution of the motion time series [48], symmetry characteristics of the legs [49], and gait patterns [50] are some of the cues used to detect and verify pedestrian candidates.

Table. 2.1 Comparison between different sensor for pedestrian detection

Sensor type	Field view	Detection range	Hardware cost	Algorithmic complexity	Illumination
Rectilinear camera	Medium	Low	Low	High	Passive reflective, needs ambient light
Omni camera	Large	Low	Medium	High	Passive reflective, needs ambient light
Near IR	Medium	Medium	Low	High	Active, work in dark
Thermal IR	Medium	Low	High	Medium	Emissive, work in dark
PMD sensor	Medium	Medium	Medium	Medium	Modulated light source
RADAR	Small	High	Medium	Low	Active, work in dark, fog, rain
LASER scaner	Large	Medium	High	Low	Active, work in dark

Stereo cameras as well as time-of-flight sensors return information from which the distance of the object from the camera can be computed. This information is very useful for disambiguating pedestrians from background, handling occlusion between pedestrians, and eliminating extraneous features based on the image size. Disparity discontinuities can be used to aid segmentation, as in [27] and [31], to divide the image of the scene into layers. In [43], stereo is used to guide the active contour model for pedestrians.

LASER scanners output radial distance at discrete azimuth angles in the scanning plane. These data are clustered into objects based on range discontinuities [49] and grouping measurements near each other in the 3-D space [50], [51]. The objects are tracked and classified into number of classes using models of object outlines and their dynamic behavior. The system also warns the driver or activates automatic braking in case of imminent collision.

2.1.2 Sensing technology for pedestrian detection

Various types of sensors have been employed for vehicles as well as infrastructure-based pedestrian detection systems. Commonly used sensors for detecting pedestrians are imaging sensors in various configurations using visible light and

infrared (IR) radiation, as well as the "time-of-flight" sensors such as RADARs and LASER scanners. Imaging sensors can capture a high-resolution perspective view of the scene, but extracting information involves substantial amount of processing.

Every sensor has its advantages and limitations (see **Table2.1**). In order to enhance the advantages and overcome the limitations, one can use a combination of multiple sensors that give complementary information.

2.1.2.1 Sensor mounting

Sensors are mounted on vehicles or embedded within the infrastructure. Vehicle-mounted sensors are very useful in detecting pedestrians and other vehicles around the host vehicle. However, they often cannot see dangerous objects that are occluded by other vehicles or stationary structures. Sensors mounted in infrastructure would be able to see many of these objects and help to get a better view of the entire scene from the top.

Sensors mounted in the front are used for detecting pedestrians ahead of the vehicle. On the other hand, side-mounted sensors cover blind spots. RADARs are often mounted in front of the vehicle to estimate the distance to the pedestrians. LASER scanners have a wide field of view and can be mounted in the front or sides to observe ahead of the vehicle as well as in blind spots.

2.1.2.2 Stereo segmentation

Stereo imagery is a natural approach to segment the object. For example, stereo is one of the main cues used to detect obstacles in human vision system. Thus, at a first glance, stereo segmentation seems to be promising method to generate candidates. However, from our experience, the single use of stereo is far from being useful for this task.

The existed some papers that claim the use of stereo-based segmentation as standalone method to perform the candidate generation [33, 34, 35], but not many of them give details regarding the algorithm employed.

In this thesis we do not consider stand-alone stereo techniques as a reliable algorithm to extract candidates, so it will not be evaluated. More details about sensing technology for pedestrian detection please refer to the appendix.

2.1.3 Features and learning algorithms

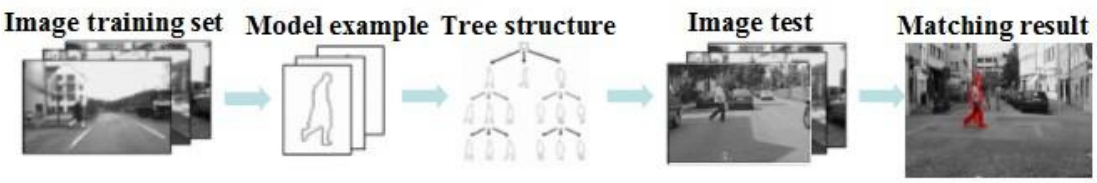


Fig. 2.1 Shape modeled for pedestrian detection

Pedestrian classification involves pedestrian appearance models, using various spatial and temporal cues. Following a rough categorization of such models into generative and discriminative models [64], we further introduce a delineation in terms of visual features and classification techniques. In both the generative and discriminative approaches to pedestrian classification, a given image is to be assigned to either the pedestrian or non-pedestrian class, depending on the corresponding class posterior probabilities. The main difference between generative and discriminative models is how posterior probabilities are estimated for each class.

Generative approaches to pedestrian classification model the appearance of the pedestrian class in terms of its class conditional density function. In combination with the class priors, the posterior probability for the pedestrian class can be inferred using a Bayesian approach.

2.1.3.1 Shape model

Shape cues are particularly attractive because of their property of reducing variations in pedestrian appearance due to lighting or clothing. At this point, we omit discussion of complex 3D human shape models [52] and focus on 2D pedestrian shape models that are commonly learned from shape contour examples. More details about 3D human shape model please refer to the appendix. In this regard, both discrete and continuous representations have been introduced to model the shape space.

Discrete approaches represent the shape manifold by a set of exemplar shapes [53], [54], [55], [56]. On one hand, exemplar-based models imply a high specificity since only plausible shape examples are included and changes of topology need not be explicitly modeled. On the other hand, such models require a large amount of example shapes (up to many thousands) to sufficiently cover the shape space due to transformations and intra-class variance. From a practical point of view, exemplar-based

models have to strike a balance between specificity and compactness to be used in real-world applications, particularly with regard to storage constraints and feasible online matching. Efficient matching techniques based on distance-transforms have been combined with pre-computed hierarchical structures, to allow for real-time online matching of many thousands of exemplars [53], [54], [55].

Continuous shape models involve a compact parametric representation of the class-conditional density, learned from a set of training shapes, given the existence of an appropriate manual [57], [58], [59] or automatic [60], [61], [62], [63], [64] shape registration method. Linear shape space representations which model the class-conditional density as a single Gaussian have been employed by Baumberg [60] and Bergtholdt et al. [65]. Forcing topologically diverse shapes into a single linear model may result in many intermediate model instantiations that are physically implausible. To recover physically plausible regions in the linear model space, conditional density models have been proposed [57], [62]. Further, nonlinear extensions have been introduced at the cost of requiring a larger number of training shapes to cope with the higher model complexity [57], [58], [59], [62], [64]. Rather than modeling the nonlinearity explicitly, most approaches break up the nonlinear shape space into piecewise linear patches. Techniques to determine these local sub-regions include fitting a mixture of Gaussians via the EM-algorithm [57] and K-means clustering in shape space [58], [59], [62], [64].

Fig.2.1 shows the example of shape model for pedestrian detection. This method using a forward rendering model to predict the images which is expensive and requires a good initialization, the computational cost is extremely high.

2.1.3.2 Generative model

Compared to discrete shape models, continuous generative models can fill gaps in the shape representation using interpolation. However, online matching proves to be more complex since recovering an estimate of the maximum-aposteriori model parameters involves iterative parameter estimation techniques, i.e., Active Contours [57], [64].

Recently, a two-layer statistical field model has been proposed to increase the robustness of shape representations to partial occlusions and background clutter by representing shapes as a distributed connected model [66]. Here, a hidden Markov field layer to capture the shape prior is combined with an observation layer, which associates shape with the likelihood of image observations.

One way to enrich the representation is to combine shape and texture information

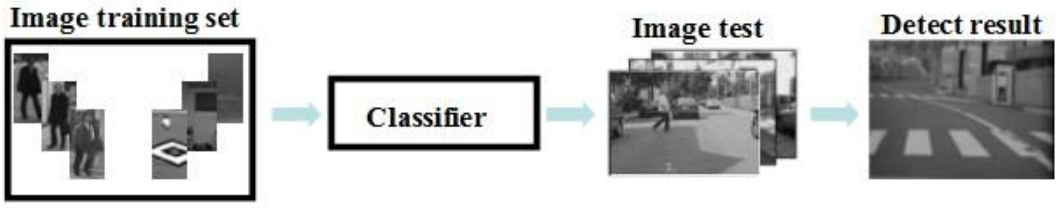


Fig. 2.2 Discriminative model for pedestrian detection

within a compound parametric appearance model [57], [62], [63], [67], [68]. These approaches involve separate statistical models for shape and intensity variations. A linear intensity model is built from shape-normalized examples guided by sparse [65], [62], [68] or dense correspondences [63], [67]. Model fitting requires joint estimation of shape and texture parameters using iterative error minimization schemes [63], [68]. To reduce the complexity of parameter estimation, the relation of the fitting error and associated model parameters can be learned from examples [57].

2.1.3.3 Discriminative model and features

In contrast to the generative models, discriminative models approximate the Bayesian maximum-a-posteriori decision by learning the parameters of a discriminant function, that is, decision boundary between the pedestrian and non-pedestrian classes from training examples. Fig.2.2 shows an example of discriminative model for pedestrian detection. We will discuss the merits and drawbacks of several feature representations and continue with a review of classifier architectures and techniques to break down the complexity of the pedestrian class.

Local filters operating on pixel intensities are a frequently used feature set [69]. Haar wavelet features have been popularized by Papageorgiou and Poggio [38] and adapted by many others [70], [71], [44]. This over complete feature dictionary represents local intensity differences at various locations, scales, and orientations. Their simplicity and fast evaluation using integral images [44], [72] contributed to the popularity of Haar wavelet features. However, the many-times redundant representation, due to overlapping spatial shifts, requires mechanisms to select the most appropriate subset of features out of the vast amount of possible features. Initially, this selection was manually designed for the pedestrian class, by incorporating prior knowledge about the geometric configuration of the human body [38], [70], [71]. Later, automatic feature

selection procedures (i.e., variants of AdaBoost [73]) were employed to select the most discriminative feature subset [44]. Because the Haar-Like features are valued as pixel values directly based the rectangular features, it show a good ability to describe the features of the various parts of the human body and Calculation of Haar-Like features is very fast.

The automatic extraction of a subset of non-adaptive features can be regarded as optimizing the features for the classification task. Likewise, the particular configuration of spatial features has been included in the actual optimization itself, yielding feature sets that adapt to the underlying data set during training. Such features are referred to as local receptive fields [54], [74], [75], [36], [76], in reference to neural structures in the human visual cortex [77]. Recent studies have empirically demonstrated the superiority of adaptive local receptive field features over non adaptive Haar wavelet features with regard to pedestrian classification [75], [36].

Another class of local intensity-based features is codebook feature patches, extracted around interesting points in the image [78], [79], [80], [81]. A codebook of distinctive object feature patches along with geometrical relations is learned from training data followed by clustering in the space of feature patches to obtain a compact representation of the underlying pedestrian class. Based on this representation, feature vectors have been extracted including information about the presence and geometric relation of codebook patches.

Others have focused on discontinuities in the image brightness function in terms of models of local edge structure. Well-normalized image gradient orientation histograms, computed over local image blocks, have become popular in both dense [82], [28], [83], (HOG, histograms of oriented gradients) and sparse representations [84] (SIFT, scale-invariant feature transform), where sparseness arises from preprocessing with an interest-point detector. Initially, dense gradient orientation histograms were computed using local image blocks at a single fixed scale [82], [28] to limit the dimensionality of the feature vector and computational costs. Extensions to variable-sized blocks have been presented in [83], [85], [86]. Results indicate a performance improvement over the original HOG approach. Recently, local spatial variation and correlation of gradient based features have been encoded using covariance matrix descriptors which increase robustness toward illumination changes [87]. This approach has become increasingly popular in the domain of pedestrian classification in both linear and non-linear variants. However, resulting performance boosts are paid for with a significant increase in computational costs and memory.

Yet others have designed local shape filters that explicitly incorporate the spatial

configuration of salient edge-like structures. Multi-scale features based on horizontal and vertical co-occurrence groups of dominant gradient orientation have been introduced by Mikolajczyk et al. [88]. Manually designed sets of edgelets, representing local line or curve segments, have been proposed to capture edge structure [89]. An extension to these predefined edgelet features has recently been introduced with regard to adapting the local edgelet features to the underlying image data [90]. So-called shapelet features are assembled from low-level oriented gradient responses using AdaBoost, to yield more discriminative local features. Again, variants of AdaBoost are frequently used to select the most discriminative subset of features.

As an extension to spatial features, spatiotemporal features have been proposed to capture human motion [91], [92], [65], [44], especially gait [93], [94], [95]. For example, Haar wavelets and local shape filters have been extended to the temporal domain by incorporating intensity differences over time [65], [44]. Local receptive field features have been generalized to spatiotemporal receptive fields [93], [76]. HOGs have been extended to histograms of differential optical flow [103]. Several papers compared the performance of otherwise identical spatial and spatiotemporal features [91], [44] and reported superior performance of the latter at the drawback of requiring temporally aligned training samples.

2.1.3.4 Discriminative classification

Discriminative classification techniques aim at determining an optimal decision boundary between pattern classes in a feature space. Feed-forward multilayer neural networks [96] implement linear discriminant functions in the feature space in which input patterns have been mapped nonlinearly (e.g., by using the previously described feature sets). Optimality of the decision boundary is assessed by minimizing an error criterion with respect to the network parameters (i.e., mean squared error [96]). In the context of pedestrian detection, multilayer neural networks have been applied particularly in conjunction with adaptive local receptive field features as nonlinearities in the hidden network layer [74], [54], [75], [36], [76]. This architecture unifies feature extraction and classification within a single model.

Support Vector Machines (SVMs) have evolved as a powerful tool to solve pattern classification problems. In contrast to neural networks, SVMs do not minimize some artificial error metric but maximize the margin of a linear decision boundary (hyper plane) to achieve maximum separation between the object classes. Regarding pedestrian classification, linear SVM classifiers have been used in combination with various

(nonlinear) feature sets [82], [91], [97], [83], [71], [85], [86].

Nonlinear SVM classification (e.g., using polynomial or radial basis function kernels as implicit mapping of the samples into a higher dimensional (and probably infinite) space), yielded further performance boosts. These are, however, paid for with a significant increase in computational costs and memory requirements [98], [70], [75], [89], [38], [36].

AdaBoost [73], which has been applied as automatic feature selection procedure, has also been used to construct strong classifiers as weighted linear combinations of the selected weak classifiers, each involving a threshold on a single feature [90], [28]. To incorporate nonlinearities and speed up the classification process, boosted detector cascades have been introduced by Viola et al. [44] and adopted by many others [88], [89], [83], [87], [101], [85], [86]. Motivated by the fact that the majority of detection windows in an image are non-pedestrians, the cascade structure is tuned to detect almost all pedestrians while rejecting non-pedestrians as early as possible. AdaBoost is used in each layer to iteratively construct a strong classifier guided by user-specified performance criteria. During training, each layer focuses on the errors the previous layers make. As a result, the whole cascade consists of increasingly more complex detectors. This contributes to the high processing speed of the cascade approach, since usually only a few feature evaluations in the early cascade layers are necessary to quickly reject non-pedestrian examples. Adaboost algorithm has been proved very efficient.

Besides introducing new feature sets and classification techniques, many recent pedestrian detection approaches attempt to break down the complex appearance of the pedestrian class into manageable subparts. First, a mixture-of-experts strategy establishes local pose-specific pedestrian clusters, followed by the training of a specialized expert classifier for each subspace [54], [97], [28], [88], [85]. Appropriate pose-based clustering involves both manually [97], [28], [89] and automatically established [71] mutually exclusive clusters, as well as soft clustering approaches using probabilistic assignment of pedestrian examples to pose clusters, obtained by a preprocessing step, e.g., shape matching [54]. Matching method paid for with a significant increase in computational costs and memory requirements.

An additional issue in mixture-of-experts architectures is how to integrate the individual expert responses to a final decision. Usually, all experts are run in parallel, where the final decision is obtained as a combination of local expert responses using techniques such as maximum selection [97], [89], majority voting [71], AdaBoost [28], trajectory based data association [85], and probabilistic shape-based weighting [54].

Second, component-based approaches decompose pedestrian appearance into parts. These parts are either semantically motivated (body parts such as head, torso, and legs) [98], [88], [70], [28], [65], [89] or concern codebook representations [81], [82], [89], [83]. A general trade-off is involved at the choice of the number and selection of the individual parts. On one hand, components should have as small spatial extent as possible, to succinctly capture articulated motion. On the other hand, components should have sufficiently large spatial extent to contain discriminative visual structure to allow reliable detection. Part-based approaches require assembly techniques to integrate the local part responses to a final detection, constrained by spatial relations among the parts.

Approaches using partitions into semantic sub-regions train a discriminative feature-based classifier (see above), (specific to a single part), along with a model for geometric relations between parts. Techniques to assemble part-based detection responses to a final classification result include the training of a combination classifier [98], [70], [28] and probabilistic inference to determine the most likely object configuration given the observed image features [88], [65], [89]. Codebook approaches represent pedestrians in a bottom-up fashion as assemblies of local codebook features, extracted around salient points in the image, combined with top-down verification [81], [82], [83].

Component-based approaches have certain advantages compared to full-body classification. They do not suffer from the unfavorable complexity related to the number of training examples necessary to adequately cover the set of possible appearances. Furthermore, the expectation of missing parts due to scene occlusions or inter object occlusions is easier addressed, particularly if explicit inter object occlusion reasoning is incorporated into the model [81], [82], [83], [101]. However, these advantages are paid for with higher complexity in both model generation and application. Their applicability to lower resolution images is limited since each component detector requires a certain spatial support for robustness.

2.2 Human pose estimation

Estimating human pose is an extremely challenging problem. A successful tracking system can find applications in motion capture, human computer interaction, and activity recognition.

2.2.1 Low-level image human pose observations

Prior to 3D human pose estimation, low-level image observations must be extracted from images or video frames. Ideally, the extracted image observations should encode salient information for subsequent use in high-level image understanding tasks. In this context, high-level understanding tasks here refer to 3D body pose estimation. Although in theory, original images or video sequences could be used as image observations, they are too "noisy" to be of any use for high level understanding tasks. Such noisy information originated from lighting variations, different clothing, cluttered background can seriously undermine high level understanding tasks. Therefore, low-level image observations to be extracted during feature extraction process is highly task-specific. It is common computer vision practice to make assumptions about the environment, acquisition, image generation process in order to simplify feature extraction process. In statistical perspectives, image observations extracted must have strong correlation to the problems at hand so that variations of estimation can be minimized. Often in 3D body pose estimation problem, commonly used image observations are silhouette, shapes, edges, motions, colors, and recent approaches are combinations of them.

Enzweiler et al. presented a novel integrated framework method for single-frame pedestrian classification and orientation estimation, and showed a significant performance between pedestrian classification and orientation estimation research areas. However, they only classified four orientations: front/back and left/right. In the real traffic world, not only the left/right view, but also diagonal views of 45º, 135º, 225º, and 315º are dangerous situations for collisions between vehicles and pedestrians. Furthermore, for developing a video based pedestrian protection system, eight orientations are more valuable. For example, when a pedestrian is walking forward and wants to turn right to cross the road, the pedestrian orientation should be changed from 90º to 0º; if the driver notices that the pedestrian is changing his orientation to the intermediate 45º, he will realize that the possibility is getting higher that the pedestrian will cross the road. Estimation of eight orientations will help efficiently to relay the information to the driver to take preventive actions. That is our reason and motivation of tackling a more difficult eight orientations problem. Comparison with the Enzweiler's method will be introduced in Chapter 4.

2.2.2 3D human pose estimation

Human body pose estimation has recently received great interest from the computer vision community. However, most researches consider 3D human pose estimation techniques [101]. Agarwal and Triggs [110] developed a learning-based method for estimating the 3D body poses of people from monocular images as well as video sequences. The approach uses the histogram of shape context descriptors as feature and various regression methods for estimating the 3D pose. Besides, work in the domain of 3D human pose estimation, a few approaches have recovered an estimate of pedestrian orientation based on 2D lower-resolution images [91], [97]. Cucchiara et al. [141] distinguished among various human postures such as standing, crouching, sitting, and lying down using Probabilistic Projection Maps on 2D silhouettes. However, most of the systems use an accurate silhouette of the pedestrian, which may not always be available, especially from a moving vehicle. More details about 3D human pose estimation please refer to the appendix.

2.3 Pedestrian walking trajectory estimation

For a complete safety system, detection should be followed by prediction of the possibility of collision. The system should relay the information to the driver in efficient and non-distracting manner or to the control system of the vehicle in order to take preventive actions.

Some researches try to estimate the pedestrian trajectory by tracking method. There has been work on the tracking of pedestrians to infer trajectory-level information. One line of research has formulated tracking as frame-by-frame association of detections based on geometry and dynamics without particular pedestrian appearance models [110], [62]. Other approaches utilize pedestrian appearance models coupled with geometry and dynamics [90], [72], [119], [73], [74]. Some approaches furthermore integrate detection and tracking in a Bayesian framework, combining appearance models with an observation density, dynamics, and probabilistic inference of the posterior state density. For this, either single [64], [101] or multiple cues [72], [119], [73] are used.

The integration of multiple cues [134] involves combining separate models for each cue into a joint observation density. The inference of the posterior state density is usually formulated as a recursive filtering process [135]. Particle filters [136] are very popular due to their ability to closely approximate complex real-world multimodal posterior densities using sets of weighted random samples. Extensions that are especially relevant for pedestrian tracking involve hybrid discrete/continuous

state-spaces [72] and efficient sampling strategies [136], [137], [138], [139].

An important issue in real-world pedestrian tracking problems is how to deal with multiple targets in the image. Two basic strategies with regard to the tracking of multiple objects have been proposed. First, the theoretically most sound approach is to construct a joint state-space involving the number of targets and their configurations which are inferred in parallel. Problems arise regarding the significantly increased and variable dimensionality of the state space. Solutions to reduce the computational complexity have involved grid-based or pre-calculated likelihoods [138] and sophisticated re-sampling techniques such as Metropolis-Hastings sampling [138], partitioned sampling [139], or annealed particle filters [136]. Second, some approaches have been proposed to limit the number of objects to one per tracker and employ multiple tracker instances instead [72]. While this technique simplifies the state-space representation, a method for initializing a track along with rules to separate neighboring tracks is required. Typically, an independent detector process is employed to initialize a new track.

Incorporating the independent detector into the proposal density tends to increase robustness by guiding the particle re-sampling toward candidate image regions. Competition rules between multiple tracker instances have been formulated in terms of heuristics [72]. In contrast to joint state-space approaches, the quality of tracking is directly dependent on the capability of the associated object detector used for initialization.

In this thesis we try to ignore dynamics and find pedestrian in each frame independently, and estimate these pedestrian orientation in each frame independently. This approach is attractive because it self-starts and is robust to drift, since it essentially re-initializes itself at each frame.

Chapter 3

Advanced Driver Assistance Systems

Pedestrian is the most vulnerable involving the traffic accidents. Our long-term goal is to develop systems which, if not avoid these accidents altogether, at least minimize their severity by employing protective measures in case of upcoming collisions.

Initiatives of pedestrian protection systems have been started to improve the safety of vulnerable road pedestrians. As introduced in Chapter 1, most projects aimed towards the development of sensor-based solutions for the detection of vulnerable road pedestrians, in order to facilitate preventive measures the use of warning or to avoid or minimized the impact of collisions. In our research, we try to realize the pedestrian protection system without any sensor, but using one monocular camera.

3.1 The architecture of whole system

To achieve the goal, we propose a three-stage method: pedestrian detection for single-frame stage, orientation estimation for single-frame stage and walking direction estimation for multi-frame stage. **Fig.3.1** shows the architecture in a general pedestrian protection system. The processing is organized as a pyramid, with base having large quantity of raw data. As one climbs up the pyramid, the useful information is distilled in successive stages, until finally, one takes action based on a yes/no decision. In this thesis, we put emphasis on the base of three steps: pedestrian detection, pedestrian orientation estimation, and pedestrian walking direction estimation. Collision prediction and action process will be future issues.

3.1.1 Stage 1: Pedestrian detection

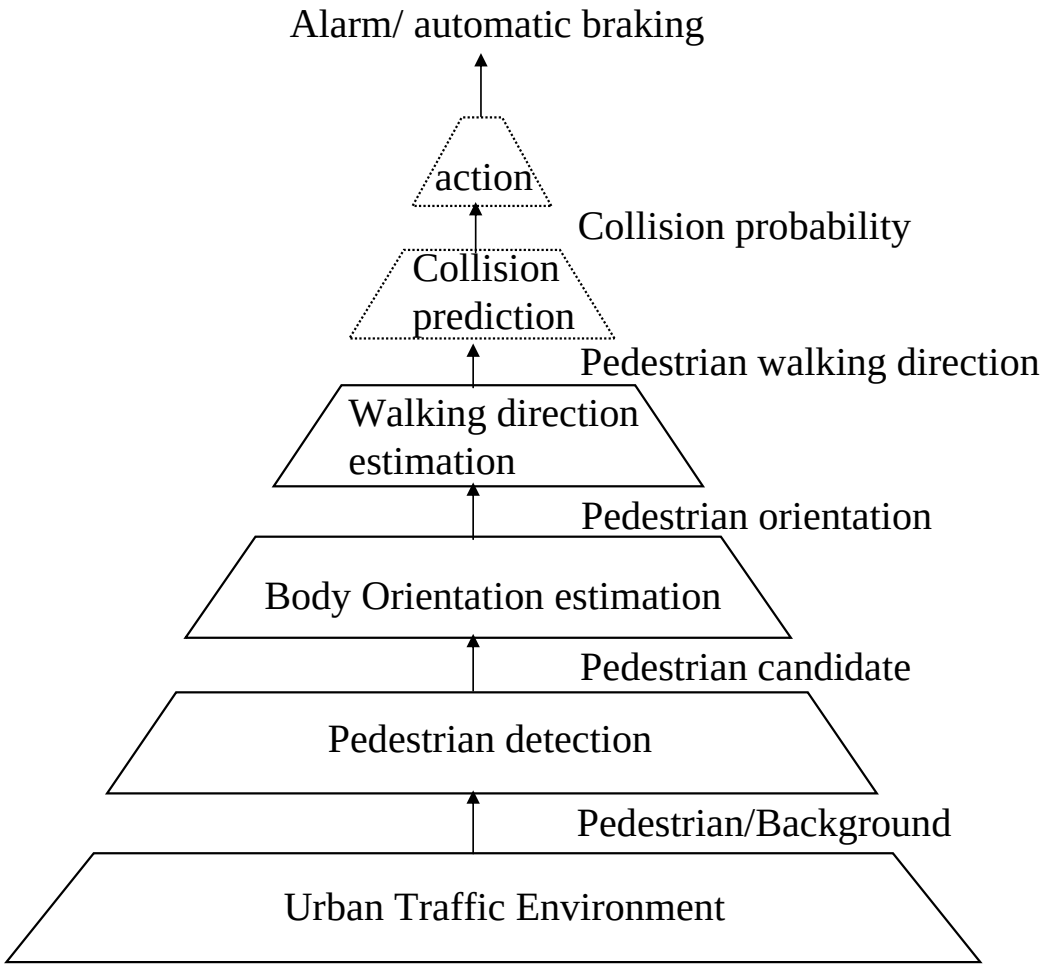


Fig. 3.1 The architecture of whole system.

Pedestrian detection is an important component of the pedestrian protection systems. This stage processes raw data using simple cues and fast algorithms to identify potential pedestrian candidate; in another word, the task of this stage is to find and locate the pedestrian in motion from the complicated traffic environment. The stage requires high detection rate even at the expense of allowing false alarms. If we inform the driver all the pedestrians that were detected in this stage, the driver needs to decide the situation which pedestrian has the most dangerous and probability of the collision to the vehicle by driver himself. It easily causes the problem of that driver cannot pay attention to driving, while exceeding information processing of danger decisions. Therefore, stage applies more complex algorithms to the candidates given by stage 1 to separate genuine pedestrians from false alarms is needed. The pedestrian orientation information also can potentially sort out less-danger candidate and improve the prediction of future

Frame1 **Frame 51**

Fig. 3.2 Urban traffic scene

trajectories that the pedestrian may take.

A major complication is that, because of the moving vehicle, one does not have the luxury to use simple background subtraction methods to obtain a foreground region containing the human (**see Fig.3.2**). There are hard real-time requirements for the vehicle application which rule out any brute-force approaches. Furthermore, because of the moving vehicle, simple tracking method such as optical-flow cannot estimate the pedestrian walking direction efficiently in image plane. When the pedestrian is moving, her/his motion direction can be inferred as well as her orientation from past trajectories using a tracking approach. However, when the pedestrian is static, the motion direction is not defined, but the person is more likely in the future to move in the direction she is facing. Therefore, orientation information can improve the prediction of future trajectories that the pedestrian may take efficiently.

3.1.2 Stage 2: Body orientation estimation

The second stage is to estimate the pedestrian orientation for single-frame based on the result from previous pedestrian detection stage. In this stage, more complex algorithms are used to estimate the orientation of each pedestrian candidate obtained from stage 1. In this stage we aim to differentiate eight orientations which is evenly divided a turn

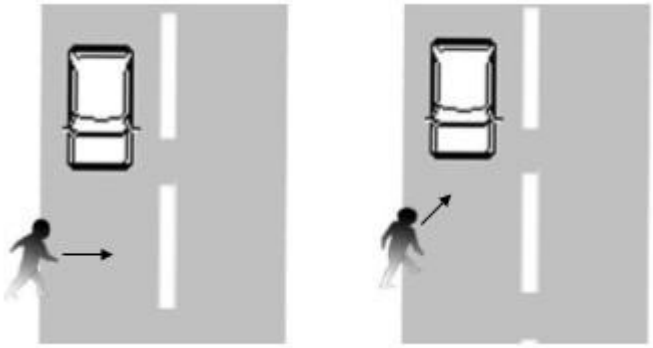


Fig. 3.3 More dangerous diagonal direction.

with every 45 degree. Enzweiler et al. [140] only estimated front/back and left/right four orientations. In the real traffic world, not only the left/right view, but also diagonal views of 45º, 135º, 225º, and 315º are dangerous situations for collisions between vehicles and pedestrians (see **Fig.3.3**).

Furthermore, for developing a video based pedestrian protection system, eight orientations are more valuable. For example, when a pedestrian is walking forward and wants to turn right to cross the road, the pedestrian orientation should be changed from 90º to 0º; if the driver notices that the pedestrian is changing his orientation to the intermediate 45º, he will realize that the possibility that is getting higher the pedestrian will cross the road (see **Fig.3.4**). Assume that estimate 16 orientation every 22.5º, it is obvious that more sudden orientation change can be estimated earlier which is benefited from such 16 orientation estimation. But at the same time, it makes the problem more difficult and should be paid for with a significant increase in computational costs and memory. Multiple orientation estimation is one of multi-class classification problems. In general, the more the number of class increase, the less the accuracy becomes. We assume that estimating of eight orientations will help efficiently relay the information to the driver to take preventive actions. That is our reason and motivation of tackling a sufficiently difficult eight orientations problem. However, increasing the classes of orientation is a scientifically interesting issue in the future.

3.1.3 Stage 3: Pedestrian walking direction estimation

Single-frame orientation estimation allows us to recover pedestrian headings without integration over time and static pedestrians can also be estimated without posing any

Fig. 3.4 Intermediate (diagonal) orientation provide early warning.

problem. But for practical application, the video-based pedestrian protection systems are necessary. So the stage 3 is to estimate the walking direction for multi-frame. In general moving object recognition, usually using the method which exploits the object tracking algorithm and gives a trace on the image coordinate space; however, we need to compute the trace's projection onto the driving surface coordinate system from a moving camera image space. Therefore, in this stage, we estimate the pedestrian walking direction for multi-frame using the result of orientation estimation in a single frame which was obtained from stage 2. It can correct the result in stage 2 effectively and can improve the prediction of future trajectories that the pedestrian may take efficiently.

The following will describe stage 1 as the important component of pedestrian protection systems. Stage 2 and stage 3 will be introduced in chapter 4 and chapter 5.

3.2 Pedestrian detection

Pedestrian detection is a difficult task from a machine vision perspective as described in

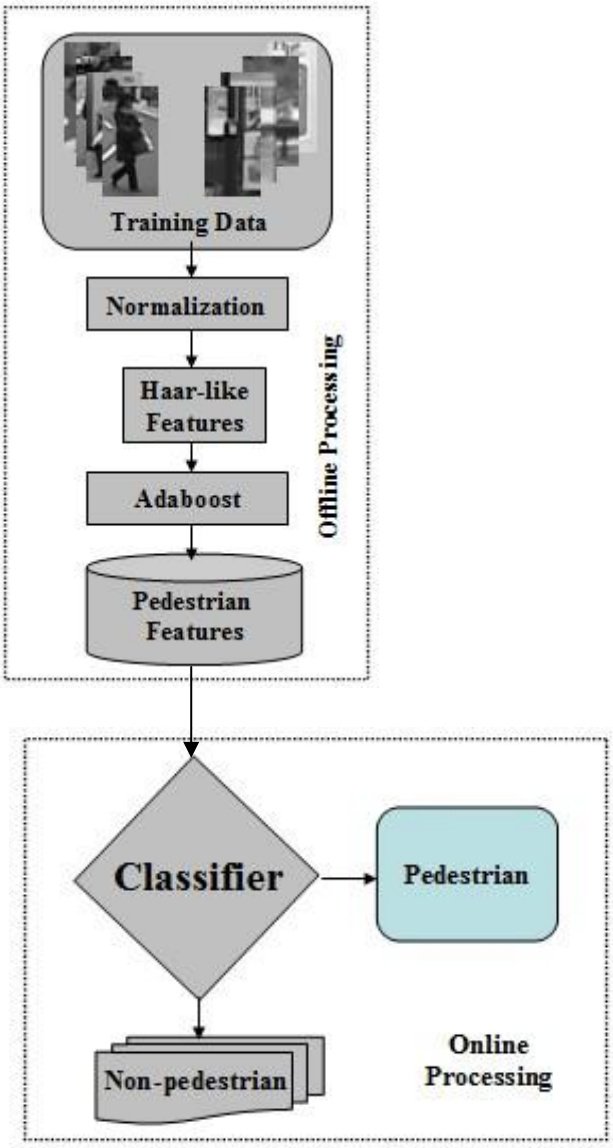


Fig. 3.5 Flowchart of the pedestrian detection system.

Chapter 2. The lack of explicit and general human walking models leads to the use of machine learning techniques, where an implicit representation is learned from examples. As such, it is an instantiation of the multiclass object categorization problem. Yet the pedestrian detection task has some of its own characteristics, which can influence the methods of choice. Foremost, there is the wide range of possible pedestrian appearance, due to changing articulated pose, clothing, lighting, and background. The detection component is typically part of a system situated in a physical environment, which means that prior scene knowledge (camera calibration, ground plane constraint) is often available to improve performance.

Finding people in images is a key ability in a variety of important applications. In this section, we are concerned with where the human body to be detected covers a smaller portion of the image (i.e., is visible at lower resolution). This implies outdoor settings with the moving vehicles, where an onboard camera watches the road ahead of possible collisions with pedestrians. More detailed recognition tasks such as human orientation estimation and walking direction estimation will be presented later.

3.2.1 Cost efficient approach

As introduced in Chapter 2, many researches for pedestrian detection have been proposed. Shape cues are widely used because of their robustness to variation in pedestrian appearance due to lighting or clothing, but these approaches require data normalization methods and, involve a significantly increased feature space.

Histogram of Oriented Gradient based features and Support Vector Machines as classifier architectures for building a pedestrian detector. This approach has become increasingly popular in the domain of pedestrian classification in both linear and non-linear variants. However, resulting performance boosts are paid for with a significant increase in computational costs and memory.

The framework for object detection proposed by Viola and Jones [42]has also proved to be very efficient especially for human faces. This method's basic idea is to select weak features, e.g., Haar wavelet, by adaboost to build a cascade structured strong classifier.

To improve pedestrian classification performance, several approaches have attempted to use video sequences and apply background subtraction to reduce clutter. Human motion information can be a rich source, as shown by Viola et al. Their motion features proved the most discriminative. However, their work is restricted to a static camera setting; we would like to realize a system with a moving vehicle.

Several approaches have attempted to establish local pose-specific clusters, followed by the training of specialized classifiers for each subspace. The final decision of the classifier ensemble involves maximum-selection. Approaches that perform object classification using multiple cameras at different view points are also relevant to our current work. In this paper, we only use a single camera.

Considering our application that detects anomalous events to warn drivers in real-time, we employed a variant of adaboost both to select a small set of features and to train the classifier because this approach's computation, which is built on the detection work of Viola and Jones, is highly efficient. Sliding-window methods have also been

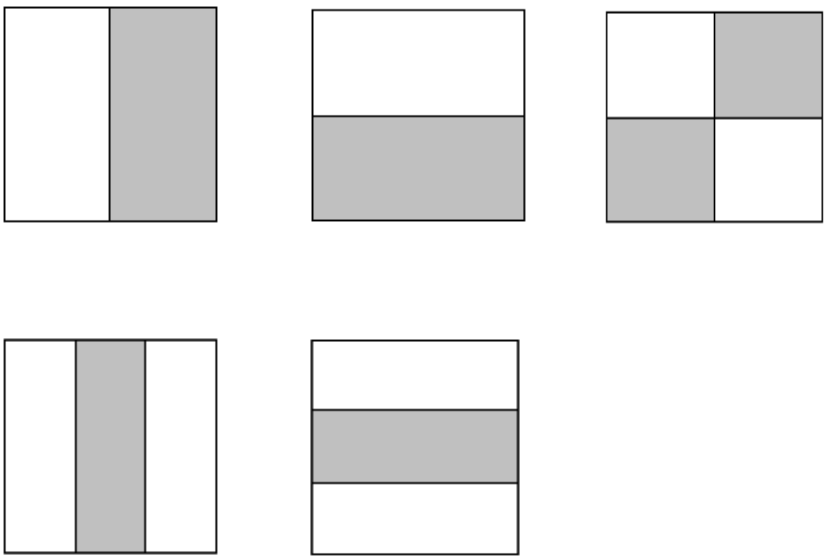


Fig. 3.6 Example rectangle features shown relative to the enclosing detection window. The sum of the pixels values which lie within the white rectangles are subtracted from the sum of pixels values in the grey rectangles.

used to scan the image at all relevant positions and scales to detect pedestrians. **Fig. 3.5** shows the Flowchart of the detection pedestrian system.

3.2.2 Adaboost with haar-like feature

Many approaches have been proposed for pedestrian detection. Considering our application that detects anomalous events to warn drivers in real-time, we employed a variant of adaboost both to select a small set of features and to train the classifier because this approach's computation, which is built on the detection work of Viola and Jones, is highly efficient. Sliding-window methods have also been used to scan the image at all relevant positions and scales to detect pedestrians.

Given feature and training sets of positive and negative samples, a classification function can be learned. The weak learning algorithm is designed to select the single rectangle feature that best separates the positive and negative examples. For each feature, the weak learner determines the optimal threshold classification function, such that the minimum number of examples is misclassified.

The approach of the training cascade classifier uses the method of Viola and Jones. Boosted classifiers can be constructed that reject many of the negative sub-windows while detecting almost all positive instances. A positive result from the first classifier

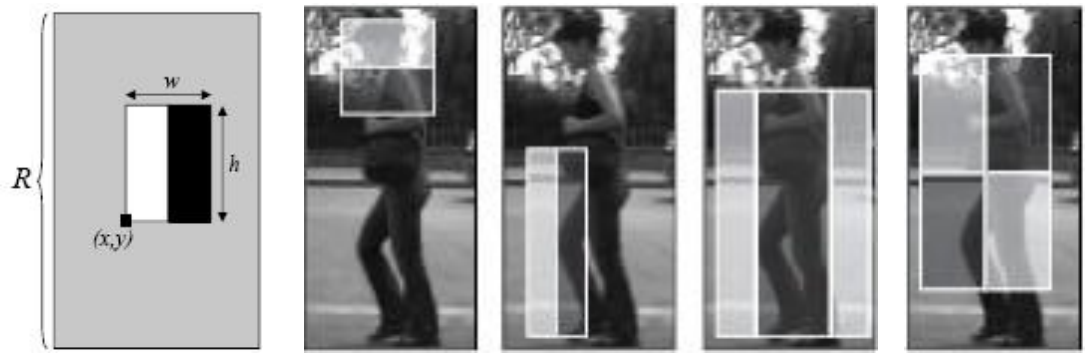


Fig. 3.7 Examples of filters that give high response in regions containing pedestrians.

triggers the evaluation of a second classifier that has also been adjusted to achieve very high detection rates. A positive result from the second classifier triggers a third classifier, and so on. A negative outcome at any point leads to the immediate rejection of the sub-window.

3.2.2.1 Features

Our pedestrian detection procedure classifies images based on the value of simple features. There are many motivations for using features rather than the pixels value directly. The most common reason is that features can act to encode ad-hoc domain knowledge that is difficult to learn using a finite quantity of training data. There is also a second critical motivation for features: the feature based system operates much faster than a pixel-based system.

More specifically, we use three kinds of features. The value of a two-rectangle feature is the difference between the sum of the pixels values within two rectangular regions. The regions have the same size and shape, and are horizontally or vertically adjacent (see **Fig.3.6**). A three-rectangle feature computes the sum of values within two outside rectangles subtracted from the sum in a center rectangle. Finally a four-rectangle feature computes the difference between diagonal pairs of rectangles. Note that unlike the Haar basis, the set of rectangle features is over complete. **Fig.3.7** shows some examples of filters that give high response in regions containing pedestrians. This figure is referred from David Geronimo. Calculation of Haar-Like features is achieved through the integral image. As indicated in **Fig. 3.8 (a)**, a gray image *I* is defined with its integral image *I* as:

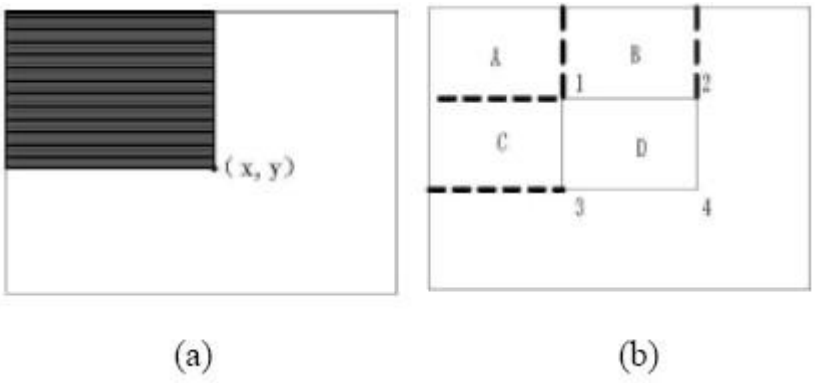


Fig. 3.8 Integral image

$$\tilde{I}(u,v) = \int_{x=0}^{u} \int_{y=0}^{v} I(x,y)dxdy \qquad (1)$$

Where, (x, y) are the pixel point.

In **Fig 3.8(b)**, the sum of pixel gray values in the rectangle D can be achieved from the four points in the integral image. S_1 is the sum of pixel gray values in the rectangle A , S_2 is the sum of pixel gray values in the rectangles $A+B$, S_3 is the sum of pixel gray values in the rectangles $A+C$, S_4 is the sum of pixel gray values in the rectangles $A+B+C+D$. So, S_1, S_2, S_3 and S_4 combined to produce the sum of pixel gray values in their bounded rectangular area D as $S_1+S_4-S_2-S_3$.

For an input image I, the integral image can be obtained by the computation based on the one-time point by point scanning of the original image, which is the sum of all the pixel gray values of the original image (x, y) with the column where the vertical axis does not exceed the point, and the recurrence formula is as below:

$$\begin{cases} s(x,y) = s(x,y-1) + i(x,y) \\ ii(x,y) = ii(x-1,y) + s(x,y) \end{cases} \qquad (2)$$

Multi-scale testing should be implemented by way of the scaling feature template. Even though the search occurs at any scale, this integral image is available. In other

words, the whole testing process requires only scanning the original image once.

Rectangle features are somewhat primitive when compared with alternatives such as steerable filters. Steerable filters, and their relatives, are excellent for the detailed analysis of boundaries, image compression, and texture analysis. In contrast rectangle features, while sensitive to the presence of edges, bars, and other simple image structure, are quite coarse. Unlike steerable filters the only orientations available are vertical, horizontal, and diagonal. The set of rectangle features, however, provides a rich image representation which supports effective learning. In conjunction with the integral image, the efficiency of the rectangle feature set provides ample compensation for their limited flexibility.

3.2.2.2 Learning Classification Functions

Given a feature set and a training set of positive and negative images, any number of machine learning approaches could be used to learn a classification function. In our system a variant of AdaBoost is used both to select a small set of features and train the classifier. In its original form, the AdaBoost learning algorithm is used to boost the classification performance of a simple learning algorithm. There are a number of formal guarantees provided by the AdaBoost learning procedure. A number of results were later proved about generalization performance. The key insight is that generalization performance is related to the margin of the examples, and that AdaBoost achieves large margins rapidly. A very small number of features can be combined to form an effective classifier. The weak learning algorithm is designed to select the single rectangle feature which best separates the positive and negative examples. For each feature, the weak learner determines the optimal threshold classification function, such that the minimum numbers of examples are misclassified.

The key to achieving pedestrian detection is training to acquire the pedestrian detection classifier. Regarding a training set given n training samples, these samples x_i fall in a certain distribution X, while the sample y_i in certain set Y. For pedestrian detection as one or two classification problems, set $\{Y = -1, 1\}$ to have the fake and real samples. The samples to be classified with k simple features is expressed as f_j , where, $1 \le j \le k$; as for x_i , the No i training sample, it is featured with $f_i(x_i)$. The weak classifier $h_j(x)$ for No. j feature consists of a feature value f_j , a threshold θ_j and a bias value p_j for the direction of inequality (given two cases as ±1):

$$h_j(x)=\begin{cases}1 & p_j f_j < p_j \theta_j \\ 0 & otherwise\end{cases} \quad (3)$$

Train process is as follow:

1. Establish the initial weight. Set $\omega_{t,i}$, (i=1,2,…,n) as weighted error of No *i* sample in No *t* times of cycle. Regarding a training set containing *n* samples, in which there are *u* fake samples, and *v* real samples, respectively, corresponding to y_i =0 and y_i =1. In the first cycle, each of the training samples is given the same weight. That is, y_i =0, 1 $\omega_{1,i}$=1/2u; y_i =1, $\omega_{1,i}$ =1/2v.

2. for t=1,…,T

① Normalized weight, enables $\omega_{t,i}$ to be a probability distribution;

$$\omega_{t,i} = \omega_{t,i} / \sum_{j=1}^{n} \omega_{t,j} \quad (4)$$

② Each of the features j is trained with its weak classifier h_j, and calculated on its weighted error;

$$\varepsilon_j = \sum_{j=1}^{n} \omega_{t,j} \left| h_j(x_i) - y_i \right| \quad (5)$$

③ The smallest classifier h_j is selected for weighted error ε_j;

④ Weight is updated, where, if x_i is correctly classified, then e_i =0; otherwise, e_i =1. Whereas $\beta_t=\varepsilon_t$ / 1-ε_t

$$\omega_{t+1,i} = \omega_{t,i} \beta_t^{1-e_i} \quad (6)$$

⑤ Output of the final strong classifier

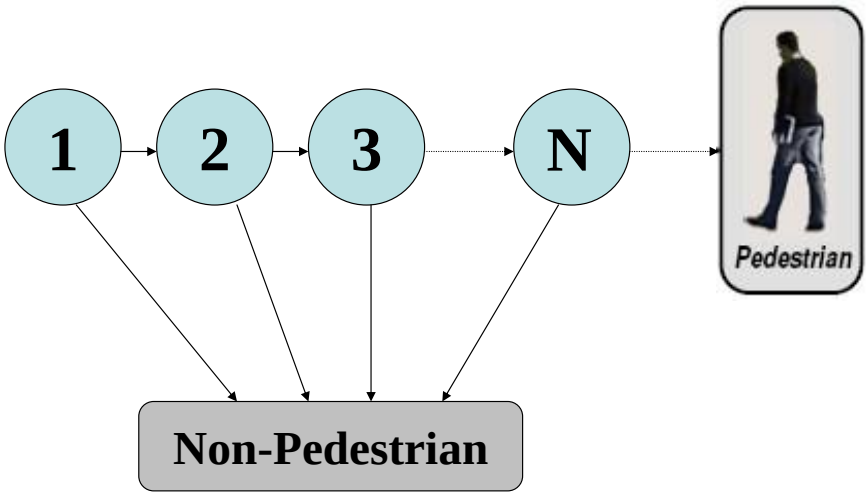


Fig. 3.9 Schematic representation of a detection cascade of classifiers with *N* stages.

$$h_j(x) = \begin{cases} 1 & \sum_{t=1}^{T} a_t h_t(x) \geq \frac{1}{2}\sum_{t=1}^{T} a_t \\ 0 & otherwise \end{cases} \quad (7)$$

Where $a_t = \log\frac{1}{\beta_t}$

3.2.2.3 Cascade classifiers

Cascade of classifiers achieves increased detection performance while radically reducing computation time. The key insight is that smaller, and therefore more efficient, boosted classifiers can be constructed which reject many of the negative sub-windows while detecting almost all positive instances. Simpler classifiers are used to reject the majority of sub-windows before more complex classifiers are called upon to achieve low false positive rates. The overall form of the detection process is that of a degenerate decision tree which called cascade. A positive result from the first classifier triggers the evaluation of a second classifier which has also been adjusted to achieve very high detection rates. A positive result from the second classifier triggers a third classifier, and so on. A negative outcome at any point leads to the immediate rejection of the sub-window.

Fig. 3.10 Positive training samples (top two rows) and negative training samples (bottom two rows)

Stages in the cascade are constructed by training classifiers using AdaBoost and then adjusting the threshold to minimize false negatives. The structure of the cascade reflects the fact that within any single image an overwhelming majority of sub-windows are negative. As such, the cascade attempts to reject as many negatives as possible at the earliest stage possible. While a positive instance will trigger the evaluation of every classifier in the cascade, this is an exceedingly rare event (see **Fig. 3.9**). Subsequent classifiers are trained using those examples which pass through all the previous stages. As a result, the second classifier faces a more difficult task than the first. The examples which make it through the first stage are harder than typical examples. The more difficult examples faced by deeper classifiers push the entire receiver operating characteristic curve downward. At a given detection rate, deeper classifiers have correspondingly higher false positive rates.

3.2.3 Experimental results

In this step, we generate the pedestrian detector; it just used to detect and locate the pedestrian in a still image. The training samples were obtained from the pictures of walking pedestrian in real world with a single camera on a moving vehicle. All training

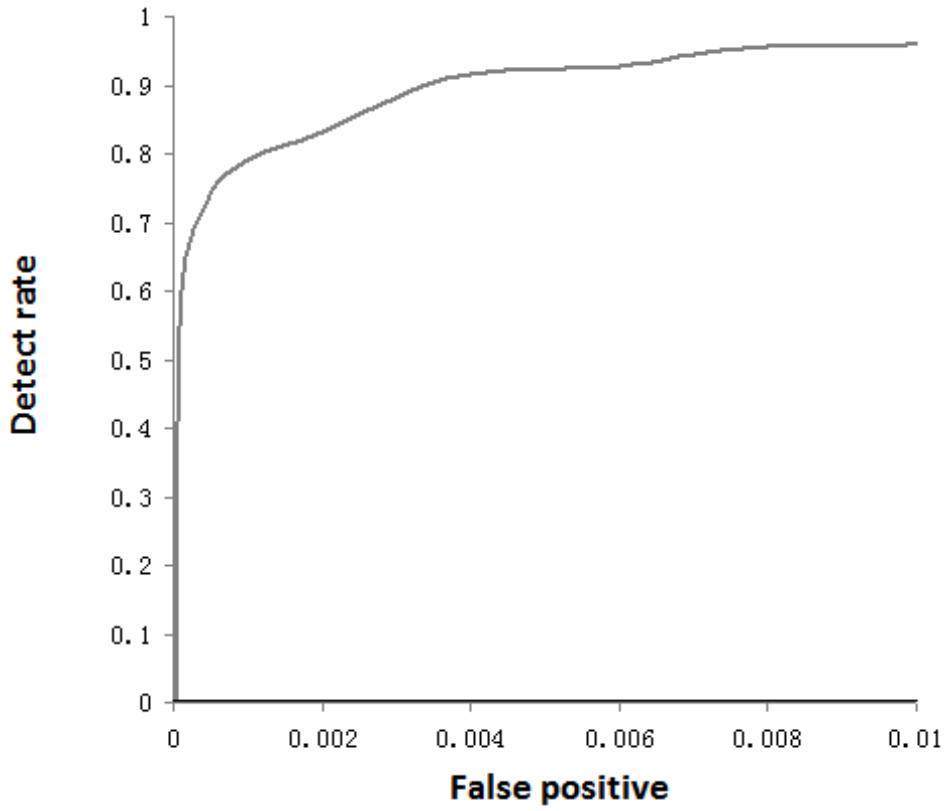


Fig. 3.11 ROC curves of the detector for general pedestrian extraction

samples were normalized to the same size. The size of each training image was 24 x 58 pixels. The general pedestrian classifier was trained with 2800 positive samples and 4100 negative samples. **Fig. 3.10** shows some positive and negative samples.

In the training session above, the general pedestrian classifier was trained in relatively good performance. The learning curve is shown in **Fig. 3.11** in the format of ROC curve. To create the ROC curve the threshold of the final layer classifier is adjusted from negative infinity to infinity. Adjusting the threshold to infinity will yield a detection rate of 0 and a false positive rate of 0. Adjusting the threshold to negative infinity, however, increases both the detection rate and false positive rate, but only to a certain point. Neither rate can be higher than the rate of the detection cascade minus the final layer. In effect, a threshold of negative infinity is equivalent to removing that layer. Further increasing the detection and false positive rates requires decreasing the threshold of the next classifier in the cascade. Thus, in order to construct a complete ROC curve, classifier layers are removed. We use the number of false positives as

opposed to the rate of false positive for the x-axis of the ROC curve to facilitate for the comparison with other systems.

3.3 Conclusion

In this stage, we generated the pedestrian detector based on the conventional Adaboost classifier using Haar-like image features; it just used to detect and locate the pedestrian in a still image. Considering our application that detects anomalous events to warn drivers in real-time, we employed a variant of Adaboost both to select a small set of features and to train the classifier because this approach's computation, which is built on the detection work of Viola and Jones, is highly efficient.

Sliding-window methods have also been used to scan the image at all relevant positions and scales to detect pedestrians.

From the experiment, the general pedestrian classifier was trained in relatively good performance.

There is no originality about employing Adaboost at this stage. However, we extend the algorithm and fully exploiting its simplicity and fast learning and run-time performance, in order to compose an eight-class classifier of body orientation estimation. The detail is introduced in Chapter 4.

Chapter 4

Pedestrian Orientation Estimation

Pedestrian protection systems not only need to detect pedestrians, but also to predict the possibility of collisions between pedestrians and vehicles. The system must efficiently relay information to drivers so that they can take preventive actions.

When the pedestrian is moving, her motion direction can be inferred from past trajectories using a tracking approach. However, when the pedestrian is static, the motion direction is hard to tell specifically, but the person is more likely in the future to move in the direction she is facing. Therefore, orientation information can potentially improve the prediction of future trajectories that the pedestrian may take and improve collision prediction as seen in **Fig. 4.1**. Tracking pedestrian from moving camera images will produce a trajectory on image plane dependent to car-pedestrian relative movement. Tracking approaches also require a certain amount of time. Quick adaptation to sudden changes in movement is crucial. Particularly in intelligent vehicle systems, time is precious, and fast reactions are necessary.

In this Chapter, we present a novel discriminative model based approach to estimate their orientation for single-frame by observing pedestrian movements in a natural environment. We first employ a two-class object recognition algorithm to construct two naïve multi-class recognition algorithms. Second, we derive a novel cascade algorithm that combine a global and a local classifiers by taking account the performances of each sub-class classifier. Finally, we propose an integrated approach of pedestrian body and head orientations.

4.1 Discriminator for pose direction

We describe the algorithm for estimating pedestrian orientations. We use Haar-like

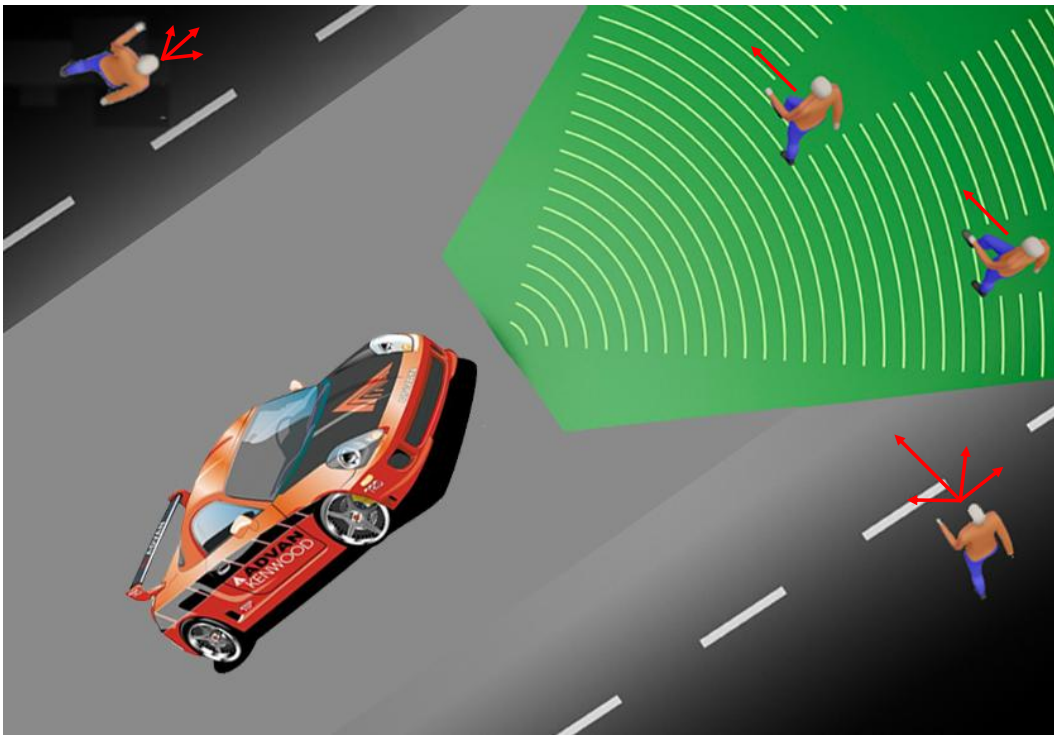


Fig. 4.1 Pedestrian is more likely to move in direction in which she is oriented. Arrows denote probabilities of anticipated movement

features to generate the feature vectors of the input images. Adaboost is used as a classifier in order to estimate the pedestrian orientation. The typical form of adaboost is a two-category classifier that forms a decision boundary between classes. Here, we design an eight-category classifier for eight orientations such as 0º, 45º, 90º,…, 315º.

As shown in **Fig. 4.2**, the orientation is differed counter-clockwise by 45º, and each is labeled by an orientation number 1 to 8. Lateral orientations 1 (0º) and 5 (180º) are often observed on the road crossing in front of a car, while orientations 2, 4, 6, and 8 are on the sidewalk. We experimentally collected training samples of diagonal directional orientations 2, 4, 6 and 8 from scenes of pedestrians on sidewalks without rigid attention to walking direction.

When the pedestrian walks alone the roadside, no doubt, he/she is consider to be safety, like the diagonal orientations 2, 4, 6, and 8. And the lateral orientations 1 and 5 is usually observed on the road crossing in front of the car, we consider the two situations is the most dangerous of the collision between the vehicle and the pedestrian. This issue is addressed later in this Chapter in body-head integration. As all known, pedestrian changes his/her orientation from orientation 1(right) to 3(back), he/she must change to the orientation 2(right-back) first. To improve the prediction of collision in real time, we consider 8 orientations classifying is necessary.

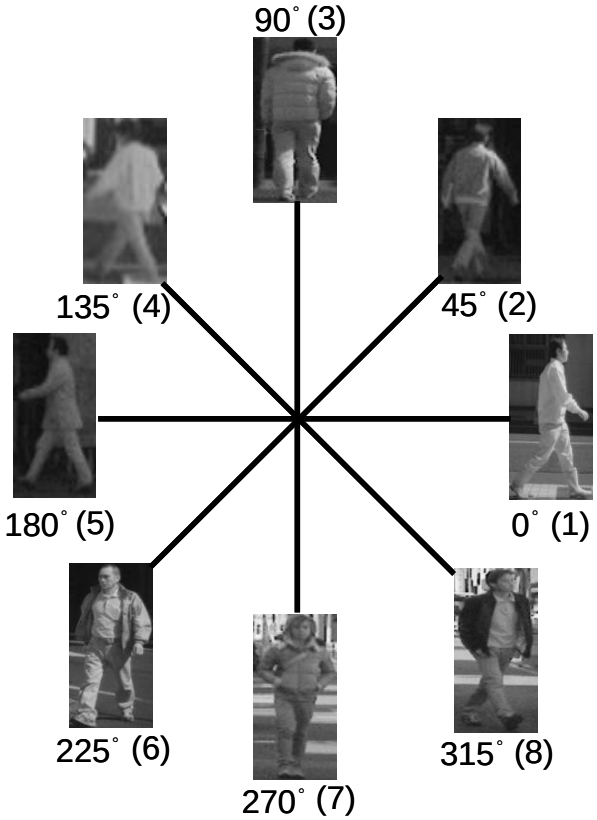


Fig. 4.2 Eight-category classifier for eight walking directions that differ counter-clockwise by 45º. Each number 1 to 8 is given to each orientation as its label

4.2 Multi classification for orientation estimation

Multi-category classification is a difficult problem in machine learning. Even though two-class (binary) classification methods are relatively well-developed, how to effectively extend them for multi-classification is an important on-going research issue.

4.2.1 One versus one classification

One naïve scheme that is particularly worthy of attention is the "all-pairs", or one versus one classification. In this approach, binary classifiers are trained; each classifier separates a pair of classes. This classification has a simple conceptual justification, and can be implemented to train faster and classify quickly.

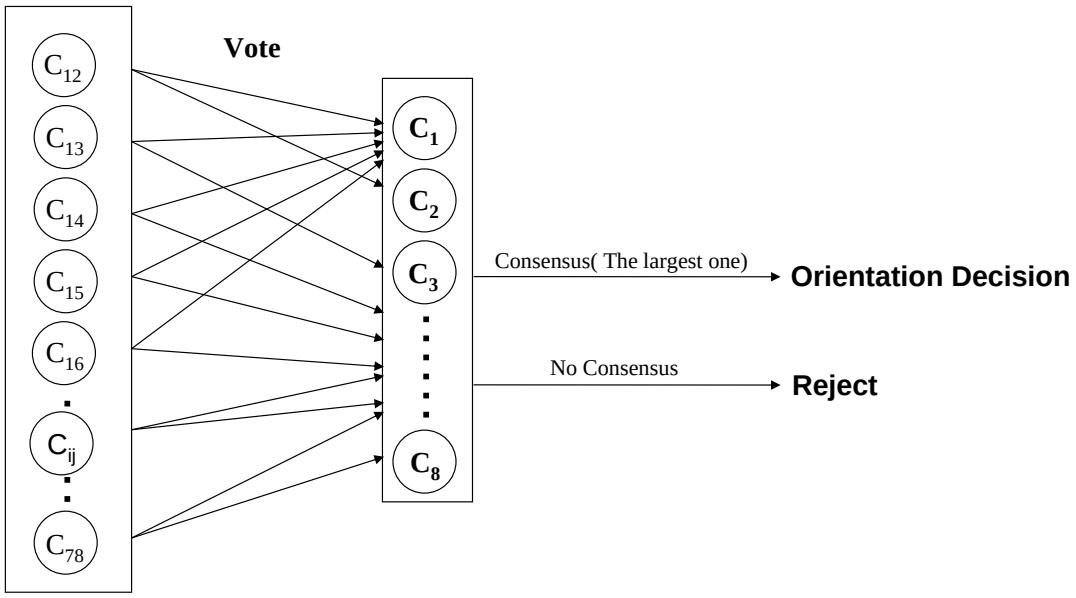


Fig. 4.3 One versus one classification

The one-versus-one method, which is usually implemented using a majority voting strategy, constructs one two-category classifier for every pair of distinct classes for a total of M(M-1)/2 two-category classifiers where M is the number of classes. The classifiers are trained with examples from one class as positive samples and another class as negative samples. For an example, if classifier C_{ij} says "in class i", it votes for class C_i, and otherwise, the vote for class C_j is counted. After M(M-1)/2 classifiers vote, this method classifies it by class with the largest number of votes (see **Fig.4.3**) .

This method has an advantage: we can reject the result when the votes are scattered among categories, but it should employ a large number of classifiers. In the training phase, first, we use one group of images as positive samples and another group as negative samples. For example, C_{ij}(i=1, j=2), that is, C_{12} is the classifier of 0º orientation as positive samples and with 45º direction as negative samples. Total 28 classifiers employing the adaboost algorithm were trained. 28 classifiers vote, and we choose the classifier with the largest number of votes that correspond to the orientation as the final estimated orientation.

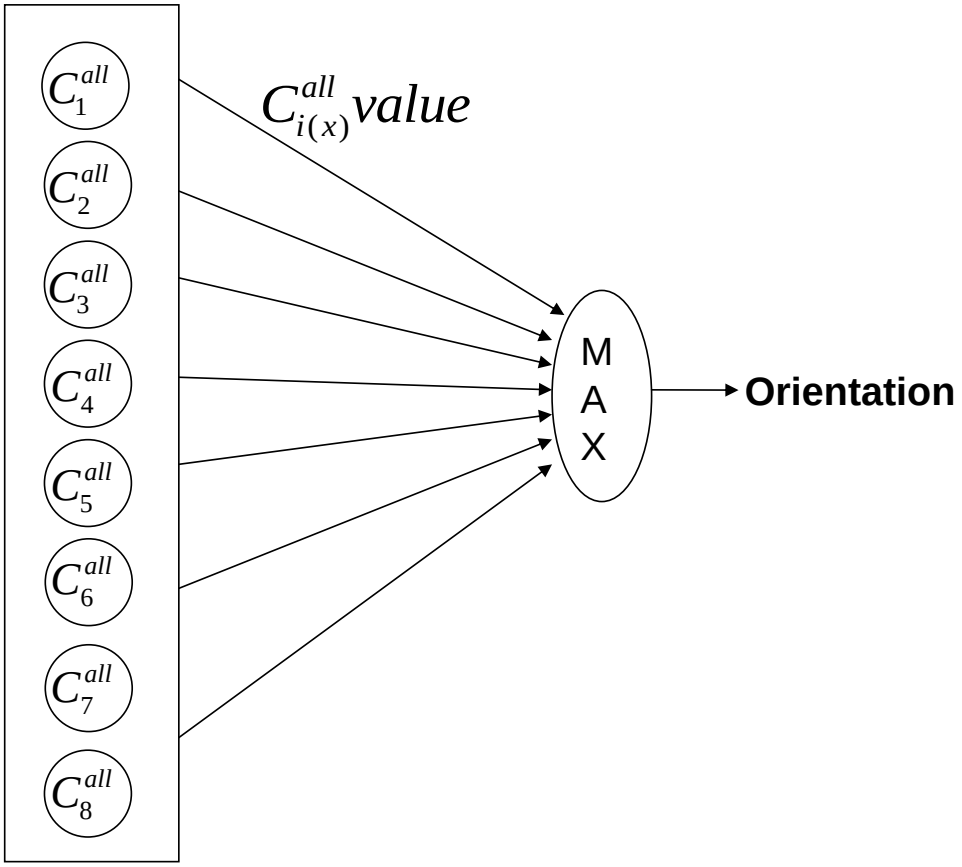


Fig. 4.4 One versus all classification

4.2.2 One versus all classification

One of the simplest multiclass classification schemes that is built on top of real-valued binary classifiers is to train M different binary classifiers, each one trained to distinguish the examples in a single class from the examples in all remaining classes. When it is desired to classify a new example, these M classifiers run, and the classifier which outputs the largest (most positive) value is chosen as the final decision. This scheme will be referred to as the one versus all classification. One might argue that one versus all classification is the first thing thought of when asked to come up with an approach for combining binary classifiers to solve multiclass problems. Although it is simple and obvious, it is extremely powerful (see **Fig.4.4**).

In the training phase, first, we use one group images as the positive samples and the other seven group images as the negative samples for a corresponding orientation classifier. For example, the classifier of class0 (0º) as positive samples and with those of the other seven orientations (45º, 90º, 135º,…,315º) as negative samples. Each classifier employing the adaboost algorithm is trained corresponding to the appropriate

0°	45°	90°	135°	180°	225°	270°	315°
0.52	0.49	0.47	0.54	0.52	0.53	0.55	0.58

One vs. one

Fig. 4.5 Performance of one vs. one approach

orientation.

This method has higher accuracy, but disadvantages include complicated implementation and slow training. At run time, each of the eight classifiers produces eight outputs. We select the classifier that gives the highest value corresponding to the orientation as the final orientation.

We exploit the adaboost decision function as the measure of the strength of class i for pattern x:

$$C_i(x) = \sum_{t=1}^{T} \alpha_t h_t(x) - \frac{1}{2}\sum_{t=1}^{T} \alpha_t \qquad (1)$$

$$\alpha_t = \log\frac{1}{\beta_t}, \ \beta_t = \varepsilon_t / 1 - \varepsilon_t \qquad (2)$$

$$k = \arg\max_i \left(C_i : i = 1,...,8\right) \qquad (3)$$

where h_t is the weak classifier and ε_t is the error rate. We select label k that corresponds to the highest value as the estimation result of the final orientation.

4.2.3 Preliminary experiment

As a baseline of orientation estimation evaluation we conducted preliminary experiments using the mentioned one vs. one and one vs. all algorithms

Fig.4.5 and **Fig.4.6** shows the performance of one vs. one and one vs. all approach. From the **Fig.4.6**, we found that some orientations are easily confused with each other, especially, symmetrical or adjacent one. For example, cell as seen in fig.4.6 with bold rectangle according to 0º orientation classifying, the symmetrical orientation 180º is easily confused which has the second largest classify rate. Besides, the adjacent orientation 45º and 315º is also easily confused with 0º orientation.

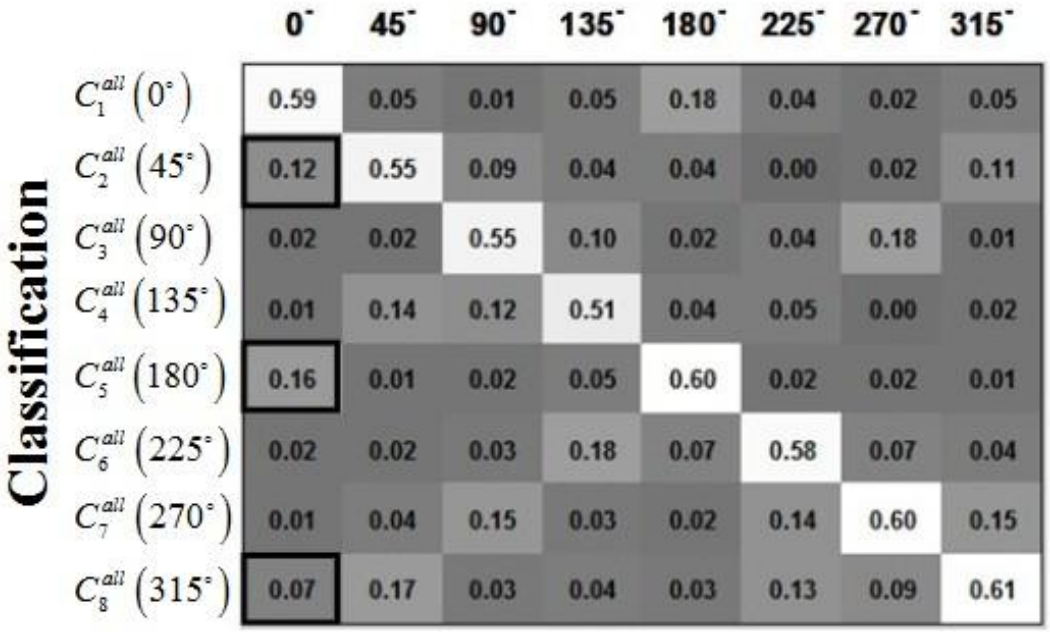


Fig. 4.6 Performance of one vs. all approach

4.3 Cascade orientation estimation

4.3.1 Naive cascade

We propose to generate a cascade orientation classifier that is different from adaboost cascade. First, the easily confused orientations are put into one positive group, and the other orientations are labeled as negative group to train a large two-class classifier called a global classifier. Next, the rather smaller classifier called a local classifier is trained for the corresponding orientations. The cascade architecture is illustrated in **Fig. 4.7**. For example, to obtain the classifier for orientation 1, a global classifier is trained with the easily confused class groups including orientations 1, 2, 5, and 8 as positive classes and with those of orientations 3, 4, 6, and 7 as negative classes. Then, local classifier is trained with the class of orientation 1 as positive samples and orientations 2, 5, and 8 as negative. We used eight global classifiers that correspond to each orientation

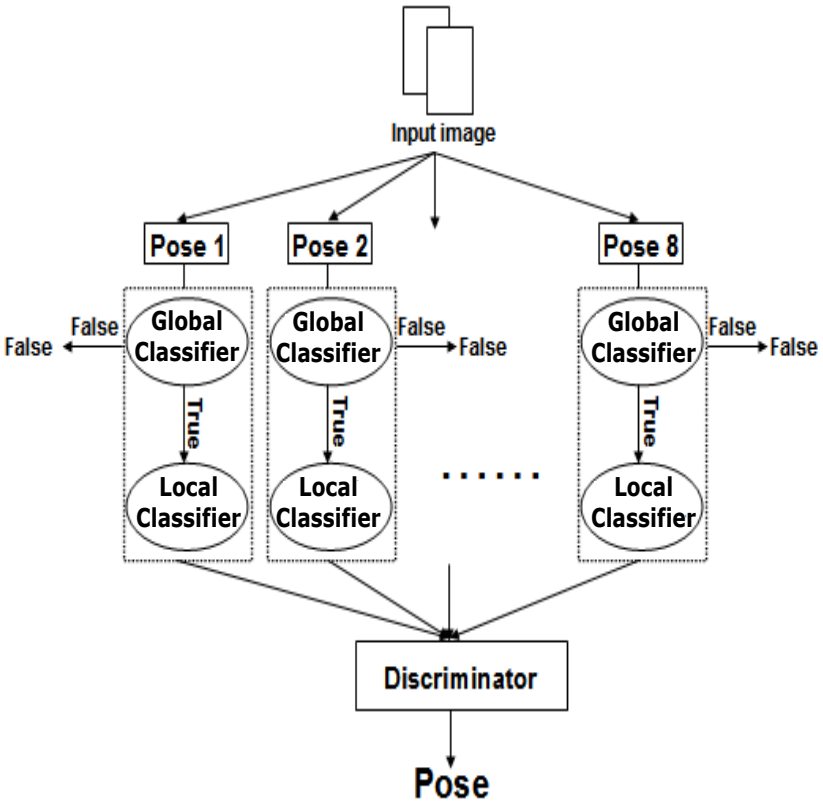


Fig. 4.7 Architecture of cascade orientation classifier

Global Classifier	
Positive orientation	**Negative orientation**
1,2,5,8	3,4,6,7
2,1,4,8	3,5,6,7
3,2,4,7	1,5,6,8
4,3,5,6	1,2,7,8
5,1,4,6	2,3,7,8
6,4,7,8	1,2,3,5
7,3,6,8	1,2,4,5
8,1,2,7	3,4,5,6

Local Classifier	
Positive orientation	**Negative orientation**
1	2,5,8
2	1,4,8
3	2,4,7
4	3,5,6
5	1,4,6
6	4,7,8
7	3,6,8
8	1,2,7

Fig. 4.8 Training sample classification of cascade orientation classifier

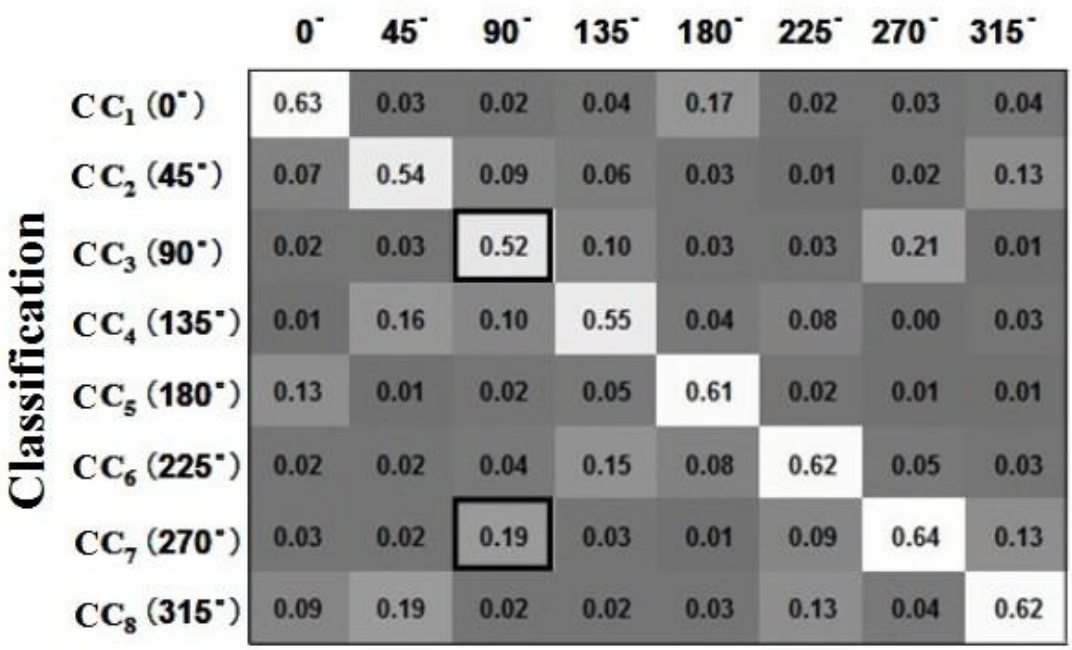


Fig. 4.9 Performance of cascade orientation approach
(Bold rectangles highlight large misclassification of 90º (0.52) to 270º)

	0°	45°	90°	135°	180°	225°	270°	315°	Average
One vs. all	0.59	0.55	0.55	0.51	0.60	0.58	0.60	0.61	0.57
Cascade	0.63	0.54	0.52	0.55	0.61	0.62	0.64	0.62	0.59

Fig. 4.10 Comparison performance between one vs. all and cascade approach

to classify the input images. Positive results from the global classifier trigger local classifiers. Finally, we selected the local classifier that gives the highest value corresponding to the orientations as the final orientation.

Fig.4.9 shows the performance of cascade orientation approach. From the preliminary experiments, the cascade orientation classifying approach has slightly better classification accuracy than the simple one vs. all classifying approach (see **Fig.4.10**). But the orientation of the opposite direction is still easily misclassified. For example, the orientations of 90º and 270º are easily misclassified. The problem of left/right orientation should be classified more accurately using a motion information based multi-frame even if the vehicle and the pedestrian are both moving. But front/rear

Fig. 4.11 Head classifier and front/rear samples

orientation remains hard to classify because the speed of the pedestrian and vehicle is not fixed. For example, a pedestrian seems likely to move back when the vehicle is moving faster than the pedestrian, although they move in the same direction. The front and back views are more confused because the front and back views of the pedestrians are highly similar both in shape and texture. The main distinguishing factor is the head/face area, which is very small compared to the torso/leg area.

In this research, in the following section, we proposed a novel approach to exploit the head orientation information (see **Fig.4.11**) to improve the performance of pedestrian orientation estimation in a single-frame.

4.3.2 Cascade orientation estimation combined head orientation

Single-frame orientation estimation allows us to recover pedestrian headings without integration over time and static pedestrians can also be estimated without posing any problem. Besides body orientation, head orientation can also provide important information for predicting pedestrian trajectories, because a person is likely to move in the direction he is facing.

The body and head orientations may be different, at such scenarios as pedestrian crossings, where they usually look to the left and right sides, but their body orientation is forward. Head orientation is more important when orientation estimation becomes difficult because of the wide range of possible pedestrian appearances, due to changing articulated poses, clothing, lighting, and backgrounds.

For head orientation classification, we separated the samples into two groups (45º, 90º and 135º as rear orientation and 225º, 270º and 315º as front orientation) and,

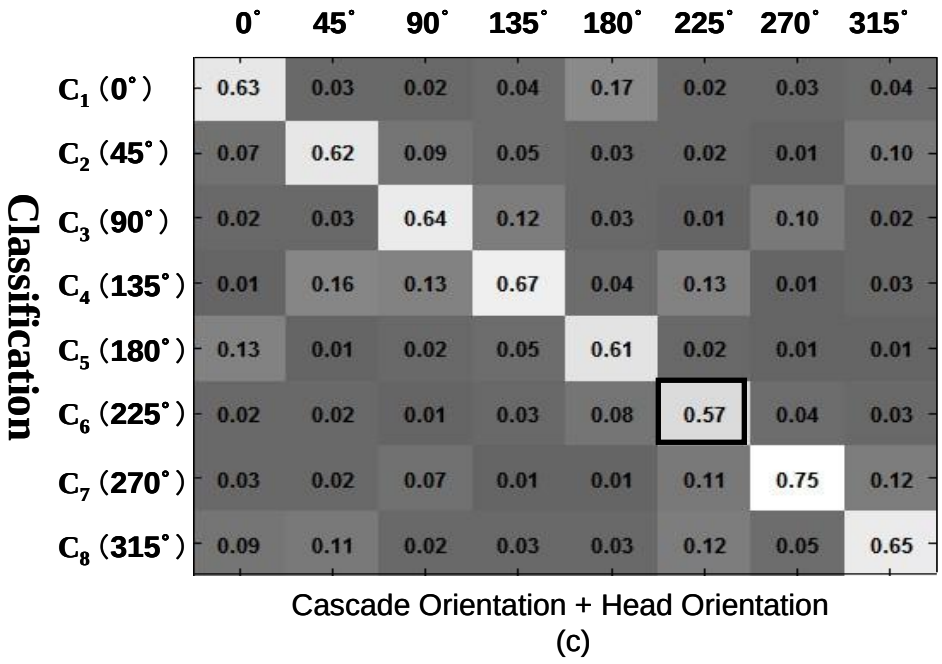


Fig. 4.12 Performance of cascade orientation combined head orientation approach

	0°	45°	90°	135°	180°	225°	270°	315°	Average
One vs. all	0.59	0.55	0.55	0.51	0.60	0.58	0.60	0.61	0.57
Cascade	0.63	0.54	0.52	0.55	0.61	0.62	0.64	0.62	0.59
Cascade+Head	0.63	0.62	0.64	0.67	0.61	0.57	0.75	0.65	0.64

Fig. 4.13 Comparison performance between three approaches

reduced our task to a two-class problem. Note that we here ignored left (180º) and right (0º) directions, as described above, so that the left/right orientations can be classified or corrected based on motion information in the later step.

Since the head and body orientations sometimes can be different, in this paper, we address the problem of pedestrian orientation by referring to a constrained combination of the head and body orientations. Since the head orientation classifier is only used to classify the head's front/rear orientation, there are two factors: body and head orientation classifiers. We proposed an approach with a multi-Bayesian model.

The input to our framework is training set D of six groups pedestrian samples that correspond to six pedestrian orientations every 45º $x_i \in D$, associated with each body

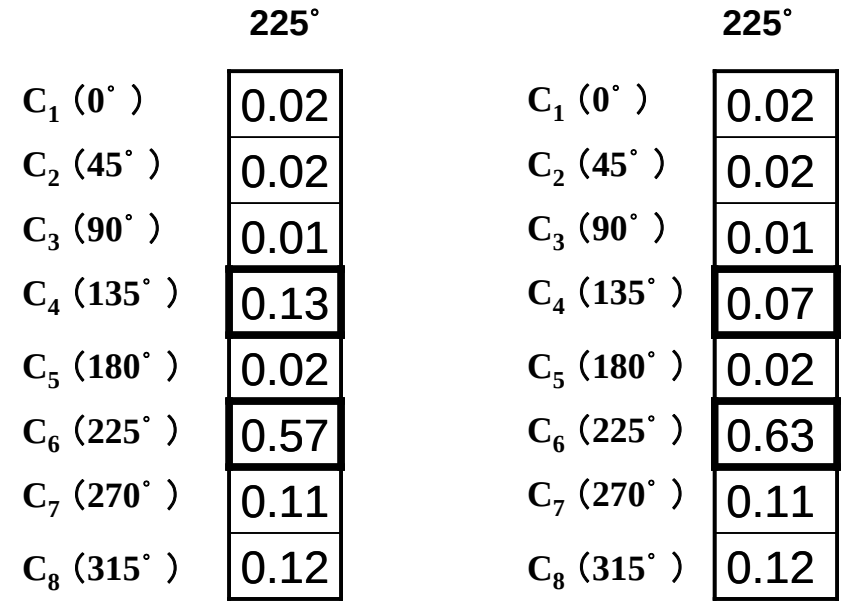
Table. 4.1 Aposteriori probabilities

θ_i, ω_i $\mathbf{x}_i$	$\theta_i = 45^\circ$ $\omega_i = F$	$\theta_i = 90^\circ$ $\omega_i = F$	$\theta_i = 135^\circ$ $\omega_i = F$	$\theta_i = 225^\circ$ $\omega_i = F$	$\theta_i = 270^\circ$ $\omega_i = F$	$\theta_i = 315^\circ$ $\omega_i = F$
45°	0.377	0.017	0.127	0.006	0.004	0.047
90°	0.040	0.191	0.051	0.007	0.027	0.003
135°	0.039	0.053	0.404	0.039	0.006	0.005
225°	0.033	0.082	0.302	0.833	0.094	0.153
270°	0.073	0.629	0	0.074	0.731	0.051
315°	0.438	0.028	0.115	0.041	0.138	0.741
θ_i, ω_i $\mathbf{x}_i$	$\theta_i = 45^\circ$ $\omega_i = R$	$\theta_i = 90^\circ$ $\omega_i = R$	$\theta_i = 135^\circ$ $\omega_i = R$	$\theta_i = 225^\circ$ $\omega_i = R$	$\theta_i = 270^\circ$ $\omega_i = R$	$\theta_i = 315^\circ$ $\omega_i = R$
45°	0.726	0.039	0.176	0.058	0.047	0.472
90°	0.137	0.762	0.125	0.132	0.511	0.056
135°	0.09	0.144	0.677	0.486	0.079	0.056
225°	0.002	0.006	0.014	0.280	0.033	0.050
270°	0.005	0.046	0	0.025	0.259	0.017
315°	0.040	0.003	0.007	0.020	0.070	0.349

orientation is class label θ_i , (θ_0 for the body orientation 45°, θ_1 for 90°, θ_2 for 135°, θ_3 for 225°, θ_4 for 270° and θ_5 for 315°, respectively). We also labeled these pedestrian samples with class label ω_i, (ω_0 for the front head orientation and ω_1 for the rear head orientation). For pedestrian direction estimation, our goal is to determine the class label of previously unseen samples. We make a multi-Bayesian decision and assign pedestrian direction $\mathbf{x}$ to the class with highest aposteriori probability:

$$f(I) = \arg\max_i P(\mathbf{x}_i \mid \theta_i, \omega_i) \qquad (4)$$

Since we assume that the body and head orientations are independent of each other, we decompose $P(\mathbf{x}_i \mid \theta_i, \omega_i)$ as, the probability of pedestrian orientation $\mathbf{x}_i$ with a given



a. Multi-Bayesian model **b.** manually define

Fig. 4.14 Effect of head orientation estimation
(improvement 0.57 to 0.63 in 225° orientation)

body orientation θ_i and a head orientation ω_i .

$$P(\mathbf{x}_i \mid \theta_i, \omega_i) = \frac{P(\theta_i \mid \mathbf{x}_i)P(\omega_i \mid \mathbf{x}_i)P(\mathbf{x}_i)}{\sum_i P(\mathbf{x}_i)P(\theta_i \mid \mathbf{x}_i)P(\omega_i \mid \mathbf{x}_i)} \qquad (5)$$

The body and head orientations of each training sample are manually labeled as ground truth. The aposteriori probabilities are calculated from the training samples using equation (5) in the learning phase. **Table 4.1** shows the obtained aposteriori probabilities of all conditions. For example, based on the maximum criteria, when the result of the body orientation classification is 45º and the result of the head orientation classification is the front, the probability (0.438) of 315º (red rectangle cell in **table 4.1**) is maximum in the first column and is chosen. This means that the factor of the head orientation classification takes a more important role than the body orientation.

Fig.4.12 shows the performance of cascade orientation combined head orientation approach. Shown as **Fig.4.13**, cascade orientation combined head orientation approach has the best accuracy compared with simple one vs. all and cascade orientation approach. The only fly in the ointment is the classify rate of the orientation 315º is change into lower than before. We trace it to the cause from the table.1, when the result of the body orientation classification is 135º and the result of the head orientation

classification is the front ($\omega_i = F$), the probability (0.404) of 135º (bold rectangle cell in **table 4.1**) is maximum in the third column and is chosen. As the same time, when the result of the body orientation classification is 135º and the result of the head orientation classification is the rear ($\omega_i = R$), the probability (0.677) of 135º is maximum in the third column and is chosen. It means, whether the head orientation is front or rear, it always is considered as the pedestrian orientation is 135º.

We are interested in the above case of probability 0.404, because all other combinations rely on head orientation than body orientation result. We tested to make a little change manually, in order to see what would be happening here. That is, when the result of the body orientation classification is 135º and the result of the head orientation classification is the front, we chose the 225º orientation as the final pedestrian orientation. The classify rate of the 225º reached to 63% which is increased more than 6%. In this case, the head orientation takes a more important role than the body orientation, and manually change learning result seems better (see **Fig.4.14**). However, for all that, we still insisted on respecting the data. Determine the pedestrian orientation using cascade orientation by exploiting head orientation by a multi-Bayesian model is better choice when based a large number learning data and we will extend the algorithm toward more general cases where no clear heuristics would not be given such as head orientation.

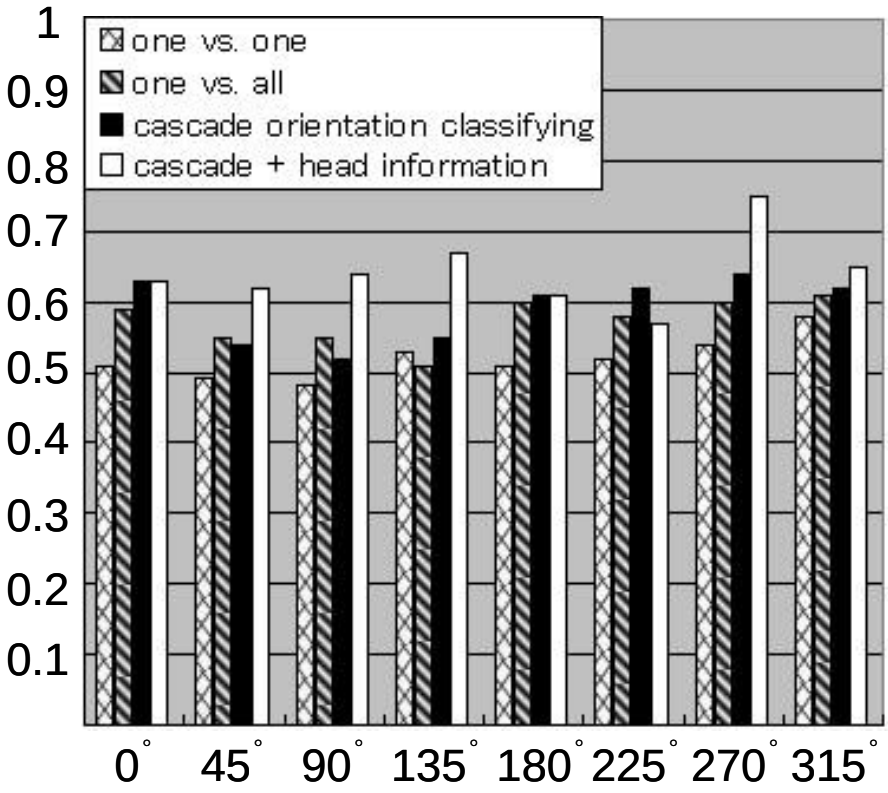


Fig. 4.15 Performance of Orientation Estimation

	0°	45°	90°	135°	180°	225°	270°	315°	Average
One vs. one	0.52	0.49	0.47	0.54	0.52	0.53	0.55	0.58	0.52
One vs. all	0.59	0.55	0.55	0.51	0.60	0.58	0.60	0.61	0.57
Cascade	0.63	0.54	0.52	0.55	0.61	0.62	0.64	0.62	0.59
Cascade+Head	0.63	0.62	0.64	0.67	0.61	0.57	0.75	0.65	0.64

Fig. 4.16 Comparison performance between each approach

4.4 Experiments

4.4.1 Eight orientation performance

In this section, we test our proposed approach in realistic conditions using images obtained from a single camera attached in a moving vehicle. To measure its performance, we performed experiments to estimate the pedestrian orientation

compared with the one-versus-one and one-versus-all methods using the same training and test samples.

All of the training and test samples were obtained manually from pictures of walking pedestrians in the real world with a single camera on a moving vehicle. All samples were normalized to the same size: 24×58.

In the pedestrian orientation estimation, training was performed using 3200 samples: a collection of 400 samples for eight orientations. The test was performed with an extra 800 pedestrian samples (100 samples corresponding to one pose direction) independently prepared to the 3200 training samples. The ground truth orientation in eight directions was manually determined.

In the head orientation estimation, the head samples were obtained manually from the pedestrian samples. The training was performed using 2400 samples that consisted of 1200 front and 1200 rear samples. The 2400 samples set is a subset of the 3200 sample set. The test was performed with 600 samples independently prepared for training samples: 300 for front head orientation and 300 for rear head orientation. The samples were all normalized to the same size: 20×20.

In our experiment, we evaluated the orientation estimation performance using the best performance approach: cascade orientation combined with the head orientation approach.

We compared our approach to our own implementations of three approaches to estimate the orientation, using the same data and evaluation criteria. First, we considered one vs. one approach. Second, we evaluated the orientation using the one vs. all approach. Third, we compared the cascade orientation approach which combined head orientation and naive cascade orientation. Each classifier learned in the orientation estimation was constructed from 200 features, and the performances are shown in **Fig. 4.12** and **Fig. 4.15**. **Fig. 4.16** shows that the one vs. all approach outperforms the one vs. one approach for the implementations of the one vs. all approach. The cascade orientation classification approach has slightly better classification accuracy than the simple one vs. all approach. The average performance of the orientation estimation based the on one vs. one classification approach reached to 52% accuracy, up from the 57% accuracy using the one vs. all classification approach and the 59% accuracy using the cascade orientation classification approach. Our approach, which is cascade orientation combined with the head orientation approach and reached 64% accuracy, increased more than 5% of the cascade orientation classifying approach, and more than 7% of the one vs. all classifying approach and more than 12% of the one vs. one classifying approach. The benefit is more significant for exploiting the head information

by a multi-Bayesian model.

4.4.2 Four orientation performance

Finally we additionally compared our approach to the approach of Enzweiler, whose technique uses HOG features and Gaussian mixture-model formulation. Because Enzweiler [140] estimated four orientations, we used our approach to estimate the same four orientations, left/right and front/back, to compare the performance based on our own data that are different to Enzweiler's. Although each experiment used different proprietary datasets, both contain fairly general situations, and we explain the potential performance of each method.

Fig.4.17 compares the performance between our approach and Enzweiler's. The result shows that, the average performance of the cascade orientation approach is 67%, but when combined with the head orientation, the average performance of our approach is 77.5%, which is slightly better than Enzweiler's 75.7%. In addition, for the back/front orientation estimation, our approach outperforms Enzweiler's, because of the benefit of exploiting the head orientation information by a multi-Bayesian model. For the left/right orientation estimation, the performance of our approach is worse than Enzweiler's, because the HOG feature is better than the Haar-like features, especially for flipped

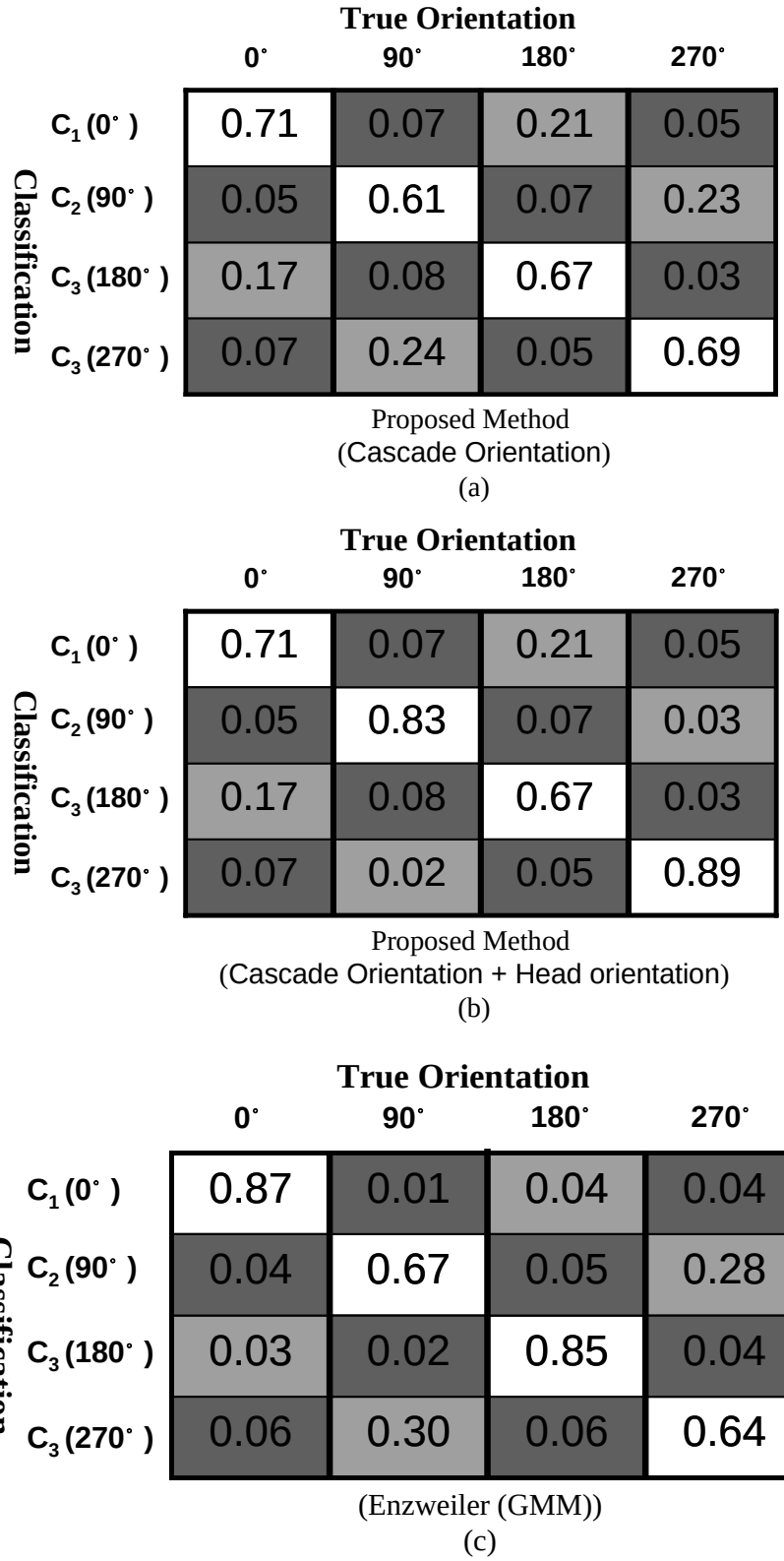


Fig. 4.17 Performance comparison of each approach for four orientation case: (a) cascade orientation (proposed), (b) cascade orientation combined head orientation (proposed), (c) GMM

Fig. 4.18 Examples of orientation estimation using single-frame

shape discrimination. **Fig.4.18** shows orientation example using single-frame estimation.

4.5 Discussion

Pedestrian protection is an important application of intelligent driver support systems. A robust system that detects pedestrians and predicts the collision probability between the vehicle and pedestrians can reduce accidents. This Chapter presented an important problem: orientation estimation.

We outlined the problem of estimation into one of eight directions. We used cascade orientation estimation that integrated the head orientation estimation by a multi-Bayesian model and reached total 64% accuracy, compared to the one vs. one, one vs. all, and cascade orientation approaches. We also compared the performance between our approach and Enzweiler's who challenge to classify four orientations. The average performance of our approach is 77.5%, which is slightly better than Enzweiler's 75.7%. Because the comparison is made based on different datasets, it is just an example. However, the size and variety of our dataset is considered general, therefore, it is worth to show the comparison. As for the characteristics of both algorithms, for the back/front orientation estimation, our approach outperforms Enzweiler's, because of the benefit of exploiting the head orientation information by a multi-Bayesian model. For the left/right orientation estimation, the performance of our approach is worse than Enzweiler's, because the HOG feature is better than the Haar-like features, especially for flipped shape discrimination. But the computational cost of HOG is bigger than Haar-like, in

order to realize the system in real-time, the sooner react, the better. We confirmed the benefit of exploiting the head orientation information, which supports the strength of our proposed approach. Since our main proposal separately exploits head orientation information, the idea can be combined with existing methods. We still believe that less computation cost is important for this application in image feature selection, as far as the total performance is acceptable, as we could demonstrate in this Chapter.

Chapter 5

Pedestrian Walking Direction Estimation

For a complete safety system, detection should be followed by prediction of the possibility of collision. The system should relay the information to the driver in efficient and non-distracting manner or to the control system of the vehicle in order to take preventive actions. Walking direction information can potentially improve the prediction of future trajectories that the pedestrian may take and improve collision prediction. This chapter introduces the approach for pedestrian walking direction estimation.

In this Chapter, we estimate the walking direction in the situation that the pedestrian walking straight, that is, the pedestrian walking toward one direction, and not change his/her walking direction. The case when the pedestrian change his/her walking direction will be researched as our future work. We consider that the performance of estimation of the pedestrian walking straight direction can improve the estimation performance even when the pedestrian walking curve direction because of the curve trajectory can be divided into some straight trajectory. Here, we estimate the pedestrian walking direction by employing an average method to integrate the pedestrian orientation which is obtained from stage 2 (presented in Chapter 4) frame by frame during a video sequence segment.

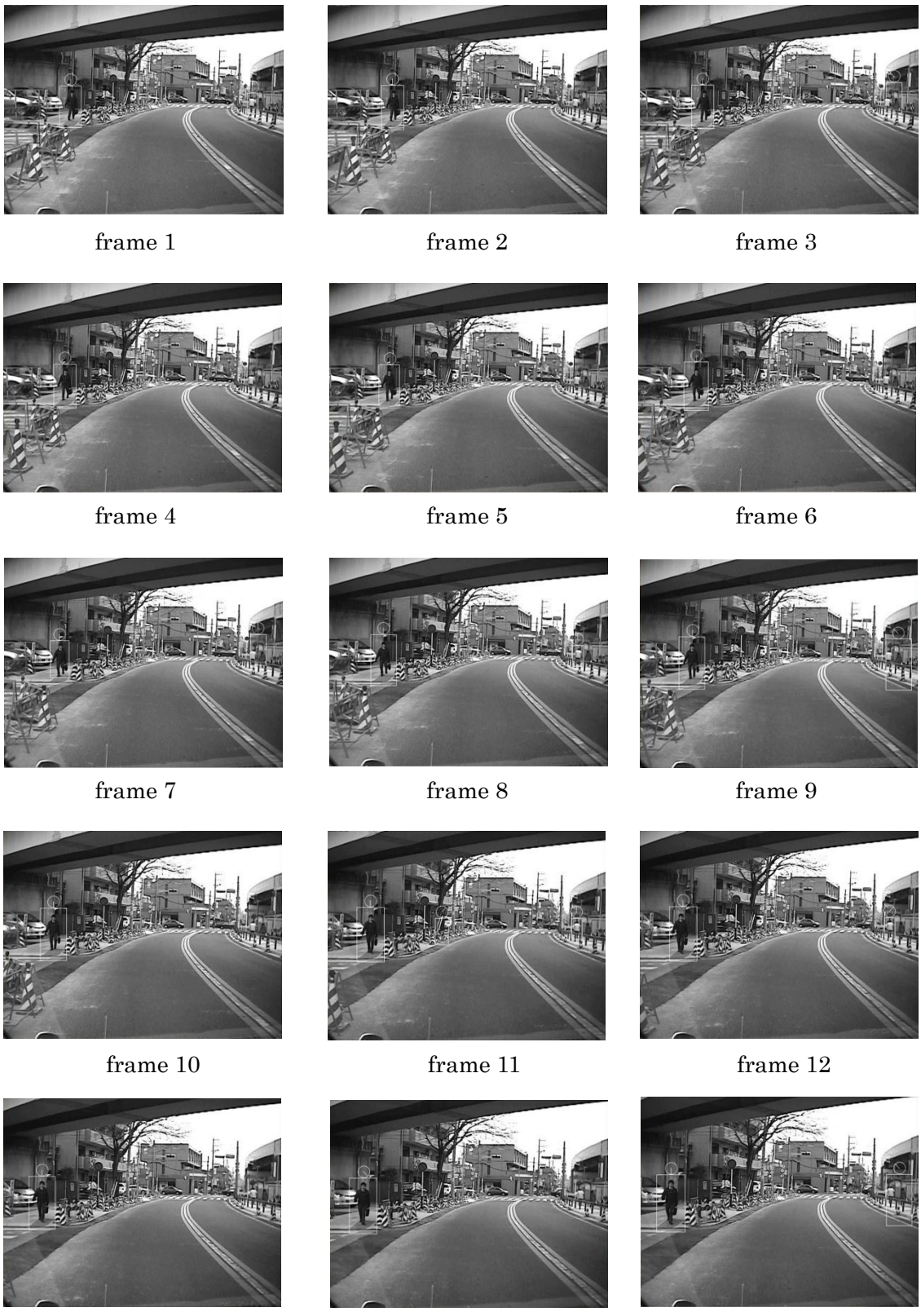


Fig. 5.1 Orientation estimation result for multi-frame

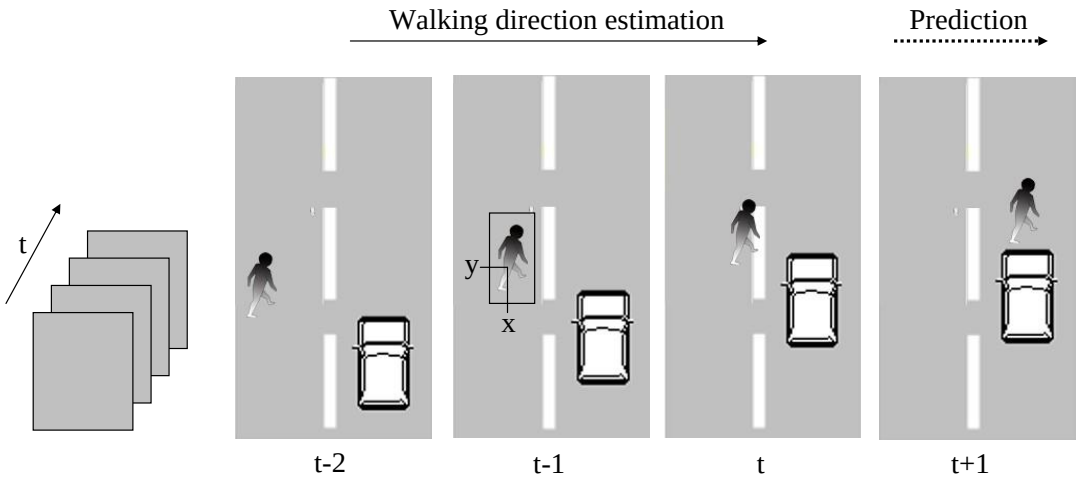


Fig.5.2. Frame by frame pedestrian direction estimation

5.1 Question study on walking direction estimation

The orientation of the pedestrian body, gives useful information about the future direction of motion. Hence, estimating the pedestrian orientation can potentially improve the motion prediction and give better estimating of collision probability.

Our long term goal is to estimate the walking direction of pedestrians in the real-world to identify situations as quickly as we can where pedestrians might be hit by car. In general moving object recognition, we often exploit an object tracking algorithm and give a trace on the image coordinate space. However, we need to compute the trace's projection onto the driving surface coordinate system from a moving camera image space. From the stage 2 which were introduced in Chapter 4, we estimate the pedestrian orientation in a single frame and get rather good estimation result. It provides more valuable information to improve the prediction of future trajectories that the pedestrian may take. Therefore, we propose algorithm of the pedestrian walking direction using the result of orientation estimation in a single frame which is presented in Chapter 4.

Fig. 5.1 is shown to introduce the orientation estimation for multi-frame, find that not every frame orientation estimation result is correct. It is to be regretted that the orientation estimation step does not reach a hundred percent performance. For example,

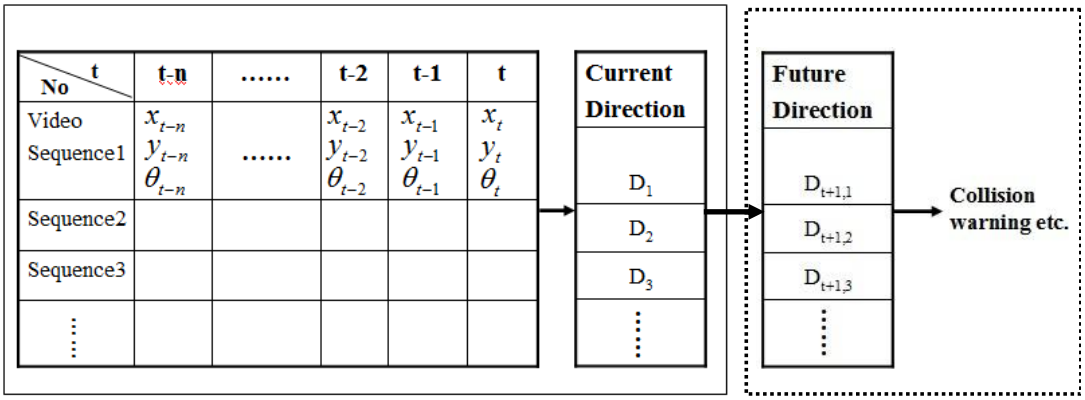


Fig.5.3. architecture of pedestrian direction estimation

the orientation estimation in frame 4 is 225º and the orientation estimation in frame 14 is 315º, although the ground-truth is 270º. As introduced in Chapter 4, the neighbor orientation is easily confused each other; we impute the failure to the feature of those orientation that is similar with each other. We even find that the most of orientation estimation result is correct. We think here the wrong results as noise, so the problem becomes to a simple smoothing problem with noise reduction.

In order to improve the reliability of estimation, we estimate the orientation for a sequence video which the single frame orientation is integrated frame by frame. The evolution of pedestrian orientation over time is modeled with an average orientation estimation method between different orientation states. We consider the orientation estimation for sequence video as the pedestrian walking direction in this duration. In this thesis, we just estimate one walking direction during a duration of video sequence, not consider when the pedestrian change the walking direction.

5.2 Sequence segmentation

In the process above, the object tracking is performed at the same time. A simple assumption is employed here to construct a sequence. That is, the neighbor regions across the consecutive frames are concatenated to construct a sequence. If any detection error occurs for a moving object, the sequence is divided into two or more separated sequences.

The walking direction estimated by integrating t image frames, the orientation of

pedestrian is obtained from each frame, using the average method to estimate the current walking direction during the segment see **Fig.5.2**. In this thesis, we estimate the pedestrian walking direction during a duration (time=t), predicting the walking direction (t+1) as the future work (see **Fig.5.3)**.

5.3 Walking direction estimation method

As the first approach of exploiting temporal consistency, we introduce the idea of average orientation over a certain duration T. We propose two methods and evaluate their performances in the experiment.

5.3.1 Most frequent method

The first method is to use a most frequent orientation estimation result as the walking direction D in the following equation.

$$D_f = \text{argmax}_i\ (f_{o1}, f_{o2},..., f_{o8}) \quad (6)$$

Where f_{oi} is the occurrence frequency of orientation i.

5.3.2 Average method

The second method is to use a rounded average of estimated orientation as in the following equation.

$$D_a=1/n\sum i \quad (7)$$

Where n is the total number of video frames for the duration T and i is the number of orientation.

Because the number of orientation is cyclic, we implement the summation by using the D_f as the bias and setting it to new label 0, and counter-clockwise 1 to 4 from it and clockwise -1 to -3 from it. For example, if the most frequent orientation D_f is 2 as shown as **Fig. 5.4**, the orientation 2 is set to new label 0 and other orientation from -3 to 4. The average D_a is computed then. The final result is rounded to the nearest direction.

5.3.3 An example of direction estimation

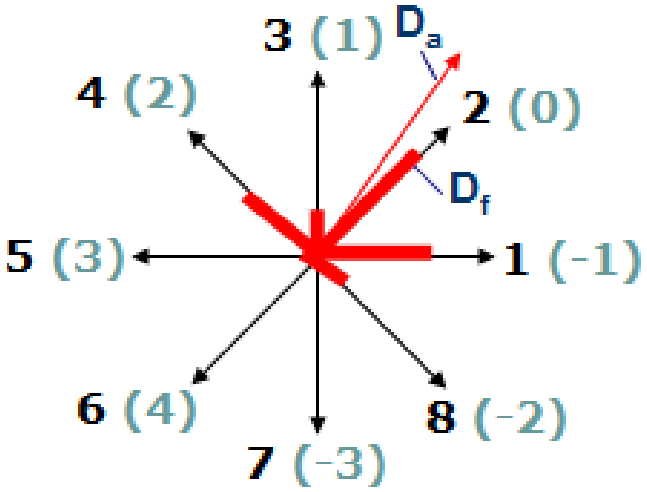


Fig. 5.4. Assign 'label 0' to the direction which has the largest frequency. counter-clockwise from it to the reverse side orderly add 1, clockwise from it to the reverse side orderly add -1. Finally, calculate the average.

Using above approach, we estimate the orientation in the sequence of **Fig. 5.1.** First we calculated the most frequent orientation; the orientation 270º is occurred 13 times in all of 15 frames. We implemented the summation by using the orientation 270º as the bias and setting it to new label 0. Label the counter-clockwise 315 º, 0º, 45º and 90º as 1 to 4, and clockwise 225º, 180º, 135º as -1 to -3. The average is the orientation 270º; we consider 270º as this pedestrian walking direction in this duration of 15 frames. In this example, it is so lucky because the only two occurred wrong orientation is the opposite orientation.

But there are two other cases which lead the walking estimation to a wrong result.

One case is when the most frequent orientation is not the ground-truth. Using the proposed average method, it causes a fatal mistake which the estimated orientation is the most frequent orientation. For example, as shown in Table 2, the sequence No. 4 whose most frequent direction was already incorrect. The ground-truth is orientation 6 (225º), but the orientation 8 (315º) occurred at the most. Although the calculated average is 0.46 which is below 0.5, the estimated walking direction is 8.46; it means the pedestrian walking direction is between the 315º and 0º; it is obviously a wrong direction estimation. But from the experiment, occurs with 4 sequences out of 212 and the error rate is 1.9%. It is benefit from the approach which is the orientation estimation for single-frame, the most frequent orientation is usually correct, and we think if the sequence is long enough, the most frequent orientation will be correct.

No	Ground -Truth	Result	Most Frequent	Averaged Residual
1	2	222422822422	2	0.17
2	3	3233374332333	3	0.23
3	7	7777734776377777	7	0.25
4	6	6868886648876	8	0.46
5	5	55455115656555	5	0.64

Fig. 5.5. Sample of the calculated estimation of walking direction

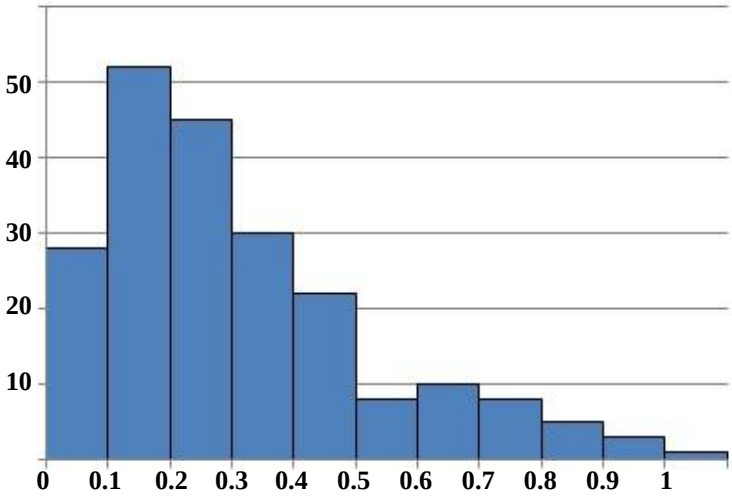


Fig. 5.6. Distribution of calculated estimation of walking direction (Distribution of the result of **Fig.5.5**)

The other case is when the opposite orientation of the ground-truth orientation occurred frequently. For example, as shown in Table 2, the sequence 5 shows this estimation error. The ground-truth orientation is orientation 5 (180º); although the most frequent orientation is correct, the frequent occurrence of opposite orientation 1 (0º) lead the calculated average which is 0.64 exceeding 0.5; the estimated walking direction is 5.64; it is more close to orientation 6 (225º) rather than orientation 5. From the experiment these types of errors occurs with 35 sequences out of 212 and the error rate is 16.5%. Most of errors follow into the left/right orientation and front/back

	Correct	Error	Recognition rate
Most Frequent	**208**	**4**	**98.1%**
Average method	**177**	**35**	**83.5%**

Fig. 5.7. Performance of walking direction estimation

direction express well performance with low error rate. We confirmed the benefit of exploiting the head orientation information, which supports the strength of our proposed approach. As to mistake for left/right orientation estimation, we think it can be solved using motion information in the future.

5.4 Experiment

In the walking direction estimation step generated 212 object sequences by concatenating spatio-temporal neighbors, discarding the short sequences. The samples results of walking direction is given in the **Fig.5. 5**. The sequences No. 1-3 are correct estimation examples for both criteria (the most frequent and the rounded average) since their averaged residuals are less than 0.5. The sequence No.4 is an incorrect example for either criterion. It's most frequent direction was already incorrect. **Fig.5.6** shows the distribution of the result of **Fig.5.5**, from the distribution, we can see the most of the results are less than 0.5. It proves the approach for estimating the walking direction has been reduced the noise efficiently, and the result of the estimation for walking direction is close with the ground-truth.

Fig.5.7 shows the performance of walking direction estimation. The error occurs with 4 sequences out of 212 and the accuracy is 98.1% by the most frequent method (see **Fig.5.9**). It achieved a surprisingly good result with the most frequent criterion. This kind of errors occurred with 35 sequence out of 212 and the accuracy falls into 83.5%.

The sequence No.5 is another incorrect example for rounded average criterion. Although the most frequent direction was correct, the averaged residual was 0.64 and exceeding 0.5(see **Fig.5.8**). Thus the final estimated direction was not right. This case was unfortunate because the pose classifiers made more errors with opposite direction estimation.

It has corrected the walking directions efficiently after the step 3 and will be benefit to predict the pedestrian's path in the future.

5.5 Conclusion

Pedestrian protection is an important issue for intelligent vehicle, pedestrian protection need not only detect the pedestrian but also predict the collision between the pedestrian and the vehicle. This Chapter presented an important problem: pedestrian walking direction estimation.

We outlined the problem of estimating the pedestrian walking direction in a duration time. It provides more valuable information to improve the prediction of future trajectories that the pedestrian may take efficiently. In this Chapter, we estimate the walking direction in the situation that the pedestrian walking straight, that is, the pedestrian walking toward one direction, and not change his/her walking direction. We used average method to integrate the pedestrian orientation frame by frame during a video sequence segment. Experiments on a large amount of real-world data show a significant performance improvement of 83.5% in estimating the walking direction against 212 targeted objects, and reached a significant performance which is 98.1% accuracy. It has corrected the walking directions efficiently after the step 3 and will be benefit to predict the pedestrian's path in the future.

Many approaches to estimate the pedestrian walking direction by using estimate the walking trajectory with tracking method. One line of research has formulated tracking as frame-by-frame association of detections based on geometry and dynamics without particular pedestrian appearance models. Other approaches utilize pedestrian appearance models coupled with geometry and dynamics; some approaches furthermore integrate detection and tracking in a Bayesian framework, combining appearance models with an observation density, dynamics, and probabilistic inference of the posterior state density. For this, either single or multiple cues are used. The integration of multiple cues involves combining separate models for each cue into a joint observation density. The inference of the posterior state density is usually formulated as

a recursive filtering process. Particle filters are very popular due to their ability to closely approximate complex real-world multimodal posterior densities using sets of weighted random samples. Extensions that are especially relevant for pedestrian tracking involve hybrid discrete/continuous state-spaces and efficient sampling strategies. But using tracking method causes the reason of computation cost, and the reaction in real time is difficult.

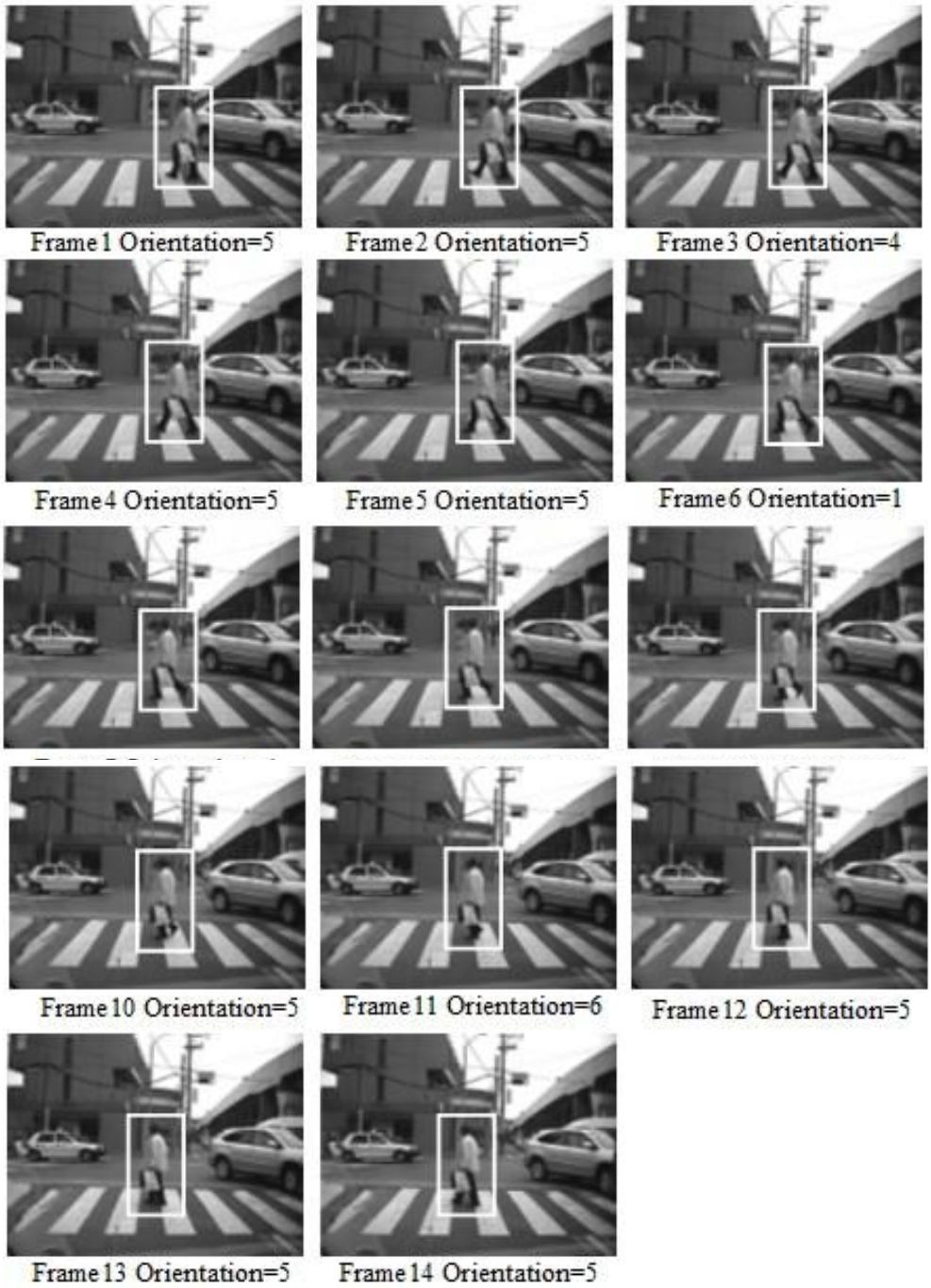


Fig.5.8. Example of pedestrian walking direction estimation
(The result of No.5 in Fig.5.5)

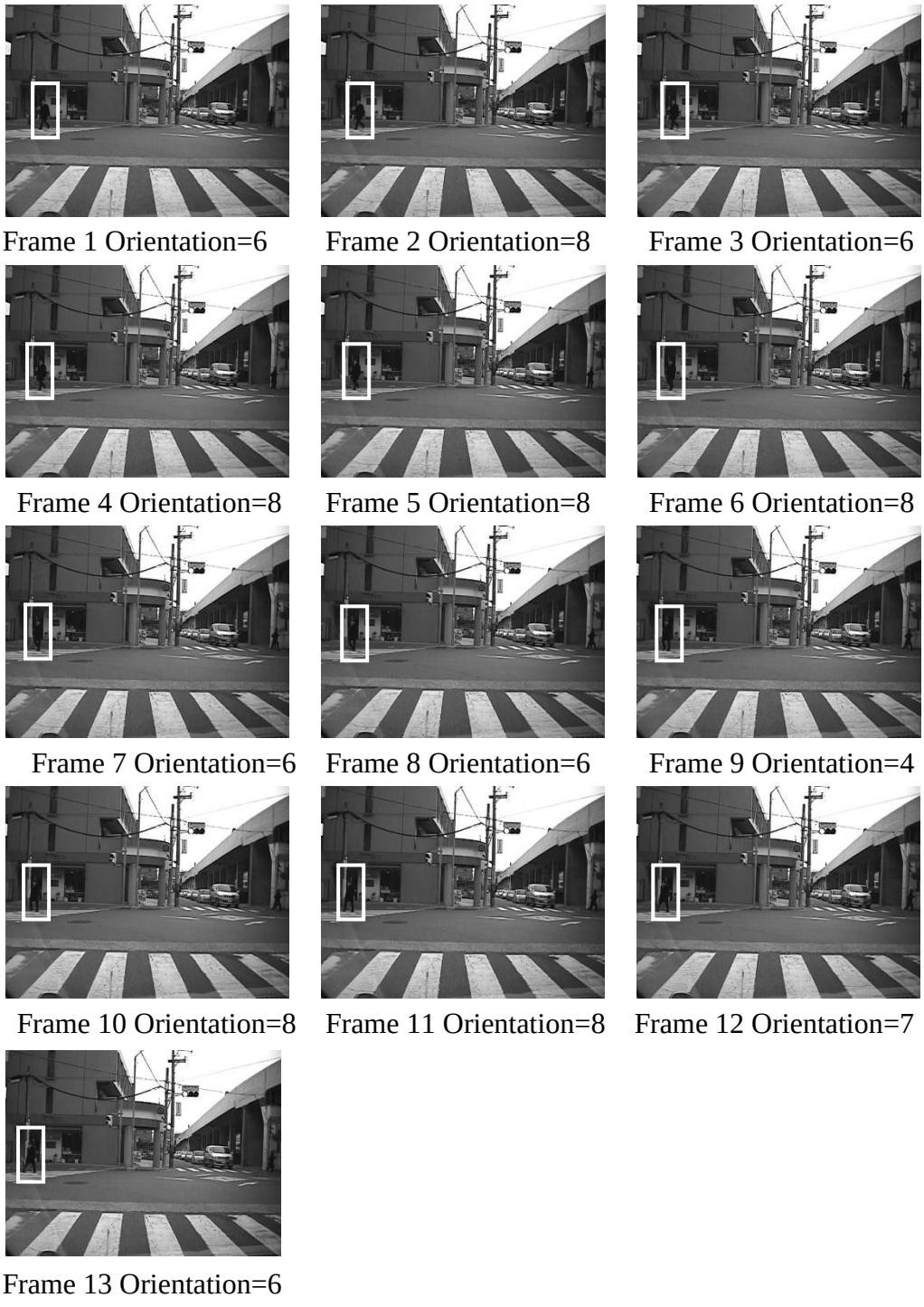


Fig.5.9 Example of pedestrian walking direction estimation
(The result of No.4 in Fig.5.5)

Chapter 6

Conclusion

This chapter summarizes the topics and results of the research covered in this thesis, and provides some directions for future research.

6.1 Summary of research and results

Pedestrians are the most vulnerable road users, and therefore, they require maximum protection on the road. A large number of fatalities and injuries show the importance of developing pedestrian protection systems. This paper discussed the global nature of the pedestrian safety problem and the initiatives taken to address it. It focused an approach and challenge in improving pedestrian safety with image processing by extracting pedestrians and estimating body orientation and eventually walking direction.

We outlined the problem of estimation into one of eight directions. We used cascade orientation estimation that integrated the head orientation estimation by a multi-Bayesian model and reached 64% accuracy, compared to the one vs. one, one vs. all, and cascade orientation approaches. We also compared the performance between our approach and Enzweiler's who challenge to classify four orientations. The average performance of our approach is 77.5%, which is slightly better than Enzweiler's 75.7%. In addition, for the back/front orientation estimation, our approach outperforms Enzweiler's, We confirmed the benefit of exploiting the head orientation information, which supports the strength of our proposed approach. Since our main proposal separately exploits head orientation information, the idea can be combined with existing methods. We still believe that less computation cost is important for this application in image feature selection, as far as the total performance is acceptable. Furthermore, we did experiments on a large amount of real-world data show a significant performance improvement of 83.5% in estimating the walking direction against 212 tageted objcets.

It has corrected the walking directions efficiently after the step 3 and will be benefit to predict the pedestrian's path in the future.

Chapter 1 presented the background and motivations of driver assistance systems and pedestrian protection systems. Technique issue and current trends of research and development was addressed.

Chapter 2 described related researches in pedestrian detection, pedestrian orientation estimation and pedestrian walking trajectory estimation.

Chapter 3 presented the proposed whole system architecture consisting of three pyramid stages, and introduced the conventional approach for pedestrian detection for the stage 1. Pedestrian detector was trained using Adaboost algorithm and Haar-like feature based on a large number training data which were obtained from the pictures of walking pedestrian in real world with a single camera on a moving vehicle. We confirmed the performance of the detector is good enough for pedestrian detection problem.

Chapter 4 presented our proposed approach for the estimation of pedestrian orientation based on the same feature extraction for the stage 1 with a newly designed multi-class classifier and describes the experimental results. In this chapter, we outlined the problem of estimation into one of eight orientations. We proposed a cascade orientation estimation that integrated the head orientation estimation by a multi-Bayesian model and reached 64% accuracy, compares the performance of our proposed approach to one vs. one and one vs. all multi-class classification approach, and also did a experiment to compare our approach with Enzweiler's who challenged to classify four orientation.

Chapter 5 introduced a new approach on the study of the problem of pedestrian walking direction estimation. We proposed an average orientation estimation method to estimate the pedestrian walking direction using the result of orientation estimation in a single frame which is obtained from stage 2 and reached 83.5% accuracy. The pedestrian walking direction estimation can provide more valuable information to improve the prediction of future trajectories that the pedestrian may take.

Chapter 6 summarized the results and identifies other possible application areas and topics for future research.

Pedestrian protection systems offer many important research problems to work on, such as development of different types of sensors, processing of sensor information to extract relevant features, analysis and classification of these features to detect and track pedestrians, behavior and intent analysis of drivers and pedestrians, as well as human factors and interfaces. In the last five years, we have seen a considerable research

activity throughout the world, particularly in Europe and Japan. This is a very positive development. Such research has already produced a lot of important results and also produced a clearer and better understanding of the remaining challenges to work on.

Much of the current research on pedestrian protection systems is toward improving and characterizing their performance. New types of non visible light sensors such as thermal IR and LASER scanners show promise of improving the detection in situations where visible light sensors would be less effective. Although these sensors are expensive at present, mass production is likely to reduce the costs of these devices. Research on sensor fusion and registration would also be very important to ensure performance enhancement using the combination of sensors. Infrastructure-mounted sensors are also likely to complement vehicle-mounted sensors in generating a complete picture of surroundings by filling blind spots of vehicles. Furthermore, detection from infrastructure-based sensors is less complex due to the static background. Research on detecting pedestrians and vehicles for surveillance as well as traffic analysis would therefore be very valuable in the development of infra structure based collision avoidance systems. For a complete system, effective communication between infrastructure and vehicles would be essential.

Pedestrian protection is an important application of intelligent driver support systems. A robust system that detects pedestrians and predicts the collision probability between the vehicle and pedestrians can reduce accidents. In this paper we proposed a method to estimate the orientation and walking direction.

6.2 Future work

Many works are possible that can further develop the research results presented here.

Driver assistance systems (particularly pedestrian protection systems), are a very young area of research. Hence, the future research possibilities are so numerous and diverse. We condense the lines we consider of key importance in a few general points.

6.2.1 Short range issues

We will continue to do the research which is described below as our short range issue.

1. To improve the accuracy of stage 2. As introduced in Chapter 4, from the experiment, the performance for the left/right orientation estimation is not good enough. We thought the haar-like feature is weak when the feature is similar. We need to

perform the experiment using other efficient feature for solve this problem in the future.

Comparison experiment between our approach with Enzweiler's has been done, but it's a pity that the experiment has not based on the same dataset. We need to try an experiment using the same dataset.

2. For pedestrian walking direction estimation stage, our approach has been proved efficient, but it was restricted to one straight walking directions. We must propose an approach for estimating the walking direction even when the pedestrian change his/her walking direction.

Moreover, in order to realize the collision prediction system, a predicting method of the pedestrian walking direction needs to be developed. A pedestrian walking direction method using some probability model such as particle filter or HMM is a promising approach.

3. We will make a further analysis of the computation cost and the accuracy when the system runs as a whole.

4. In this thesis, we consider that the system should be realized in real time. So we choice the Haar-like feature and Adaboost algorithms which is proved fast and efficient. In the future, it is an alternative approach to pursue high accuracy. It is worth to try using other feature or classifier, such as HOG features and SVM classifier.

5. A traffic-monitoring system is designed to predict various behaviors, including collision between vehicles and pedestrians. The possibility of collision is determined using the zone of interaction, which is defined as an elliptic region with the same orientation as the target but with larger size. Events where targets are close and have dangerously high relative velocities trigger a potential collision event. But this approach need that the camera is stationary and is not suitable to attach on the vehicle.

6. A stochastic model of the pedestrian dynamics is most appropriate for predicting the collision probability.

Monte Carlo simulations can then be used to generate a number of possible trajectories based on the dynamic model. The collision probability is then predicted based on the fraction of trajectories that eventually collide with the vehicle. Particle filtering is a natural framework for simultaneously tracking the object and predicting the collision probability.

For developing a robust system for pedestrian protection, a thorough evaluation of these models, the conditions under which they work, and their performance in real world is required. Using these approaches is necessary to build such a model by machine learning or manual labor using samples of the trajectories of pedestrians collected in advance. However, it is difficult to collect enough samples to build a

behavioral model of dangerous pedestrians, since the collision of a pedestrian with a vehicle rarely happens in the real world and moreover, the samples of the dangerous behavior are hardly collected. In addition, nothing but the pedestrian whose behavior deviates from the model of pedestrian behavior is dangerous. Thus, predicting the collision possibility of a pedestrian with a vehicle based on the prediction of the future trajectory of a pedestrian is fraught with difficulty.

6.2.2 Long range issues

6.2.2.1 Pedestrian behavior modeling

An effective pedestrian protection system needs to not only detect pedestrians but also predict the possibility of collision, which is based on modeling of pedestrian behaviors. The improvement of the classification results using motion information based on multi-frames is necessary. One of future interesting issues is to find a more feasible model to evaluate the walking direction and predict the pedestrian walking direction a few second after. For this purpose, such scene context as whether pedestrians are on the sidewalk, in the crosswalk, or in the middle of the road as well as the traffic signal state should be incorporated.

Behavior modeling and prediction is an active area of research. In particular, Monte Carlo method in particle filtering framework is a promising approach for integrating pedestrian detection with collision prediction. One of the challenges in behavior modeling, specifically for collision predictions, is the scarcity of real-world data, since accidents are rare events, and performing the experiments to collect data would involve human subjects in potentially dangerous situations. Hence, a large number of experiments using trajectory simulation in addition to the available real-world accident data would be the only acceptable method in developing and characterizing such systems.

It is seen that there are various models that are developed for pedestrian behavior analysis. Some of these models have been applied for collision prediction. The “discrete choice model” is used in which a pedestrian makes a choice at every step about the speed and direction of the next step. It is assumed that a pedestrian would normally move toward the destination direction, avoid frequent direction changes, and try to adjust speed to a desired speed. The discrete choice behavioral model can be also integrated with person detection and tracking from static cameras based on image processing in order to improve performance. This approach differs from the

conventional tracking since it uses behavior rather than appearance for detection. Also, instead of making hard decisions about target presence on every frame, it integrates the evidence from a number of frames before making a decision. The pedestrian dynamics is modeled using a hidden Markov model with four states corresponding to standing still, walking, jogging, and running. For each state, the probability distributions of absolute speed as well as the change of direction are modeled as truncated Gaussians. a model of pedestrian behavior in crowds is developed based on "stress" that the pedestrians experience while walking in crowd, including pedestrian stress from other pedestrians, which would push them away, and destination stress, which pulls them toward their destination.

6.2.2.2 Driver's state modeling

Finally, in addition to the extraction of information about surrounding objects, it is also important to ascertain the driver's state in order to generate appropriate warnings or actions so that the system would help the driver rather than cause distraction. For example, if a driver has already seen a pedestrian and is taking appropriate action, one may not want to alarm the driver unnecessarily. For this purpose, it is important to not only look outside the vehicle to detect dangerous situations but also look inside the vehicle in order to assess the state and intent of the driver.

6.2.2.3 Infrastructure design

Another interesting and useful research for pedestrian protection is vehicle to vehicle communication. Nowadays it exists a shared conjecture on the benefits of vehicle to vehicle communications. It is expected that vehicle to vehicle communications will enhance the safety and efficiency of human driven vehicles. Multiple initiatives have been deployed to explore these ideas. If benefits are possible for human driven vehicles, the same applies to driverless vehicle.

Again it is expected that by exchanging data the vehicles will increase their knowledge of the situation, and that more information will allow better decisions. Vehicle to vehicle communication for driverless vehicles can operate as an extension or replacement of the driving rules for humans, which are mainly used to allow one human to predict the behavior of another human. The exchange of data could replace such rules based predictions.

It is seen that the research on pedestrian protection systems is still young and in the

long process of reaching maturity. The success of this research should eventually find systems in future automobiles, and help in saving lives and reducing injuries to pedestrians on the road.

Reference

[1] N. Sharkey. The programmable robot of ancient greece. New Scientist, (2611):32-35, 2007.

[2] S.Nedevschi, R. Danescu, D. Frentiu, T.Marita, F. Oniga, C. Pocol, T. Graf, and R. Schmidt. High accuracy stereovision approach for obstacle detection on non-planar roads. Proc. IEEE intelligent Engineering Systems, PP:211-216, 2004.

[3] Organization international des construceurs automobiles. http://www.oica.net.

[4] I. Fallon and D. O'neill. The world's first automobile fatality. Accident Analysis and Prevention, 35, 2005.

[5] M. Peden, R. Scurfield, D. Sleet, D. Mohan, A.A. Hyder, E. Jarawan, and C. Mathers. World Report on road traffic injury prevention. World Health Organization, Geneva, Switzerland, 2004.

[6] Various Authors. The World Health Report 2002. Reducing risks, promoting healthy life. World Health Organization, Geneva, Switzerland, 2002.

[7] E.D. Dickmanns and A. Zapp. A curvature-based scheme for improving road vehicle guidance by computer vision. In Proceedings of the SPIE Conference on Mobile Robots, 727,PP:161-168, 1986.

[8] L. Vlacic, M. Parent, and F. Harashima. Intelligent Vehicle Technologies. Butterworth-Heinemann, 2001.

[9] IVsource. Adaptive cruise control arrives in the USA. http://www.ivsource.net/archivep/2000/sep/a000929_USacc.html, 2000.

[10] IVsource. Iteris' lane departure warning system now available on Mercedes trucks in Europe, 2000.

[11] http://www.worldbank.org/html/fpd/transport/roads/safety.htm

[12] Traffic Safety Facts 2004: A Compilation of Motor Vehicle Crash Data From the Fatality Analysis Reporting System and the General Estimates System, Washington, D.C.: Nat. Highway Traffic Safety Assoc., U.S. Dept. Transp. http://wwwnrd.nhtsa.dot.gov/pdf/nrd-30/NCSA/TSFAnn/TSF2004.pdf

[13] J. R. Crandall, K. S. Bhalla, and N. J.Madeley, "Designing road vehicles for pedestrian protection," Brit. Med. J., vol. 324, no. 7346, pp. 1145–1148, May 11, 2002.

[14] S. K. Singh, "Review of urban transportation in India," J. Public Transp.,vol. 8, no.

1, pp. 79–97, 2005.
[15] D. Mohan, “Traffic safety and health in Indian cities,” J. Transp. Infrastruct., vol. 9, no. 1, pp. 79–94, 2002.
[16] C. Mock, R. Quansah, R. Krishnan, C. Arreola-Risa, and F. Rivara, “Strengthening the prevention and care of injuries worldwide,” Lancet, vol. 26, no. 363, pp. 2172–2179, 2004.
[17] D. Mohan, “Work trips and safety of bicyclists,” Indian J. Transp.Manag., vol. 26, no. 2, pp. 225–232, Apr.–Jun. 2002.
[18] http://www.tfhrc.gov/safety/pedbike/pedbike.htm
[19] http ://www.walkinginfo.org/pedsmart
[20] http://www.path.berkeley.edu/
[21]http://prevent.ertico.webhouse.net/en/prevent_subprojects/vulnerable_road_users_collision_mitigation/apalaci/
[22] http://www.gavrila.net/Computer_Vision/Smart_Vehicles/smart_vehicles.html
[23] Advanced Highway Systems Program, Japanese Ministry of Land, Infrastructure and Transport, Road Bureau. [Online]. Available: http://www.mlit.go.jp/road/ITS/index.html
[24] 39th Tokyo Motorshow News, 2005. [Online]. Available: http://www.tokyo-motorshow.com/eng/release/news/index.html
[25] S. J. Ashton and G. M. Mackay. Benefits from changes in vehicle exterior design. Proceedings of the Society of Automotive Engineers, PP:255-264, 1983.
[26] D.M. Gavrila, J. Giebel, and S. Munder. Vision-based pedestrian detection: The PROTECTOR system. In Proc. IEEE Intelligent Vehicles Symp., PP:13-18, Parma, Italy, 2004.
[27] A. Shashua, Y. Gdalyahu, and G. hayun. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. In Proc. IEEE Intelligent Vehicles Symp., PP:1-6, Parma, Italy, 2004.
[28] Z. Sun, G. Bebis, and R. Miller. On-road vehicle detection: A review. IEEE Trans. On Pattern Analysis and Machine Intelligence, 28(5):695-711,2006.
[29] R. Bishop, Intelligent Vehicle Technology and Trends. Norwood, MA: Artech House, 2005.
[30] B. Fardi, J. Douˇsa, G. Wanielik, B. Elias, and A. Barke, “Obstacle detection and pedestrian recognition using a 3D PMD camera,” in Proc.IEEE Intell. Veh. Symp., Jun. 2006, pp. 225–230.
[31] T. Gandhi and M. M. Trivedi, “Vehicle mounted wide FOV stereo for traffic and pedestrian detection,” in Proc. Int. Conf. Image Process., Sep. 2005, vol. 2, pp.

121–124.
[32] L. Zhao and C. Thorpe. Stereo and neural network-based pedestrian detection. IEEE Trans. On Intelligent Transportation Systems, 1(3): 148-154, 2000.
[33] M. Soga, T. Kato, M. Ohta, and Y. Ninomiya. Pedestrian detection with stereo vision. In Proc. IEEE Int. Conf. On Data Engineering Workshop, PP: 1200, Tokyo, Japan, 2005.
[34] I. Parra, D. Fernandez, M. A. Sotelo, L.M. Bergasa, P. Revenga, J. Nuevo, M. Ocana, and M.A. Garcia. Combination of feature extraction method for SVM pedestrian detection. IEEE Trans. On Intelligent Transportation Systems, 8(2):292-307, 2007.
[35] M. Szarvas, A. Yoshizawa, M. Yamamoto, and J. Ogata, "Pedestrian detection with convolutional neural networks," in Proc. IEEE Intell. Veh. Symp., Jun. 2005, pp. 224–229.
[36] A. Broggi, A. Fascioli, M. Carletti, T. Graf, and M. Meinecke, "A multiresolution approach for infrared vision-based pedestrian detection," in Proc. IEEE Intell. Veh. Symp., Jun. 2004, pp. 7–12.
[37] C. Papageorgiou and T. Poggio, "A trainable system for object detection," Int. J. Comput. Vis., vol. 38, no. 1, pp. 15–33, 2000.
[38] H. Cheng, N. Zheng, and J. Qin, "Pedestrian detection using sparse Gabor filters and support vector machine," in Proc. IEEE Intell. Veh.Symp., Jun. 2005, pp. 583–587.
[39] L. Havasi, Z. Szlávik, and T. Szirányi, "Pedestrian detection using derived third-order symmetry of legs," in Proc. Int. Conf. Comput. Vis. Graph., 2004.
[40] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi, "Pedestrian detection using infrared images and histograms of oriented gradients," in Proc. IEEE Intell. Veh. Symp., Jun. 2006, pp. 206–212.
[41] C. Hilario, J.M. Collado, J. M. Armingol, and A. De la Escalera, "Pedestrian detection for intelligent vehicles based on active contour models and stereo vision," in Proc. Int. Workshop Comput. Aided Syst. Theory, Feb. 2005, pp. 537–542.
[42] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," Int. J. Comput. Vis., vol. 63, no. 2, pp. 153–161, 2005.
[43] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," IEEE Trans. Intell. Transp. Syst., vol. 6, no. 1, pp. 63–71, Mar. 2005.
[44] Y. Fang, K. Yamada, Y. Ninomiya, B. K. P. Horn, and I. Masaki, "A shape-independent method for pedestrian detection with far-infrared images," IEEE Trans. Veh. Technol., vol. 53, no. 6, pp. 1679–1697, Nov. 2004.
[45] X. Liu and K. Fujimura, "Pedestrian detection using stereo night vision," IEEE Trans. Veh. Technol., vol. 53, no. 6, pp. 1657–1665, Nov. 2004.

[46] Y. Zhang, S. J. Kiselewich, W. A. Bauson, and R. Hammoud, "Robust moving object detection at distance in the visible spectrum and beyond using a moving camera," in Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshop, 2006, p. 131.
[47] T. Hasiyama, D. Mochizuki, Y. Yano, and S. Okuma, "Active frame subtraction for pedestrian detection from images of moving camera," in Proc. IEEE Int. Conf. Syst., Man, Cybern., Oct. 2003, vol. 1, pp. 480–485.
[48] B. Fardi, I. Seifert, G.Wanielik, and J. Gayko, "Motion-based pedestrian recognition from a moving vehicle," in Proc. IEEE Intell. Veh. Symp., Jun. 2006, pp. 219–224.
[49] K. Fuerstenberg and V. Willhoeft, "Object tracking and classification using laserscanners—Pedestrian recognition in urban environment," in Proc. IEEE Intell. Transp. Syst. Conf., Aug. 2001, pp. 451–453.
[50] K. C. Fuerstenberg, K. C. J. Dietmayer, and V. Willhoeft, "Pedestrian recognition in urban traffic using a vehicle based multilayer laserscanner," in Proc. IEEE Intell. Veh. Symp., Jun. 2002, pp. 31–35.
[51] K. C. Fuerstenberg, "Pedestrian protection using laserscanners," in Proc.IEEE Intell. Transp. Syst. Conf., Sep. 2005, pp. 437–442.
[52] D.M. Gavrila, "The Visual Analysis of Human Movement: A Survey," Computer Vision and Image Understanding, vol. 73, no. 1, pp. 82-98, 1999.
[53] D.M. Gavrila, "A Bayesian Exemplar-Based Approach to Hierarchical Shape Matching," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 8, pp. 1408-1421, Aug. 2007.
[54] D.M. Gavrila and S. Munder, "Multi-Cue Pedestrian Detection and Tracking from a Moving Vehicle," Int'l J. Computer Vision, vol. 73, no. 1, pp. 41-59, 2007.
[55] B. Stenger, A. Thayananthan, P.H.S. Torr, and R. Cipolla, "Model-Based Hand Tracking Using a Hierarchical Bayesian Filter," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 9, pp. 1372-1385, Sept. 2006.
[56] K. Toyama and A. Blake, "Probabilistic Tracking with Exemplars in a Metric Space," Int'l J. Computer Vision, vol. 48, no. 1, pp. 9-19, 2002.
[57] T.F. Cootes and C.J. Taylor, "Statistical Models of Appearance for Computer Vision," technical report, Univ. of Manchester, 2004.
[58] T. Heap and D. Hogg, "Improving Specificity in PDMs Using a Hierarchical Approach," Proc. British Machine Vision Conf., pp. 80-89, 1997.
[59] T. Heap and D. Hogg, "Wormholes in Shape Space: Tracking through Discontinuous Changes in Shape," Proc. Int'l Conf. Computer Vision, pp. 344-349, 1998.

[60] A. Baumberg, “Hierarchical Shape Fitting Using an Iterated Linear Filter,” Proc. British Machine Vision Conf., pp. 313-323, 1996.
[61] M. Bergtholdt, D. Cremers, and C. Schno¨ rr, “Variational Segmentation with Shape Priors,” Handbook of Math. Models in Computer Vision, N. Paragios, Y. Chen, and O. Faugeras, eds., Springer, 2005.
[62] M. Enzweiler and D.M. Gavrila, “A Mixed Generative-Discriminative Framework for Pedestrian Classification,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2008.
[63] M.J. Jones and T. Poggio, “Multidimensional Morphable Models,” Proc. Int’l Conf. Computer Vision, pp. 683-688, 1998.
[64] S. Munder, C. Schno¨ rr, and D.M. Gavrila, “Pedestrian Detection and Tracking Using a Mixture of View-Based Shape-Texture Models,” IEEE Trans. Intelligent Transportation Systems, vol. 9, no. 2, pp. 333-343, June 2008.
[65] H. Sidenbladh and M.J. Black, “Learning the Statistics of People in Images and Video,” Int’l J. Computer Vision, vol. 54, nos. 1-3, pp. 183-209, 2003.
[66] Y. Wu and T. Yu, “A Field Model for Human Detection and Tracking,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 5, pp. 753-765, May 2006.
[67] T.F. Cootes, S. Marsland, C.J. Twining, K. Smith, and C.J. Taylor, “Groupwise Diffeomorphic Non-Rigid Registration for Automatic Model Building,” Proc. European Conf. Computer Vision, pp. 316-327, 2004.
[68] L. Fan, K.-K. Sung, and T.-K. Ng, “Pedestrian Registration in Static Images with Unconstrained Background,” Pattern Recognition, vol. 36, pp. 1019-1029, 2003.
[69] T. Randen and J.H. Husøy, “Filtering for Texture Classification: A Comparative Study,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 4, pp. 291-310, Apr. 1999.
[70] A. Mohan, C. Papageorgiou, and T. Poggio, “Example-Based Object Detection in Images by Components,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 4, pp. 349-361, Apr. 2001.
[71] H. Shimizu and T. Poggio, “Direction Estimation of Pedestrian from Multiple Still Images,” Proc. IEEE Intelligent Vehicles Symp., pp. 596-600, 2004.
[72] R. Lienhart and J. Maydt, “An Extended Set of Haar-Like Features for Rapid Object Detection,” Proc. Int’l Conf. Image Processing, pp. 900-903, 2002.
[73] Y. Freund and R.E. Schapire, “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,” Proc. European Conf. Computational Learning Theory, pp. 23-37, 1995.

[74] K. Fukushima, S. Miyake, and T. Ito, “Neocognitron: A Neural Network Model for a Mechanism of Visual Pattern Recognition,” IEEE Trans. Systems, Man, and Cybernetics, vol. 13, pp. 826-834, 1983.
[75] S. Munder and D.M. Gavrila, “An Experimental Study on Pedestrian Classification,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 11, pp. 1863-1868, Nov. 2006.
[76] C. Wo¨hler and J. Anlauf, “An Adaptable Time-Delay Neural-Network Algorithm for Image Sequence Analysis,” IEEE Trans. Neural Networks, vol. 10, no. 6, pp. 1531-1536, Nov. 1999.
[77] B.E. Goldstein, Sensation and Perception, sixth ed. Wadsworth, 2002.
[78] S. Agarwal, A. Awan, and D. Roth, “Learning to Detect Objects in Images via a Sparse, Part-Based Representation,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, no. 11, pp. 1475-1490, Nov. 2004.
[79] B. Leibe, N. Cornelis, K. Cornelis, and L.V. Gool, “Dynamic 3D Scene Analysis from a Moving Vehicle,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2007.
[80] B. Leibe, E. Seemann, and B. Schiele, “Pedestrian Detection in Crowded Scenes,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, pp. 878-885, 2005.
[81] E. Seemann, M. Fritz, and B. Schiele, “Towards Robust Pedestrian Detection in Crowded Image Sequences,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2007.
[82]N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, pp. 886-893, 2005.
[83] V.D. Shet, J. Neumann, V. Ramesh, and L.S. Davis, “Bilattice-Based Logical Reasoning for Human Detection,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2007.
[84] D.G. Lowe, “Distinctive Image Features from Scale Invariant Keypoints,” Int’l J. Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.
[85] L. Zhang, B. Wu, and R. Nevatia, “Detection and Tracking of Multiple Humans with Extensive Pose Articulation,” Proc. Int’l Conf. Computer Vision, 2007.
[86] Q. Zhu, S. Avidan, M. Yeh, and K. Cheng, “Fast Human Detection Using a Cascade of Histograms of Oriented Gradients,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, pp. 1491-1498, 2006.
[87] O. Tuzel, F. Porikli, and P. Meer, “Human Detection via Classification on Riemannian Manifolds,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2007.

[88] K. Mikolajczyk, C. Schmid, and A. Zisserman, “Human Detection Based on a Probabilistic Assembly of Robust Part Detectors,” Proc. European Conf. Computer Vision, pp. 69-81, 2004.
[89] B. Wu and R. Nevatia, “Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet Based Part Detectors,” Int’l J. Computer Vision, vol. 75, no. 2, pp. 247-266, 2007.
[90] P. Sabzmeydani and G. Mori, “Detecting Pedestrians by Learning Shapelet Features,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2007.
[91] N. Dalal, B. Triggs, and C. Schmid, “Human Detection Using Oriented Histograms of Flow and Appearance,” Proc. European Conf. Computer Vision, pp. 428-441, 2006.
[92] M. Enzweiler, P. Kanter, and D.M. Gavrila, “Monocular Pedestrian Recognition Using Motion Parallax,” Proc. IEEE Intelligent Vehicles Symp., pp. 792-797, 2008.
[93] B. Heisele and C. Wo¨ hler, “Motion-Based Recognition of Pedestrians,” Proc. Int’l Conf. Pattern Recognition, pp. 1325-1330, 1998.
[94] S. Lee, Y. Liu, and R. Collins, “Shape Variation-Based Frieze Pattern for Robust Gait Recognition,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2007.
[95]R. Polana and R. Nelson, “Low-Level Recognition of Human Motion,” Proc. IEEE Workshop Motion of Non-Rigid and Articulated Objects, pp. 77-92, 1994.
[96] A.K. Jain, R.P.W. Duin, and J. Mao, “Statistical Pattern Recognition: A Review,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4-37, Jan. 2000.
[97] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio, “Full-Body Recognition System,” Pattern Recognition, vol. 36, pp. 1997-2006, 2003.
[98] I.P. Alonso et al. “Combination of Feature Extraction Methods for SVM Pedestrian Detection,” IEEE Trans. Intelligent Transportation Systems, vol. 8, no. 2, pp. 292-307, June 2007.
[99] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe, “A Boosted Particle Filter: Multitarget Detection and Tracking,” Proc. European Conf. Computer Vision, pp. 28-39, 2004.
[100] C. Sminchisescu, “3D Human Motion Analysis in Monocular Video: Techniques and Challenges,” in Human Motion - Understanding, Modelling, Capture and Animation, vol. 36, A. Elgammal, B. Rosenhahn, K. Reinhard, Eds. New York: Springer, 2008, pp. 185-211.
[101] T.B. Moeslund, A. Hilton, and V. Krüger, “A survey of advances in vision-based human motion capture and analysis,” Computer Vision and Image

Understanding(CVIU), vol. 104 (2-3), pp. 90-126, 2006.
[102] R. Poppe, "Vision-based human motion analysis: An overview," Computer Vision and Image Understanding(CVIU), vol. 108, pp. 4-18, 2007.
[103] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan, "Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis," in Foundations and Trends® in Computer Graphics and Vision, vol. 1 (2-3), 2006, pp. 77-254.
[104] D. Ramanan and D. A. Forsyth, "Finding and Tracking People from Bottom Up," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2003, pp. 467-474.
[105] D. Ramanan, D. A. Forsyth and A. Zisserman, "Strike a Pose: Tracking people by Finding Stylized Poses," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2005, pp. 271-278.
[106] B. Wu and R. Nevatia, "Tracking of Multiple, partially occluded humans based on static body part detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2006, pp. 951-958.
[107] R. Urtasun, D. J. Fleet and P. Fua, "3D People Tracking with Gaussian Process Dynamical Models," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2006, pp. 238-245.
[108] J. M. Wang, D. J. Fleet and A. Hertzmann, "Gaussian process dynamical models," in Neural Information Processing Systems, 2006, pp. 1441-1448.
[109] C. S. Lee and A. Elgammal, "Modeling view and posture manifolds for tracking," International Conf. on. Computer Vision, 2007, pp. 1-8.
[110] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," IEEE Transactions on Pattern Analysis and Pattern Recognition(PAMI) vol. 28, pp. 44-58, 2006.
[111] J. Deutscher, and I. Reid, "Articulated Body Motion Capture by Stochastic Search," International Journal of Computer Vision(IJCV), vol. 61(2), pp. 185-205, 2005.
[112] M. W. Lee and R. Nevatia, "Human Pose Tracking in Monocular Sequence using Multilevel Structured Models," IEEE Transactions on Pattern Analysis and Pattern Recognition(PAMI), vol. 31(1), pp. 27-38, 2009.
[113] R. Kehl and L. V. Gool, "Markerless tracking of complex human motions from multiple views," Computer Vision and Image Understanding(CVIU), vol. 104(2-3), pp. 190-209, 2006.
[114] D. Gavrila, "Vision-based 3D Tracking of Humans in Actions," PhD thesis, Department of Computer Science, University of Maryland, 1996.

[115] J. Carranza, C. Theobalt, M. A. Magnor and H-P. Seidel, “Freeviewpoint video of human actors,” ACM Transactions on Computer Graphics, vol. 22(3), pp. 569-577, 2003.
[116] A. Doucet, N. D. Freitas and N. Gordon, Sequential Monte Carlo Methods in Practice. Springer-Verlag, 2001.
[117] M. S. Arulampalam, S. Maskell, N. Gordon and T. Clapp, “A tutorial on particle filtering for online nonlinear/non-Gaussian Bayesian tracking,” IEEE Transactions on Signal Processing, vol. 50(2), pp. 174-188, 2002.
[118] H. Siddenbladh, M.J. Black, L. Sigal, “Implicit Probabilistic models of human motion for synthesis and tracking,” in Proc. European Conf. on Computer Vision, 2000, pp. 784-800.
[119] H. Sidenbladh, M. Black and D. Fleet, “Stochastic tracking of 3D human figure using 2D image motion,” in Proc. European Conf. on Computer Vision, 2000, pp. 702-718.
[120] C. Sminchisescu and B. Triggs, “Covariance scaled sampling for monocular 3D body tracking,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2001, pp. 447-454.
[121] C. Sminchisescu and B. Triggs, “Kinematic jump processes for monocular 3D human tracking,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2003, pp. 69-76.
[122] Z. W. Tu, S. C. Zhu and H. Y. Shum, “Image segmentation by data driven markov chain monte carlo,” in Proc IEEE International Conf. on Computer Vision, 2001, pp. 131-138.
[123] M. Lee and I. Cohen, “Proposal maps driven MCMC for estimating human body pose in static images,” in Proc IEEE Conf. on Computer Vision and Pattern Recognition, 2004, pp. 334-341.
[124] A. Elgammal and C. S. Lee, “Inferring 3D body pose from silhouettes using activity manifold learning,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2004, pp. 681-688.
[125] N. R. Howe, M. E. Leventon and W. T. Freeman, “Bayesian reconstruction of 3D human motion from single-camera video,” in Neural Information Processing Systems, 2000, pp. 800-826.
[126] A. Agarwal and B. Triggs, “Tracking articulated motion using a mixture of autoregressive models,” in Proc. European Conf. on Computer Vision, 2004, pp. 54-65.
[127] L. Sigal, S. Bhatia, S. Roth, M. J. Black and M. Isard, “Tracking looselimb people,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2004, pp.

421-428.
[128] E. B. Sudderth, A. T. Ihler and W. T. Freeman, “Nonparametric belief propagation,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2003, pp. 605-615.
[129] C. Sminchisescu and A. Jepson, “Generative modeling for continuous non-linearly embedded visual inference,” in Proc. ACM International Conf. on Machine Learning, 2004, pp. 759-766.
[130] N. D. Lawrence, “Gaussian process latent variable models for visualisation of high dimensional data,” in Neural Information Processing Systems, 2004, pp. 329-336.
[131] N. R. Howe, “Silhouette lookup for automatic pose tracking,” in IEEE Workshop on Articulated and Non-rigid Motion, 2004, pp. 15-22.
[132] D. Ramanan and D. A. Forsyth, “Automatic annotation of everyday movements,” in Neural Information Processing Systems, 2003, pp. 329-336.
[133] N. Ikizler and D. A. Forsyth, “Searching for complex human activities with no visual examples,” International Journal of Computer Vision(IJCV), vol. 80(3), pp. 337-357, 2008.
[134] M. Spengler and B. Schiele, “Towards Robust Multi-Cue Integration for Visual Tracking,” Machine Vision and Applications, vol. 14, no. 1, pp. 50-58, 2003.
[135] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, “A Tutorial on Particle Filters for On-Line Non-Linear/Non-Gaussian Bayesian Tracking,” IEEE Trans. Signal Processing, vol. 50, no. 2, pp. 174-188, Feb. 2002.
[136] J. Deutscher, A. Blake, and I.D. Reid, “Articulated Body Motion Capture by Annealed Particle Filtering,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, pp. 126-133, 2000.
[137] M. Isard and J. MacCormick, “BraMBLE: A Bayesian Multiple-Blob Tracker,” Proc. Int’l Conf. Computer Vision, pp. 34-41, 2001.
[138] Z. Khan, T. Balch, and F. Dellaert, “MCMC-Based Particle Filtering for Tracking a Variable Number of Interacting Targets,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 11, pp. 1805-1819, Nov. 2005.
[139] J. MacCormick and A. Blake, “A Probabilistic Exclusion Principle for Tracking Multiple Objects,” Int’l J. Computer Vision, vol. 39, no. 1, pp. 57-71, 2000.
[140] Markus Enzweiler and Dariu M. Gavrila, Integrated pedestrian classification and orientation estimation. CVPR,pp.982-989, 2010.
[141] R. Cucchiara, C. Grana, A. Prati, and R. Vezzani. Probabilistic posture classification for human-behavior analysis. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 35(1):42–54, 2005.

Appendix I Detailed Survey of Review

I.1 Infrastructure design enhancements

Since parked vehicles block the vision of drivers as well as pedestrians, removing on-street parking and implementing diagonal parking in residential streets would help in reducing accidents. For single-lane or low-density roads, there has been no significant difference between fatality rates on marked and unmarked intersections. However, multilane marked crosswalks have greater rate of accidents than the unmarked ones, particularly for multilane crossings. A possible explanation is that multilane roads with heavy traffic at high speeds are difficult to cross for many pedestrians. In such cases, the presence of marked crosswalks may be encouraging people to cross there instead of using a signal-controlled intersection, therefore increasing the number of at-risk pedestrians. Marking intersections should be used in conjunction with other measures described above in order to increase safety. Flashing light warnings on pedestrian crosswalks in order to warn the drivers of the presence of pedestrians is also useful. These lights can be triggered by a button pressed by the pedestrian or an automatic detection system. In India, the solutions proposed include putting regulators on buses, trucks, and other heavy vehicles to limit their speed, segregation of fast and slow traffic, providing safe walking and road crossing facilities, and traffic calming measures such as the use of roundabouts. Measures necessary in urban areas as well as those on highways and rural roads are specifically addressed.

I.2 Project of advanced driver assistance systems

Enhancing comfort and safety of the driver and the occupants of an automobile has been a major motivator in the innovations associated with Intelligent Vehicles and Intelligent Transportation Systems. In the United States, the Turner–Fairbank Highway Research Center, which is affiliated with the Federal Highway Administration (FHWA), conducts research on various topics related to transportation. In particular, the

Pedestrian and Bicycle Safety Research Program [18] seek to enhance the safety and mobility of pedestrians and bicyclists. The Pedestrian smart program [19] has the objective of applying the ITS technology to improve pedestrian safety. They have developed various devices that provide feedback to the waiting and crossing pedestrians as well as the motorists. They have also developed the software called the Pedestrian and Bicycle Crash Analysis Tool to analyze the interactions between pedestrians, bicyclists, and motor vehicles. This tool has an application for developing and testing countermeasures for enhancing pedestrian safety. The California Partners for Advanced Transit and Highways [20] conducts research on transportation safety issues, including pedestrian protection, driver behavior modeling, and intersection collision prevention. In particular, they have performed research on analyzing the collision behavior at marked and unmarked crosswalks, automatic pedestrian detection systems at intersections, and LED signals to alert drivers to the presence of pedestrians.

The European Union has been conducting several projects in collaboration with industry and research institutes for intelligent vehicle systems in general and pedestrian safety in particular. The project PReVENT [21] deals with the development of safety technologies which help drivers prevent or mitigate the effects of an accident using sensor-based analysis of surroundings as well as the state of the driver. In particular, the sub project COMPOSE focuses on detection of pedestrians, cyclists, and other vehicles using data fusion from sensors and protection using autonomous or semiautonomous braking. The PROTECTOR project and its successor SAVE-U were particularly focused on reducing accidents involving vulnerable road users [22]. The European project PUVAME proposes an infrastructure-based solution to prevent collisions between vulnerable road users and transit buses. They use off-board cameras that observe intersections and bus stops to track the movement of buses as well as vulnerable road users.

In Japan, Infrastructure, and Transport has promoted the Advanced Safety Vehicle project, which has spanned over three 5-year phases between 1991 and 2005 [23]. The final phase of Advanced Safety Vehicle emphasized car-to-car communications in order to improve safety. Pedestrian detection was an important component of this research. Systems that warned the driver about the presence of pedestrians while making turns were demonstrated at the Tokyo Motor Show [24].

I.3 3D human pose estimation

Human body pose estimation has recently received great interest from the computer vision community. However, most researches consider 3D human pose estimation techniques. Agarwal and Triggs developed a learning-based method for estimating the 3D body poses of people from monocular images as well as video sequences. The approach uses the histogram of shape context descriptors as feature and various regression methods for estimating the 3D pose. Besides, work in the domain of 3D human pose estimation, a few approaches have recovered an estimate of pedestrian orientation based on 2D lower-resolution images. Cucchiara et al. distinguished among various human postures such as standing, crouching, sitting, and lying down using Probabilistic Projection Maps on 2D silhouettes. However, most of the systems use an accurate silhouette of the pedestrian, which may not always be available, especially from a moving vehicle. More details about 3d human pose estimation please refer to the appendix.

Human body pose estimation has recently received great interest from the computer vision community. With lower costs of cameras and advances in computing power, accurate analysis of 3D human pose from video can help surveillance operators to identify events such as running, walking, shop-lifting, wall-climbing, loitering and other abnormal human activities.

According to a survey [100], there are few technical challenges to be addressed in single camera pose estimation such as depth ambiguities, high-dimensional representation of human pose, self-occlusion, unconstrained motions, observation ambiguities, motion blurs and unconstrained lighting. Depth ambiguities arise because 3D world is projected into a 2D image, causing loss of depth information. Loss of depth information cannot be recovered using single camera only. For this reason, recovering full body human pose is an ill-posed problem since kinematic tree/skeleton is commonly used to represent 3D human body. Without depth information, it is challenging to reconstruct skeleton in 3D. Furthermore, it is common that skeleton is modeled according to human anatomy where limbs (hand, legs, elbow, arms, etc.) and torso are modeled using 30-60 joint angle variables. However, estimation of these joint angle variables is computationally expensive because of high dimensional space. Self-occlusion occurs frequently in a single camera view as one body part tends to hide another body part during motion of these articulated limbs. Unconstrained motions are the result of highly diversified human movements; withstanding human movements can not be highly structured at the same time. For instance, walking or running has repetitive human pose structure over time although it also shows large pose variations at different speed, acceleration and at different body size. Observation ambiguities occurs

because single image observation can be mapped to more than one possible 3D human pose; it is difficult to disambiguate 3D human pose without depth information. Lighting inevitably changes at different environment as a function of space and time. Lighting variations affect image observations for estimation of body pose. For instance, silhouette shape of the same human pose may appear differently at different capturing time. While capturing rapid human motion, slow camera shutter time causes blurring of image objects; this affects the quality of image observations also.

There are few related surveys published in the area of vision-based human motion analysis, all [101, 102, 103] except one [100] give broad overview of vision-based human motion analysis using the taxonomy of detection, tracking, pose estimation and recognition. T. B. Moeslund [101] review advances made in human motion analysis and capture from 2000-2006, extending his previous survey [104] to include new research directions such as detection and tracking of human in natural environment rather than laboratory environment, model-based pose estimation approaches where motion and stochastic sampling framework are employed to search for optimal state (human pose) given the image observations. Their survey divided papers into different taxonomy such as initialization, tracking, pose estimation and recognition. R. Poppe [102] review about advances human motion analysis such that human pose estimation problems are divided into two main classes: model-based approaches and model-less approaches. Sminchisescu [100] categorized single camera reconstruction of full-body 3D human motion problems into generative or discriminative approaches. Generative approaches build and optimize objective function to match image observations so that the correct human pose hypotheses should maximise observation likelihood within the probabilistic framework. Discriminative methods formulate pose estimation into recognition problem, pose estimation is predicted by trained model using training sets consisting of joint pose and image observations. Discriminative approaches use machine learning extensively to predict state distributions in the absence of depth information. D. A. Forsyth et al. [105] review methods to track human body from video focus on tracking and motion synthesis. 3D body pose can be inferred by lifting 2D pose to 3D pose. They believe that ambiguities during lifting can be partially if not completely solved whenever motion, geometric and context information are incorporated appropriately into probabilistic framework. Recent new research directions have emerged such as (1) Bottom up approach, detection of local parts (hand, legs, torso) using data-driven approach before pose estimation [101,104,105,106], (2) Learning in low-dimensional pose manifold rather than original high-dimensional appearance manifold, [107], and (3) Learning of nonlinear dynamics of human motion models for smoother 3D human pose from video

sequence [107,108,109].

I.4 3D Human Body Model

Full human body is highly articulated structure but body parts can be considered as rigid structure. To that end, there are few body models to represent articulated 3D human pose. In most cases, 3D human pose can be represented by kinematic tree model, consisting of segments linked by joints. In kinematic tree model, joints are consider non-articulated and can have maximum three degrees of freedom (DOF) corresponding to three orthogonal directions. Number of joints and DOF required depends on the degree of details required in the application. Whereas for spatial resolution of human in the image, smaller the spatial resolution of the human, smaller the number of DOF required and vice versa. Even though detailed DOF can produce more realistic human pose, it also increases computational complexity as estimation now has to be performed at higher dimensional space. Therefore trade-off must be made to balance degree of details required against computational complexity. Papers that use kinematic tree models are [110]. Besides kinematic model, human pose can also be represented by volumetric models such as elliptical cylinder. This volumetric model have all the limbs fleshed out by elliptical cylinder; papers using this model are [111, 112]. Lee [112] represents 3D human body by both kinematic tree model and volumetric elliptical cylinder. His model can simultaneously describe human shape of different body size and clothing that a person wears. Another volumetric model is super quadrics [113] and generalized cones [114]. Volumetric base models representation have limited description capabilities when comes to variations of body size. In older works, width or length of limbs in elliptical cylinder model is manually fixed during initialization for computational convenience. However, some researchers [112,115] have started to take this issue into consideration by recovering the parameters of limbs automatically during initialization. Of course, this will take additional steps thereby increasing computational complexity.

I.5 Single camera 3D human pose estimation

Pose estimation can be formulated as optimization problem since the main concern is to find the pose parameters that minimize errors between the image observations and the

2D projected 3D body pose. Human pose estimation can also be regarded as inferring the underlying kinematic structure (in the form human skeletal) from image observations. In single camera problem, it is common to impose additional constraints (kinematics, motions, or geometrical information) to reduce the ambiguities that appear in the inferred 3D body configurations. Without use of constraints, there are many possible 3D configurations that could explain a single observation manifested in multiple modes in likelihood function.

To deal with multiple mode problem in the posterior (likelihood function), randomized search in the form of stochastic sampling method particle filters are commonly employed. Few publications explaining the idea behind particle filters are given in [116].

The key idea behind particle filters is to use randomized sampling to search the posterior because the distribution of posterior is non-Gaussian in general. To that end, particle filters approximate the posterior distribution by sets of points concentrating around places where large values of likelihood are found. The posterior evolves over time using assumed underlying dynamical state space model to predict time varying configurations. The sampled representation of the new posterior are then the predicted prior of body configuration. The prior body configuration is later matched to image observations, and those sets of points that give good comparisons are given more weights, resulting in new representation of the desired posterior distribution. Particle filters are becoming important for applications that need approximate model to explain the underlying dynamics of time-varying physical system that typically shows nonlinearity/non-Gaussianity (when Kalman filtering fails) in their posterior distribution. For instance, one can use particle filtering to track multivariate data in time-series problem. A good tutorial of using particle filters on tracking problem can be found in [117].

However, one disadvantage of using particle filters in estimating 3D body pose is high data dimensionality. One 3D human body can have at least 20 degrees of freedoms according to D. Forsyth [103] (one at each knee, two at each hip, three at each shoulder, one at each elbow and six for the root). To deal with problems of high-dimensionality, researchers focus their efforts on building more efficient search methods rather than improving the core algorithm of particle filters.

Sidenbladh et al. [118, 119] use important sampling method to guide particle filtering search on likelihood either on the learnt walking model or on motion database. The importance sampling method use a proposal distribution as alternative to the prior likelihood function so that samples can be drawn in places that are more likely. Another

variant of importance sampling is to use annealed search by Deutscher et al. [120]. This method shows improvement in speed than the former method in tracking 3D pose of a subject. However, this method uses three cameras to track a single person under simple black background. Moreover, the method shows no experimental evidence that it can work in natural environment when clutters and other textures are all too common.

Sminchisescu et al. [120] present an extension to particle filtering method to recover 3D human pose from single camera image sequences. They believe that for successful single camera 3D body tracking, at least three difficulties need to be resolved such as (1) to estimate 30 joint parameters, (2) to estimate depth information to recover the unobservable 1/3 DOF (3) to match image observations with complex body model under self-occlusion and cluttered background. They design cost matching function using combination of edge, optical flow and motion boundaries while enforcing hard joint angle limits with non self-intersection constraints. Covariance scaled sampling method is used as search strategies to find good poses in high-dimensional body configuration space. From their observations, cost minima occur most likely along the local valley of cost surface where covariance has highly uncertain directions. The hypothesis distribution is determined along these highly uncertain directions while combining with certain temporal dynamic models. Good 3D human body poses are sampled using random or regular pattern from this hypothesis distribution for rescaled covariances. Results demonstrate robust tracking of entire arm and full-body 3D for those video sequences that contain self-occlusion and cluttered background. In further research [121], they improve the speed of the search for local minimum by constructing an interpretation tree that can generate many possible 3D human poses to explain the same image observations by introducing inverse kinematics. Experiments show that it can track 3D human body in short video sequences that contain fast, unpredictable and complex motion (such as dancing) under cluttered backgrounds.

In recent work, Lee et al. [112] introduced novel methods to address automatic initialization issue in monocular 3D human pose tracking by estimating multiple human positions and sizes before inferring their corresponding 3D poses. Automatic initialization is seldom addressed and is an important step to bootstrap pose estimation for many applications. In their approach, body part positions are estimated by the head, shoulders and limbs detector modules. Multiple image cues/observations such as skin color, head shoulder shapes are used in tandem with learnt detector to locate the head position. Torso location is estimated based on head position (just below the head). In the second stage, belief propagation technique is used to refine the positions of the body parts from the earlier detection. In the last stage, data-driven Markov chain Monte Carlo

(MCMC) [122, 123] algorithm is used to estimate 3D human pose in each frame. MCMC is a variant of stochastic sampling method used in particle filtering framework to explore solution spaces by carefully designing proposal function to generate optimal candidate states (body pose). In their work, proposal function to evaluate likelihood of a candidate state is formulated using four criteria (1) region consistency (2) color dissimilarity with background (3) skin color and (4) foreground matching.

Although human pose can be estimated from static images only, motion prior helps in smooth tracking of human pose by enforcing strong constraints. Recent approaches mainly concentrate on learnt motion model to obtain realizable human poses and motions [119, 124, 125]. Learnt motion models often use motion capture training data. One best known publicly available motion capture database is from CMU. Nevertheless, one weakness with learnt motion models approach is the excessive dependency on the amount and quality of motion capture training data. There must be sufficiently large amount of training data to accurately learn the representation of all possible human motions. Besides, some motion capture data is noisy because of the presence of random and systematic errors inherent during human motion measurement capture.

Howe et al. [125] presented a system to reconstruct 3D human motion using single camera. In each frame, they track the entire body using learnt 2D body parts detector. Motion capture data is assembled into snippets (motions consisting of 11 successive frames) are later used to train mixture-of-Gaussians probability density functions for few classes of 3D body configurations. The output of 2D body parts are matched to the probability density functions to find the corresponding 3D body configuration. Siddenbladh et al. [119] solved similar problem using generative model. Generative model determines the likelihood of observing certain image observations (shape, appearance and motion features) given a state (3D human pose and movement). Learnt motion models are determined using previous history of states from motion capture data. Particle filtering optimization is later used to find the approximate states. Agarwal et al. [126] presented a novel approach that is able to track unseen human pose not in motion capture training data in the presence of complex background. Their method track 2D human pose but readily extends to 3D when needed. Rather than learning the whole state space parameters, they partition the state space parameters into regions with similar dynamical characteristics. This facilitates learning of nonlinear dynamics using piece-wise linear autoregressive process for each region. Bottom-up processes are also been incorporated into top-down processes as illustrated in [127]. The advantages in this approach are automatic initialization in contrast to manual initialization and recovery from tracking failures. In [127], body is represented as graphical model. Each node

corresponds to a body part and edges between nodes represents statistical dependencies and physical constraints. Probabilistic models are learned from motion capture training sets to capture temporal evolution of each node over time. 3D human pose at any time instant is then recovered by probabilistic inference using non-parametric belief-propagation [128].

One major problem in estimating 3D human pose is high dimensionality of pose space data. This problem poses great challenge to machine learning approaches including particle filtering framework because number of available training samples are often too small to cover all possible human movement intricacies. This phenomenon is the well known "curse of dimensionality". To avoid this phenomenon, recent research trends concentrate on discovering methods to reduce the data dimensionality prior to estimation [124]. Moreover, these approaches are also motivated by the findings that human pose can be sufficiently represented by low-dimensional latent manifolds as shown in [124]. To that end, these approaches attempt to learn low-dimensional mapping functions relating human pose to the image observations. However, learning mapping functions are difficult because the manifolds are nonlinear. To recover 3D human pose from image observations, two mapping functions are learned such as mapping of image observations space to pose space and its corresponding inverse mapping.

Elgammal et al. [124] introduced a method to reconstruct 3D body from a given viewpoint from silhouettes information using single camera. Local linear embedding (LLE) is used to learn mapping of pose space to silhouette space and its corresponding inverse mapping. Unseen 3D human pose (not part of the training data) is recovered by interpolation by radial-basis function (RBF). While their earlier work [124] is strictly view dependent and limited to walking pose activity, Elgammal et al. [109] in his later work modeled 3D body pose of a person observed at different viewpoints and extensible to general human motions. Body pose and viewpoint are explicitly modeled in two separated low dimensional representations. In similar spirit, Sminchisescu et al. [129] proposed the use of spectral embedding algorithm to learn mapping of image observation space to low-dimensional manifold pose space and its inverse mapping separately. Tracking of human pose is later constrained to the learnt low-dimensional manifolds. Agarwal et al. [110] implicitly achieved data dimensionality reduction using relevance vector machines (RVM) regression. RVM selects only the "most relevant" basis function by retaining only the relevant input features. As a consequence, large training data are reduced to a minimal subset.

Although LLE and spectral embedding methods can learn low-dimensional

embedding manifolds from data, they lack probabilistic interpretation. This suggests that no straightforward learning-based method can be applied to the learnt low-dimensional manifolds. It is also difficult if not impossible to find inverse mapping between low dimensional latent space back to image observations space. Urtasun et al. [130] proposed the use of Scaled Gaussian Process Latent Variable Model (SGPLVM) [130] that admits probabilistic interpretation to learn human pose prior with continuous mapping between observation space and pose space. Human pose is recovered by finding body pose that maximizes the likelihood of the learnt SGPLVM model given sets of image observations. Results demonstrate good tracking accuracy of 3D body pose for both walking and golfing activity but in their experiments 3D body positions are manually initialized. SGPLVM has generalized well even with small available training set. In similar work, SGPLVM was extended by incorporating additional nonlinear dynamics mapping while retaining the original mapping and its corresponding inverse mapping obtained through Gaussian Process Dynamical Model (GPDM) [107]. GPDM produces smoother motion models compared to SGPLVM. In recent work, the fact that the tasks of body estimation are strongly correlated with activity recognition motivated T. Jaeggli et al. to introduce method to simultaneously track human pose and recognition of multiple action categories. In similar spirit they use LLE as mapping of body pose to low-dimensional space, while kernel regressor as inverse mapping back to original body pose. Low dimensional models are learned separately for different activities; for instance each low-dimensional manifolds are dedicated to walking and running respectively. To model activity switching, nonlinear mapping between pair of activities are also modeled. Likelihood function using Gaussian distribution is used as probability measures to determine activity transition. Experiments demonstrate that their approach can reliably track subjects while recognizing activity transition simultaneously even with low-resolution video.

Vast pose estimation literature sees major problem lies in managing multiple modes in likelihood function in high dimensional data, which explains the extensive use of particle filtering variants. Some evidence [103], though inconclusive, seems to suggest that ambiguities may not persist when short motions (snippets) are used in place of single frame. Howe et al. [131] reconstructed 3D body pose by comparing 3D motion capture data with 2D snippets via dynamic programming. Ramanan et al. [132] used similar approach to lift 2D snippets into full 3D body pose by matching them to the stored 3D motion capture database with assumption camera is in lateral view. Best matching 3D pose is also recovered by dynamic programming. The same approach is used in [133] to build viewpoint invariant human activities retrieval system. 3D body

pose is constructed by lifting the output from 2D limb detectors [105]. One disadvantage in this approach is one needs large 3D motion capture database in order to lift 2D pose into good 3D pose let alone massive computation incurred during the matching process. Therefore, this approach remains as an open research problem. It remains to be seen on how one can reconstruct 3D body pose from 2D snippets using smaller motion capture database.