8-2016

Visual Clutter Study for Pedestrian Using Large Scale Naturalistic Driving Data

Kai Yang
Purdue University

Graduate School Form 30
(Revised 08/14)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Kai Yang

Entitled: Visual Clutter Study for Pedestrian Using Large Scale Naturalistic Driving Data

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

EDWARD J. DELP, Co-Chair

MARY L. COMER

YINGZI DU, Co-Chair

BRIAN S. KING

MAHER E. RIZKALLA

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): EDWARD J. DELP, Co-Chair
~~YINGZI DU, Co-Chair~~

Approved by: V. Balakrishnan 05/31/2016

Head of the Department Graduate Program Date

VISUAL CLUTTER STUDY FOR PEDESTRIAN USING LARGE SCALE NATURALISTIC DRIVING DATA

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Kai Yang

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2016

Purdue University

West Lafayette, Indiana

For my family

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge my thesis advisor Dr. Eliza Y. Du, for her consistent assistance and guidance during all my courses and research projects related to this thesis work.

I would like to thank my co-advisor Dr. Edward J. Delp for his assistance, guidance, and supervision during my PhD study. I would also like to thank the advisory committee members, Dr. Mary Comer, Dr. Brian King, and Dr. Maher Rizkalla for their time and insight during construction of this thesis.

Finally, I express my gratitude to my family for their support and encouragement during all my life, and my friends for helping me out in time of need.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

ABSTRACT

Yang, Kai Ph.D., Purdue University, August 2016. Visual Clutter Study for Pedestrian Using Large Scale Naturalistic Driving Data. Major Professors: Eliza Du, Edward J. Delp.

Some of the pedestrian crashes are due to driver's late or difficult perception of pedestrian's appearance. Recognition of pedestrians during driving is a complex cognitive activity. Visual clutter analysis can be used to study the factors that affect human visual search efficiency and help design advanced driver assistant system for better decision making and user experience. In this thesis, we propose the pedestrian perception evaluation model which can quantitatively analyze the pedestrian perception difficulty using naturalistic driving data. An efficient detection framework was developed to locate pedestrians within large scale naturalistic driving data. Visual clutter analysis was used to study the factors that may affect the driver's ability to perceive pedestrian appearance. The candidate factors were explored by the designed exploratory study using naturalistic driving data and a bottom-up image-based pedestrian clutter metric was proposed to quantify the pedestrian perception difficulty in naturalistic driving data. Based on the proposed bottom-up clutter metrics and top-down pedestrian appearance based estimator, a Bayesian probabilistic pedestrian perception evaluation model was further constructed to simulate the pedestrian perception process.

1. INTRODUCTION

1.1 Motivation

In United States, National Highway Traffic Administration (NHTSA) reports that 4280 pedestrian were killed in the traffic crashes in 2010, with around 70000 injuries [1]. In Europe, more than 30000 people were killed on road in 2011 based on the European Commission data [2]. Pedestrian safety is a worldwide public safety and health issue.

Among them, some of the crashes are due to driver's late or difficult perception of pedestrian appearance. Perception of pedestrian appearance during driving is a complex cognitive activity. It may be affected by varied factors, such as driving scenarios, background complexity, illumination conditions, pedestrian appearance etc. Exploring and understanding the factors which may affect pedestrian perception difficulty by driver could be interesting and meaningful for both researchers and road safety practitioners. First of all, it could enable deeper insight into human visual perception process/model by providing evidences from real life visual attention task. Secondly, the results may be very valuable for safer road component design. Thirdly, a computational model with quantitative analysis methods of pedestrian perception could be the basis for more reliable pedestrian active safety system with better decision making and user experience.

Visual clutter [3, 4] has been proposed to represent the highly variable visual information that may lead to a degradation of some tasks. It may interfere with quickly and precisely gathering information and making decisions. Visual clutter is closely related to visual attention/perception ability. Visual clutter analysis can provide significant information to study and justify visual attention/perception model. Edquist [5] claimed that cluttered driving environment has been shown to impair driving performance, e.g., increasing the driver's response time to detect changes, impairing the detection of road signs, etc. Visual

clutter analysis could be used to study the factors that may affect the driver's ability to perceive pedestrian appearance. However, most of the previous study relied on conducting human subject tests using limited visual stimuli (e.g. scenery photographs, synthetic driving scene, etc) and focused on the visual search task, which may not be suitable for exploring pedestrian perception in naturalistic driving scenarios. A comprehensive study of understanding of pedestrian perception during driving is encouraged by the application of large scale naturalistic driving data in driver behavior study. Moreover, an automatic and quantitative framework of pedestrian perception analysis, including automatic pedestrian detection, visual clutter computation and pedestrian perception estimation, is meaningful for potential incorporation into current intelligent transportation system. The lacking of both the theoretical and practical analysis methods of pedestrian perception encourages the topic and exploration in this thesis.

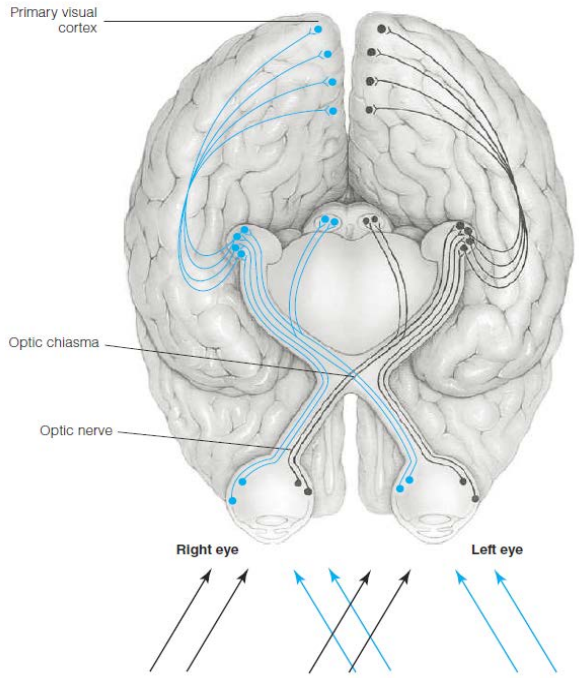
1.2 Background

We briefly introduce and review some backgrounds and researches which are closely related to the topic in this thesis, including the biological mechanism of human visual perception process, visual attention models, visual clutter analysis models and visual clutter study using naturalistic driving data.

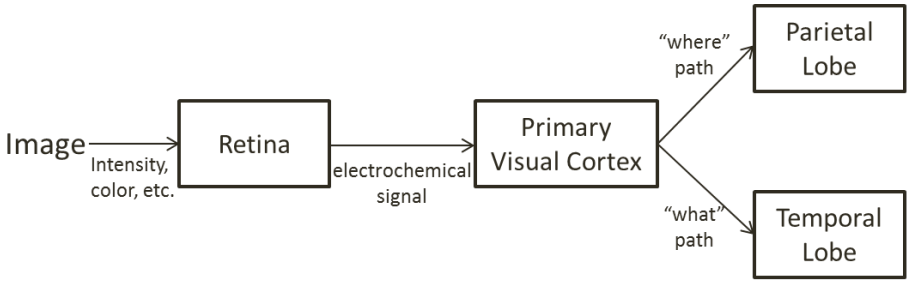
1.2.1 Human visual perception models

Visual perception is the ability to interpret the surrounding environment by processing information that is contained in visible light. Figure 1.1 (a) and (b) illustrate the primary visual path within human visual system and an oversimplified visual perception model[6]. Intensity, color, edge and other features from the visual scene are sensed by the photoreceptor and formed an image on the retina located on the back of eyeball. The light signal is then converted to electrochemical signal and transmitted to brain via the optic nerves. The signal is received by the Lateral Geniculate Nucleus (LGN) and sent to the primary visual cortex (V1) by multiple layers of neurons. Two pathways [7] can be identified within the visual cortex: a dorsal or mangocellular pathway reaches to the parietal lobes and mainly encodes the spatial and motion information ("where"), and a

pavocellular or ventral pathway leading to the temporal lobes and is concerned with detailed visual information used for the recognition of objects ("what"). While the actual functional and structural complexity of the visual system is far more than the oversimplified model in Figure 1.1(b), the two-pathway theory is valid and widely accepted.



(a)



(b)

Figure 1.1 (a) Human vision system. (b) Oversimplified flowchart of human visual perception mechanism[6][7]

During the past several decades, a wide variety of visual attention models have been proposed in psychology field to simulate human perception. The bottom-up process originates from sensory information and is driven by the physical data. It senses from individual parts to the whole images. In contrast, the top-down process originates from cognitive information and is driven by our knowledge, expectation and goals. It usually senses from the whole image to individual parts. The two types of models are further integrated and combined to explain the visual search and recognition process.

The two-stage pre-attention-recognition model [4, 8] has been widely accepted and studied. It claimed that when human vision system perceives a particular target from a complex background, a pre-attentive stage is first initiated to detect basic features in parallel and then bind those features into a selective attention area/object. During the pre-attentive stage, the visual scene parts are parallel sensed in a bottom-up way and generate a weighted representation indicating the varied levels of visual response. During the recognition stage, top-down knowledge plays the main role and helps to disambiguate the objects from the noisy bottom-up weighted representation. Here is an overview of some popularly used visual attention models.

The Feature Integration Theory (FIT) model

Treisman and Gelade [8] introduce the Feature Integration Theory (FIT) to explain the visual attention mechanism which is considered as one of the earliest seminal work of computational visual attention model. Multiple separated feature maps are computed from the low level features (e.g. intensity, color, orientation) within the entire visual field in parallel. The separated feature maps are then combined to generate a master map to guide the attention (Figure 1.2). The master map indicates the bottom-up feature saliency within the entire visual field leading to a serial scanning directs the focus of attention towards selected scene entities. Object profiles are learned from the target location within the master map and could be served as top-down knowledge for higher perception tasks.

Two different ways of object identification was discussed in the FIT model: a bottom-up focal attention process and a top-down recognition process. It is claimed that the two routes may act together while they could be independent during extreme cases. The first route performing object identification depends on focal attention from different locations to integrate the features registered within the same spatio-temporal “spotlight” on to a particular object. The second route may act when focused attention are blocked by overloading. Top-down process achieves the identification by predicting the context of the environment and matching the disjunctive features to those in the scene.

The FIT model also provided important insights into the preattentive processing by studying the possible preattentive features and how the preattentive process is performed by human visual system. The possible preattentive features were found by conducting experiment in which the subjects were asked to find target among distractors. If the predefined response time and accuracy thresholds can be achieved regardless of the number of the distractors, the task is said to be preattentive. The FIT model well explained the preattention mechanism: one can access the individual feature maps, which were believed to be preattentive features, and quick complete the search/perceive task; while a conjunction target of multiple features cannot be detected by accessing individual feature maps, therefore requiring longer response time with lower accuracy.

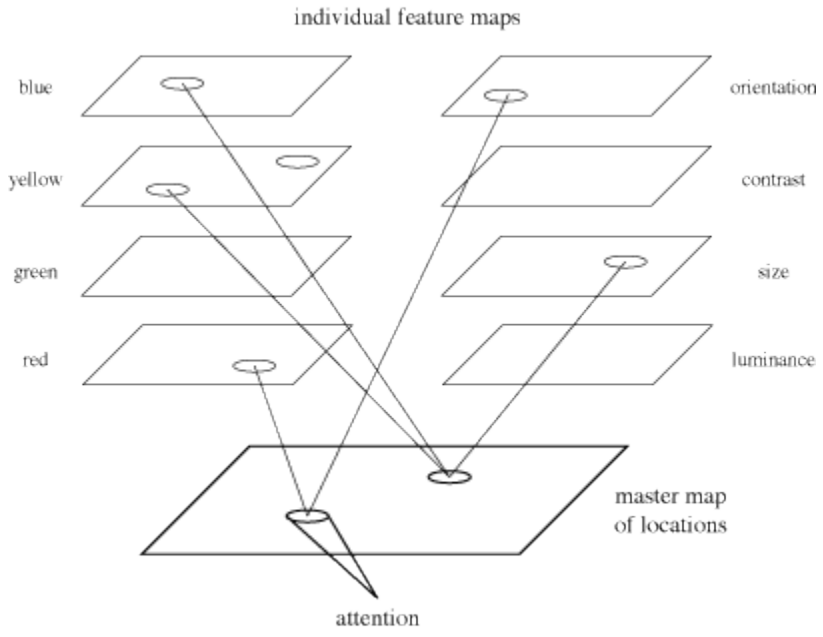


Figure 1.2 Feature Integration Theory [8]

The "Spotlight" and "Zoom lens" model

Attention was initially compared to a "*spotlight*" indicating its selective mechanism [9, 10]. The visual process will be enhanced within the illuminated spotlight area of a few degrees of visual angle. Later, Eriksen and James [11] proposed to modify the *spotlight* into a *zoom lens model* to explain the visual attention process. They claimed that the visual attention area size could be varied depending on the task similar to a zoom lens. This model relies on the natural low level visual features therefore do not take high-level object appearance into consideration. The zoom lens analogy suggests that the density of the visual processing resource may decrease as the size of the attention area increases.

The Guided Search model

Wolfe [4, 12, 13] proposed the Guided Search theory which shares a lot of concepts with FIT. Feature maps are computed from different types of low-level features in parallel and a master activation map is combined by summing all the computed feature maps. In contrast with FIT, the activation map emphasizes top-down knowledge to weigh the

relative bottom-up feature map by selecting the feature type that best distinguishes the target from its distractors.

Figure 1.3 shows the model of Guided Search theory. The activation map based on both bottom-up and top-down information is constructed during visual search. Wolfe believes the early vision divides the image into different feature maps. Each feature type (*e.g.* color, orientation) has one corresponding feature map. Different feature maps may have different relationship with each other. Bottom-up activation measures how different an element is from its neighbors and such difference is computed and combined. Top-down activation is driven by user, including searching purpose, knowledge, searching experience and so on. The final activation map is a combination of bottom-up and top-down activations, with task dependent weights assigned to each feature map.

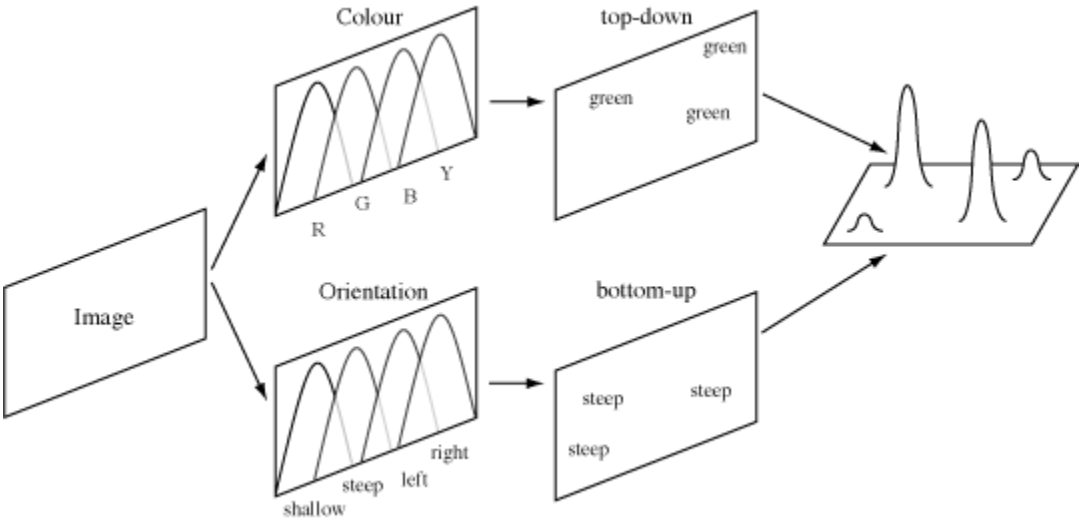


Figure 1.3 The guided search model [4]

The Biased Competition Model

Desimone and Duncan [14] proposed a similar visual attention model which combined both bottom-up feature maps and top-down priors to guide attention. Competitions are involved when two or more bottom-up stimuli are exciting the attention. The bottom-up stimuli are influenced by a top-down modulation and the relative responses are biased.

The biased competition model implies the prioritizing of task relevant visual information. The visual system only have limited bandwidth available for processing therefore a mechanism to select relevant information and ignore irrelevant stimuli is reasonably built by the model.

Bottom-up Saliency Map

Koch and Ullman [15] proposed a pure bottom-up computational architecture of visual attention. Based on FIT, this model relies on computing conspicuities from several types of low-level features and constructing a bottom-up saliency map to guide attention. A *winner-takes-all* (WTA) neural network was proposed to determine the most salient location within the entire visual field. The selected most salient location is then routed to a central presentation containing only features within the routed region which simulates the fixation process of human vision system.

Based on Koch and Ullman's theory, Itti *et al.* [16] proposed a detailed bottom-up computational model (Figure 1.4) which is one of the most popularly used models. Multiscale features, including intensity, color and orientations are computed using image pyramid and combined into a topographical saliency map. A dynamic neural network which involves the global inhibition of WTA and local inhibition effect are then activated to select the attention locations based on the computed saliency map.

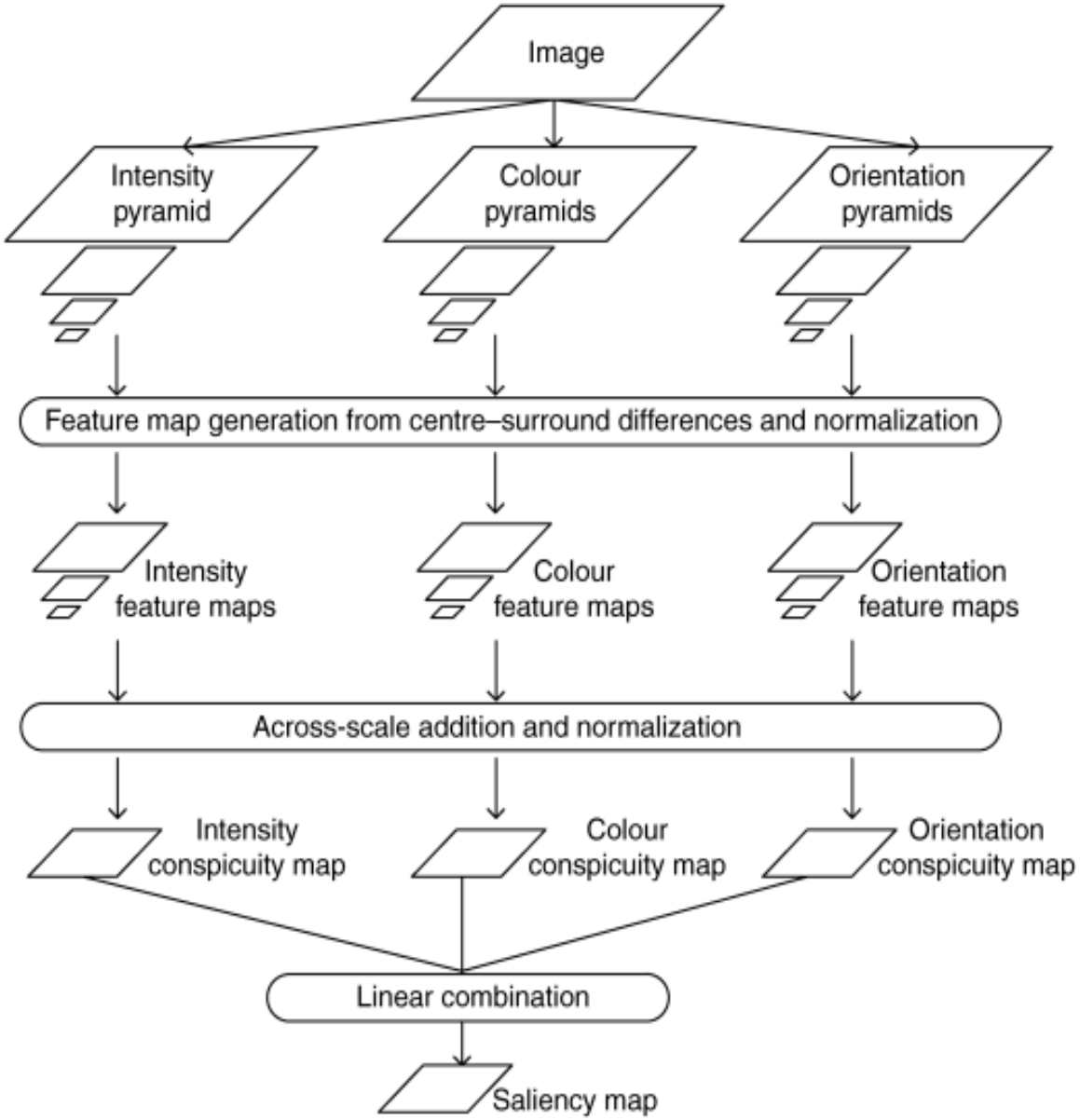


Figure 1.4 Bottom-up saliency map [16]

The Connectionist Models

Besides the aforementioned computational models which compute the feature maps using linear filters, connectionist models (Figure 1.5) rely mainly on neural networks and claim to be more biologically plausible than linear filter based models. Tsotsos et al.[17] proposed the *Selective Tuning Model* which constructs a pyramid architecture with passing

zone and inhibit zone. The passing zone selects the interest location for further process and the inhibit zone inhibits all the other locations that are not belonged to the pass zone. Cave [18] proposed the *FeatureGate* model which is implemented in a neural network consisting of a hierarchy of spatial maps. Attentional gates controlled by both bottom-up and top-down features are designed to control the flow between each level of the hierarchy.

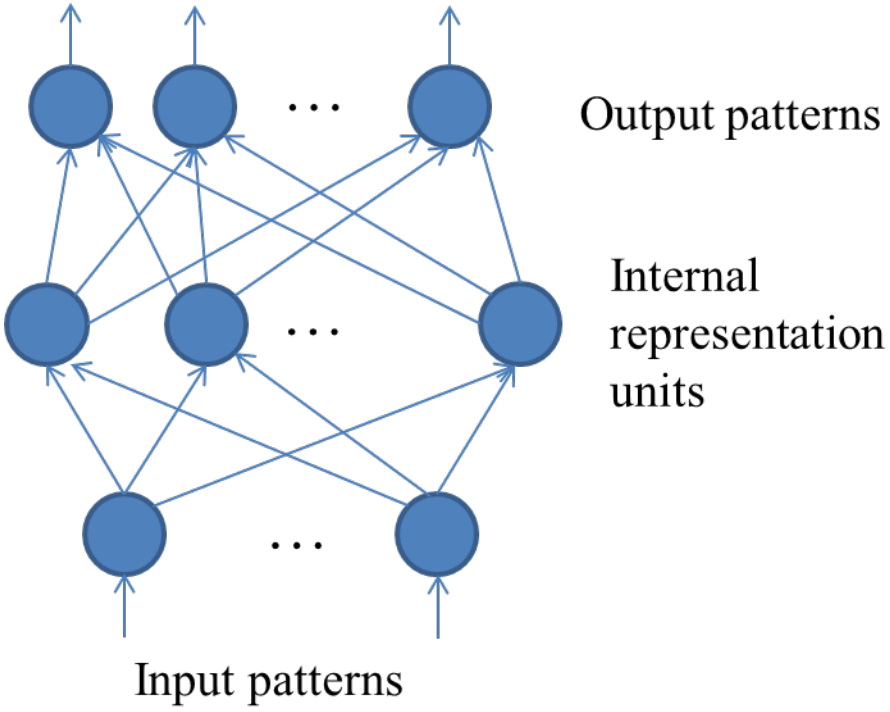


Figure 1.5 Connectionist model

The Probabilistic attention guiding framework

The probabilistic attention guide model assumes that attention can be modeled as the likelihood function of target presence given the image feature and location. Torralba *et al.* [19] proposed the contextual guidance probability model which splits the target presence likelihood into bottom-up saliency, top-down object knowledge and contextual prior using Bayesian rules to calculate the target presence probability at any location. Zhang *et al.* [20] proposed a similar Bayesian framework to guide free-viewing attention with variation that derives the saliency measure from natural scenes. Later, Kanan *et al.* [21] extended this

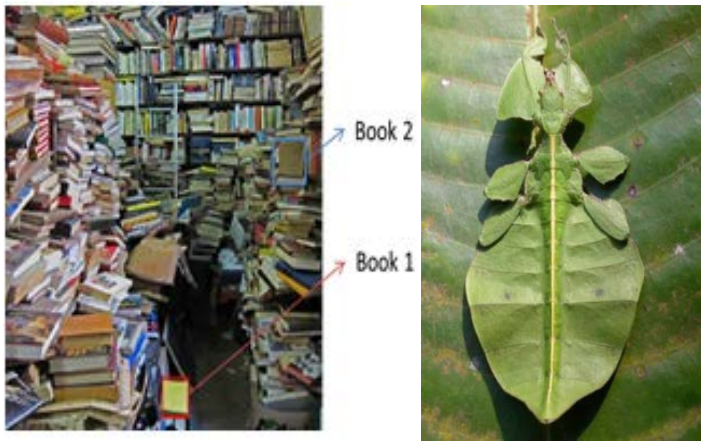
framework to incorporate the top-down object appearance information into their bottom-up saliency map which achieves better results than pure bottom-up saliency model.

Rao [22] proposed another type of probability model which interpret visual attention and perception as estimating the posterior probability of object features and locations. The belief propagation Bayesian algorithm was applied to prescribe the "message" (probability) transmission process from one node to another which simulates the feature encoding process within the visual cortex. This idea was later extended by Yu and Dayan [23], and Chikkerur *et al.*[24] to a Bayesian inference model of attention.

1.2.2 Visual clutter analysis models and study

Visual clutter has been shown close related to human visual attention/perception ability. Researchers have studied the factors that may impair the visual search efficiency for human machine interface. Treisman and Gelade [8] proposed to use the set size, i.e. the number of items in scene, to study the visual search efficiency. Wolfe *et al.* [25] proposed to study the visual search clutter by measuring the background complexity. Bravo and Farid [26] studied the effect of occlusion on search efficiency. Duncan and Humphreys [27] proposed to measure the target saliency by comparing its visual feature to the background. Several models of quantitatively measuring clutter have been proposed and shown well correlation with visual search efficiency by conducting subject tests [3, 26, 28, 29]. The two-stage attention-perception model [4, 8] has been widely accepted and studied. It claimed that when human vision system search for a particular target from a complex background, a pre-attentive stage is first initiated to detect basic features in parallel and then bind those features into a selective attention area/object. Global features are extracted first and attention is guided into local features [30]. Based on this theory, Reddy and VanRullen [31] showed that there are two limitations on human attention that may cause inefficient search: attention for recognition and attention against competition. Attention for recognition refers to the feature detection and binding stage which is affected by global features and attention against competition happens when the search target is close to similar items or background which is closely related to local features. Beck *et al.* [28] then proposed global and local

clutter measure methods to measure global clutter and local clutter which are closely related to the above limitations on attention. Global clutter measures the overall amount of visual information while the local clutter measures the visual information surrounding the search target. The global clutter and local clutter is believed to be interactive or additive to each other to determine the difficulty of target search [28].



(a) (b)

Figure 1.6 Examples of global clutter and local clutter (a) Hundreds of books in clutter (b) Camouflaged insect on a green leaf

The effect of global and local clutter on visual attention/perception ability can be shown by a simple example. Figure 1.6 (a) shows a globally cluttered image. The bright yellow book (book 1) in the red box has much lower local clutter level (high saliency) than the brown book (book 2) in the blue box, which makes it much easier to search the bright yellow book. On the other hand, Figure 1.6 (b) shows a relatively less globally cluttered image with very small color variation, however, the local clutter of the insects should be high due to its low saliency and contrast to the surroundings, which makes it even more difficult to be noticed than the bright yellow book placed on a much cluttered background in Figure 1.6 (a) if they are in the same scale. The global-local clutter representation has shown reasonable well correlation with human visual search performance. This example shows that global clutter may indicate the search efficiency in general in the image. But local clutter is the key for search efficiency for a particular object/target.

Rosenholtz et al. [3] defined visual clutter as a situation where excessive visual information with high variability may lead to the degradation of visual task performance. Treisman and Gelade [8] proposed to use set size, i.e. the number of items and target-distractor dissimilarity in an image to measure the clutter level. The corresponding set size-reaction time function was used as a cue to decide the search difficulty. Wolfe [13] proposed to use features including contrast, orientation, color and motion to measure clutter. Voicu et al. [32] proposed a clutter model to measure infrared images. Global features and local features are computed and applied to train a genetic model to classify the clutter level.

Later, Mack and Oliva [33] proposed to use edge density to measure the image complexity. This measure has been proven to have good correlation with the influence of background on visual search performance by multiple human subject experiments. Rosenholtz et al. [3] proposed two clutter measure methods: Feature Congestion and Subband Entropy. The Feature Congestion model [34] relies on calculating the target saliency and the local variability at multiple scales. Color, orientation and luminance contrast are selected as the features to measure the target saliency versus the local variability. Subbanding Entropy is based on the notion that clutter level should be reflected by the bits required for subband image coding. To compute the subbanding entropy, the image is first converted into Lab and then decomposed into wavelet subbands using steerable pyramid [35]. The generated wavelet coefficients are binned and the entropy is calculated within each subband. The final score is a weighted sum of the entropies computed in luminance and chrominance channels.

Based on the attention limitation model, Beck *et al*. [28] proposed global clutter and local clutter measure respectively and studied the interaction between these two clutters. Color-cluster clutter (C3) algorithm [36] was applied as a predictor to measure the clutter level. The algorithm selects color variability as the main feature and computes a clutter score based on the color density and color saliency due to the characteristics of the geospatial images.

1.2.3 Clutter study using the naturalistic driving data

Naturalistic driving study has been increasingly conducted during the recent decade to fill the gap in traditional driver/road user behavior/interaction study, which normally relied on simulator and test track studies. The naturalistic driving data are collected by a variety of sensors installed in the subject vehicle in an unobtrusive and simultaneous way. Although the traditional data collection methods were valuable for building the baseline of the driving data study, they are not suitable enough for the real behavior within the complex driving environment, especially for the pedestrian behavior study [37].

In this section, we introduce the collected large scale naturalistic driving dataset for this thesis. The data used in this research is collected from an on-going naturalistic driving pedestrian data analysis project sponsored by Toyota North America. In this study, we recruited 110 cars and their drivers in the greater Indianapolis area for a one year naturalistic driving study starting in March 2012. The Transportation Active Safety Institute (TASI) at IUPUI is located in the heart of downtown Indianapolis. In addition, within the 30 mile radius around Indianapolis, where many people commute daily, there is a variety of urban streets, highways, freeways, suburban areas, and rural areas. This makes it possible to collect driving and vehicle data from very diverse driving conditions. We used off-the-shelf vehicle black boxes for data recording, which are installed at the front windshield behind the rear-view mirrors, which record high-resolution forward-view videos (recording driving views outside of front windshield), GPS information, and G-sensor information. We designed and developed a suite of tools to process the data, perform automatic pedestrian detection, and pedestrian behavior analysis.

In this project, we installed a DOD GS600 DVR in each vehicle to collect the naturalistic driving data that consists the driving scene video, GPS information, and vehicle acceleration in X,Y, and Z directions. The DOD GS600 DVR can collect data continuously and save the data into a micro SD card. We used 32GB micro SD cards which can hold up to 10 hours of driving data. The SD card can be easily accessed and switched at the bottom of the camera. Figure 1.7 shows the specification of the DOD GS600 DVR. It includes

one 120^{o} wide angle lens video camera, a GPS with internal antenna, and G sensor. We set the DOD GS600 DVR to record video 30 frames per second with 1280x720 resolution.

Sensor	5 Mega pixels CMOS
Wide Angle Lens	120◦
Focus	10cm - infinity
GPS	Internal antenna
Video output	MINI HDMI : Support 1080i/720p output mode Composite : NTSC/PAL
Recording format	mov
Recording resolution	1920 x 1080(30fps) 1280 x 720(30fps) 848 x 480(60fps)
Video Code	H.264
Recording media	Micro Card(Upto 64G SDHC Card)

Figure 1.7 The specification of DOD GS600

Figure 1.8 shows the example installation to the subject's vehicle. It is installed behind the rear mirror on the front windshield via its suction cup to record the driving scene. The power cable of the DVR is connected to the vehicle's cigarette charger. The camera will be turned on when vehicle is on; and will be off when the vehicle is off.

Figure 1.8 An example installation

Figure 1.9 shows an example collected video frame, GPS and G sensor data. Video data in .mov format which is encoded using H.264. In the generated video, the GPS location and vehicle speed is displayed on the top left corner of the video. At the same time, it outputs a separate data file in text format with GPS location, speed, and G sensor information. Each second, it would output the GPS information along with calculated moving speed. Every 0.1 second, it would record the G sensor information.

Figure 1.9 Example collected data.

While more and more researchers have linked the visual clutter within the driving environment directly to the degradation of the visual task performance during driving, such as vehicle/pedestrian/road sign detection, there is very limited research has been done using the naturalistic driving data. Jenkins [38] firstly studied the effect of "visual clutter" using photographs of various road scenes. Each subject was asked to rank the photographs from most cluttered to least cluttered and to detect synthetic disc targets from the photographs. Ho *et al.* [39] studied the clutter of traffic scene and its effect on traffic sign detection by conducting a series of human subject tests. Edquist [40] systematically studied the effect of clutter on driving performance, especially focus on the road sign detection performance affected by the advertising billboards. All the above studies suggested impair of traffic signs detection ability related to the visual clutter of the traffic scene. However, none of these studies proposed a reasonable computational model to quantify the effect of visual clutter on driver's perception. Moreover, there is no previous study focus on exploring the effect of visual clutter on pedestrian perception using naturalistic driving data. The study in this thesis aims to fill this gap.

1.3 Limitations and Challenges

The existing visual attention/perception models and related clutter measure methods can reasonably predict the true human attention and provide information to multiple tasks, such as object searching, human machine interface design etc. However, there are several limitations. Currently, psychological exploratory experiments have been conducted to

study the visual clutter effect. There lacks computational models which can automatically evaluate the visual clutter. Furthermore, the existing visual clutter measure approaches are not designed for pedestrian clutter evaluation and may not be applicable to pedestrian visual clutter study. Some of the proposed models are correlated to a well-controlled human subject test, which is conducted using artificial stimuli, synthetic images [12, 26] or scenery photographs. The naturalistic driving scenes have very different characteristics that are associated with pedestrian appearance perception difficulty. In addition, some of those models are tested and applied on clutter measure from a specific category, such as geospatial displays [28], infrared images [32]. Most of these clutter measure models require manual parameters adjustment based on each image's characteristics. This would be inefficient and won't be applicable to real-life driving data analysis.

On the other hand, given the fact that large scale naturalistic driving data was used in this study, an efficient pedestrian localization within the large dataset is required. Unlike the synthetic images or scenery photographs with limited number used in previous visual clutter study, pure manually selection/localization of the target (pedestrian in this study) from the test data is not an option for the TASI 110-car naturalistic driving data with billions of video frames collected. An effective pedestrian detection algorithm can work well in the collected naturalistic driving data, which is very challenging, is the preliminary requirement for later pedestrian clutter analysis.

1.4 Contribution and Organization

There are mainly three contributions in this study. First, an efficient categorization-based pedestrian detection for large scale naturalistic driving dataset, which is very challenging, was proposed and state-of-the-art detection results were achieved on the TASI 110-car naturalistic driving dataset. The same framework was later extended to bicyclist detection and explored with feature learning using deep networks. Second, the factors which affect the pedestrian perception within naturalistic driving scene were studied and two types of visual clutter metrics were proposed to measure the driving environment complexity and pedestrian perception difficulty. The proposed computational clutter metrics were justified

by human subject tests using naturalistic driving data, which are, to our best knowledge, the first clutter measurements particularly designed for naturalistic driving data and pedestrian perception. Third, with the proposed pedestrian detection and clutter metrics, we proposed a computational pedestrian perception evaluation model to quantify the perception difficulty of pedestrians appeared within naturalistic driving scene. The computational model could mimic the human visual perception and provide quantitative measurement of the pedestrian perception difficulty, which could be potentially incorporated into the current advanced driver assistance system (ADAS) for better decision making and user experience.

The main contributions of the categorization based pedestrian/bicyclist detection framework for large scale naturalistic driving data are:

- We proposed a novel categorization-based detection strategy which integrated the information collected from camera, GPS and G-sensor.
- We developed a two-stage detection scheme which efficiently detects pedestrians/bicyclists from large scale naturalistic driving data.
- We explored and investigated the possible best bicyclist features using feature learning and constructed a deep network for multi-pose bicyclist detection.
- We collected the 110-car TASI naturalistic driving dataset.

The main contributions of the proposed computational clutter metrics for pedestrian within naturalistic driving data are:

- We proposed two clutter metrics which are particularly designed for naturalistic driving data and pedestrian perception.
- We conducted several human subject tests using naturalistic driving data to justify the proposed clutter metrics.
- We proposed a bottom-up pedestrian perception predictor.
- We compared the proposed clutter metrics and predictor with existing methods

The main contributions of the computational pedestrian perception evaluation model are:

- We proposed a computational pedestrian perception estimator which extended the Bayesian framework for visual attention
- We conducted several human subject tests using naturalistic driving data and qualitative test to justify the proposed computational pedestrian perception evaluation model by comparing with existing computational visual attention/perception models.

The rest of thesis is organized as follow. The designed automatic pedestrian detection system to locate pedestrian in large scale naturalistic driving data will be introduced in chapter 2. The proposed pedestrian clutter measure approaches will be illustrated in detail in chapter 3 with the experimental results of both human subject tests and naturalistic driving data. The proposed computational pedestrian perception evaluation model for naturalistic driving data will be illustrated in chapter 4 with experimental results followed by the conclusions in chapter 5.

1.5 Publications Resulting from This Thesis

Journal Publications

1. **K. Yang**, E. J. Delp and E. Y. Du, “A Pedestrian Perception Evaluation Model for the Advanced Driver Assistant System”, IEEE Transaction on Intelligent Transportation System (submitted).

2. **K. Yang**, E. J. Delp and E. Y. Du, “Bicyclist Detection in Large Scale Naturalistic Driving Video Comparing Feature Engineering and Feature Learning”, IEEE Transaction on Intelligent Transportation System (submitted).

3. **K. Yang**, E. J. Delp and E. Y. Du, "Categorization-based two-stage pedestrian detection system for naturalistic driving data", *Signal, Image and Video Processing* 8, no. 1,pp.135-144, 2014.

Conference Publications

4. L. Dong, S. Chien, **K. Yang**, Y. Chen, D. Good, R. Sherony and H. Takahashi, Determination of Pedestrian Mannequin Clothing Color for the Evaluation of Image Recognition Performance of Pedestrian Pre-Collision Systems, ESV 2015 (accepted)

5. **K. Yang**, Liu, C., Zheng, J. Y., Christopher, L., & Chen, Y. (2014, October). Bicyclist detection in large scale naturalistic driving video. In Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on (pp. 1638-1643). IEEE.

6. R. Tian, L. Li, **K. Yang**, Y. Chen and R. Sherony, Estimation of the Vehicle-Pedestrian Encountering Risk in the Road Based on TASI 110-Car Naturalistic Driving Data Collection, 2014 IEEE Intelligent Vehicles Symposium (IV'14), Dearborn, Michigan, USA, 2014.

7. Tian, R., Li, L., **K. Yang**, Chien, S., Chen, Y., & Sherony, R. (2014, June). Estimation of the vehicle-pedestrian encounter/conflict risk on the road based on TASI 110-car naturalistic driving data collection. In Intelligent Vehicles Symposium Proceedings, 2014 IEEE (pp. 623-629). IEEE.

8. **K. Yang**, E. Y. Du, E. J. Delp, P. Jiang, Y. Chen, R. Sherony, and H. Takahashi, “A New Approach of Visual Clutter Analysis for Pedestrian Detection”, IEEE International Conference on Intelligent Transportation Systems (IEEE ITSC), 2013

9. **K. Yang**, E.Y. Du, P. Jiang, Y. Chen, R. Sherony and H. Takahashi, “In-depth Analysis of HOG based Pedestrian Detection using Naturalistic Driving Data”, FAST-zero'13.,2013

10. E.Y. Du, **K. Yang**, P. Jiang, Y. Chen, R. Sherony and H. Takahashi, “Pedestrian Behavior Analysis Using Naturalistic Driving Data in USA”, the 23rd ESV Conference, Seoul, Korea.,2013

11. **K. Yang**, E. Y. Du, E. J. Delp, P. Jiang, F. Jiang, Y. Chen, R. Sherony, and H. Takahashi, "An Extreme Learning Machine-based Pedestrian Detection Method", IEEE Symposium on Intelligent Vehicle(IEEE IV), 2013.

12. **K. Yang**, E. Y. Du, P. Jiang, Y. Chen, R. Sherony, and H. Takahashi, “Automatic Categorization-based Multi-stage Pedestrian Detection,” IEEE International Conference on Intelligent Transportation Systems (IEEE ITSC), 2012

2. THE PROPOSED PEDESTRIAN DETECTION SYSTEM FOR LARGE SCALE NATURALISTIC DRIVING DATA

Extensive research interest from both vehicle manufacturers and road safety practitioners has been focused on protecting vulnerable road users; such as pedestrians, bicyclists, and wheelchairs. Pre-collision systems (PCS), with vulnerable road user detection capability are becoming a standard feature of active safety systems in the market. Understanding the road user (pedestrians, bicyclists) behavior is important to the pre-collision system design and testing. Large scale naturalistic driving data analysis can provide valuable and objective information on how road users behave in real life. Analyzing road user behavior within large scale naturalistic driving data requires efficient detection methods.

In order to extract the local pedestrian clutter feature, accurately locating pedestrians within the entire naturalistic driving scene is the very first and important step. Given the fact that huge amount of naturalistic driving data are collect and for this research, it is unaffordable to manually label each pedestrian. Furthermore, the output detection probability of the pedestrian detection algorithm will be served as top-down pedestrian based knowledge for the probabilistic pedestrian clutter evaluation model. In this chapter, the proposed automatic pedestrian detection system for large scale naturalistic driving data is introduced. The pedestrian detection system automatically locates pedestrians within large scale naturalistic driving frames collected from the TASI 110-car Naturalistic Driving Dataset [41].

2.1 Related Work

Pedestrian detection has achieved advances in recent years. Different types of sensor techniques have been proposed to perform both on-board pedestrian detection and offline

pedestrian data analysis, including vision (monocular camera [42-45], stereo camera [46, 47], NIR camera [48], TIR camera [49]), Radar [50, 51] and Lidar [52-54]. Among all these sensor modalities, vision-based pedestrian detection is popularly used for its low cost and high compatibility with other tasks, such as lane detection [55-58]. Even though an enormous effort has been made in object detection, or specifically, human detection in the past decade, it is still not ideal enough for pedestrian detection in naturalistic driving data sets. Detecting pedestrians from a large scale naturalistic driving data set collected by a monocular in-car camera could be a very challenging problem due to the following reasons:

- The pedestrians appearing in the naturalistic data are of high variance in size, location, gait, pose, clothes. The quality of the video data may vary a lot due to the limitation of the acquisition system. This makes this task more difficult than detecting people from a well-focused dataset taken from a photographic camera.
- The constantly changing background of naturalistic driving data, the weather and illumination change and the cluttered urban scene makes the foreground segmentation very difficult, especially for a monocular vision system.
- The size of the naturalistic driving data is large; therefore, the accuracy and efficiency should be well balanced to achieve satisfactory detection results.

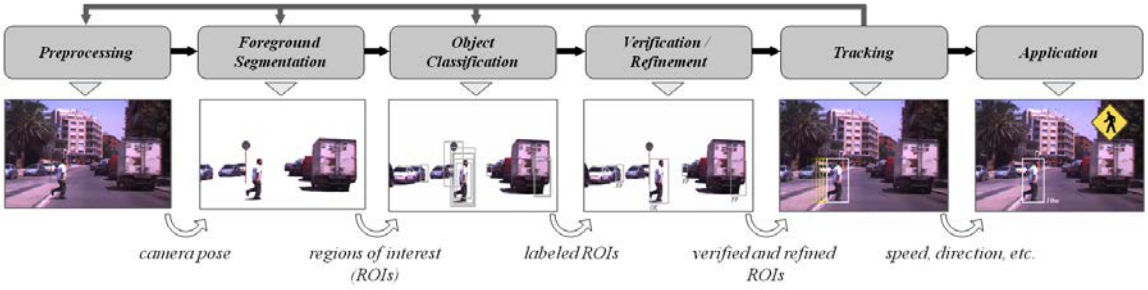


Figure 2.1 A schematic overview of different modules of a pedestrian detection system

A typical pedestrian detection system may include the following modules: preprocessing, foreground segmentation, object classification, verification, tracking and the related applications [56]. (Figure 2.1) Some of the modules may be optional or combined together.

Preprocessing

Preprocessing typically includes tasks such as gain adjustment, camera calibration, image enhancement etc. low-level adjustment, such as exposure and dynamic range, are normally not included in pedestrian detection related publications due to the difficulty of real-time adjustment. Solutions exploiting image enhancement and high dynamic range images have gained increasing interest in dealing with low saturated data, especially videos/images collected in complicated urban environment. Camera calibration is another main step in the preprocessing module. Depending on the type of camera used in the system, the calibration can be divided into monocular-based and stereo-based. Stereo-based methods have provided more robust results in spite of increasing computational cost.

Foreground segmentation

Foreground segmentation is also known as region of interest (ROI) generation, which extracts the meaningful regions of the image and removes as many background regions as possible. One of the most common procedure is the sliding window search, which is an exhaustive scanning approach. Pedestrian size constraints are considered when constructing the search window. Computational ROI generation can be divided into 2D-based, stereo and motion-based method.

Object classification

The extracted ROIs are sent to the object classification to classify as pedestrian or non-pedestrian aiming minimizing the false positive rate and false negative rate. Most effort has been made in classification module. The object classification module can be roughly separated into two steps: feature extraction and classification. A number of pedestrian features and classifiers have been explored in the past two decades. We will review some of them in detail later.

Verification/refinement

A typical system usually contains a step to verify and refine the classified pedestrian windows. The verification uses a different set of criteria than the classifier to filter out false positives. The refinement usually performs a fine segmentation around the set of detected pedestrian to remove the overlapping windows and to find the best fit. The module sometimes can be integrated to the classification module.

Tracking

Pedestrians tracking has been increasingly integrated into the pedestrian detection system. It has been applied to follow the detected pedestrians over time to further avoid false positives. Kalman filter and partical filter have been heavily used to provide prediction based on various cues such as sill silhouette, location, color, texture,etc.

Application

The last module receives pedestrian information from the previous modules and makes high-level decision. Typical applications include areas such as environmental perception, human-machine interaction, etc. Most of them are out of the scope of this thesis. We will focus our review on the previous detection methods.

Extensive research has been explored in monocular vision system based pedestrian detection. Due to the difficulty of foreground segmentation [59] and keypoint selection [60] in naturalistic driving data with dynamic background and low resolution, a sliding window search based strategy is generally applied to locate the possible regions of interest (ROIs) which may contain pedestrians. Basically, a set of visual features are extracted and encoded from the image patch inside each sliding window, and then the encoded feature is classified by a pre-learned classifier as pedestrian or non-pedestrian. We separate our literature review into feature extraction methods and classification methods.

A. Pedestrian feature extraction

Pedestrian appearance features, such as shape, texture, color and other information, have been tested to be relatively robust pedestrian cue. In the past decades, many researchers have designed pedestrian appearance feature-based detection method. These appearance based features can be used in a holistic way or in a part-based model. For example, in [46], Gavrila and Munder proposed shape-based silhouette template matching method to detect a pedestrian with existing templates. Similar idea can be found in [61], Lin and Davis used hierarchical part-template matching approach. The variations in pedestrians make it difficult to directly apply the template matching without further exploiting the appearance feature of the pedestrian. In [62], Papageorgiou and Poggio first proposed to use Haar wavelets (HWs) to extract the local feature of regions of interest, HWs works as a large scale derivative, which computes the difference between two rectangular regions. Similarly, Viola and Jones extended their successful HW like detector for face detection to pedestrian as a fast computing local representation which built a foundation for future pedestrian detector. Later, Dalal and Triggs [42] proposed a human classification algorithm that uses Histogram of Oriented Gradients (HOG) relying on the dense representation of histogram of gradient within a detection window. It computes local gradient histogram within multiple overlapping blocks and generates a concatenated descriptor of all blocks within each detection window. Similarly gradient histogram based feature extraction methods can be found in shapelet [63], edgelet [64], Edge Oriented Histogram(EOH) [65] and etc. HOG achieved promising result on human detection and was popularly used in combination with texture and color features to further improve the accuracy, such as Local Binary Pattern (LBP) [66], Local Tinery Pattern [67], color histogram [68, 69]. Statistical features such as covariance matrix [70] and co-occurrence matrix [71] are also explored by other researchers as different feature representations. In addition, motion cues are also explored in video-based pedestrian detection methods [72]. However, it is challenging to incorporate motion cues with appearance feature when the camera is in motion.

With all the above review pedestrian features, one main trend in the field is to incorporate multiple types of features or to explore best feature from a large feature pool. These techniques have shown considerable improvement over a single feature detector. Wojek and Schiele [73] proposed to combine Haar-like features, shapelets and HOG to achieve improvement over any single feature. Munder *et al.* proposed to combine the shape and texture information to apply to the multi-cue pedestrian detection and tracking system. Dollar *et al.* [74] proposed to extract fast computing Haar-like features from multiple channels which can best separate pedestrian and non-pedestrian windows. Extension of this work includes [75-77]. Xu *et al.* [78] proposed to combine LBP based motion feature, HOG+Haar features and temporal information in a cascaded way to efficiently detect sudden crossing pedestrians. Enzweiler and Gavrila [79] proposed a multi-level mixture-of-experts framework which utilizes HOG and LBP features from depth, intensity and motion channels to increase pedestrian classification accuracy.

B. Classification

Learning pre-trained classifier from a full labeled training set is still the dominant way to generate pedestrian classifier. Support Vector Machine (SVM) is popularly used in object detection so as to pedestrian detection task. Linear SVM is applied in [42] to classify the extracted HOG feature. In [80], Mohan *et al.* proposed to independently classify four human parts by using HWs and a quadratic SVM. The classifications of these parts are then combined with a linear SVM. Felzenszwalb et al. [81] proposed to use latent SVM to model the unknown positions of pedestrian parts in their part-based model. Lin and Davis [82] used a radial basis function (RBF) kernel SVM to classify the computed HOGs from matched silhouette with higher accuracy but slower speed than linear SVM. Recently, Maji et al. [83] proposed to use Additive Kernel SVMs for efficient pedestrian classification which achieved better accuracy but the same speed as linear SVM. Nevertheless, the speed of SVM is still a concern when applied in real time situations.

Another big family of classifiers popularly used in pedestrian detection consists of different types of boosting methods. Boosting methods are powerful when the dimension of features

is high. It relies on selecting best features from a large candidate feature pool and these selected features are served as weak classifiers. Normally a cascaded style is applied to group these weak classifiers into a strong classifier for fast classification purpose. Viola and Jones [72] proposed to use Adaboost [84] to learn a cascade of weak classifiers which can reject non-pedestrian windows at very early stage. Similar framework can be found in [70, 85] using different boosting methods including Realboost [84], logitboost [86] and etc. While boosting methods have shown advantage in classification time, the training process is not trivia. The configuration parameters needs to be tuned over and over and when the training sample number and candidate feature number are huge, the training itself may take very long time.

Other classification methods, such as conventional Neural Networks [87, 88], chamfer distance [89], have also been explored by researchers with related feature extraction methods. We also observed a trend which combines multiple classification techniques in a cascade [90] or parallel [91] way aiming to make a good tradeoff between performance and efficiency.

2.2 Proposed Pedestrian Detection Approach

The proposed categorization based two-stage pedestrian detection scheme is designed to effectively and efficiently find frames with pedestrian appearance from a very large scale naturalistic driving data. Given the fact that the huge number of frames to process and the various driving scenarios, a well balance between the accuracy and efficiency should be achieved. It is very challenging to design a specific algorithm that can work with all kinds of real-life driving scenarios. It would be very rational to categorize the driving scenario first and apply a corresponding detection algorithm for different categories, or at least a few preprocessing. Moreover, in our situation, frames with different pedestrian appearance probability should be treated differently to maximize the processing efficiency. On one hand, some frames may contain little or no information of pedestrians, which would “eat up” the processing speed. Therefore, it is unnecessary to apply the most accurate and sophisticated algorithm to such category frame by frame. On the other hand, some data

may provide critical information of pedestrian behavior, which requires high recognition accuracy and a smaller detection interval.

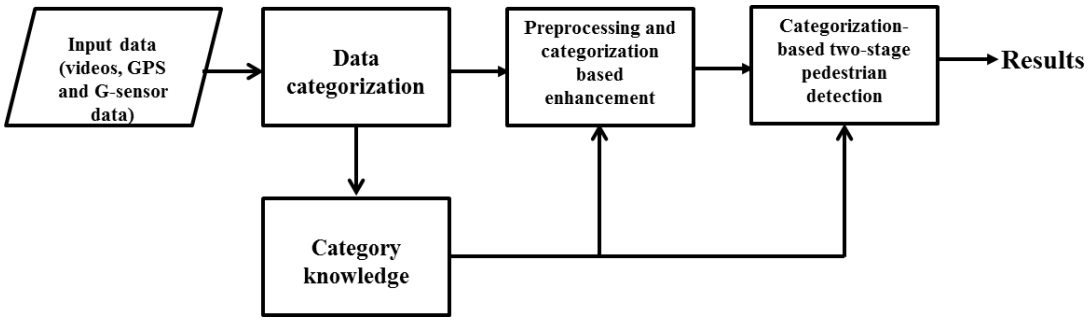


Figure 2.2 Overview of the proposed pedestrian detection system

The overview flow chart of the proposed scheme is shown in Figure 2.2. The naturalistic driving video, GPS and G-sensor data collected from a monocular in-car recording device is first categorized into different driving scenarios based on the GPS, G-sensor data and other available database, for example, the weather database etc. The category knowledge includes the driving location, illumination and weather condition, driving speed and an estimated pedestrian appearance probability by the category information. Based on the driving speed of the vehicle, two levels of pedestrian detection algorithm are applied to vehicle stop and slow moving period and vehicle fast moving period respectively. The categorization based preprocessing and enhancement utilizes the prior knowledge collected from the data categorization module to efficiently perform necessary image enhancement on certain categories. The pedestrian detector is then incorporate with the category knowledge to efficiently generate the ROIs and decide optimal detection and classifier parameters. We will illustrate each module in detail in the rest of this section.

B. Naturalistic driving data categorization

As we mentioned before, for efficiency purpose, it is very rational to categorize the driving data based on the driving condition, driving location, weather and illumination condition and applied appropriate algorithm for each category. There are several benefits from the categorization module. First, categorization module can provide more information about

the data which could be utilized to process the video data more efficiently. For example, if we know that the vehicle is moving very slow or even stop, background subtraction can be applied to locate pedestrian regions of interest (ROIs) much faster than applying a whole frame sliding window search. Second, the categorization information can provide more characteristics of the video data so that a category-related preprocessing can be automatic performed specifically. For example, image enhancement can be performed on a video taken at dawn or dust to mitigate the low illumination issues. Third, a Bayesian model can be learned from a small part of the categorized data to estimate the pedestrian appearance probability given a set of category labels. The estimated probability will provide important information to the parameter and threshold selection of the detection system. The collected data will be first categorized based on the analysis results of the video content and other sensory information (GPS, G-Sensor, date, weather database, etc.). The goal of this module is to improve the accuracy and efficiency of data analysis and, at the same time, to provide statistical information about driving scenarios. Therefore the categorization will focus on classifying the status of driving scenarios, location conditions and the current time and weather environment which will help to improve the efficiency and adaptability in pedestrian detection.

Each frame is automatically categorized based on its vehicle status, location, time and weather. Vehicle status can be directly categorized by the speed of vehicle calculated from GPS module. Location categorization classifier is learned from kernel-based clustering of the GPS and G-sensor data taken at different locations and K nearest neighborhood based method is applied to classify each video. The time and weather information can be retrieved from weather database.

With the category information, a Bayesian model can be constructed to estimate the probability of pedestrian appearance:

$$P(w_i|X, C = C_j) = \frac{P(X|w_i, C = C_j) \cdot P(w_i, C = C_j)}{P(X)}, \tag{2.1}$$

where w_i is the classified label (pedestrian or non-pedestrian), $\boldsymbol{X}$ is the extracted image feature vector and $\boldsymbol{C_j}$ is the category vector.

$$P(X) = \sum_{i=1}^{c} P(X|w_i, C_j)P(w_i, C_j) \tag{2.2}$$

is the evidence learned from the training set and c is the number of training samples. The estimated pedestrian appearance probability will be used as an indicating weight to determine the categorization based pedestrian detection algorithm parameters.

Table 2.1 Time categories and necessary preprocessing or modifications

Category	**Characteristics**	**Preprocessing or modification**
Clear daytime	Good illumination and very few noise	N/A
Heavily cloudy daytime	Moderate illumination	Contrast stretching
Bright sunshine	Strong illumination, may contain glare, pedestrian normally darker than background	Glare detection and remove, select appropriate threshold to generate ROI
Clear night	Low illumination, pedestrian brighter than background	select appropriate threshold to generate ROI

C. Preprocessing and categorization based enhancement

Preprocessing and enhancement of the video data is necessary due to the various illumination and weather situations. We list the necessary preprocessing or modifications for different time and weather categories we generated from last module in Table 2.1.

D. Categorization-based two-stage pedestrian detection

a. ROI prescreening

The TASI 110 car naturalistic driving dataset [39, 44] is collected by installing a video camera recorder on each of the 110 subject cars. The recorded videos are generally uncalibrated due to the different height of the subject vehicles and accidental adjustment of the camera angle/positions made by the subjects. This can result in a substantial variation of camera viewpoint (Figure 2.3). Therefore a motion based automatic ROI prescreening step is designed to mitigate this problem. Moreover, by accurately locating the vehicle hood/control panel and the skyline, both the detection speed and false positives will be reduced for the refined ROIs.

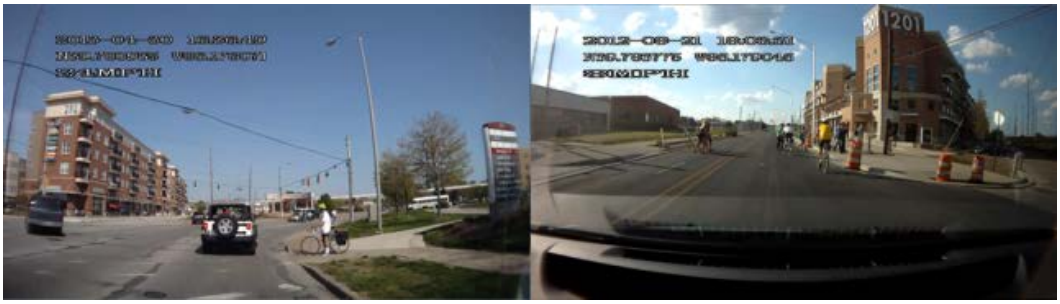


Figure 2.3 Videos with different viewpoints recorded by different subject vehicles

To detect the horizon with a moving camera, motion flow field is evaluated to determine a Focus of Expansion (FOE) of the flow vectors. Sun's method [114] is applied to calculate the optical flow of consecutive frames with the energy function written as:

$$E(u,v) = \sum_{i,j} \{\rho_D \left(I_1(i,j) - I_2\left(i + u_{i,j}, j + v_{i,j}\right)\right) + \lambda[\rho_s\left(u_{i,j} - u_{i+1,j}\right) + \rho_s\left(u_{i,j} - u_{i,j+1}\right) + \rho_s\left(v_{i,j} - v_{i+1,j}\right) + \rho_s\left(v_{i,j} - v_{i,j+1}\right)]\} \quad (2.3)$$

where i and j are the horizontal and vertical pixel coordinates, u and v are the horizontal and vertical components of the optical flow field, ρ_D and ρ_s are the data and spatial penalty functions, which are set by experience or trial and error, in order to create the desired flow field effects. The value of λ is a balancing factor between the data term and spatial smoothness term. By minimizing this energy function, the optical flow for each pixel is generated and combined into the flow field for the whole image.

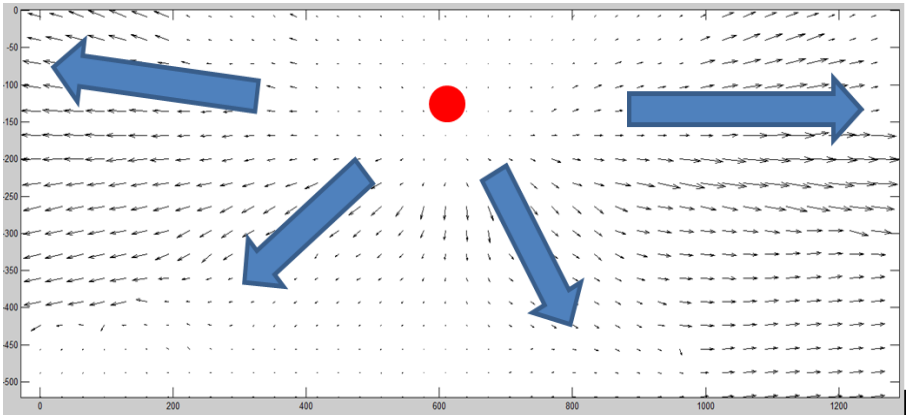


Figure 2.4 Optical flow field from video motion (a red spot overlaid with four blue arrows to show the general direction of flows and the FOE).

The generated motion field is shown in Figure 2.4. The FOE is calculated based on the optical flow field in the following way: several groups of two consecutive video frames are first selected from the longer video sequences in which there is car motion (found by speed data from the log file recorded along with the video). The optical flow field is then divided into left half and right half. The highest motion regions in both left and right flow fields are isolated. Based on the optical flow vectors of each pixel in these regions, we calculate the crossing points of each pair of pixels by extending the flow lines in left and right regions. Finally, by sorting the vertical axis of these crossing points, we first delete the highest and

lowest vertical axis and then average the remaining ones to find the final vertical axis of vanishing point. The hood/vehicle control panel part can be found in the flow field as “no motion” part down at the bottom if exists. The final refined ROI is determined as the region between the two detected lines as shown in Figure 2.5.

Figure 2.5 Example of Prescreened ROI

b. ROI refinement

With the vehicle status category information, ROI generation will be performed more effectively and efficiently. In particular, for frames categorized as vehicle slow moving or stop, the background is relatively constant. A fast background subtraction algorithm can be applied to generate possible binary foreground ROIs which may contain pedestrians. In particular, we locate the possible ROIs in each frame k by comparing the overall variation between frame k and the synthetic average frame generated from the N previous frames within each detecting window with a threshold T:

$$ROI_k = \begin{cases} 1, if \sum_{x=1}^{W} \sum_{y=1}^{H} G(x,y) * (F_k(x,y) - \frac{1}{N} \sum_{n=1}^{N} F_{k-N}(x,y)) > T \\ 0, \quad otherwise \end{cases} \tag{2.4}$$

where W and H are the width and height of the detecting window, $F_k(x,y)$ is the pixel value of frame k at (x, y), $G(x,y)$ is a Gaussian smooth filter and T is a pre-defined threshold learned from training images. For each generated ROI, a set of pedestrian constraints which can be computed quickly are applied to check whether the ROI contains a pedestrian, including the shape, size, the ratio of height to width, the orientation, etc. so that the possible ROIs are further refined.

For vehicle fast moving cases, sliding window based detection is necessary. However, a prescreening step is still necessary for both efficiency purpose and reducing false positives. We observed that most appearance based detector, such as HOG, normally generate false positives mainly from objects and/or complex background with shape close to human body. A large portion of the false positives come from trees, pole structured objects and building outlier (Figure 2.6):

- Trees sometime may appear to have similar local shape and edge response as pedestrians while encoded in HOG. Especially when the top tree shape is close to the pedestrian head-shoulder ratio and the bottom tree contains the trunk, it is likely to be recognized as human. The color frame is used to eliminate tree regions within the frame. The ratio of green component to the other two channels and the ratio of green part to the whole area of detecting window are two cues to separate tree regions.

- Pole-like objects have similar overall shape as a standing pedestrian. The strong vertical edge response will be emphasized by the positive weight of the classifier.

- Vehicles have strong vertical edges on the two rims of the wheels and the two rims could appear close to the leg part of pedestrian. Moreover the rigid edges generated

by the vehicle frame can also generate high positive score after weighted by the pedestrian classifier near the head or shoulder area.

- Building outliers also have very strong vertical edge response in HOG representation. The positive score of the classifier will be emphasized by the dominant histogram contributed by vertical edge. Pole-structured objects, vehicles, building outliers and road components are considered having more rigid and longer vertical edges or horizontal than pedestrian shape. A direct template matching between the edge map and a long rigid line template can eliminate a certain amount of pole-structured objects, vehicle and building outliers.

Figure 2.6 Common false positives of appearance based pedestrian detector. The corresponding gradient representation of each patch is shown on the right side.

Based on the rationale and the false positive analysis above, prescreening step is shown in Figure 2.7. The goal of the prescreen step is to eliminate regions where pedestrians are of low probability to locate within the whole frame so that the number of sliding windows can be greatly reduced. The rationale is that pedestrian will have relatively strong vertical edge response therefore ROIs are only detected at certain regions of the whole frame. The color frame and grayscale frame are both generated during the preprocessing and prescreening step. The pedestrian location constraints are first applied to narrow down the sliding window scanning region. The top part of the frame containing mostly the sky and the bottom part of the frame containing mostly the panel will be excluded. Edge detection and tree color detection are performed in parallel on grayscale image and color image respectively. A mask map containing the possible pedestrian ROI will be generated by the prescreening step. The sliding window search will only be performed on the regions containing vertical edges determined by the mask map since a standing pedestrian is considered to have relatively strong vertical edges compared to the background. In particular, ROIs in each frame will only be selected by the equation:

$$\mathrm{ROI_k} = \begin{cases} 1, \text{if} \sum_{\mathrm{x=1}}^{\mathrm{W}} \sum_{\mathrm{y=1}}^{\mathrm{H}} \dfrac{\partial \mathrm{G(x,y)} * \mathrm{f}}{\partial \mathrm{x}} (\mathrm{x,y}) > \mathrm{T} \\ 0, \text{otherwise} \end{cases} \tag{2.5}$$

where W and H are the width and height of the detecting window, $\frac{\partial f}{\partial x}(x, y)$ is the vertical edge map value of frame at (x, y) and $G(x, y)$ is a Gaussian smooth filter. T is a pre-defined threshold decided by evaluating the training images. In this way, the number of detection windows will be greatly reduced while at the same time maintaining a high detection rate.

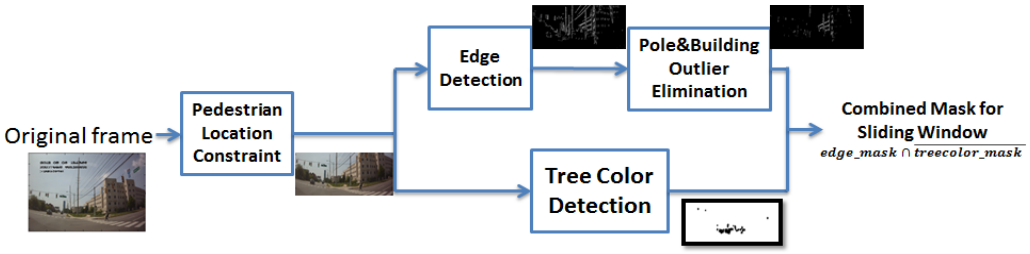


Figure 2.7 Prescreening step of vehicle fast moving cases

c. multi-stage pedestrian detection

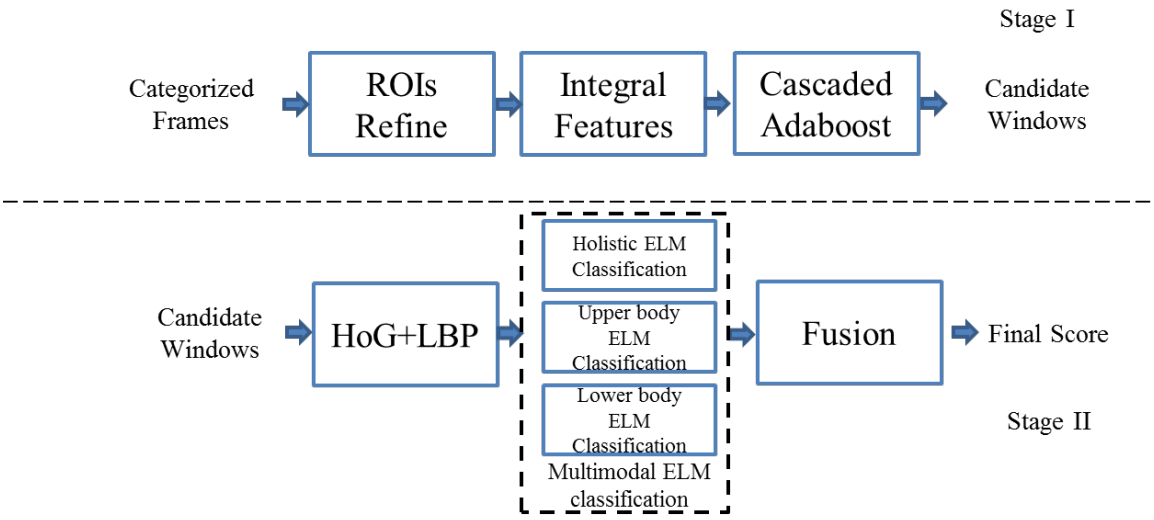


Figure 2.8 Two-stage pedestrian detection scheme

As we reviewed in section II, pre-trained SVM generally achieves good results and can be more easily applied in part-based model to achieve better accuracy. However, sliding window based SVM classification is very challenging to achieve real time processing speed. It takes seconds to process a whole 1280×720 image in our implementation even with greatly refined ROIs. Considering the huge number of frames of the naturalistic driving data we collected, this is unaffordable. On the other hand, cascaded boosting based classifiers can eliminate most of the non-pedestrian windows at very early stage. A combination of these two types of classifiers can achieve reasonably optimal tradeoff for our purpose.

Stage I: Cascaded boosting based detection

The flowchart of the two-stage detection is shown in Figure 2.8. On stage I, integral features [74] are extracted from each sliding window for its compromising performance and computation efficiency. The integral feature makes use of integral image aiming at reducing the computation cost of filtering operation from $O(n^2)$ to $O(n)$. The integral image is computed rapidly from an input image and is used to speed up the calculation of any

upright rectangular area. The integral image is generated by summing the entire pixel values between each pixel and the origin. For example, give an image I and a point (x, y), the value at (x, y) of the integral image I_Σ is calculated by the formula:

$$I_\Sigma = \sum_{i=0}^{i \le x} \sum_{j=0}^{j \le y} I(x, y) \tag{2.6}$$

The convolution of an image I with an n × n box filter with value f at point (x, y) can be implemented by only four operations using integral image I_Σ:

$$I_{conv}(x, y) = f * ((A + D) - (B + C)) \tag{2.7}$$

where A, B, C, D is the value of the four corners of the convolved regions in integral image I_Σ : (Figure 2.9)

$$\begin{aligned} A &= I_\Sigma(x - \left\lfloor \frac{n}{2} \right\rfloor, y - \left\lfloor \frac{n}{2} \right\rfloor) \\ B &= I_\Sigma\left(x + \left\lfloor \frac{n}{2} \right\rfloor, y - \left\lfloor \frac{n}{2} \right\rfloor\right) \\ C &= I_\Sigma\left(x - \left\lfloor \frac{n}{2} \right\rfloor, y + \left\lfloor \frac{n}{2} \right\rfloor\right) \\ D &= I_\Sigma(x + \left\lfloor \frac{n}{2} \right\rfloor, y + \left\lfloor \frac{n}{2} \right\rfloor) \end{aligned} \tag{2.8}$$

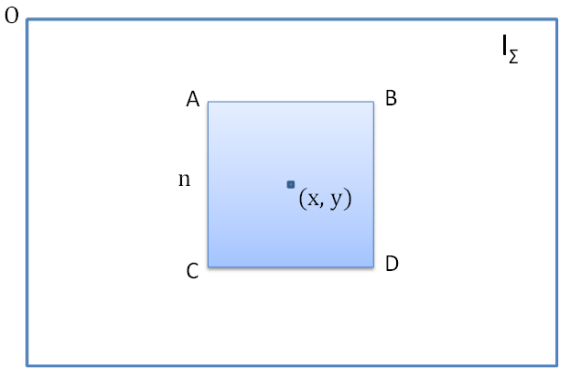


Figure 2.9 Filter Convolution using Integral Image

Integral features are extracted by applying box filter on integral image. The 10 channels breakdown is shown in Figure 2.10, including three color channels, one gradient magnitude channel and six gradient histogram channels. A pre-trained cascaded Adaboost classifier is applied to fast eliminate non-pedestrian windows and generate refined candidate windows. We follow the implementation in [74] for the feature extraction step, but instead of training a 2000 stage cascaded classifier, only 100 features are selected from the integral feature pool. Note that we emphasize the fast elimination non-pedestrian windows on this stage instead of an accurate classification. 100 stage cascaded classifier is enough for this purpose and much faster to train and process compared to the 2000 stage classifier in [74]. A certain amount of false positive is allowed in this stage and will be further eliminated by later stages.

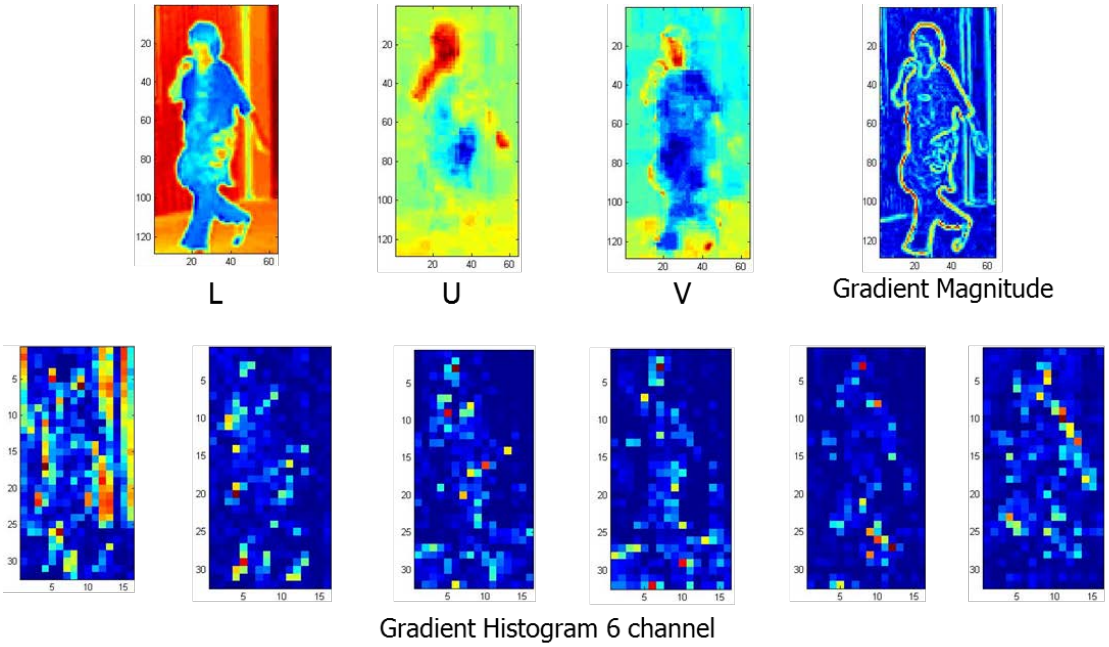


Figure 2.10 Integral features from 10 channels

Stage II: ELM based multimodal detection

On stage II, the candidate windows will be encoded into the HOG+LBP representation. The traditional HOG method relies on stably computing the overlapping local histogram

of edge direction in a dense way. The detection window is first normalized into a 128×64 image patch. Each detection window is divided into 15×7 overlapping blocks and each block is further divided into 2×2 cells. The 9-orientation histogram of gradients is generated within each cell. The locally computed distribution vector is then concatenated into a 3780 dimensional descriptor (Figure 2.11).

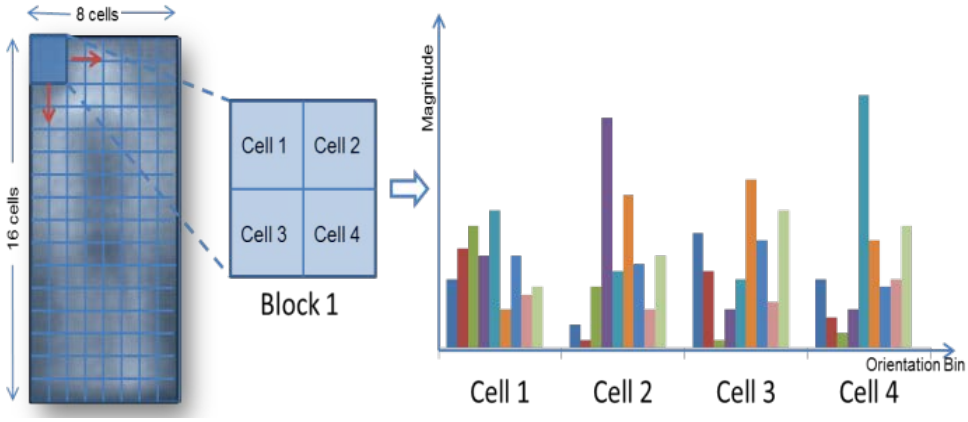
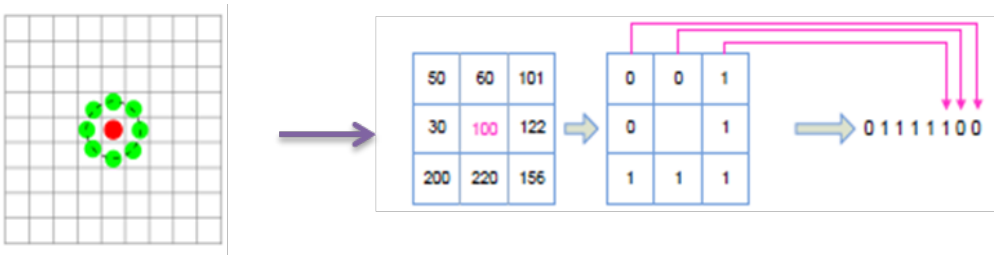


Figure 2.11 HOG descriptor generation

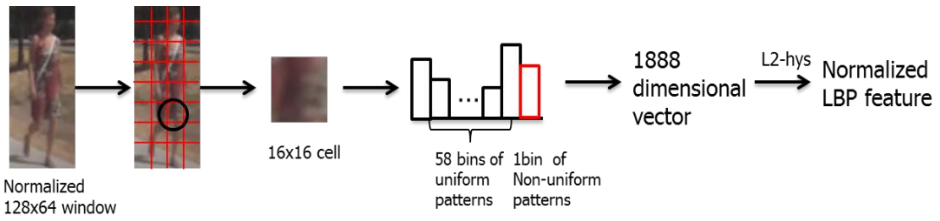
It is shown in [58] that incorporating LBP into HOG can provide more texture information and achieve considerable improvement over HOG representation, especially when the resolution of the classifier ROI is relatively good. LBP_8^2 is selected for its relatively better performance on pedestrian data than other forms. The binary pattern is computed by comparing the neighbor pixels with the central pixel and arranged as a binary sequence. The histogram of the binary sequences within each cell is calculated and concatenating as a vector. In our implementation, the LBP feature is extracted as a 1888 dimensional vector and concatenated with the HOG feature (Figure 2.12). A dimension reduction algorithm is optional to apply on the concatenated HOG+LBP feature vector to reduce the classification time for our efficiency without impairing too much accuracy.

An ELM multimodal detection scheme is shown in the lower part of Figure 2.8. The upper body and low body classifiers are trained simultaneously with the holistic classifier. During the detection, the three classifier outputs are fused to generate the final detection score. We found that the upper body and lower body can provide additional cue for the traditional

whole body model in an affordable extra overhead and it is effective in reducing false positives. For example, a road sign or a tree may have very similar upper shape and whole shape to a pedestrian, however, a large portion of such hard examples with shape close to pedestrians in whole still have considerably large differences in the leg part, and the difference could possibly strengthened by a lower body representation therefore fewer false positives are expected. Similarly, a vehicle or a building outlier false positive window may have very similar lower part shape to a pedestrian and an upper body classifier can strengthen the head-shoulder representation therefore rejecting more false positive windows of those two categories. We compromises the false positive rate and the processing speed by adopting such a part-based model which can be computed parallel with the holistic features without adding too much processing time. A set of classifier fusion methods will be tested on our pilot test set to ensure a high detection rate while at the same time reducing more false positive windows. We will introduce the fusion details and results later.



(a)



(b)

Figure 2.12 LBP feature extraction (a) the binary sequence generation (b) the feature vector generation

ELM has been applied to many different areas including biometrics, image segmentation, human action recognition and etc. It shows advantages over traditional classifier such as SVM, SLFN both on performance and efficiency. It is the first time ELM is applied to pedestrian detection and a considerable improvement over the traditionally used SVM classifier is observed during our pilot experiments shown in chapter 2.4.

Huang et al. [92] theoretically and experimentally proved that ELM can be used as a unified learning platform which does not need to tune the hidden layer parameters as traditional Single layer neural networks (Figure 2.13) do. Instead of using the time-consuming gradient descent based learning method; ELM relies on computing the Moore-Penrose generalized inverse of the hidden layer matrix [93].

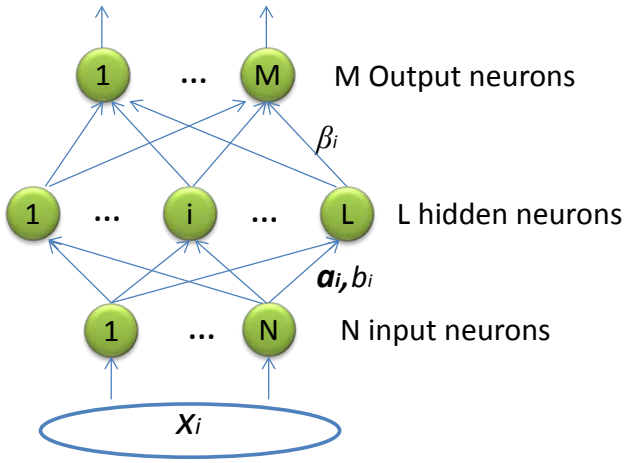


Figure 2.13 Single layer neural networks

In general, ELM maps any given SLFN hidden layer into a matrix form:

$$h(x) = [G(a_1, b_1, x_1), G(a_2, b_2, x_2), \ldots, G(a_L, b_L, x_L)] \quad (2.9)$$

where **a**, **b** are the random initialized hidden layer parameter matrix, L is the number of nodes in hidden layers, G is the node activation function, which could be additive, radial basis function (RBF) or etc. In particular, the additive node activation has the form:

$$G(\boldsymbol{a_i}, b_i, \boldsymbol{x}) = g(\boldsymbol{a_i x} + b_i) \tag{2.10}$$

where a_i is the weight vector connecting the *i*th hidden node and the input nodes and b_i is the bias of the *i*th hidden node. The RBF node has the form:

$$G(\boldsymbol{a_i}, b_i, \boldsymbol{x}) = g(b_i \|\boldsymbol{x} - \boldsymbol{a_i}\|) \tag{2.11}$$

where $\boldsymbol{a_i}$ is the center of the *i*th hidden node and b_i is the impact factor of the *i*th hidden node. Therefore the output function of the SLFN can be written as

$$f(\boldsymbol{x}) = \sum_{i=1}^{L} \beta_i G(\boldsymbol{a_i}, b_i, \boldsymbol{x}) = \boldsymbol{h(x)\beta} \tag{2.12}$$

where $\boldsymbol{h(x)}$ is the hidden layer output corresponding to input sample *x* and $\boldsymbol{\beta}$ is the output weight vector between the hidden layer and the output layer. With calculated hidden layer matrix of N input samples:

$$\boldsymbol{H} = [\boldsymbol{h(x_1)^T}, \boldsymbol{h(x_2)^T}, \dots, \boldsymbol{h(x_N)^T}]^T$$

and the target matrix:

$$\boldsymbol{T} = [t_1, t_2, \dots, t_N]^T,$$

then $\boldsymbol{\beta}$ can be directly calculated as:

$$\boldsymbol{\beta} = \boldsymbol{H^{\dagger}T} \tag{2.13}$$

In this way, the input layer and hidden layer parameters $\boldsymbol{a_i}, b_i$ do not need to be tuned and the network can be trained very efficiently.

Huang et al. [94] shows that dual optimization objective functions of ELM is consistent with that of SVM while ELM searches optimal solution in a greater domain with faster implementation. Therefore ELM achieves better performance in general and multiple tests have also proved it [94]. In particular, for a binary case, the decision function of ELM classifier can be written as

$$f(\boldsymbol{x}) = sign(h(\boldsymbol{x})\boldsymbol{H}^T\left(\frac{\boldsymbol{I}}{C} + \boldsymbol{H}\boldsymbol{H}^{\boldsymbol{T}}\right)^{-1}\boldsymbol{T}) \tag{2.14}$$

where $\boldsymbol{H}$ is the hidden layer matrix calculated from the training samples, $\boldsymbol{T}$ is the target matrix of training samples and $\frac{\boldsymbol{I}}{C}$ is a positive constant matrix for a stabler inverse result.

Kernel formed ELM is applied to learn the holistic classifier, upper body classifier and lower body classifier. The output function of extended kernel based generalized SLFNs has the form:

$$f(\boldsymbol{x}) = [K(\boldsymbol{x}, \boldsymbol{x}_1), K(\boldsymbol{x}, \boldsymbol{x}_2), \dots, K(\boldsymbol{x}, \boldsymbol{x}_N)]\left(\boldsymbol{\Omega} + \frac{I}{C}\right)^{-1} T \tag{2.15}$$

where T is the target label vector, c is the positive constant, $\boldsymbol{\Omega}$ is the kernel matrix with $\Omega_{ij} = K(x_i, x_j)$and K is the kernel function. In our application, RBF kernel is used which has the form:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\frac{\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2}{d}) \tag{2.16}$$

Where d is the kernel parameter controls the width of the function.

The generated ELM scores from the holistic, upper and lower body classifiers are fused to generate the final decision. Several score fusion methods can be used to fuse matching results: simple-average (SA), minimum-score (MIN), maximum-score (MAX), classifier weighting (CW) and Dempster Shafer method (DS).The first four are commonly used fusion method and the DS method is based on DS theory [95]

(1). Simple-Average (SA): the normalized scores S from different modalities with score S_i are averaged directly using $S = \frac{1}{M}\sum_{i=1}^{M} S_i$, where M is the number of the modalities.

(2). Product (Pro): the normalized score S is the product of score S_i from different modalities$S = \prod_{i=1}^{M} S_i$.

(3). Minimum-Score (MIN): select the minimal score as the fusion score $S = \min\{S_1, \dots, S_M\}$.

(4). Maximum -Score (MAX): select the maximal score as the fusion score $S = \max\{S_1, \dots, S_M\}$.

(5). Classifier Weighting (CW): each modality classifier is assigned a weight based on its Equal Error Rate (EER). The weights for more accurate matchers are higher than those of less accurate matchers. The fusion score is calculated as: $S = \sum_{i=1}^{M} w_i S_i$, where w(i) is the weight for classifier *i* calculated as $w_i = \frac{\frac{1}{\sum_{i=1}^{M} 1/E_i}}{E_i}$ where $E_i$is the Equal Error Rate (EER) of classifier *i*.

(6). Dempster Shafer method (DS): this information fusion method is based on Dempster Shafer theory. The belief of each event is initialized as 0 (uncertainty is 1) and updated based on incoming evidences. The theory assumes that the incoming evidences are independent pairwisely and their emerging order is unimportant. However, the evidences here are from the same pedestrian therefore assuming them independent is invalid. We adopt the modified Dempster's rule by Murphy and Kalka, the fusion score D_i is calculated as:

$$D_i = \frac{(D_{i-1} * D_i)^n}{(D_{i-1} * D_i)^n + \left((1 - D_{i-1}) * (1 - D_i)\right)^n}, \quad i = 2,3 \tag{2.17}$$

D_1 is initialized as the smallest score, the scores are sort in ascending order, here n = 0.5 which gives equal weight to all evidences.

We tested each fusion method by the pilot experiments and the one with the best performance is applied to the large scale naturalistic driving data.

2.3 Experimental Results of the Proposed Pedestrian Detection System

A. Naturalistic driving data

In this study, we recruited 110 cars and their drivers in the greater Indianapolis area for a one year naturalistic driving study starting in March 2012. The drivers were selected based on their geographic, demographic, and driving route representativeness. We used off-the-shelf vehicle black boxes for data recording, which are installed at the front windshield behind the rear-view mirrors, which record high-resolution forward-view videos (recording driving views outside of front windshield), GPS information, and G-sensor information. Some examples of the collected naturalistic driving data are shown in Figure 2.14. Over the one-year period, it collects about 80 Terabytes (TB) of data which covers over 1.3 million miles and 36,000 hours of driving data.

Figure 2.14 Examples of collected naturalistic driving data

B. Implementation details

A training set including 1487 positive samples and 2857 negative samples cropped from the collected naturalistic driving data are used to train all the test and baseline classifiers. Each training sample is normalized into 128×64 image patch. For the cascaded boosting classifier in stage I, 30000 integral features are randomly generated from each training patch and 100 stage cascaded classifier is learned. For the multimodal ELM classifier in stage II, HOG+LBP feature is generated from holistic, upper body and lower body patches respectively. For HOG feature, we compute the fast HOG using integral image [96]. 8×8 block and 2×2 cell is applied as in [42]. For LBP features, LBP_8^2 is used for its tested

optimal performance for pedestrian data [66]. Each 16x16 block is encoded into a 59 dimensional feature vector and all the encoded LBP vectors are concatenated and normalized by L2-hys [42]. Therefore the concatenated feature vectors of holistic, upper body and lower body have dimension of 5668, 2708 and 2708 respectively. For speed reason, a dimension reduction algorithm is performed on each of the three feature vectors before ELM classification and score fusion. The ELM multimodal classifier will output a fused score as the final result. The best suitable dimension reduction and fusion algorithm and relative parameters are determined by our designed pilot experiments.

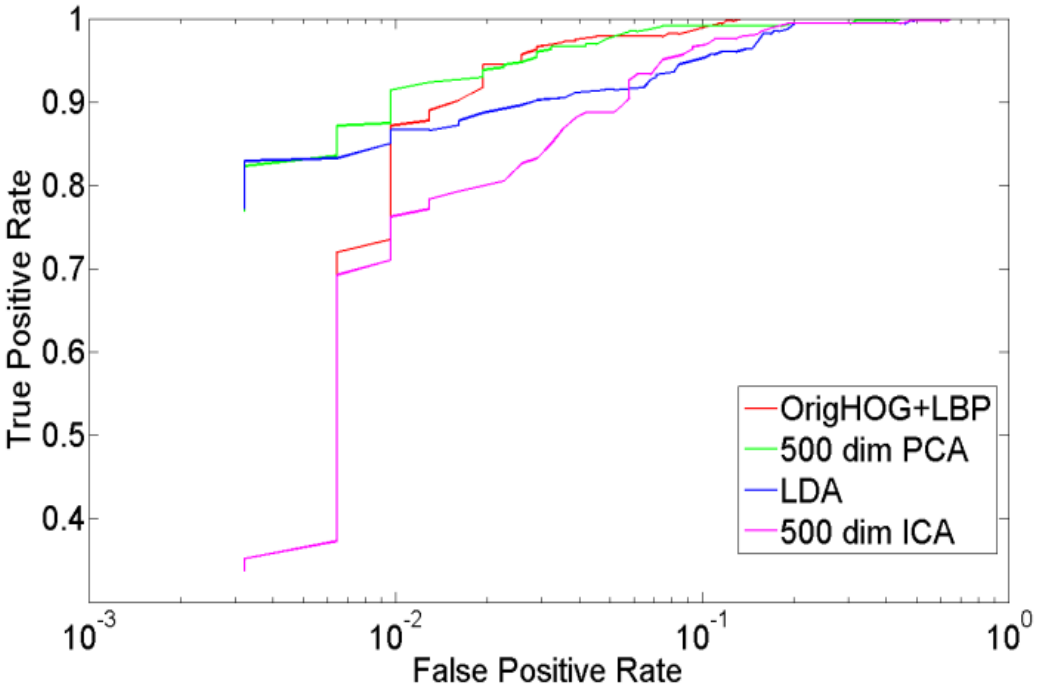
C. Pilot experimental results on test samples

We generate a set of test samples cropped from the naturalistic driving videos which are not overlapped with the training set. The goals of this experiment are (1) find the suitable dimension reduction method and parameters; (2) find the optimal fusion method of ELM multimodal classifier. (3) compare the performance of the proposed ELM multimodal classifier with traditional SVM classifier.

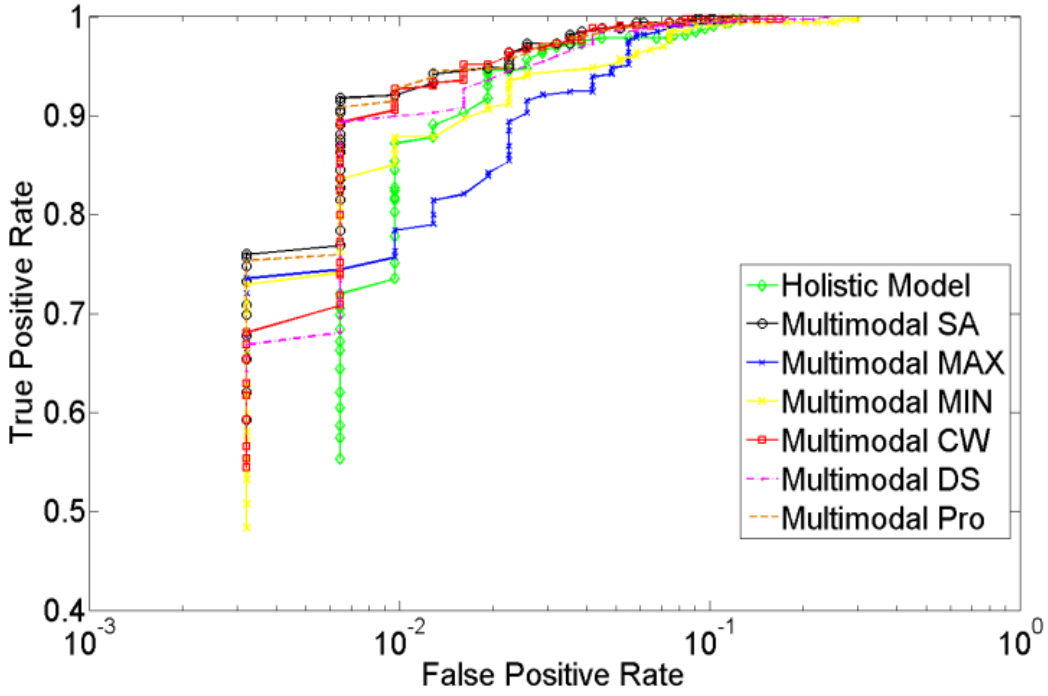
The test samples with 639 pedestrian samples and 1029 non-pedestrian samples are cropped from our naturalistic driving data which are very challenging. The pedestrian samples vary in illumination, pose, clutter, etc. and the non-pedestrians include a lot of hard examples like trees, pole-structured objects, etc. Some of the test samples are shown in Figure 2.15, the left four pedestrian samples have different pose, shape and illuminations with cluttered background and the right four non-pedestrian samples are considered to be hard example as they have very similar shape to a pedestrian.

(a) (b)

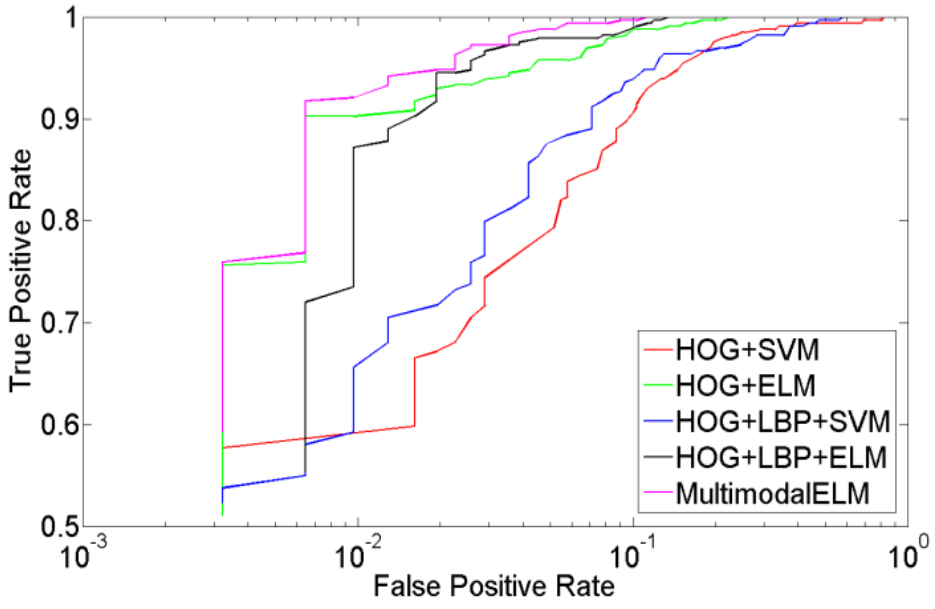
Figure 2.15 Examples of pilot test samples (a).Pedestrian test samples (b).Non-pedestrian test sample



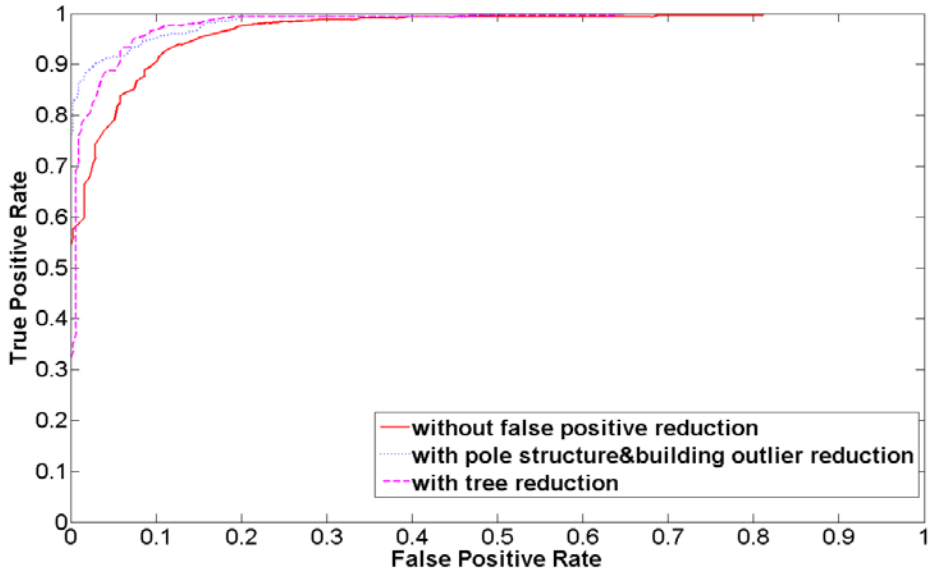
(a)


True Positive Rate
False Positive Rate
Holistic Model
Multimodal SA
Multimodal MAX
Multimodal MIN
Multimodal CW
Multimodal DS
Multimodal Pro

(b)


True Positive Rate
False Positive Rate
HOG+SVM
HOG+ELM
HOG+LBP+SVM
HOG+LBP+ELM
MultimodalELM

(c)



(d)

Figure 2.16 Results of the pilot experiments. (a) Comparison of different dimension reduction methods. (b) Comparison of different fusion methods. (c) Comparison of classifiers. (d) Comparison of with or without false positive reduction

Classification speed is directly related to the dimension of input feature vector. Effectively applying dimension reduction methods could sufficiently reduce the classification computation time while at the same time not impairing the accuracy very much. The reduction of classification time of each window will dramatically increase the processing efficiency for the large scale naturalistic driving data. For this purpose, we tested several dimension reduction methods including principle component analysis (PCA), linear discriminant analysis (LDA) and independent component analysis (ICA). For PCA and ICA, we both use 500 components. The comparison results in shown in Figure 2.16 (a) and we can obviously see that PCA outperforms other two dimension reduction methods and achieves almost the same accuracy as the original HOG+LBP feature. Moreover, as we can see from the results, PCA shows better performance at the low false positive rate region, which is exact the region where select the thresholds for large scale naturalistic driving data processing.

Figure 2.16 (b) shows the results of the multimodal ELM classifier using different fusion methods. SA, Pro and CW outperformed the holistic ELM classifier and DS achieved similar performance as the single modal HOG+LBP feature. In addition, the multimodal classifier had considerable improvement over the single modal at the low false positive rate (FPR) region and overall it effectively reduced the false positives, which is very meaningful for our naturalistic driving data detection.

To better justify the advantage of ELM as pedestrian classifier, we directly compare its performance with traditional SVM which is very popularly used in pedestrian detection systems. (Figure 2.16 (c)) We used exactly the same feature extraction and training process for the two types of classifiers. Both HOG only and HOG+LBP feature classifiers are tested and compared. We observed dramatically improvement from the pilot experimental results. In particular, ELM achieves more than 15% improvement of detection rate at 0.01 FPR. Moreover, ELM shows classification speed advantage over SVM both in training time and test time.

Figure 2.16 (d) shows the results of the trained multimodal ELM detector with and without the prescreening step we illustrated in section IV.D.a to reduce the possible false positives from trees, pole-like structures and vehicles. The prescreen step effectively reduced the false positives, which are considered as hard examples in naturalistic driving scenarios. To better show the individual false positives from trees, poles and vehicle wheels reduced by the prescreening step, we ran the test set without prescreening, with tree elimination, with pole structure and building outlier elimination respectively. The false positive reduction step was shown to effectively eliminate the typical “hard examples” in pedestrian detection.

Table 2.2 shows the processing time comparison between ELM and SVM where ELM takes only 2-3 seconds to train a HOG+LBP classifier on the training set aforementioned comparing with minutes of SVM. For each window, ELM only takes 1/3 processing time of SVM. Note for the speed test, we use the 5668 dimensional holistic HOG+LBP feature as input and implement the classifier on Matlab environment. On the other hand, HOG+LBP feature also achieves better performance than HOG only feature on both ELM

and SVM classifiers. The proposed multimodal ELM classifier using SA fusion is also shown for comparison.

Table 2.2 Computational time comparison of ELM and SVM

	SVM	The proposed ELM
Training time (sec)	87.83	2.23
Test time per window (sec)	0.215	0.078

D. Experimental results on INRIA person dataset.

We further tested the proposed method with an empirically selected dimension reduction and multimodal fusion algorithm on INRIA person dataset. We used the 288 test images with pedestrians to compare the proposed detector with existing methods. The full image was evaluated and a standard sliding window scanning scheme was performed. Non-maximum suppression (NMS) was implemented to combine nearby and overlapping detections. The comparison result is shown in Figure 2.17. We compared the proposed detector with the traditional HOG detector and the state-of-the-art FPDW detector. The HOG+LBP feature was extracted and reduced to 500 dimensions using PCA and fused by SA rules. The proposed multimodal ELM outperformed the HOG detector and achieved comparable result with the FPDW detector.

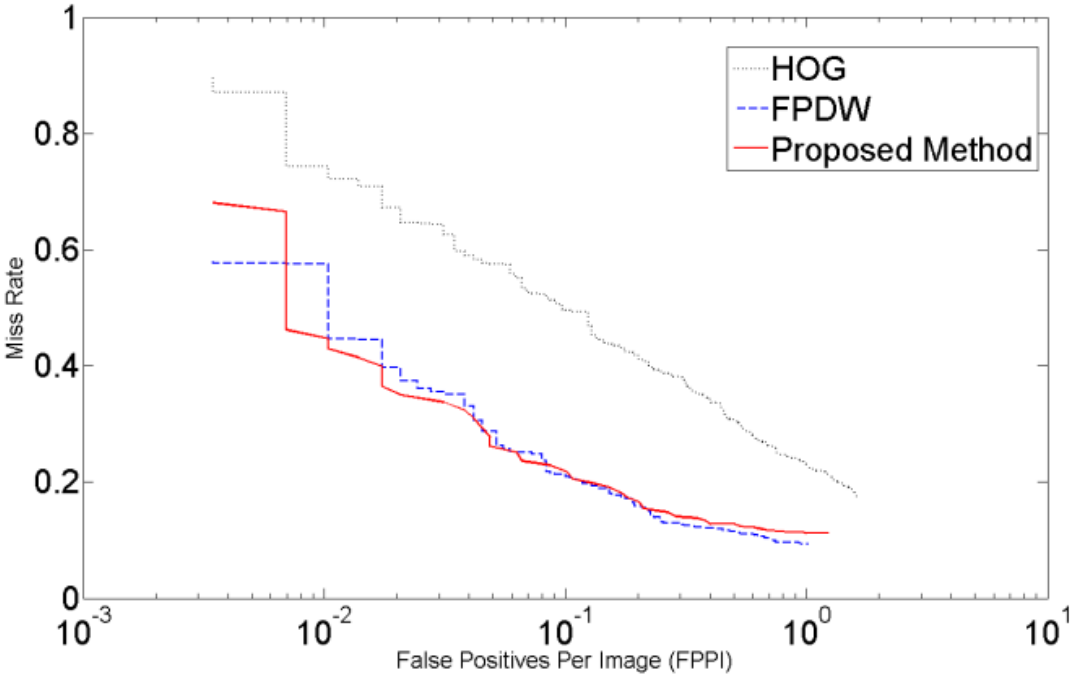


Figure 2.17 Comparison of experimental results on INRIA person dataset.

E. Experimental results on naturalistic driving data

With the comparison results on dimension reduction and fusion methods from the pilot test, we applied the proposed method on naturalistic driving data using the optimal methods and parameters tested from the cropped test set. The HOG+LBP feature was extracted and reduced to 500 dimensions using PCA and fused by SA rules. The tested video content involved different driving scenarios, including different road types, weather conditions, illuminations, etc. Twelve five-minute test videos are randomly selected from our large scale dataset with over 3600 seconds of data including over 100,000 frames. Each frame was 1280×720 high resolution. Similar to experiments on INRIA person dataset, we applied false positive per image versus the miss rate metric which is popularly used to measure the performance of the pedestrian detection system using the standard PASCAL measurement [97]. A standard multi-scale window based technique was incorporated with the proposed preprocessing and ROI generation to minimize the sliding window number. A non-maximum suppression method similar to [42] was applied to combine multiple

overlapping detections. The parameters are listed in Table 2.3. For our pedestrian behavior analysis purpose, we only annotated and detected pedestrians of size greater than 48 pixels in height in the test set, since pedestrians from too far away are considered to have no potential conflicts with the vehicle. We compared the proposed multimodal ELM classifier using 500 PCA components and SA fusion rule with two popularly used baseline methods: HOG+SVM and HOG+LBP+SVM (Figure 2.19). The proposed method outperformed both baseline methods. To better illustrate the improvement of the multimodal ELM detector applying HOG+LBP features, Figure 2.18 shows the detection results of the proposed detector and classic HOG detector. The proposed detector achieved 0.3 false positive per image (fppi) compared to 1.3 fppi of the HOG detector at the same detection rate. The ROC curve and the computational time breakdown are shown in Figure 2.19 and Table 2.4. The proposed detector with categorization and preprocessing achieved slightly better performance than the detector without category specific preprocessing and ROI refinement. Moreover, the computation time was greatly reduced due to the implementation of the categorization based ROI refinement. Compared to traditional HOG+SVM, the proposed classifier had approximately five times improvement in speed. A tracking by detection example result of a five-second video clip is shown in Figure 2.20, where the pedestrian within was detected at different distances with different gaits.

HOG detector Multimodal ELM detector

Figure 2.18 False positives comparison of HOG detector and the proposed detector

Table 2.3 Parameter value used in experiments

Parameter	value
Normalized window size	128×64
Sliding window stride	4
Window scale number	5
HOG block size	8×8
HOG cell	2×2
LBP block size	16×16
ELM regularization (c)	1
Kernel width (d)	10
Normalized detection score	[-1,1]

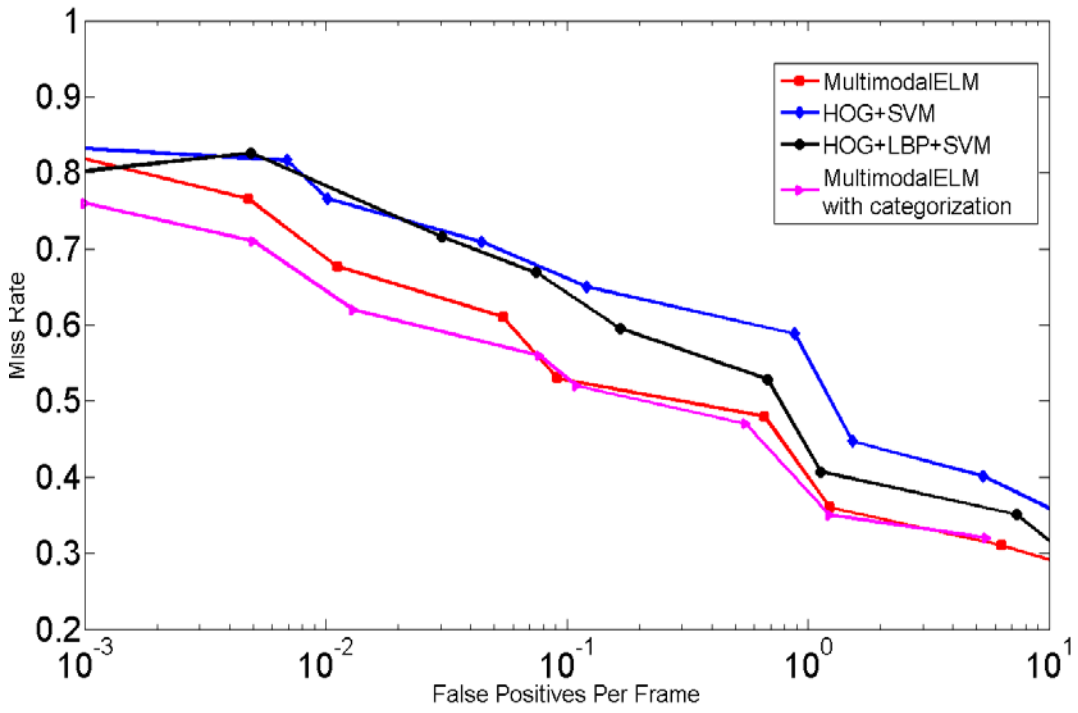


Figure 2.19 Comparison of experimental results on naturalistic driving data

Table 2.4 Process time breakdown for per second video (in average)

	Categorization (sec)	**Frame Generation from video(sec)**	**Pedestrian Detection (sec)**
HOG+ SVM	-	0.39	3.6
Multi-modal ELM	-	0.39	1.77
Multi-modal ELM with categorization	0.008	0.39	0.62

Figure 2.20 A tracking by detection example of a five-second video

F. Effect of categorization based prescreening and enhancement

To justify the effect of categorization based prescreening and enhancement on improving the detection efficiency in large scale naturalistic driving data, we sampled a total number of 20 5 minute-long naturalistic driving video from the entire TASI 110-car dataset. The selected data covered all the categories and were selected based on the actual statistic of the entire TASI 110-car statistics. The statistics of the sampled data is shown in Table 2.5. 66 pedestrians were labeled within the 20 5-minute videos with about 180,000 frames. We focus the experiments and analysis on each categorization-based prescreening method individually.

Location

The vehicle location information was provided by the GPS data recorded along with the videos. Highway, rural and urban areas have very different background clutters therefore different prescreening methods could be applied. In our implementation, highway videos were considered to have the lowest background clutter and only vehicle structure elimination was applied. Tree reduction was further applied to rural videos and pole structure/building outlier reduction step was further applied to urban videos. In addition, only roadside regions were considered as ROI for highway and rural scenarios for efficiency purpose. We ran the multimodal ELM detector using the same parameter set in Table 3 and set the threshold to 0.2. The comparison results are shown in Table 2.6. We compared the pedestrians detected versus the total number of false detected frames. The location category information provided refined ROIs and prescreened windows to the detection module therefore substantially reduced the false positives. The computational time was also greatly reduced due to the reduction of ROIs and window numbers.

Time/illumination

Illumination has substantial effect on the pedestrian appearance and detection performance. Necessary preprocessing was applied to the video frame according to its category. Cloudy videos with moderate illuminate were enhanced with contrast stretching. For some night

videos with strong backlight, glare removal step was implemented to eliminate possible false positives introduced by the strong backlight. For dark night videos, due to the constraints of the camera, only windows brighter than a pre-determined threshold were considered as ROIs. Very dark patches were ignored as background to save computational power. The results in Table 2.6 show that the illumination based enhancement and prescreening substantially reduced false positives in night videos.

Vehicle status

Moving pedestrians in videos with constant background can be quickly separated with the still background while the vehicle is stopped or moving slowly. Therefore a fast background subtraction method was applied to quickly generate ROIs for moving pedestrians and to further refine the ROIs based on the size, height-width ratio and orientation. A certain amount of computation time reduction was observed while maintaining the detection rate in the experiment.

Table 2.5 Statistics of the test data

	By location		By Time/illumination				By status	
	Urban/suburban	Rural/highway	Clear daytime	Cloudy daytime	Backlight	night	Moving	Stop/slow
Percentage	40%	60%	40%	20%	5%	35%	90%	10%

Table 2.6 Comparison results with category information vs. without category information

	Detection rate	False detected frames	False positive rate	Computational time(s/frame)
Without category info	77.3%	2192	1.22%	1.51
With location info	77.3%	1322	0.73%	0.68
with time info	80.3%	977	0.54%	0.76
With vehicle status info	77.3%	1788	1.0%	1.35

2.4 Bicyclist Detection in Large Scale Naturalistic Driving Video Comparing Feature Engineering and Feature Learning

In addition to pedestrian detection, the proposed detection system was further developed and explored for bicyclist detection in large scale naturalistic driving videos. Compared to pedestrian detection, real time on-board bicyclist detection is even more challenging due to the following reasons:

- Bicyclists in driving video have higher appearance variance than pedestrians. In particular, as shown in Figure 2.21, bicyclists with five different poses are largely varied in shape and appearance, which cannot be easily represented by a single model as the traditional pedestrian detector does.

- Bicyclists normally move much faster than pedestrians, which requires the faster response PCS with a more efficient detection algorithm.

Figure 2.21 Bicyclists with different poses in naturalistic driving videos

In the literature, two types of strategies deal with the high intra-class variance of the bicyclists. One solution [98] is to introduce several different holistic models for different poses and detect bicyclists with corresponding poses in parallel. The computational cost for this method is increased since multiple passes of sliding window detections are performed. This could be the bottleneck of a large scale detection or real-time on-board system. Therefore a more efficient feature extraction and classification method is needed. The other method [99] is to use a part-based model [81] to handle the variance of poses, gestures, clothing and bicycle types. However, the performance of part-based models could be degraded due to the low-resolution representation of the objects with a small scale. Moreover, part-based model usually requires higher computational cost.

Recently, deep learning networks have been extensively studied and applied to computer vision tasks, such as object detection, sematic learning, etc. Deep networks has shown significant improvement over traditional neural networks on a number of applications. The primary advantage is that it can compactly represent a significantly larger set of functions than shallow networks. In particular, Deep networks also provide an end-to-end framework to traditional object detection task. It relies on learning features by the network itself instead of designing the hand-engineered features.

The two-stage detection scheme was applied to bicyclist detection and a multi-modal bicyclist detector which efficiently detects bicyclists with varied poses from large scale naturalistic driving data was proposed. Motion based region of interest or bounding box

detection was designed and first applied to the entire video to refine the region for sliding-window detection. Then an efficient integral feature [74] based detector is applied to quickly filter out the negative windows. The remaining candidate windows are then encoded and tested by three pre-learned pose-specific detectors. On the other hand, we also explored the possibility of applying state-of-the-art deep networks on bicyclist detection from naturalistic driving data. A multi-layer auto-encoder (AE) based deep network was learned. The extracted features are directly learned from the dataset in contrast to the integral features and HOG detector we used in the proposed two-stage multi-modal bicyclist detection scheme. The two methods are illustrated and compared using a subset of our TASI 110-car naturalistic driving dataset.

A. Feature engineering

Similar to the two-stage pedestrian detection, during stage I detection, Integral features [74] are extracted from each sliding window balancing the performance and the computational efficiency. The extracted features are from color and gradient channels. A pre-trained cascaded Adaboost classifier is applied to quickly eliminate non-bicyclist windows and generate refined candidate windows. Most negative windows can be rejected in this early stage. A 100-stage cascaded classifier is adequate for this purpose and it is much faster to train and process, compared to the 2000 stage classifier in [74]. A certain amount of false positives are allowed in this stage as a trade-off.

To handle the high intra-class variation of bicyclists with different poses, we treat bicyclists with different poses as different classes. For each pose, pose-specific classifiers are trained by the categorized samples cropped using the training set collected and sampled from the TASI 110-car naturalistic driving dataset. However, naïvely assuming five pose-specific classifiers (as illustrated in Figure 2.21) will add five times the detection burden, since each pose-specific detector needs to scan the entire ROI. For the best efficiency, we train only three poses: side view, front-side view and rear-side view (Figure 2.22). The two reasons that we choose only these three poses are: (1) these three models are the most dominant cases of bicyclists spotted in the sampled naturalistic driving data, which can be extended

into the general real-life situations, where these three poses are mostly observed; and (2) the front and rear poses will change to one of these three poses as the position relationship between the observing vehicle and bicyclists changes. In other words, there will eventually be a moment when a front/rear posed bicyclist changes its pose to one of the three chosen poses so it can be captured by these three detectors in the video. To further accelerate the scanning, the front-side view detector will be only performed on the left half of the ROI and the rear-side view detector will be only performed on the right half of the ROI for the TASI 110-car dataset.

Because of the changeable background and bicyclist colors and intensities, our feature relies on the outline or edge of bicyclists. HOG features can loosely describe global shape but provide flexible changes locally to the shape. During stage II detection, the candidate windows output from stage I were encoded into the HOG representation. The traditional HOG method relies on stably computing the overlapping local histogram of edge direction in a dense way. The detection window is normalized into a 128×64 pixel image patch. Each window is divided into 15×7 overlapping blocks, and each block is further divided into 2x2 cells. The 9-orientation histogram of gradients is computed within each cell. The locally computed distribution vector is then concatenated into a 3780 dimensional descriptor.

In our implementation, we use 128×64 pixel normalized windows for the two slanted view poses and 128×128 pixel normalized windows for the side view poses. For the HOG features, we compute the fast HOG using integral image [96]. 8×8 pixel blocks and 2×2 cells are applied [42]. The average gradient representation of the three poses is generated from the training set. The resulting HOG is shown in Figure 2.22, where the dominant parts from many bicyclist training windows are used to distinguish bicyclists from other objects.

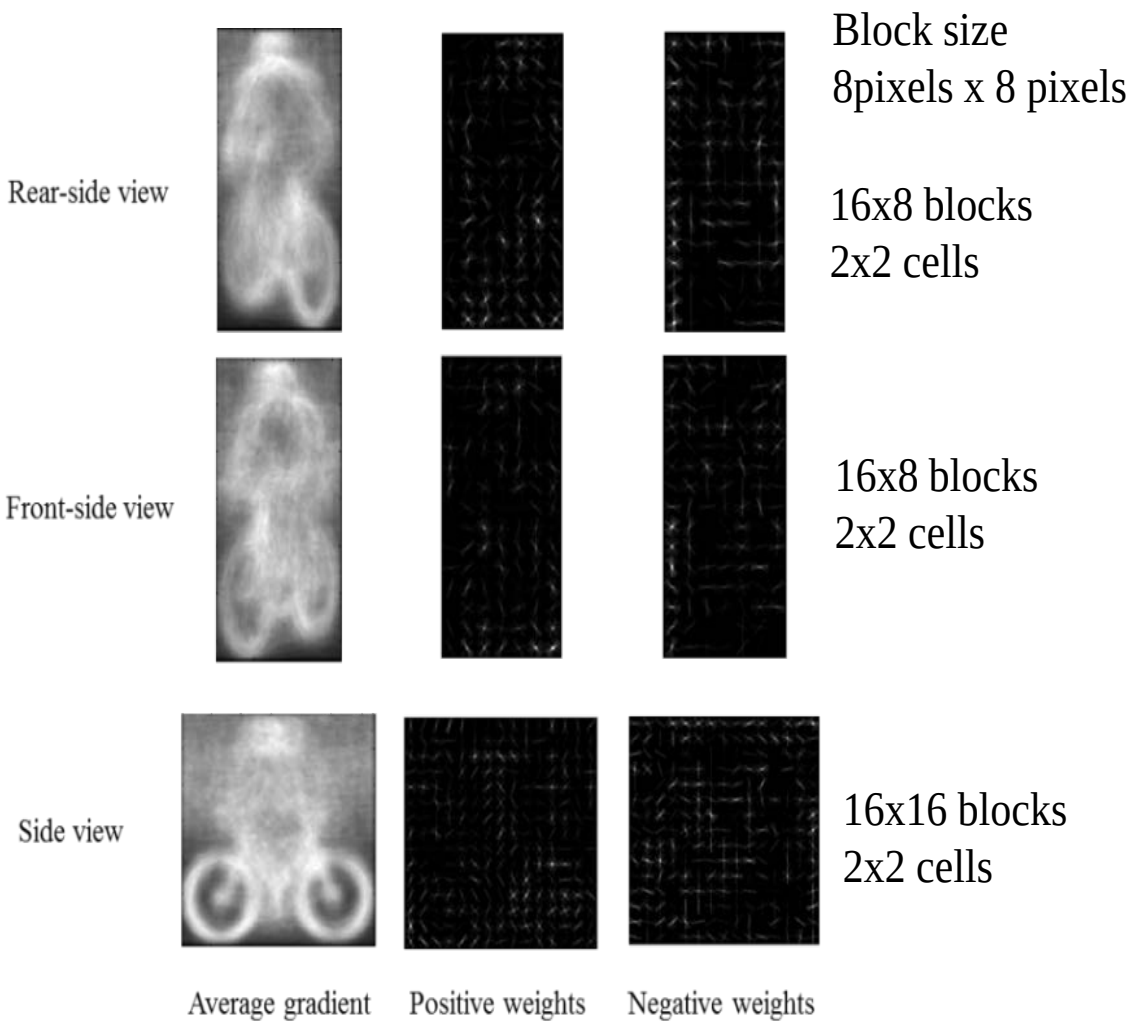


Figure 2.22 HOG representation and trained classifier weights (on intensity and orientation) of three pose-specific classifiers for bicyclists.

B. Feature learning

Beyond human engineered features based on our environment and target analysis, we want to further confirm if there is any better way in classifying the bicyclists because our designed features and learned models may not be optimal. In this section, the possibility of applying deep networks to bicyclist detection was explored purely based on the large dataset. A deep learning framework was formed to extract first-order features from bicyclist patches that is used for extracting feature. A fine-tuning step was followed by a

supervised learning using RBF kernel-based ELM (Extreme Learning Machine). Note the deep network relies on its representation ability to automatically extracted low level features from an unlabeled dataset therefore can be used as an unsupervised learning algorithm. The low-level features of bicyclist patches can be extracted by a single layer auto-encoder and used as substitute or compliment for the raw representation or other hand-engineered features, such as HOG. The layer learning can be repeated and stacked to learn high-level representations.

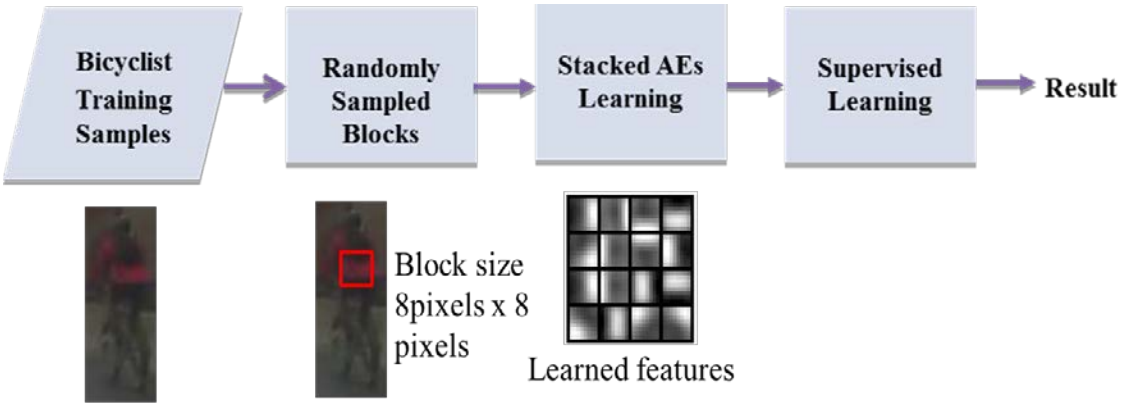


Figure 2.23 The proposed framework to learn bicyclist features using deep learning

Figure 2.23 shows the proposed framework of applying the idea of deep network to learn features of bicyclist using naturalistic driving data. Unsupervised learning and supervised learning are mixed. The raw pixels of the color and gradients of the randomly sampled blocks are directly used as the input to a single layer AE to learn the hidden layer parameters. The learned result served as the first layer filter of the convolution network. The AE learning are repeated and stacked to learn from low-level features to high-level features. Sparse constraints are applied to the AE and a single activation is constrained on the hidden layer. The features are learned layer by layer without supervision. The outputs of the learned, stacked AEs are then input into a multi-layer perceptron network (MLP) and fined tuned in a supervised fashion.

The fine-tune process is shown in Figure 2.24. Four-layer stacked AEs including two convolution layers and two pooling layers are constructed and learned using color images

and gradient images of bicyclist patches. More layers can be stacked to get higher level representation. The final output of all the sub-blocks are concatenated into a single vector and a two-layer fully connected ELM is learned with supervision.

The extracted features from the hidden layer of each single layer AE can be used as the input to a next layer of AE and the multiple-layer AE can be stacked and higher level features will be extracted in later stages of AE. The high level features can be directly used as extracted features or combined with hand engineered features such as HOG to form a final round of supervised learning, or "fine-tuning" to improve the final detection result. The round of supervised learning is shown to be very useful in improving the classification performance.

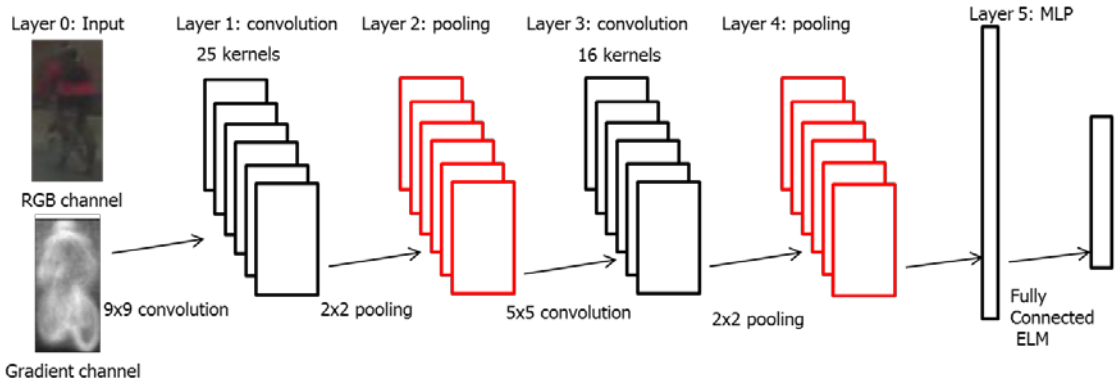


Figure 2.24 Fine-tune of the learned stacked AEs for bicyclists

2.5 Experimental Results of the Bicyclist Detection

A. Experimental results on test sample frames using two stage multi-modal bicyclist detector

During stage I detection, Integral features are extracted from each sliding window balancing the performance and the computational efficiency. The extracted features are from color and gradient channels. A pre-trained cascaded Adaboost classifier is applied to quickly eliminate non-pedestrian windows and generate refined candidate windows. Most

negative windows can be rejected in this early stage. In this section, we will discuss the building process of the cascaded classifier.

To train the two-stage multimodal bicyclist detector and test the performance, we generated a very challenging test sample set which contains 160 frames with bicyclists, randomly selected from the TASI 110 naturalistic driving video. Bicyclists within the test set varied greatly in size and appearance, with height ranging from 30 pixels to 250 pixels. Three pose-specific classifiers have been trained using a manually cropped training set. The training set is not overlapping with the test set. The positive set contains 922 cropped patches for front-side view cases, 1628 cropped patches for rear-side view cases, and 733 cropped patches for side view cases. The negative set was randomly generated from the naturalistic videos without bicyclists. Three rounds of bootstrapping have been implemented using a selected bootstrapping training sets and hard examples to retrain the classifier. For best performance, we intentionally kept a certain amount of margin around the cropped training samples.

Three pose specific cascaded classifiers were trained using the cropped training samples. The integral features are randomly generated and computed. The parameter of each layer was selected by ensuring no false rejection and detect highest false windows. We kept adding more cascaded layers until the object false window reduction rate was attained. Remember the goal of stage I classifier was to quickly remove most false windows. We set the false window reduction rate as 99% and use 10 as increment when adding the cascaded layers. The results on test set is shown in Figure 2.25. The reduction rate was observed saturated as layer number increased. We selected 100 as the layer number since it achieved adequate reduction rate while still ensuring no false rejection.

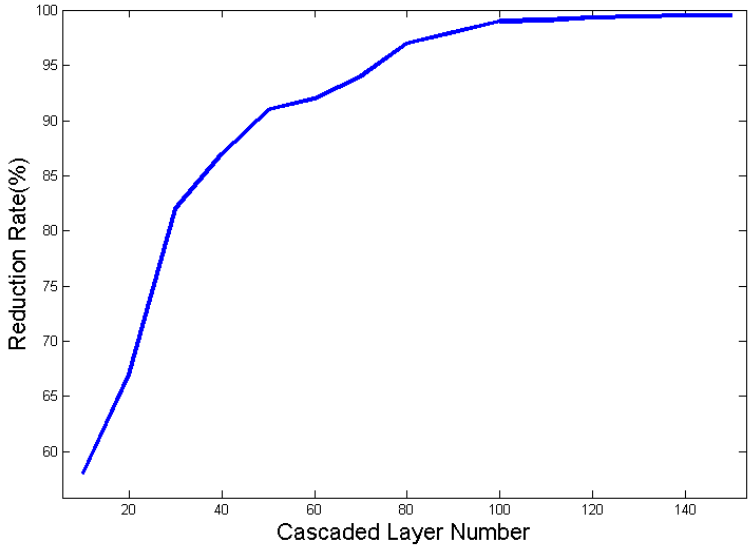


Figure 2.25 False window reduction rate by cascade classifier layer number

The full image was evaluated and a standard sliding window scanning scheme was performed. Non-maximum suppression (NMS) was implemented to combine nearby and overlapping detections. We applied false positive per image versus the miss rate metric which is popularly used to measure the performance of pedestrian detection system using the standard PASCAL measurement. The ROC curves of the three separate pose-specific classifiers are shown in Figure 2.26. We observed that the side view detector performs best among the three due to the unique and explicit bicycle appearance of the side view bicyclist. The traditional HOG+SVM detector [2] was also implemented and performed on the test sample frames as a baseline result. The same three pose-specific HOG+SVM detectors were trained using the same training set we generated from the naturalistic driving video. The same parameter setting as the HOG encoding in stage II was applied and the linear SVM was trained to classify the sliding window. The same evaluated metric was used and the comparison result is shown in Figure 2.27. The proposed two-stage detector outperforms the traditional HOG detector on the test sample frames. The proposed detector also achieved over 10 times improvement in terms of the computational time compared with the time-consuming HOG+SVM detector.

B. Experimental results on large scale naturalistic driving videos

The comparison results of the proposed detector with motion-based ROI prescreening and the whole frame sliding-window based detector is shown in Table 2.7. The false positive rate is calculated as the number of false detected frames divided by the total number of frames without bicyclists. The prescreening step efficiently reduces the sliding window searching region while maintaining the detection rate as the whole frame scanning. The resulted false positive rate is also reduced from 4.7% to 3.1%. Some of the detection results are shown in Figure 2.28. The horizontal lines are the detected bound of the ROI using the motion based prescreening. The red, green and blue bounding boxes stand for detections of front-side view, rear-side view and side view respectively. Two examples of tracking-by-detection are shown in Figure 2.29, where side view and rear-side view bicyclists are captured at different distances.

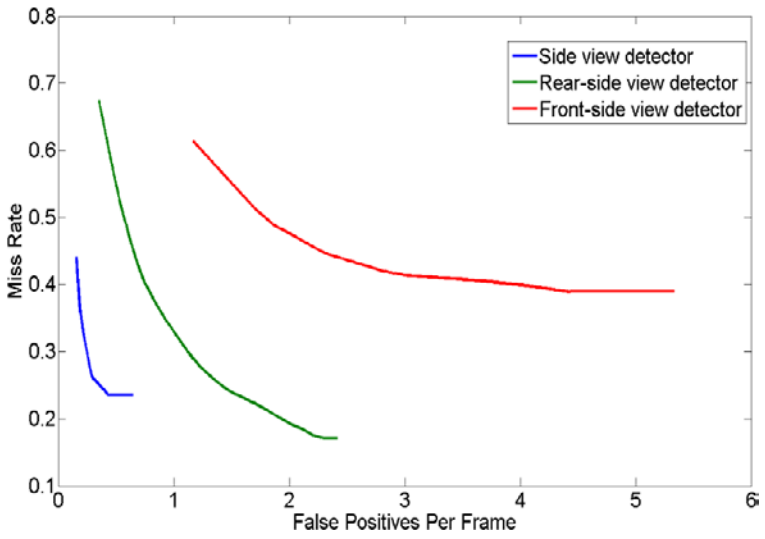


Figure 2.26 ROC curves of the pose-specific bicyclist detectors. Blue: Side view detector, Green: Rear-side view detector, Red: Front-side view detector.

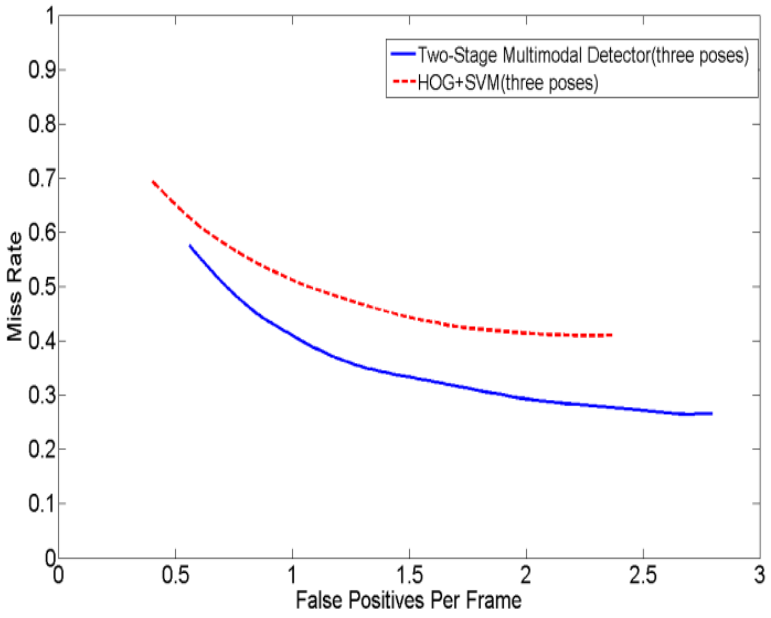


Figure 2.27 Comparison result between the proposed detector and traditional HOG detector

Table 2.7. Comparison result of the proposed method with and without prescreening

	Detection Rate	False Positive Rate	Computation time (seconds per frame in average)
Proposed method without prescreening	88.1%	4.7%	0.21
Proposed method with prescreening	88.1%	3.1%	0.16

Figure 2.28 Detection examples of pose-specific bicyclist detector on naturalistic driving videos

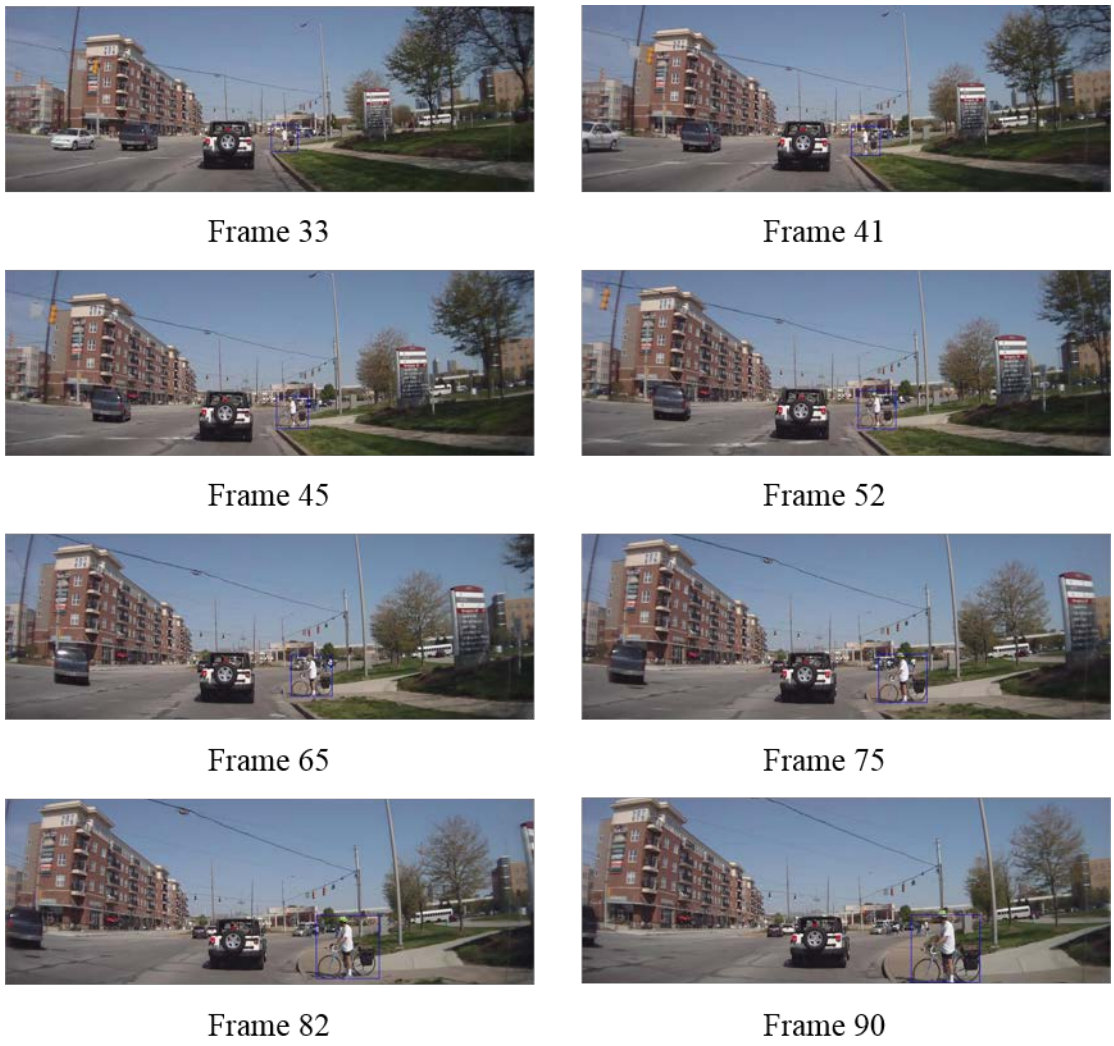
Frame 45 Frame 48

Frame 54 Frame 61

Frame 69 Frame 76

Frame 83 Frame 90

(a)



(b)

Figure 2.29 Examples of tracking-by-detection of five-second bicyclist videos (a) rear-side view (b) side view

To show how the proposed bicyclist detector interact with pedestrian detector, we ran experiment on a test set including both bicyclists and pedestrians. We used the pedestrian detector proposed in chapter 2.2 to detector pedestrians. We retrained the pedestrian detector with added bicyclist samples in the negative training set. We also retrained the multi-pose bicyclist detector with only the lower body to potentially reduce false acceptance from pedestrians.

The test set in last section was used again in this experiment. 42 bicyclists and 74 pedestrians are labeled in the test set. We ran the retrained bicyclist detection and pedestrian detector frame by frame respectively and integrated based on the detection score. The detection results is shown in Table 2.8. The detection rate of the bicyclist reduced slightly compared to the previous experiment due to the mis-detection as pedestrians. No mis-detection of pedestrians as bicyclists was reported due to the lower body of the bicyclist was used for training.

Table 2.8 Results of the experiment with bicyclists and pedestrians combined

	Detection rate	False positive rate	Mis-detected as bicyclist/pedestrian
Pedestrian detector	60/74	1.52%	0
Bicyclist detector	33/42	5.4%	4

C. Experimental results of the learned deep network

The proposed deep network and learned features using bicyclists in naturalistic driving data is explored and compared with the feature engineering methods. For simplicity, only the rear-side view and the front-side view samples were used as training and test set.

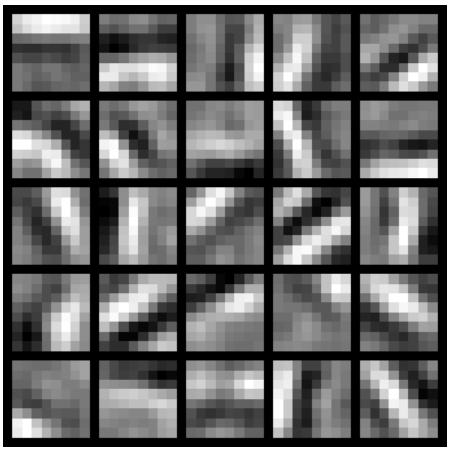


Figure 2.30 The first layer AE learned features using natural images

The randomly sampled blocks from natural images are directly used as the input to a single layer AE to learn the hidden layer parameters. The learned result served as the first layer filter of the network. Sparse constraints were applied to the AE and a single activation was constrained on the hidden layer. The trained weights of the learned 25 filters are visualized in Figure 2.30, where we can see the outputs are wavelet form representations which implies edge detection is important.

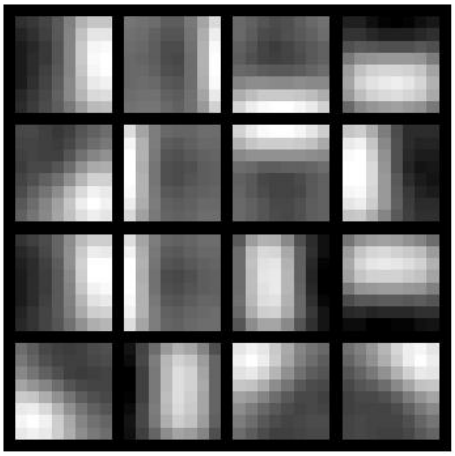


Figure 2.31 The second layer AE learned features using bicyclist images

A second layer AE was learned on top of the first layer node trained before. The second layer activation is shown in Figure 2.31, where each of the 16 node outputs is actually a linear combination of the first layer output. The second layer node weights are learned using patches from naturalistic driving data and the positive bicyclist training sample. It shows higher level representation of the image, such as the part of the bicycle wheel.

Based on the 2nd Layer node activation weights trained above, a 2-stacked convolution network is trained and a round of fine-tuning using the labeled training set is carried out. The learned features from this trained network were served as the feature extractor of the bicyclists and a supervised ELM is learned on top of the 2-stacked convolution network. The test results on the test set are shown in Figure 2.32. We compared the convolution network with different layers and the proposed two-stage bicyclist detector using hand engineered features. We also compared to results with gradient image and without gradient

image as initial input. The gradient image achieves much better results than just using raw image and 2 layer network outperforms the single layer counterpart respectively. The result of the 2-stack deep learning using the gradient image is close to the HOG method, which is promising, in that just two layers of AEs are learned. For simplicity, only rear-side view and front-side view are used for the test so that no warping is needed. We also trained a 2-stacked convolution network on top of HOG features. The resulting detector achieved comparable performance to the two-stage detector, which on the other side shows the effectiveness of HOG in representing bicyclist. Table 2.9 shows a breakdown of the two stage detector and the convolution network. The running time of the deep network is acceptable due to the reducing number of windows to classify after the stage I false window reduction.

To further evaluate the 2-stacked convolution network, the entire test set sampled from TASI 110-car naturalistic driving dataset contains about 900,000 frames with 42 labeled bicyclists was tested using both the proposed two-stage detector in chapter 2.4 and the learned 2-stacked convolution network. The gradient image was used for the convolution network for optimal results. A frame-by-frame detection with window-based evaluation metric was applied. The window based true positive rate versus false positive rate was reported. A “hit” window was counted when the ratio between the intersection of the detection window and the labeled window and the union of the detection window and the labeled window is greater than 50%. The comparison result is shown in Figure 2.33.

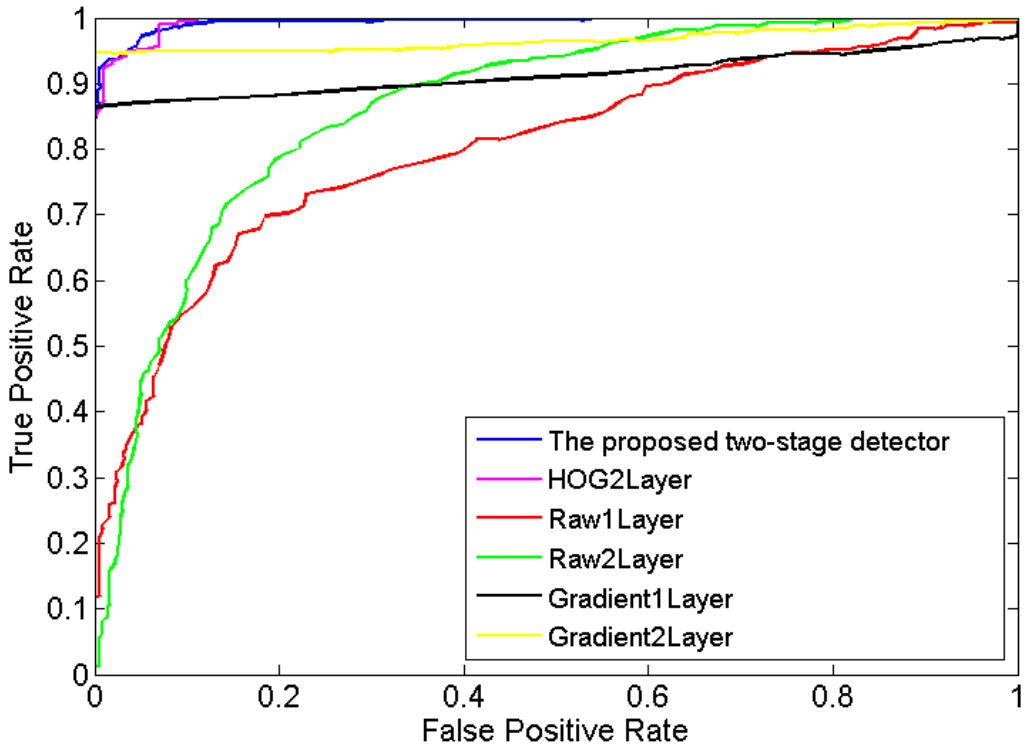


Figure 2.32 Comparison results on test set using only rear-side and front side bicyclists

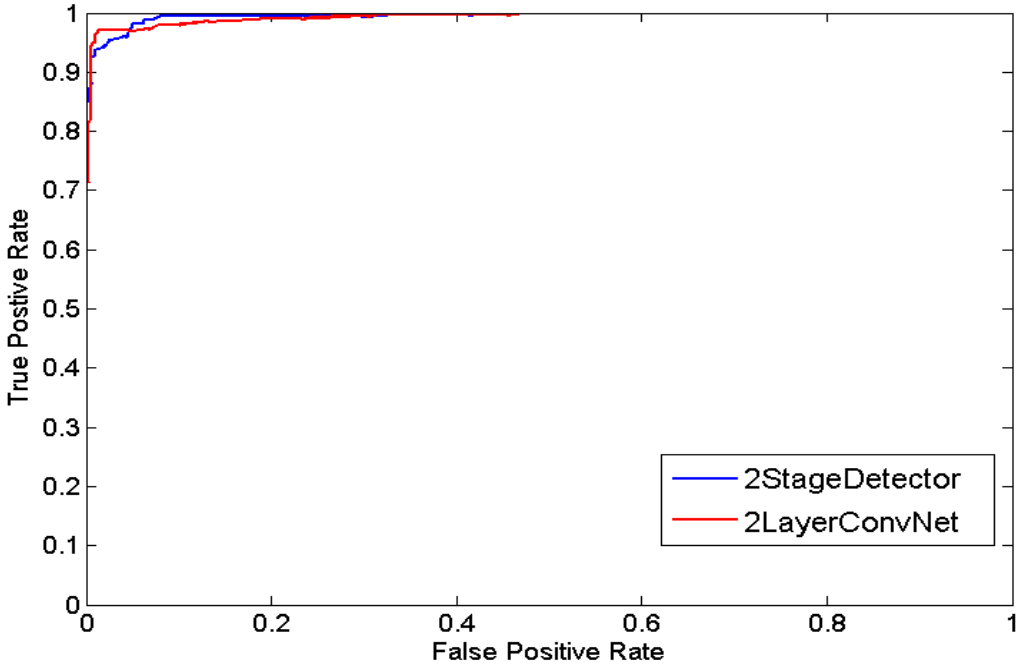


Figure 2.33 Comparison results on naturalistic driving data

Table 2.9. Computation time breakdown of the proposed two-stage detector and the convolution network

	Stage I cascaded classifier (seconds per frame in average)	Stage II Multi-pose detector(seconds per frame in average)
The proposed two-stage detector	0.05	0.16
2-stack convolution network	0.05	0.41

Because of the nature of problems to recognize several poses of bicyclists against changeable background, the HOG based features coupled with strong contrast results in a high ROC, with the area under the curve at 0.983. A deep network has not reached this level and the best area under the ROC curve is 0.968, though a controlled process with supervised learning begins to converge to the HOG results, as seen in Figure 15. The deep network middle layer output shows reasonable features, close to edges and bicyclist primitives in the recognition. The deep network middle layer output shows reasonable features, close to edges and bicyclist primitives in the recognition. In deep learning, edge based data gains better results than color data while convolution network built on HOG features achieved comparable results to the proposed detector, which reflect the correctness of using HOG features to detect variety of clothes of bicyclists in a changeable background. It could be helpful to improve the performance of the deep network by stacking more convolution layers and pulling layers. However, due to the limited number of the bicyclist samples in our study, the very deep network is unlikely to converge. In future, we plan to collect and generate more bicyclists training samples to train deeper nets.

3. THE PROPOSED BOTTOM-UP IMAGE-BASED PEDESTRIAN CLUTTER METRIC

3.1 Definition of Image Clutter Metric

Many researchers have studied visual clutter and given their understandings. Clutter is a term borrowed from radar image referring to any signal in a scene that is of no interest to the observer [40]. The definition of clutter is mostly related to the visual search/detection task therefore Bhanu [100] first defined clutter as an object which resembles the target. While the consensus of background clutter is still unclear and varied from task to task, the effect of clutter on target acquisition performance has been widely studied. There are generally two types of operational definition for visual clutter. The first category of definition relates the clutter with the degradation of visual task performance. It is believe that the clutter acts as a distractor during the target search phase and reduce the accuracy during the target detection task. Among them, Rosenholtz et al. [3] particularly studied the effect of clutter on degradation of visual tasks and defined visual clutter as a situation where excessive visual information with high variability may lead to the degradation of visual task performance. The second type of definition relates the clutter with set size [13] or the "crowdedness" [101] of the scene by building the correlation between the "object" number in a scene with searching efficiency.

Based on the aforementioned two different types of understanding of visual clutter. The computational visual models can be roughly divided into two categories: feature space-based model and set size-based model. The first category relies on building a mapping from image-based metric extracted from multiple feature spaces to the clutter level. The input signal is decomposed into multiple feature spaces and a subset of them is selected to measure the clutter intensity. The second category relies on counting the number of the objects in a scene and build a mapping from the set size to clutter intensity. Although it is

argued that the set size is difficult to quantify in realistic scenes, computer vision aided segmentation methods are usually applied to calculate the “object” numbers.

Pedestrian detection within naturalistic driving scene is a complicated process which combines the vision perception and brain cognition, not a simple visual search task tested in building the above clutter metric. The clutter intensity in this case should be conditional. In another word, it should be both feature space based and target related. For example, a pedestrian may completely merge into the background if he/she has low contrast no matter how much information the background feature space may contain. A pedestrian with high local contrast may still be able to be detected promptly given a highly variant background. However, most existed clutter metric for visual search task does not consider the searching/detection target itself. Moreover, the feature space selection and weighting for general image may not applicable to naturalistic driving scene and a customized clutter metric is need for pedestrian clutter modeling.

As we mentioned before, the limitation of the existed clutter metric and computational model for general visual task does not suitable for modeling the visual clutter effect on pedestrian detection from naturalistic driving. We split the clutter metric into a complexity-based global environmental clutter measure and a contrast-based local pedestrian clutter measure.

3.2 Global Environmental Clutter (GEC) Measure

3.2.1 Existing Global Clutter Metrics

Global clutter metrics were developed to measure the overall complexity of the scene from physical image property without considering the cognitive assessment of the observer. The subjective ratings from human observer were usually compared with the objective clutter metric to build a reasonable model.

Many global clutter metrics have been proposed during the past two decades. There is no agreement on which metric is best yet, therefore we explored several popularly used metrics before we proposed our customized metric for naturalistic driving data.

SW metric

Schmieder and Weathersby [102] proposed to measure the scene complexity by computing the root mean square of the image intensity. In particular, SW metric computes the average of the variance within consecutive image blocks:

$$\mathrm{SW} = \sqrt{\frac{1}{\mathrm{M} * \mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{M}} \sum_{\mathrm{j}=1}^{\mathrm{N}} \sigma^2{}_{\mathrm{i,j}}} \tag{3.1}$$

where M and N is the divided grid number of horizontal and vertical directions within the entire image. $\sigma_{\mathrm{i,j}}$ is the variance of the pixel intensity computed within block $\mathrm{i,j}$.

POE metric

The probability of edge (POE) metric [103] emulates the human vision system which is sensitive to edges. It calculates the edge map using image preprocessed by difference of offset Gaussian filters. Canny edge detector is used with predetermined thresholds. The POE clutter is the average of edge point numbers counted from the edge map block. Given threshold T and block numberi, the POE metric is calculated as

$$\mathrm{POE} = \sqrt{\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{POE}_{T,i}^2} \tag{3.2}$$

where $\mathrm{POE}_{T,i}$ is the count of edge number within block i given threshold T.

Feature Congestion (FC)

Rosenholtz et al. [3] studied the visual clutter by assuming the clutter in a local part of a display should be determined by the local variability of several key features. The Feature Congestion model [34] relies on calculating the target saliency and the local variability at multiple scales. Color, orientation and luminance contrast are selected as the features to measure the target saliency versus the local variability.

Subbanding Entropy (SE)

Subbanding Entropy [3] is based on the notion that clutter level should be reflected by the bits required for subband image coding. To compute the subbanding entropy, the image is first converted into Lab and then decomposed into wavelet subbands using steerable pyramid [35]. The generated wavelet coefficients are binned and the entropy is calculated within each subband. The final score is a weighted sum of the entropies computed in luminance and chrominance channels.

C3 metric

More recently, Lohrenz et al. [36] proposed their C3 (Color-Cluster Clutter) model of clutter, which derives clutter estimates by combining color density with global saliency. Color density is computed by clustering into polygons those pixels that are similar in both location and color. Global saliency is computed by taking the weighted average of the distances between each of the color density clusters.

3.2.2 The Proposed Global Environmental Clutter Metric

Figure 3.1 Region of interest (ROI) of the GEC measure

Construct the feature space

As we mentioned before, the image clutter metrics relies on building the feature spaces and generating a mapping from the extracted feature set to clutter intensity. However, the mapping of the existing metrics are usually empirical selected and case sensitive. Moreover, the feature selection and mapping could be varied from different types of scenes. For instance, a monochrome city map image would emphasize the edge map more than other feature spaces while the clutter of a color world map image would better measured by color variance. In another word, such mapping should be scene-specific and task-specific to best reflect the true human vision perception.

To build the specific mapping between the clutter score and naturalistic driving scene, candidate feature maps have to be constructed first. We propose the GEC metric to directly measure the overall clutter score of the entire image based on several candidate features. We then build the mapping through human perception inspired study. The regions of interest (ROI: region inside red box shown in Figure 3.1) is first selected from the full view to exclude the sky and driving panel parts which should not be the pedestrian search region during driving. The upper bound of the ROI is set at a fixed position to get rid of the sky and the driving information recorded by the camera shown on the upper left corner of the video to emulate the actually view while the driver is driving.

The global environmental clutter (GEC) feature vector is select as:

$$f_{GEC} = [\rho_{GE} \quad \sigma_{GL} \quad \sigma_{GC} \quad \rho_{GM}]^T \tag{3.3}$$

where ρ_{GE} is the global edge density, σ_{GL} is the global luminance variation, σ_{GC} is the global chrominance variation and ρ_{GM} is the global motion density. We will illustrate the rationale and implementation detail of each feature next.

- *Global Edge density (ρ_{GE})*

  Global edge density has been proven to be a good indicator of the global clutter level correlated to human vision perception. Oliva *et al.* [33] proposed to use canny edge

density as a clutter feature which achieves good correlation with human perception. Therefore we use a canny detector with fixed threshold range to detect edge to fairly compare edge density of different frames acquired in different driving scenarios, illumination and weather conditions. The low threshold is set as 0.11 and high threshold is set to be 0.27 respectively, following the parameter selection in [3]. Considering the low-pass characteristic of human vision system, a 7 by 7 Gaussian filter is applied to each image before the Canny detector to remove excess high frequency image component to which human vision system are not very sensitive. The final edge density is calculated as the ratio between the number of edge pixels and the total number of pixels in the frame.

- *Global Luminance variation (σ_{GL})*

Global luminance variation is computed in a block way on the luminance channel of L*a*b* to measure the luminance change of the entire image. A luminance variation matrix with the same size of the entire image is pre-generated. A 9 by 9 sliding window slides all over the image and the standard deviation of the luminance value within that 9 by 9 window is computed as the entry of the luminance matrix corresponding to the center pixel of the sliding window. The final luminance variation is the mean value of the luminance matrix.

- *Global Chrominance variation (σ_{GC})*

Global chrominance variation is computed on two chrominance channels *a* and *b* respectively similar to the way of computing luminance variation. The final chrominance variations is calculated as

$$\sigma_{GC} = \sqrt{\sigma_a^2 + \sigma_b^2} \tag{3.4}$$

where σ_{GC} is the final chrominance variations, σ_a is the chrominance variations of channel a and σ_b is the chrominance variations of channel b.

- *Global Motion density (σ_{GC})*

  Global motion density is optional feature for video input only. It is computed as the average magnitude of the motion vector of the entire frame.

The global environmental clutter score is a function of edge density, luminance variation, chrominance variation and optional motion density. An example of the four feature maps and computed features are shown in Figure 3.2. The higher GEC score means higher global clutter.

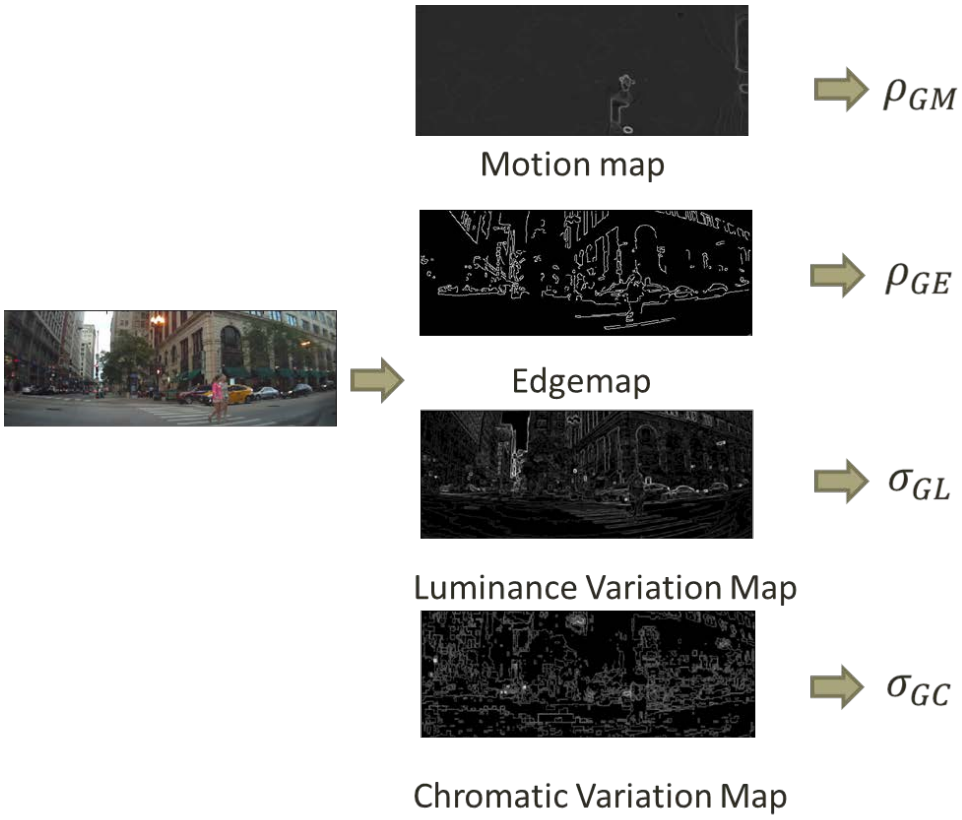


Figure 3.2 Global environmental clutter features computed from four feature maps

To derive the mapping function from the candidate feature to the GEC score which emulates the true human visual perception, a human perception inspired study was conducted. A labeled behavioral ground truth set containing naturalistic driving data is

collected through the GEC exploratory experiments. The mapping and related parameters of the GEC are learned from the objective image metric and behavioral ground truth. Therefore our method is interdisciplinary, applying the image feature extraction algorithm on naturalistic driving data for human clutter perception task. We now introduce the designed experiment of human perception inspired study for GEC metric.

Experiment 3.1: Global environmental clutter rating for naturalistic driving image

This experiment focuses on exploring how subjects perceive the overall clutter level of a given image taken from the naturalistic driving scenarios. A set of images were displayed to the human subjects and the perceived subjective ratings of global clutter were collected. The set of images were divided into a training set and a test set. The training set is used to learn the mapping function and related parameters and the test set is used to evaluate the learned GEC metric.

Method

Participants

A total of 12 subjects with age from 22 to 33 and driving experience from 2 years to 11 years participated in the GEC rating experiment for naturalistic driving images. All had normal or corrected-to-normal vision, by self-report, and were not guided with any clutter rating judgment standard before.

Stimuli

Stimuli consisted of 100 1280×720 images selected from the TASI 110-car naturalistic driving dataset. The selected image set includes data sampled under different driving scenarios, illumination conditions and weather conditions. The TASI 110-car naturalistic driving dataset is pre-labeled and the category information is shown below:

- Driving scenario: urban (downtown area)/rural/school area/shopping area/community

- Illumination condition: daytime/dusk/dawn/night with street light/night with head light
- Weather condition: clear/cloudy/fog/rain/snow.

The percentage of each category is in accordance with the distribution of the entire TASI 110-car naturalistic driving dataset.

Design

The global clutter level naturalistic driving images were rated by subjects based on their true perception and driving experience. The images were shown in random order to reduce the effects of order or potential bias. For instance, a high-cluttered image could possibly receive higher rate if shown after a series of low-cluttered images and vice versa. No definition of clutter was given to the subjects while they were asked to come up with their own definitions and be consistent through the entire experiment. The GUI of Experiment 3.1 is shown in Figure 3.3.

Procedure

The experiment uses naturalistic driving images taken from an in-car camera. Each subject was asked to sit in front of a computer monitor. A series of naturalistic driving images was shown on the monitor and the subject was asked to input his/her perceived clutter level in the designated box. The rating experiment was carried out using a graphic user interface written in MATLAB running on a Windows 7 PC with a 19-inch LCD monitor. The actual display size of the image region is 20.1×11.3 cm^2. The rating is set to be from 1 to 5, with 1 stands for the lowest clutter level and 5 stands for the highest. Before the experiment set, a practice set with baseline images from different scenarios were given to each subject. The purpose of the practice set and the baseline images was not only to ensure each subject understands the rating process, but also to help subjects build reasonable rating rules with respect to different scenarios on their own. Only the results of experiment set were recorded. During the experiments, subjects were free to go back to the images they had already rated and rerated them if they felt they had made a mistake.

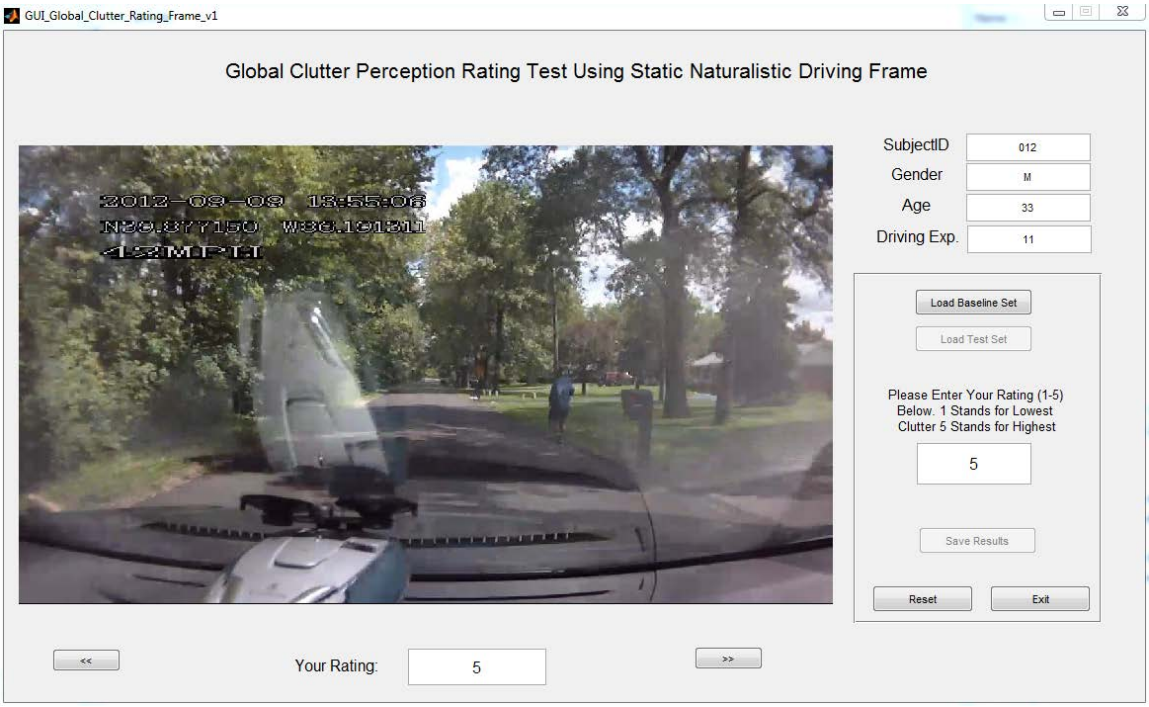


Figure 3.3 The GUI of the GEC rating experiment for naturalistic driving image

Experiment 3.2: Global environmental clutter rating using naturalistic driving videos

This experiment focuses on exploring how subjects perceive the overall clutter level of a given naturalistic driving videos, which is closer related to the true perception of driver than using images. Motion features were extracted from the naturalistic driving video and motion map was added to the feature spaces. The collected results will be served as the ground truth for our video based pedestrian clutter analysis.

Method

Participants

The same group of Experiment participated this study.

Stimuli

50 naturalistic driving video clips with 15 seconds long each were sampled from the TASI 110-car naturalistic driving dataset and similar to Experiment 4.1, the distribution of the selected set was set to match that of the entire dataset. Similarly the 50 video clips were divided into a training set and a test set for the mapping function learning.

Design

Similar to Experiment 3.1, the global clutter level naturalistic driving video clips were rated by subjects based on their true perception and driving experience. The rating experiment was carried out using a graphic user interface written in MATLAB running on a Windows 7 PC with a 19-inch LCD monitor. The actual display size of the image region is 20.1×11.3 cm^2. The 15 seconds video clips are extracted from the large scale driving data set such that the acquisition vehicles may have potential conflicts with the pedestrians. One video clip was shown in the computer monitor screen each time for subjects to rate. The subject was asked to input his/her perceived clutter level in the designated box. The videos can be paused and resumed by subjects at any time and can be played at multiple frame rate. Again, the videos were shown in random order to exclude potential bias. The GUI of Experiment 3.2 is shown in Figure 3.4.

Procedure

The 15 second potential conflict naturalistic driving video clips were shown in the computer monitor and the subjects were free to view the videos as many times as they want. The subjects were asked to input his/her perceived clutter level in the designated box below the video clip. Again, the rating is from 1 to 5, 1 stands for the lowest clutter level and 5 stands for the highest. Similar to experiment, a practice set is prepared for subjects before the test set to help them get familiar with the rating process and build initial impression about the clutter level rating using video clips. Only the clutter level ratings of the test set were recorded.

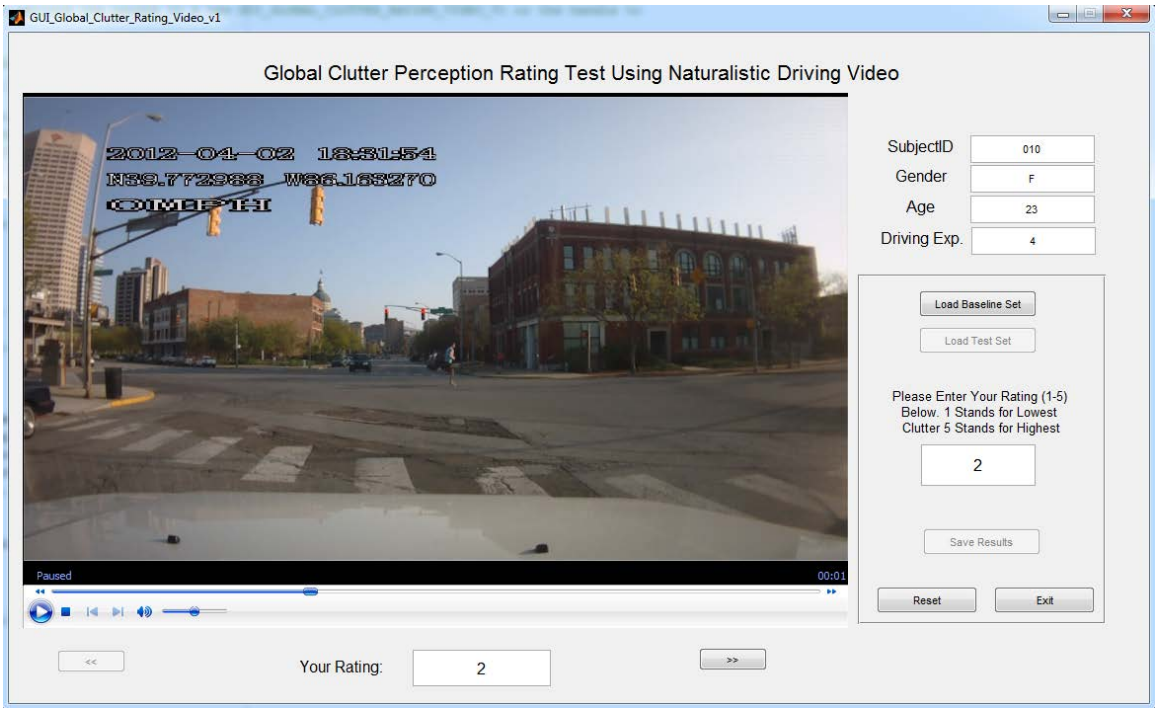


Figure 3.4 The GUI of the GEC rating experiment for naturalistic driving video

Learn the mapping function

The training set collect from the GEC rating experiments were used to learn the mapping function from bottom-up image-based feature to GEC level obtained by subjective rating. The test set was only used for model evaluation and comparison with other existing clutter metric. The parameter learning and tune were through cross-validation using the training set only. A candidate set of regression and learning techniques were selected and the corresponding mapping function or model was learned and tuned. The results were further evaluated by the test set.

Linear/non-linear Regression [104]: the most direct method is to applied regression method to the training data collected from exploratory study. Linear regression, polynomial regression and logistic regression were used to find the best regression function and parameters. Suggested by [105], non-linear regression was also tested to check the possible

fitting model using the Steven's power law [106] between the objective intensity and the perceived magnitude.

Support Vector Machine [107]: SVM is popularly used in object detection, classification and machine learning applications. It aims at finding the best hyper plane to separate different classes. Multi-class SVM was used to derive the mapping. Linear SVM and kernel-based SVM were evaluated.

*Single Layer Feed-forward Neural Network (SLFN)***:** neural network is also a well-known machine learning algorithm which has been developed into a variety of forms. Among the family of SLFNs, Extreme Learning Machine (ELM) has been proposed recently and show better performance than traditional SLFN in multiple tasks. Huang et al.[94] theoretically and experimentally proved that ELM can be used as a unified learning platform which does not need to tune the hidden layer parameters as traditional Single layer neural do. Instead of using the time-consuming gradient descent based learning method; ELM relies on computing the Moore-Penrose generalized inverse of the hidden layer matrix. Later**,** Huang et al. shows that dual optimization objective functions of ELM is consistent with that of SVM while ELM searches optimal solution in a greater domain with faster implementation.

Results

The training set was first used to learn the best mapping function and tune the parameters when necessary using cross validation. The ground truth of each image/video clip was calculated by taking the median subjective ratings of all participants. The reason that we preferred median to mean was to eliminate the effect of outliers. The ground truth was normalized into [0, 1] range as a numerical value instead of a categorical value for classification methods. The rooted mean square error (RMSE) between the predicted value and the ground truth of the validation set (i.e., residuals) were assessed for different regression methods. In addition, a better fit of the regression does not necessarily lead to a better correlation between the GEC score and the human perceived clutter level.

Therefore the correlation between the predict GEC value and the subjective ratings were also compared to find the best predictor.

We first computed the intra-class correlation coefficient (ICC) [108] by averaging the Pearson's correlation between all pairs of subjective ratings. The ICC of all 12 subjects is 0.702, which tells us that there was good agreement among all the subjects. This also shows that the subjective ratings can be served as a reliable ground truth for human perception of global clutter so that the different computational GEC metrics can be meaningfully compared.

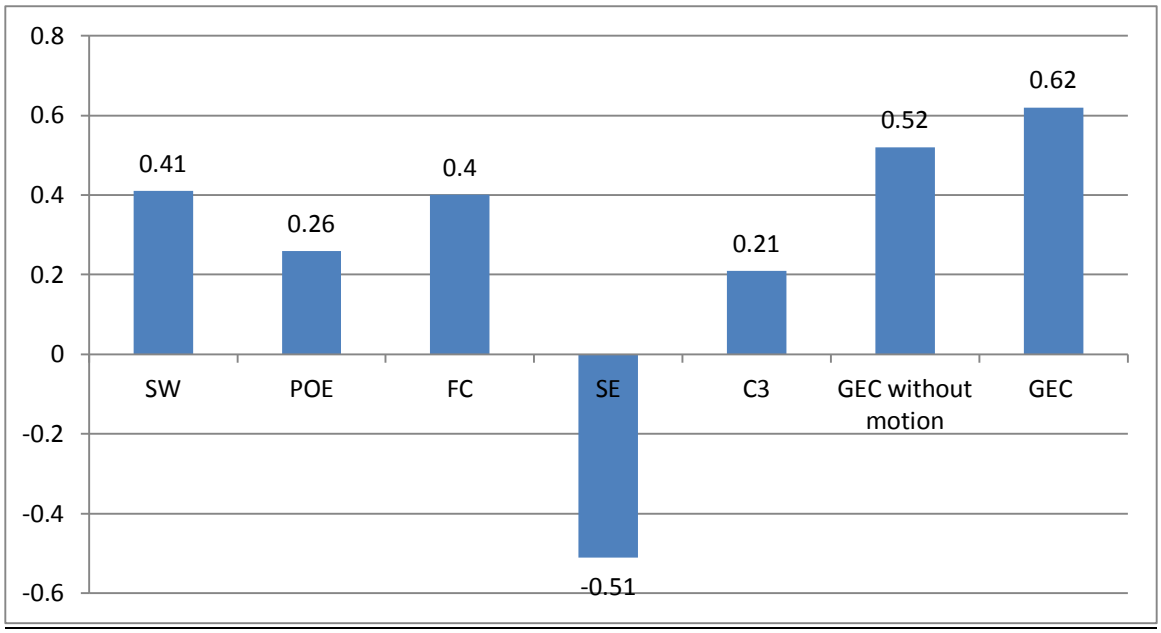


Figure 3.5 Correlation of the proposed GEC compared with other existing methods

Table 3.1 shows the results of the mapping function using different regression models. RMSE and R values are listed and compared. We also computed the *p*-value and all of the tests have *p*-value less than 0.05 which means the correlation is statistically significant. The non-linear regression using the power function achieves best fit and correlation results. We therefore selected it as the mapping function for GEC metric.

The proposed GEC metric was also compared with other well-known global clutter metrics mentioned in Chapter 3.2.1. The Pearson's correlations (R) between the image-based computed metrics and the median of subjective ratings of all test data were computed. The comparison result is shown in Figure 3.5. The SW (0.41), POE (0.26), FC (0.4) and C3 (0.21) metrics are all have weak correlations with the subjective ratings while the SE (-0.51) has a negative correlation, which means none of these existing metrics can predict the true human perception of global clutter of naturalistic driving scene very well. All the tests have *p*-value less than 0.05. In contrast, the proposed GEC (0.62) shows good correlation with the true human clutter perception and outperforms the existing global clutter metrics. Since none of the listed existing metrics considered the motion feature space, to be fair, we also computed the GEC without the motion channel (0.52). The GEC without motion is also correlated well and can better predict the global clutter perception.

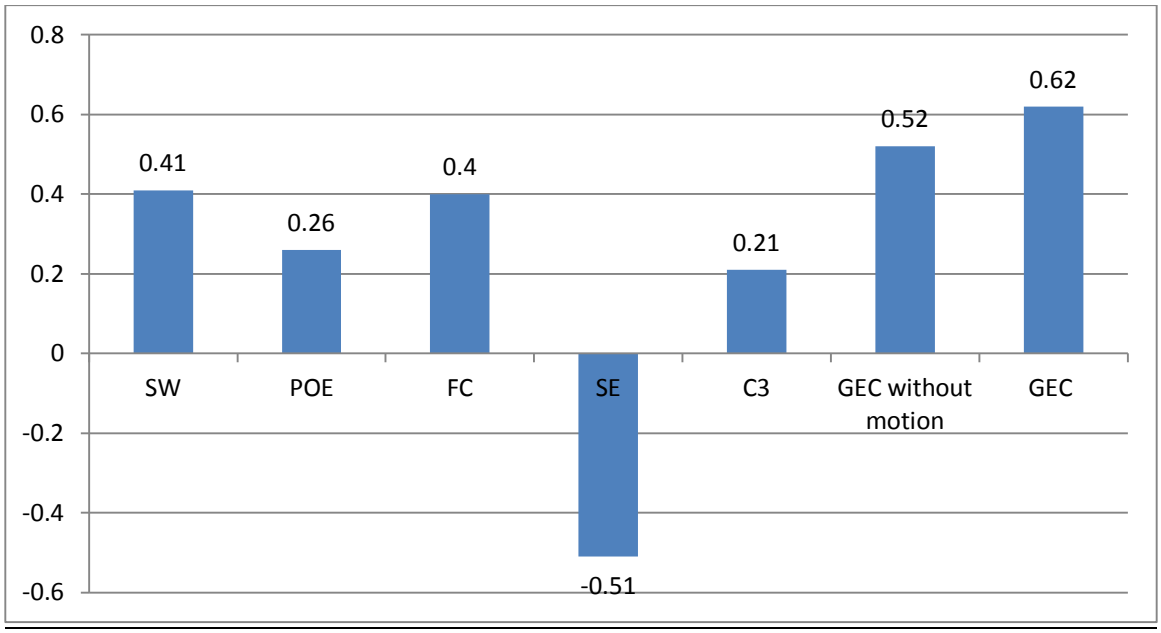


Figure 3.5 Correlation of the proposed GEC compared with other existing methods

Table 3.1 Results of regression models of GEC metrics

Regression model	RMSE	R	*p*-value
linear	0.21	0.50	0.01
logistic	0.22	0.48	0.01
Non-linear (power)	0.17	0.62	0.01
SVM	0.22	0.45	0.02
ELM	0.20	0.51	0.02

3.3 Local Pedestrian Clutter (LPC) Measure

Local clutter metric measures the clutter level in local region around the target. It is essentially measuring the difference or contrast between the target and the local background. Similar to global clutter metric, feature spaces usually are built and a difference function is designed to calculate feature contrast.

3.3.1 Existing local clutter metrics

Several popularly used local clutter metrics have been proposed for years to measure the target-to-background contrast in general target search task for both natural images and synthetic images.

Root-sum-of-squares (RSS) metric

The RSS metric [109] is defined as

$$\mathrm{RSS} = \sqrt{(\mu_T - \mu_B)^2 + \sigma_T^2} \tag{3.5}$$

where μ_T and μ_B are the mean intensity of the target and background respectively and σ_T is the standard deviation of the target intensity.

Doyle metric

In addition to RSS metric, the Doyle metric [110] measures the difference between the target and background in terms of both mean and standard deviation. The Doyle metric is defined as:

$$\text{Doyle} = \sqrt{(\mu_T - \mu_B)^2 + (\sigma_T - \sigma_B)^2} \tag{3.6}$$

where σ_B is the standard deviation of the background intensity.

Although the above existing local clutter metrics are easy to compute and widely used for visual clutter measure for target search tasks, they have several limitations when applying to the pedestrian perception task during naturalistic driving. First, most feature spaces involved in the clutter metric are selected empirically using synthetic image and target search experiments while perceiving pedestrian from naturalistic driving scene could be very different and a different set of feature spaces need to be explored. Second, the difference function treated all feature spaces equal while this may not be the case for pedestrian perception within naturalistic scene. Third, the target feature vector usually extracted from the entire bounding box around the target, which is rough and inaccurate. Refined target segmentation is required to accurately extract the feature of the target. Therefore we proposed a local pedestrian clutter (LPC) to compensate the above issues next

3.3.2 The Proposed Local Pedestrian Clutter Metric

Given the limitations of the existing local clutter metric discussed above, the proposed LPC metric for pedestrian perception within naturalistic driving scene was explored. Similar to GEC metric, the feature spaces were first constructed and the human perception inspired study is explored to learn the optimal combination of the different features. A difference function was then applied to generate the LPC level. Instead of manually label the target area as most previous work did, the pedestrian detection system we proposed for large scale naturalistic driving data was applied to automatically locate the pedestrians within the

naturalistic driving scene. To further extract the local pedestrian clutter feature accurately from the pedestrian region, an active contour based pedestrian region refinement was implemented before feature space construction and feature extraction.

Pedestrian locating

The proposed pedestrian detection system in chapter 3 is implemented to detect frames with pedestrians within billions of naturalistic driving video frames. Some of the locating examples are shown in Figure 3.6. The detected result will be verified and the coordinates of the accurately located pedestrian window (the red bounding box) will be input into the LPC measure module such that a center-surround LPC measure method could be applied to the pedestrian region.

Figure 3.6 Examples of pedestrian locating

Pedestrian contour refinement and cloth extraction

To achieve accurate center-surround LPC measure region, pedestrian cloth region needs to be extracted as accurately as possible. In order to accurately locate pedestrian cloth region, in addition to the two-stage sliding window detection illustrated above, an active contour [111] based pedestrian contour generation is further applied to the detected and verified pedestrian windows. A deformable model is initiated around the actual pedestrian contour

and energy minimization is used to evolve the contour. The energy function can be written as:

$$E(C) = \alpha \int_0^1 |C'(s)|^2 ds + \beta \int_0^1 |C''(s)|^2 ds - \gamma \int_0^1 |\nabla \mathrm{u}_0(C(s))|^2 ds \tag{3.7}$$

where the first two integrals stand for the internal energy which control the contour smoothness and the third integral is the external energy which evolves the contour to the object. $C'(s)$ is the tangent of the curve and $C''(s)$ is normal to the curve. The edge detector function can be defined as:

$$g\big(\nabla \mathrm{u}_0(x,y)\big) = \frac{1}{1 + |\nabla \mathrm{G}_\sigma(x,y) * \mathrm{u}_0(x,y)|^p} \tag{3.8}$$

where G_σ is a Gaussian smooth filter and $\nabla \mathrm{u}_0$ is the image gradient. The generated contour defines the pedestrian mask which will be used to compute pedestrian clutter features, including local luminance variation and local chrominance variation.

In general, a pedestrian has a relatively homogenous cloth region in color and luminance intensity. The color and luminance contrast between the homogenous cloth region and the surrounding background is intuitively more accurate and meaningful corresponding to human visual attention model. K-mean color clustering based cloth region segmentation [112] is then applied to the detected pedestrian window to segment the cloth region. In particular, K color subsets are generated to minimize the within-cluster distance:

$$argmin_s \sum_{n=1}^{k} \sum_{I(x,y) \in S_n} \|I(x,y) - \mu_n\|^2 \tag{3.9}$$

where $S = \{S_1, \dots, S_k\}$ is the k clusters, $I(x,y)$is the chrominance pixel value and μ_n is the mean value of each cluster. The final cloth mask is an intersection of the pedestrian mask by active contour and cloth region derived from K-mean color clustering algorithm.

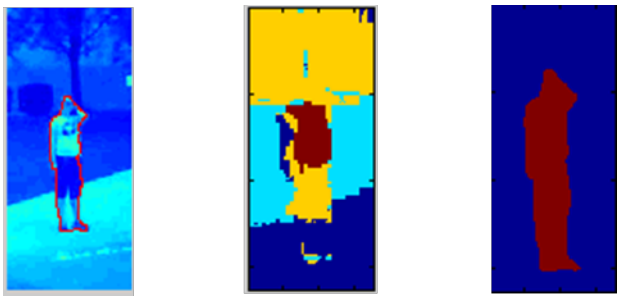


Figure 3.7 Pedestrian contour refinement and cloth extraction result. From left to right: pedestrian contour, cloth color clustering, pedestrian target mask.

One example of the refined contour and extract cloth color cluster is shown in Figure 3.7. The left image shows the result of the active contour generation. The middle image is the color cluster result. Here we use k = 4 which is determined empirically and achieve good result in general. The right image is the pedestrian-background mask which will be used for later feature extraction.

Local Pedestrian Clutter Feature Extraction

Local pedestrian clutter is measured by the contrast between the pedestrian area and the surrounding background area using low-level image based features. In particular, the contrast is represented by the distance between the feature vectors extracted from pedestrian area and background area respectively. The background window is defined as a larger surrounding window with twice the area of the detected pedestrian window (Figure 3.8). We illustrate each proposed feature in detail next. The local pedestrian clutter (LPC) score is defined as:

$$LPC = 1 - \frac{\Delta(T, B)}{\|\Delta(T, B)\|} \tag{3.10}$$

where T is the 15 dimensional feature vector $[T_1, \dots T_{15}]^T$ of pedestrian area and B is the 15 dimensional feature vector $[B_1, \dots B_{15}]^T$ of background area. Δ measures the distance between the two vectors. In our current implementation, the saliency distance [3] is used:

$$\Delta(T,B) = \sqrt{\left(f(T) - f(B)\right)^{T} \Sigma^{-1} (f(B))(f(T) - f(B))} \tag{3.11}$$

where f is the mapping function we want to learn from the human perception inspired study and Σ stands for the covariance matrix. Note that for the last four features, we use a bin size of 16 while calculating the distance, i.e. the luminance intensity and chrominance intensities are regrouped into 16 intensity levels for the entire 0 to 255 range. The local pedestrian clutter score is also a normalized value from 0 to 1. The higher the local pedestrian clutter is, the more cluttered the pedestrian is, suggesting more difficult to perceive the pedestrian from the background.

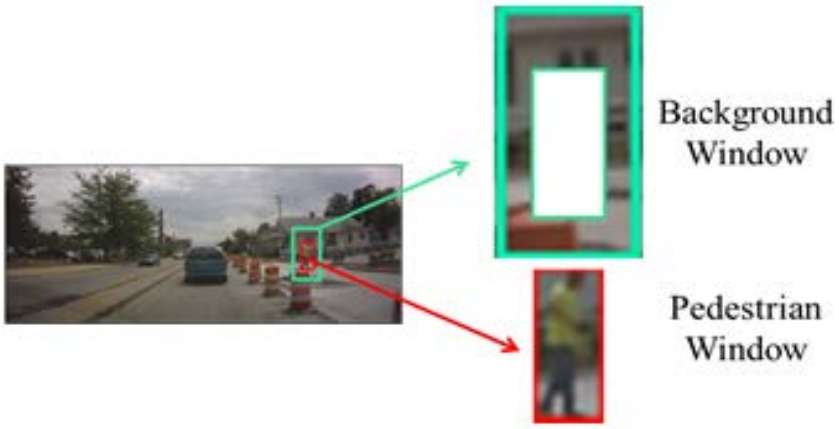


Figure 3.8 Illustration of background window and pedestrian window

- *Local Edge density (ρ_{LE})*

The local edge density is calculated the same way as computing global environmental clutter score within the pedestrian window and within the region generated by subtracting pedestrian window from the background window respectively.

- *Edge distribution (H_{LD})*

Local edge distribution is a histogram of edge magnitude binned by the edge orientation similar to the idea of HOG. Two distributions are computed within the same two regions

defined in computing edge density. Orientation bin number is empirically set as 9 as the HOG representation.

- *Local luminance variation (σ_{LL})*

Local luminance variation is computed within the pedestrian mask defined by the pedestrian contour and within the region generated by subtracting pedestrian mask region from the background window respectively. It is computed in the same way as that of the global environmental clutter score measure.

- *Local chrominance variation (σ_{LC})*

Local Chrominance variation is computed within the two regions defined in computing local luminance variation using the same way as the chrominance variation in global environmental clutter score measure.

- *Mean luminance intensity (μ_l)*

Mean luminance intensity is computed within the cloth mask region and within the region generated by subtracting cloth mask region from the background window. The average luminance intensity is calculated using the L channel of Lab representation.

- *Mean chrominance intensity (μ_C)*

Mean chrominance intensity is computed within the two regions defined in computing mean luminance intensity respectively. The average chrominance intensities are calculated using a and b channels of Lab representation respectively.

- *Mean motion magnitude (μ_m)*

Mean motion magnitude is computed within the two regions defined in computing mean luminance intensity respectively. The average magnitude of motion vector within the defined regions will be computed only for video based stimuli.

Similar to GEC metric, a mapping function f in Equation 3.11 was learned given the extract features and the results of the human perception inspired study for local pedestrian clutter. The same set of regression/learning methods of GEC was used for LPC. Both images and videos were used as stimuli for the LPC rating experiments. The LPC experiments are introduced next.

Experiment 3.3: Local pedestrian clutter rating for naturalistic driving image

The pedestrian clutter level perception test was designed to collect the true perception that how difficult pedestrians in naturalistic driving scenarios can be perceived. The pedestrian clutter result collected from the subjects will be treated as the ground truth for the mapping function learning. The most correlated features (could be extracted from local pedestrian window, global feature map and different saliency maps) with true human perception will be learned and assigned appropriate weights based on the analysis of the study results.

Method

Participants

The same group of subjects in Experiment 3.1 attended this study.

Stimuli

The stimuli in this experiment were naturalistic driving images which contain one or multiple pedestrians. The same set of test images as Experiment 3.1 was used.

Design

The stimuli in this experiment were naturalistic driving images which contain one or multiple pedestrians. The pedestrian clutter level was subjectively rated by each subject based on their perception of the pedestrians. A red box was shown around the pedestrian area to indicate the target pedestrian for rating. The red box would disappear after three seconds therefore no artifacts would affect the pedestrian clutter perception. The images were shown in random order to reduce the effects of order or potential bias.

Similarly, each subject was asked to input the ratings in the designated box. 1 stands for the lowest pedestrian clutter, i.e. easiest to detect the pedestrian by naked eyes and 5 stands for the highest pedestrian clutter, i.e. most difficulty to detect the pedestrian by naked eyes. The GUI of Experiment 3.3 is shown in Figure 3.9.

Procedure

The experiment uses naturalistic driving images taken from an in-car camera. The rating experiment was carried out using a graphic user interface written in MATLAB running on a Windows 7 PC with a 19-inch LCD monitor. The actual display size of the image region is 20.1×11.3 cm^2. A series of naturalistic driving images were shown on the monitor. Each subject was asked to rate 100 images with respect to the local target pedestrian area based on their perception and understanding of clutter. Before the test set, a practice set was given to each subject. The purpose of the practice set and the baseline images was not only to ensure each subject would understand the rating process, but also to help subjects build reasonable rating rules with respect to different scenarios on their own. Only the results of the test set were recorded.

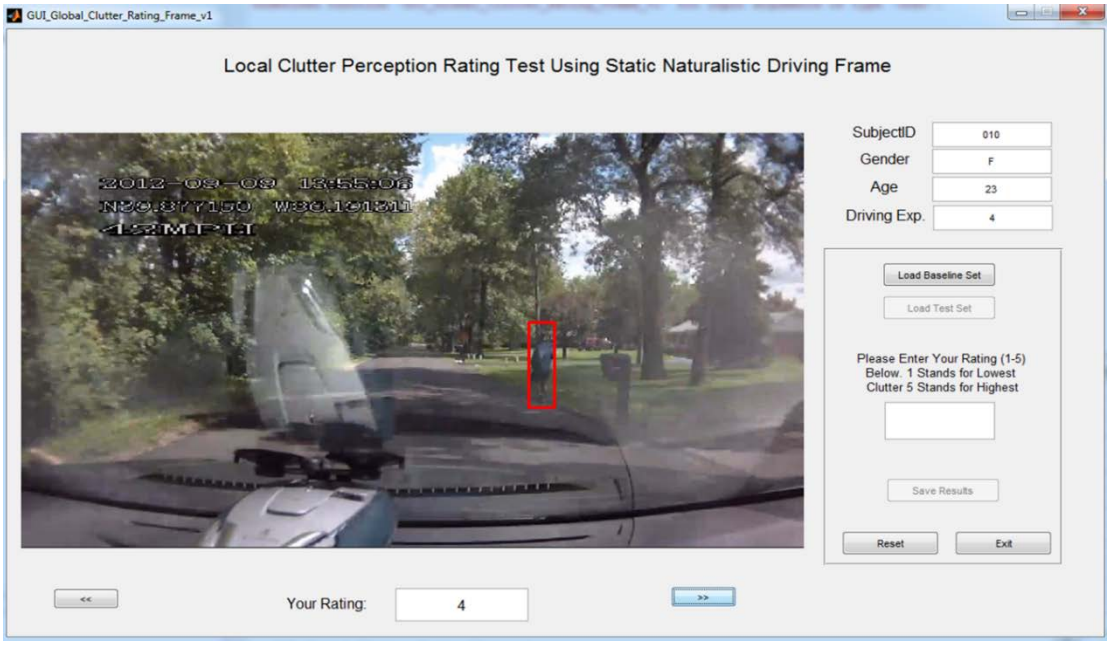


Figure 3.9 The GUI of the LPC rating experiment for naturalistic driving image

Experiment 3.4: Local pedestrian clutter rating for naturalistic driving video

This experiment focuses on exploring how subjects perceive the local pedestrian clutter level of a set of given naturalistic driving video, which is closer related to the true perception of driver than using images. Pedestrian motion and driver pedestrian interactions would be important additional factors to affect the local pedestrian perception difficulty.

Method

Participants

The same group of subjects in Experiment 3.1 participated this study.

Stimuli

The stimuli in this experiment was 15 seconds long naturalistic driving videos which contain one or multiple pedestrians. It was the same experiment set as Experiment 3.2

Design

The pedestrian clutter level was subjectively rated by each subject based on their perception of the pedestrians. One video clip was shown at one time on the monitor screen. A red box was shown around the pedestrian area to indicate the target pedestrian for rating. The red box was only last for 3 seconds and was removed after that without adding artifacts to the clutter rating. The subject can replay the videos as many times as they want to make sure they confirmed the pedestrian to rate. The videos also can be played at multiple frame rates and paused at any time. Similarly, each subject input the subjective rating from 1 to 5 for each video in designated box. 1 stands for the lowest pedestrian clutter, i.e. easiest to detect the pedestrian by naked eyes and 5 stands for the highest pedestrian clutter, i.e. most difficulty to detect the pedestrian by naked eyes. Rated videos can be acessed later and the ratings can be modified if the subject. Again the videos were shown in random order to

reduce the effects of order or potential bias. The GUI of Experiment 3.4 is shown in Figure 3.10.

Procedure

The experiment uses naturalistic driving videos taken from an in-car camera. Each subject was asked to seated in front of a computer monitor. The test videos was shown on the screen one at a time in a random order. The rating experiment was carried out using a graphic user interface written in MATLAB running on a Windows 7 PC with a 19-inch LCD monitor. The actual display size of the image region is 20.1×11.3 cm^2. Each subject was asked to rate the 50 videos with respect to the local target pedestrian area based on their perception and understanding of clutter. Similarly, before the test set, a practice set was given to each subject. Only the results of the test set were recorded.

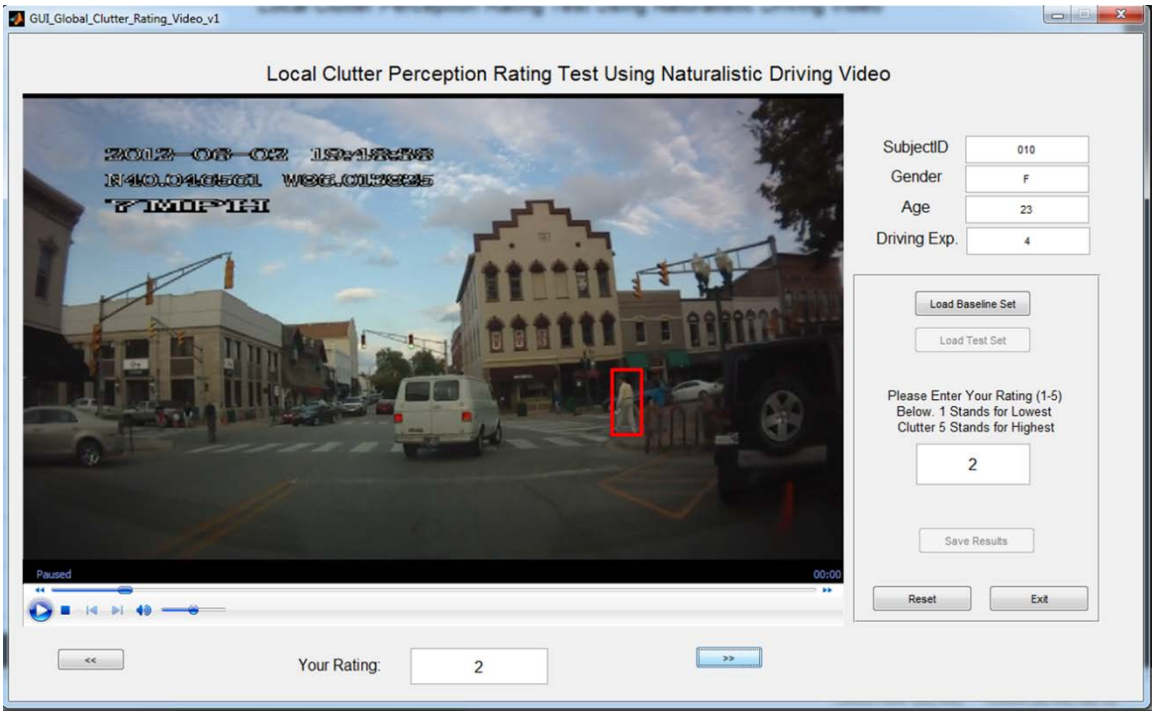


Figure 3.10 The GUI of the LPC rating experiment for naturalistic driving video

Learn the mapping function

To learn the mapping function f in Equation 3.11, the same set of regression and learning methods used in GEC mapping function learning was also tested and compared. The ground truth of the human local clutter perception was collected by Experiment 3.3 and Experiment 3.4. The training set and test set was divided the same as the GEC study. Cross validation was applied to tune the parameters of the SVM and ELM regression.

Results

Similar to GEC study, the training set was used to learn the best mapping function using cross validation. The test set was used to compare different regression methods and models. The human perceived local clutter ground truth of each image/video clip was calculated by taking the median subjective ratings of all participants to remove the effect of outliers. The ground truth was normalized into [0, 1] range as a numerical value instead of a categorical value for classification methods. The rooted mean square error (RMSE) between the predicted value and the ground truth of the validation set (i.e., residuals) were assessed for different regression methods. In addition, a better fit of the regression does not necessarily lead to a better correlation between the LPC score and the human perceived clutter level. Therefore the correlation between the predict LPC value and the subjective ratings were also compared to find the best predictor.

Similarly, for LPC study we also computed the intra-class correlation coefficient (ICC) by averaging the Pearson's correlation between all pairs of subjective ratings. The ICC of all 12 subjects is 0.802, which showed good agreement among all the subject ratings. This also shows that the subjective ratings can be served as a reliable ground truth for human perception of LPC so that the different computational LPC metrics can be meaningfully compared. Table 3.2 shows the results of the mapping function using different regression models. RMSE and R values are listed and compared. We also computed the *p*-value and all of the tests have *p*-value less than 0.01 which means the correlation is statistically significant. All the regression models show good correlation with the human perception of

LPC. The non-linear regression using the power function achieves best fit and correlation results, again showing a power relation between the physical measure and the perceived intensity, which is in accordance to Steven's power law.

The proposed GEC metric was also compared with two other popularly used local contrast metrics mentioned in Chapter 3.3.1. The Pearson's correlations (R) between the image-based computed metrics and the median of subjective ratings of all test data were computed. The comparison result is shown in Figure 3.11. The RSS (-0.48) and Doyle (-0.50) both showed a negative correlations, which means neither of these existing metrics can predict the true human perception of LPC within naturalistic driving scene well. All the tests have *p*-value less than 0.01. The LPC metric correlates well with the true human perception. To be fair, we also compared the LPC using the image stimuli without incorporating the motion channel with RSS and Doyle metrics.

Table 3.2 Results of regression models of LPC metrics

Regression model	RMSE	R	*p*-value
linear	0.18	0.60	0.004
logistic	0.17	0.58	0.005
Non-linear (power)	0.12	0.72	0.004
SVM	0.22	0.55	0.005
ELM	0.18	0.59	0.004

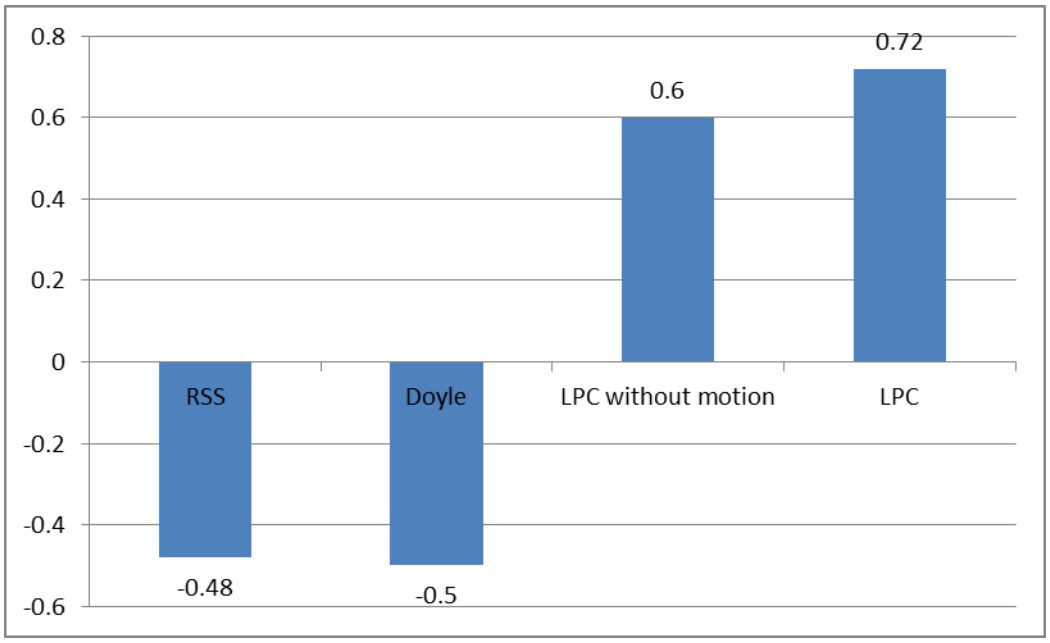


Figure 3.11 Correlation of the proposed LPC compared with existing local contrast methods

3.4 Example experimental results of GEC and LPC metrics

3.4.1 Results on natural images

We first evaluated our global clutter measure and local clutter measure on natural images. The two simple examples shown in Chapter 2 are evaluated to justify our approach. Figure 3.12 shows the clutter measure results of the book and insects image respectively. The colored boxes represent the target area we used. The book image on the left has a much higher clutter score (0.518) than the insect image on the right (0.086), which is in accordance with human perception. On the other hand, the bright yellow book (Book 1) has much lower local clutter score (0.508) than the insect (0.851), suggesting a higher local saliency and less detection difficulty than the green insects, which is also a good reference and reflection of true human perception. The bright yellow book (book 1) on the left image has much lower local clutter score (0.508) than the dark brown one (0.913) (book 2), indicating an easier attention and perception, which is also a quite reasonable reference and reflection of the true human perception.

(a) (b)

Figure 3.12 Experimental results on natural images using the proposed measures. (a) local clutter scores of two books on the same global environment, and (b) local clutter score of insect is high even when the image's global clutter score is low.

Figure 3.13 An example comparison of GEC measure and SE measure

We compared our task independent global environment clutter score with the Subband Entropy (SE) method [3] in Figure 3.13. The much larger difference between the two

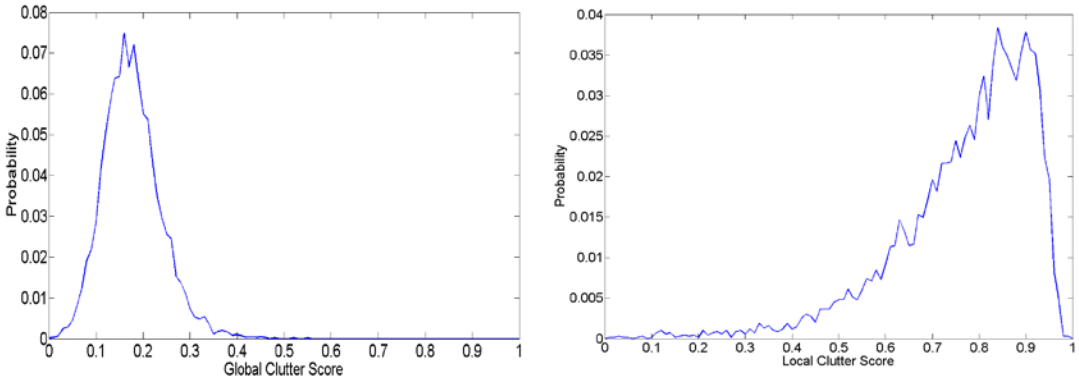
images of the GEC measure than SE measure shows that GEC score are more reasonable than SE score and more consistent with human perception.

3.4.2 Example results on naturalistic driving data

We next show the clutter measure results on naturalistic driving images. Figure 3.14 shows six examples of measured GEC and LPC using the test naturalistic driving data. Image 4 and Image 5 are the same image and have the same global background with GEC score 0.287. The GEC scores provide reasonable reference to the global clutter level although they are not very discriminative while comparing some similar driving scenes. However, the LPC score reflects the difficulty of pedestrian perception quite well compared to the GEC score. The pilot test and study on the naturalistic driving data shows that (1) low contrast image tends to have lower GEC score, such as night image (Image 1 with GEC score 0.116) and image with excessive glares and reflections (Image 2 with GEC score 0.200). (2) Color Saliency is the most important factor that may affect the LPC score, e.g. Image 6 has the lowest LPC score (0.507) due to its highly saturated and discriminative pants color compared to the neighborhood area and (3) LPC could be a better indicator and reference for pedestrian perception difficulty in real naturalistic driving scenarios. For example, even though Image 1 has the lowest GEC score (0.116), it is most difficult to detect the pedestrian in dark due to its high LPC score (0.926). Note that all these scores are currently normalized objective score computed from the image feature maps. More accurate model and evaluation approaches will be learned after the exploratory study analysis and probabilistic learning.

Figure 3.14 Clutter measure results on test naturalistic driving images. Note that Image 4 and Image 5 are the same image but we measured LPC scores for different pedestrians

We also tested the proposed GEC and LPC metric on our large scale naturalistic driving data. 1850 5-second videos containing 3418 pedestrians have been analyzed using the proposed pedestrian locating and clutter measure approach. The 1850 videos are generated and selected from TASI 110 car naturalistic driving dataset with the standard that the pedestrian may have potential conflicts with the vehicle. The global clutter score and local clutter score distribution of all the tested 1850 videos are shown in Figure 3.15.



(a) (b)

Figure 3.15 Results of the 3418 pedestrians from 1850 images in preliminary test (a) GEC score distribution (b) LPC score distribution

3.5 Bottom-up pedestrian perception predictor

With the proposed GEC and LPC metrics which are correlates well with the true clutter perception of naturalistic driving data, we now can present the combined bottom-up pedestrian perception predictor for naturalistic driving scene. Suggested by [105] using a combination of global clutter metric, local contrast metric and target size as a predictor for pedestrian detection performance in night vision system, we proposed our bottom-up pedestrian perception predictor similarly.

The proposed bottom-up pedestrian perception predictor (BUP3) is a combination of the proposed GEC, LPC metrics and the target size metric. The target size is defined as the square root of the pixel number of the target (RPOT) based on the pedestrian contour refinement results in chapter 4.3.2. BUP3 is then expressed as:

$$BUP3 = \frac{(1 - LPC) * \mathrm{RPOT}}{GEC} \tag{3.12}$$

Intuitively, the BUP3 is proportional to $1 - LPC$ and is inverse proportional to GEC, which means the higher the local contrast is and the less complex the global environment is, the easier the pedestrian can be perceived.

A similar metric was proposed in [105] and the human subject test using night vision data proved its effectiveness in predicting the pedestrian detection efficiency. To validate this predictor on naturalistic driving data, we designed another experiment to collect pedestrian perception data.

Experiment 3.5: Pedestrian perception using naturalistic driving video

This experiment aims to simulate the naturalistic driving scenarios by letting subject perceive pedestrians within naturalistic driving videos. A set of naturalistic driving videos with pedestrians sampled from the TASI 110-car naturalistic driving dataset is used to measure the pedestrian perception efficiency of each subject. The collected response time (RT) was used to measure the performance of the proposed bottom-up pedestrian perception predictor.

Method

Participants

The same group of subjects in Experiment 3.1 participated this study.

Stimuli

50 15-second naturalistic driving videos containing only one pedestrian were used as the stimuli. The selected videos are varied in driving scenario, illumination and weather condition. The percentage of each category is in accordance with the distribution of the entire TASI 110-car naturalistic driving dataset. The pedestrian within each selected video may have potential conflict with the vehicle.

Design

The stimuli in this experiment are 15 second long naturalistic driving videos containing only one pedestrian. The 15 second video includes the full interaction between the pedestrian and the vehicle. In another word, a typical potential conflict video include the first appearance of pedestrian, the potential conflict between the vehicle and the pedestrian,

and the disappearance of the pedestrian. Each human subject was asked to response when they first observed and confirmed the pedestrian. The RT between the first appearance of pedestrian and the response was recorded. In this study, the first appearance of pedestrian is defined as the point when the full body of the pedestrian was shown in the naturalistic driving scene. The point was determined through automatic pedestrian detection introduced in Chapter 2 and verified by human annotator.

Procedure

The experiment used 50 15 seconds long naturalistic driving videos containing only one pedestrians taken from an in-car camera. Each subject will be seated in front of a computer monitor. One video was shown on the screen at a time. Again the videos are played in random order to exclude bias. The video can only be played at its taken frame rate and can only be viewed once. Each human subject was asked to hit the spacebar when they observed and confirmed the pedestrian and the RT would be recorded automatically. A confirmation sound indicated that the key press had been recorded. The subjects can take a break if they want after complete a video and continue to the next video by clicking the "Next" button when they are ready. Similarly, before the test set, a practice set would be given to each subject. Only the results of the test set was recorded.

Results

We first compute the ICC of the collected RT among all 12 subjects. The ICC is 0.717 which indicates a good agreement of pedestrian perception using test data among the 12 human subjects. To evaluate the performance of the predictor, we computed the Pearson Linear correlation coefficient denoted by r_p, Spearman's rank correlation coefficient denoted by r_s and Kendall's rank correlation coefficient denoted by r_k between the inverse of the RT and the value of the predictor. We also compared the results with five popularly used bottom-up saliency metrics, including Itti's method [16], Feature congestion (FC) method [34], Difference of Gaussian (DoG) based method [20], Independent Component Analysis (ICA) based method [20] and DCT based method [113] shown in Table 3.3. The

proposed BUP3 metric achieves the best correlation with the true pedestrian perception within naturalistic driving scene with statistical significance (all p-values are less than 0.05). Note here the inverse of the RT is used therefore a higher positive correlation value indicates a better predictor.

Table 3.3 Results of bottom-up metrics for pedestrian perception predictor (correlations between the inverse of RTs and the bottom-up metrics)

	r_p	r_s	r_k	p-value
Itti's [16]	0.401	0.288	0.323	0.01
FC[34]	0.572	0.397	0.411	0.02
DCT[113]	0.466	0.381	0.350	0.01
ICA[20]	0.501	0.393	0.401	0.01
DoG[20]	0.525	0.491	0.588	0.02
The Proposed BUP3	**0.731**	**0.602**	**0.708**	**0.01**

4. PEDESTRIAN PERCEPTION ESTIMATION MODEL

4.1 Overview

In this chapter, the proposed pedestrian perception estimation model will be illustrated in detail. In the proposed model, pedestrian perception is modeled as a combination of pedestrian pre-attention process and pedestrian recognition process. Bayesian probabilistic theory is applied to derive the mathematic form of the pedestrian perception model. A Bayesian probabilistic framework based system will be learned to automatic evaluate the pedestrian clutter score which reflects the pedestrian perception difficulty. The derivation of the Bayesian framework will be first introduced and the corresponding mathematic meaning of each module and implementation details will be presented later.

4.2 Pedestrian Perception Estimator (PPE)

In the proposed pedestrian perception model, the pedestrian perception is modeled as a two-stage pre-attention recognition process. Both bottom-up stimulus-driven information and top-down task-driven knowledge will contribute to the perception result, i.e., the pedestrian clutter score. During the pre-attention stage, a stimulus-driven search model plays the main role and shifts driver's attention to the salient components within the naturalistic driving scene. During the recognition stage, a goal-driven search model takes over and driver's attention was guided by his/her knowledge, experience and assumption of pedestrian appearance, location, etc. The two stage output will be combined by to generate the pedestrian perception results.

To model the combination stage, we follow and extend the Bayesian framework for visual attention in [20]. Pedestrian perception by driver can be modeled by a Bayesian probabilistic framework, and should be determined by both global features and local features. The pedestrian perception estimator (PPE) is formulated by estimating the

likelihood of pedestrian presence given the local feature set L, global feature set *G* and location X. In particular, the probability of pedestrian presence given the local feature set l_t computed from the target area, the global feature set g_I of the entire image and the target locationx_t, $P(O = 1|L = l_t, G = g_I, X = x_t)$ can be calculated using Bayesian rules:

$$P(O = 1|L, G, X) = \frac{P(L, G, X|O = 1)P(O = 1)}{P(L, G, X)} \tag{4.1}$$

For simplicity, the location X and the extracted features L, G are considered to be conditionally independent. Eq.4.1 can be split and derived as:

$$\frac{P(L, G, X|O = 1)P(O = 1)}{P(L, G, X)}$$

$$= \frac{P(L, G|O = 1)P(X|O = 1)P(O = 1)}{P(L, G)P(X)}$$

$$= \frac{1}{P(L, G)} \frac{P(L, G|O = 1)P(X|O = 1)P(O = 1)}{P(X)}$$

$$= \frac{1}{P(L, G)} P(L, G|O = 1)P(O = 1|X) \tag{4.2}$$

The first term of Eq.4.2 can be seen as the self-information if we take log on both sides. It reflects the bottom-up saliency which is determined by the joint probability of local features of the target area and global features of the entire image. Rare probability patterns will have more saliency. The second term is the top-down knowledge containing the target based posterior probability of local features and global features. The global features here can be related to the contextual information proposed by Torralba *et al.*[19]. The third term is the location prior, i.e., the probability of pedestrian presence at a given location which reflects the location expectation and knowledge of the driver.

4.3 The Proposed Pedestrian Perception Estimation Model

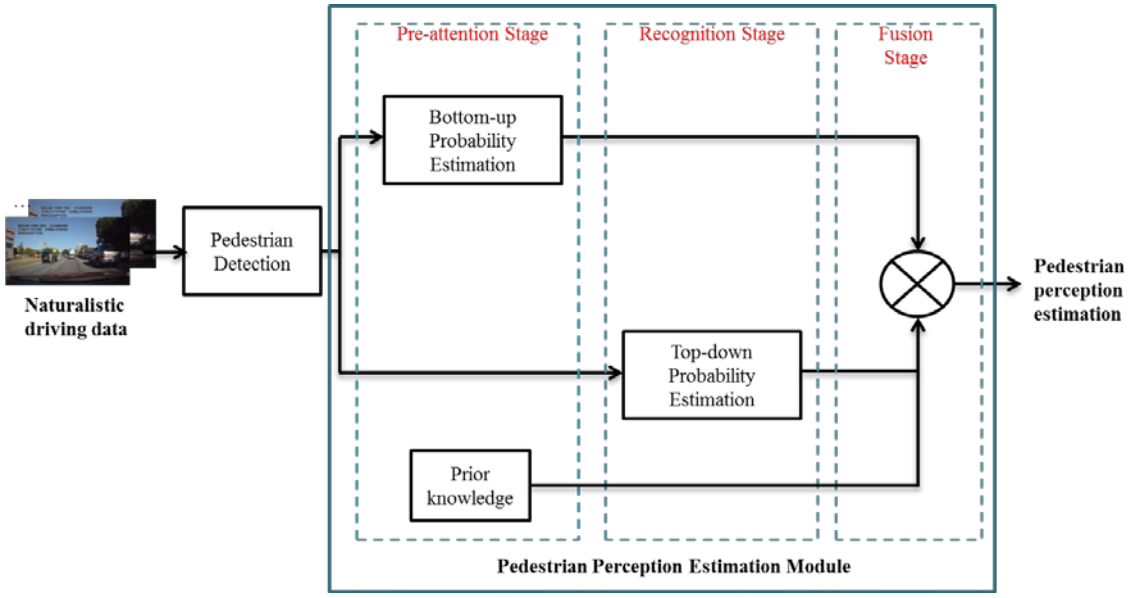


Figure 4.1 Diagram of the proposed pedestrian clutter evaluation system

Based on the hypothesis in chapter 4.2, we propose a pedestrian perception estimator which combines the bottom-up saliency term, top-down knowledge and location prior term. The overall diagram of the proposed pedestrian perception estimation system is shown in Figure 4.1. The pedestrian is firstly located in the naturalistic driving scene automatically using the proposed pedestrian detection method in chapter 2. During the pre-attention stage, the bottom-up information of the entire image is computed based on the proposed BUP3 clutter metric in chapter 3. Remember both the GEC of the entire naturalistic driving scene and the LPC in the local regions were explored in building the BUP3 clutter metric particularly designed for naturalistic driving scene and pedestrians. The location prior is learned from the large scale naturalistic driving data with the exact pedestrian locations provided by the proposed pedestrian detection method. During the recognition stage, the top-down pedestrian knowledge probability is calculated based on the sliding window based pedestrian detection probability. The top-down probability reflects the probability of the appearance based pedestrian features given the fact that the target is a pedestrian. The computation of the bottom-up and top-down probability maps will be introduced later.

During the fusion stage, all the above three terms are combined together to generate the final pedestrian perception estimation.

Generating the pedestrian perception probability map

To estimate the pedestrian perception difficulty within the entire naturalistic driving scene, a pedestrian perception probability map is required to compare the perception probability all over the given the naturalistic driving scene. In the proposed model, the perception probability map can be split into the bottom-up probability map, top-down probability map and location prior map. We now introduce how the three maps are generated respectively.

The bottom-up probability map is based on the proposed BUP3 in chapter 3 aiming at represent the bottom-up saliency of the target. The BUP3 metric was particularly built for pedestrian within naturalistic driving scene and justified by the human subject tests in chapter 3. The bottom-up probability map is generated as follows:

1. Obtain the target size from the pedestrian detection module.
2. Using sliding window and compute the BUP3 of each window, which results in a lattice with the calculated BUP3 scores. The stride of the sliding window is set to be 4 in our experiments.
3. Interpolate the resulted lattice to generate a full map with the same size of the entire naturalistic driving scene
4. Gaussian smooth the generated map and normalize to [0,1] range

Figure 4.2 shows an example of the generated bottom-up probability map. The pedestrians with high saliency are highlighted in the heat map with a high bottom-up probability, as well as the other sitting workers, traffic signs and vehicles.

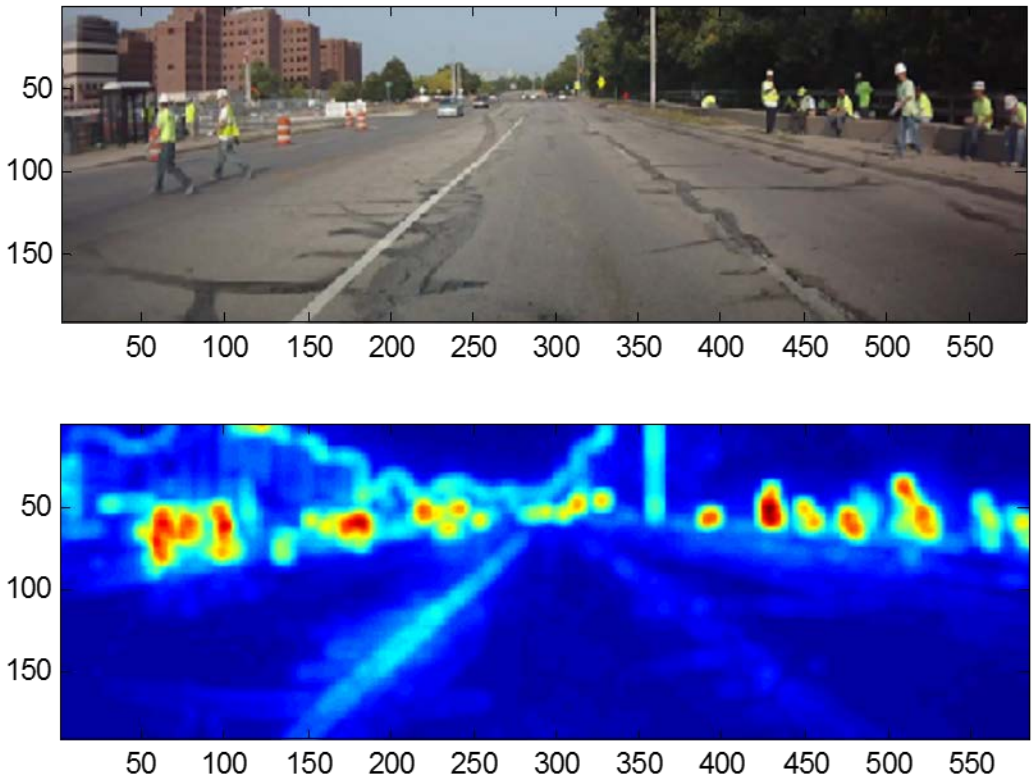


Figure 4.2 An example of bottom-up probability map

The top-down probability is generated based on the probability of the appearance feature surrounding the target given the fact that the target is a pedestrian. The probability can be directly related the pedestrian detection score which reflects the a posterior probability using Bayes' rule. In particular, the top-down probability can be written as

$$P(L,G|O=1)=\frac{P(O=1|L,G)P(L,G)}{P(O=1)}.$$

Assuming the prior term constant for pedestrian appearance probability, the top-down probability is directly proportional to the a posterior probability which can be generated by the pedestrian classifier learned in chapter 2. The entire top-down probability map is generated similar to bottom-up map as follows:

1. Obtain the target size from the pedestrian detection module.

2. Using sliding window and compute the pedestrian appearance probability of each window, which results in a lattice with the calculated pedestrian detection scores. The stride of the sliding window is set to be 4 in our experiments.
3. Interpolate the resulted lattice to generate a full map with the same size of the entire naturalistic driving scene
4. Gaussian smooth the generated map and normalize to [0,1] range

The top-down probability map of the naturalistic driving scene in Figure 4.2 is shown in Figure 4.3, where the region with pedestrian appearance has relatively high top-down probability.

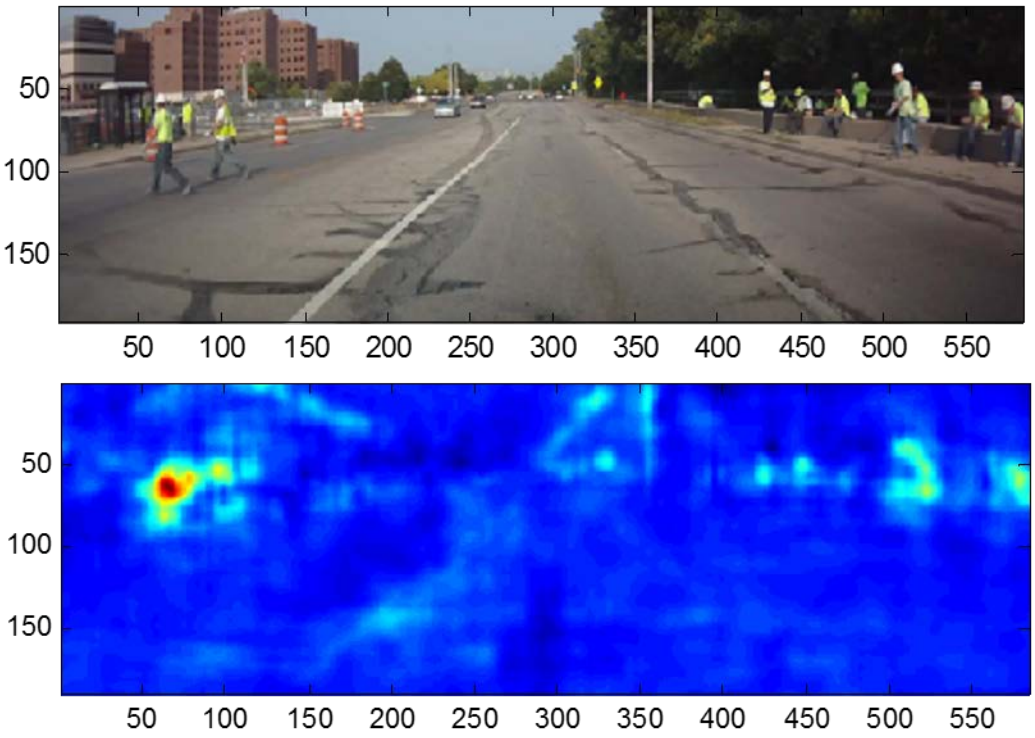


Figure 4.3 An example of top-down probability map

The location prior is learned by accumulating all pedestrian appearance locations within the aligned 1850 naturalistic driving potential conflict videos mentioned in chapter 3.4.2. It acts as a constant factor when finally generating the pedestrian perception probability map.

4.4 Experimental Results

In this section, examples of the experimental results using the proposed pedestrian perception estimator (PPE) are present and compared with other visual clutter/perception measure. The human subject test results of experiment 3.5 is correlated with the proposed pedestrian perception estimator to compare with other visual saliency methods.

The proposed pedestrian perception probability map was generated for all the 50 naturalistic driving scenes used in human subject test experiment 3.5. To compare fairly with existing visual clutter/saliency methods to predict the perception, the location prior was not included in the generation of the proposed perception map in this experiment.

An example of qualitative comparison is shown in Figure 4.4. The proposed PPE is compared to the other five existing visual perception/saliency map. The Itti's [16] map, the FC[34] map and the ICA[20] map were generated using the code provided by the authors while the DCT[113] map and the DoG[20] map were re-implemented based on their paper respectively.

A quantitative comparison was also carried out by correlating to the results of the human subject test in experiment 3.5. The mean values of saliency/perception probability within the target box were correlated with the inverse of the RT and the Pearson Linear correlation coefficient denoted by r_p, Spearman's rank correlation coefficient denoted by r_s and Kendall's rank correlation coefficient denoted by r_k were computed to evaluate the performance of the pedestrian perception of all the methods. The proposed PPE achieved the best correlation with the true visual perception of pedestrian within naturalistic driving scene. With the incorporation of top-down information, the proposed PPE outperformed all other bottom-up metrics, including the BUP3 proposed in chapter 3. Note here the inverse of the RT is used therefore a higher positive correlation value indicates a better predictor. The proposed PPE can be used as a reasonable predictor of the pedestrian perception in naturalistic driving scenes.

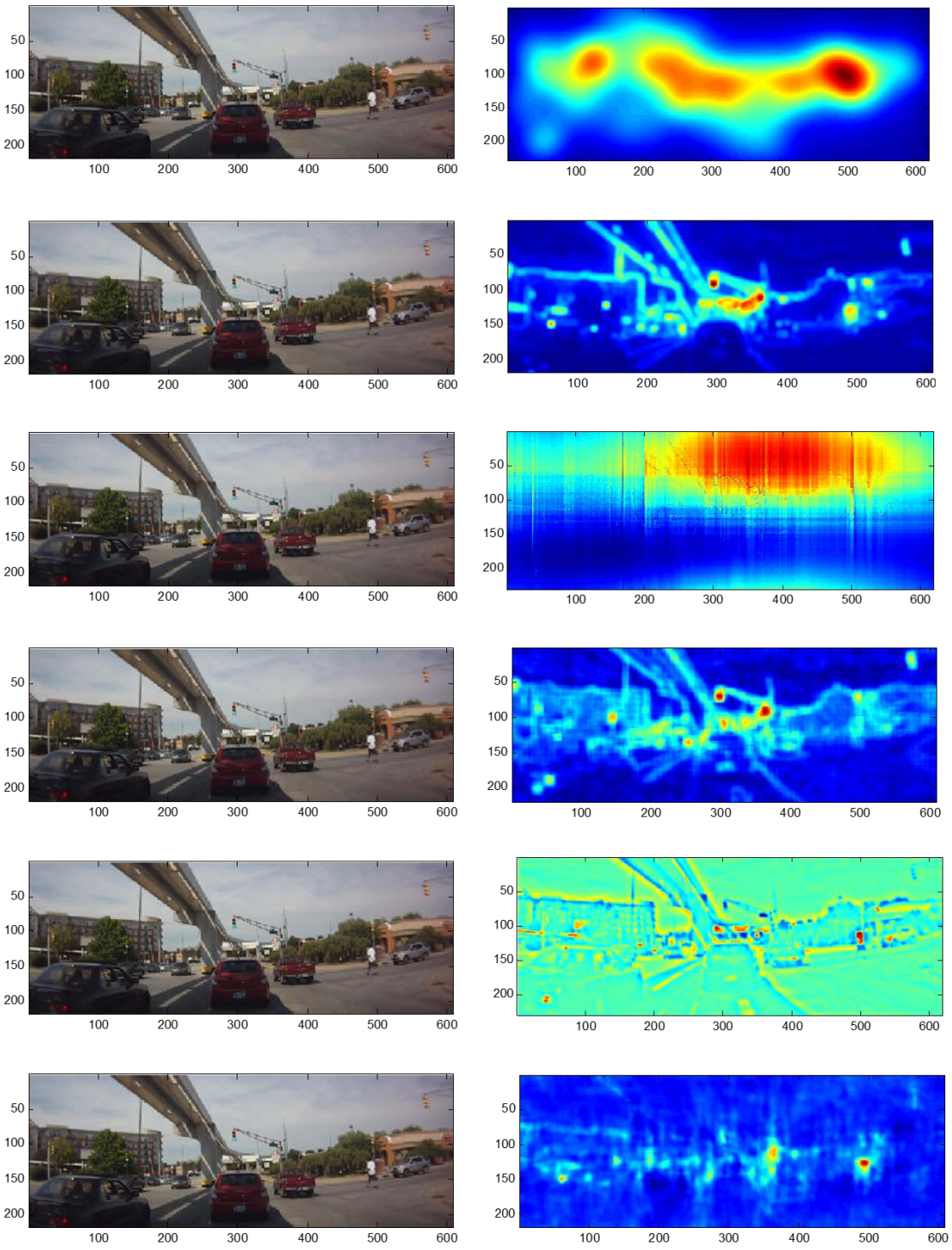


Figure 4.4 An example of qualitative comparison of saliency/perception maps. From top to bottom: Itti's [16], FC[34], DCT[113], ICA[20], DoG[20] and the proposed PPE

Table 4.1 Results of pedestrian perception predictor (correlations between the inverse of RTs and the estimated probability/saliency)

	r_p	r_s	r_k	p-value
Itti's [16]	0.401	0.288	0.323	0.01
FC[34]	0.572	0.397	0.411	0.02
DCT[113]	0.466	0.381	0.350	0.01
ICA[20]	0.501	0.393	0.401	0.01
DoG[20]	0.525	0.491	0.588	0.02
BUP3	**0.731**	**0.602**	**0.708**	**0.01**
The proposed PPE	**0.773**	**0.633**	**0.755**	**0.01**

5. CONCLUSIONS

5.1 Conclusions

In this thesis, we proposed a pedestrian perception evaluation model which can automatically and quantitatively evaluate the pedestrian clutter and analyze the pedestrian perception difficulty using naturalistic driving data. We designed the categorization-based multi-stage automatic pedestrian detection system to locate the pedestrians in large scale naturalistic driving data instead of manual labeling. Visual clutter analysis was used to study the factors that may affect the driver's ability to perceive pedestrian appearance. We designed two quantitative measures: global environment clutter (GEC) score to capture the complexity of the driving environment in terms of visual search; and local pedestrian clutter (LPC) score to evaluate the search efficiency of the pedestrian in the given driving environment. The candidate features were studied by the designed exploratory study using naturalistic driving data. The results of the exploratory study were served as the ground truth of pedestrian perception and a Bayesian probabilistic model which can quantitatively compute the pedestrian perception difficulty was proposed.

Recognition of pedestrians during driving is a complex cognitive activity. Some of the pedestrian crashes are due to driver's late or difficult perception of pedestrian's appearance. Visual clutter analysis is used to study the factors that may affect the driver's ability to perceive pedestrian appearance. This could enable us more insight into the human visual perception process by providing evidence from real-life tasks. Moreover, the results could provide road safety practitioners valuable information about road component and pedestrian safety features design. An automatic pedestrian perception valuation system could further be incorporated into pedestrian active safety systems to ide more robustness and reliability.

5.2 Publications Resulting from This Work

Journal Publications

1. **K. Yang**, E. J. Delp and E. Y. Du, “A Pedestrian Perception Evaluation Model for the Advanced Driver Assistant System”, IEEE Transaction on Intelligent Transportation System (submitted).

2. **K. Yang**, E. J. Delp and E. Y. Du, “Bicyclist Detection in Large Scale Naturalistic Driving Video Comparing Feature Engineering and Feature Learning”, IEEE Transaction on Intelligent Transportation System (submitted).

3. **K. Yang**, E. J. Delp and E. Y. Du, "Categorization-based two-stage pedestrian detection system for naturalistic driving data", *Signal, Image and Video Processing* 8, no. 1,pp.135-144, 2014.

Conference Publications

4. L. Dong, S. Chien, **K. Yang**, Y. Chen, D. Good, R. Sherony and H. Takahashi, Determination of Pedestrian Mannequin Clothing Color for the Evaluation of Image Recognition Performance of Pedestrian Pre-Collision Systems, ESV 2015 (accepted)

5. **K. Yang**, Liu, C., Zheng, J. Y., Christopher, L., & Chen, Y. (2014, October). Bicyclist detection in large scale naturalistic driving video. In Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on (pp. 1638-1643). IEEE.

6. R. Tian, L. Li, **K. Yang**, Y. Chen and R. Sherony, Estimation of the Vehicle-Pedestrian Encountering Risk in the Road Based on TASI 110-Car Naturalistic Driving Data Collection, 2014 IEEE Intelligent Vehicles Symposium (IV'14), Dearborn, Michigan, USA, 2014.

7. Tian, R., Li, L., **K. Yang**, Chien, S., Chen, Y., & Sherony, R. (2014, June). Estimation of the vehicle-pedestrian encounter/conflict risk on the road based on TASI 110-car naturalistic driving data collection. In Intelligent Vehicles Symposium Proceedings, 2014 IEEE (pp. 623-629). IEEE.

8. **K. Yang**, E. Y. Du, E. J. Delp, P. Jiang, Y. Chen, R. Sherony, and H. Takahashi, “A New Approach of Visual Clutter Analysis for Pedestrian Detection”, IEEE International Conference on Intelligent Transportation Systems (IEEE ITSC), 2012

9. **K. Yang**, E.Y. Du, P. Jiang, Y. Chen, R. Sherony and H. Takahashi, “In-depth Analysis of HOG based Pedestrian Detection using Naturalistic Driving Data”, FAST-zero'13.,2013

10. E.Y. Du, **K. Yang**, P. Jiang, Y. Chen, R. Sherony and H. Takahashi, “Pedestrian Behavior Analysis Using Naturalistic Driving Data in USA”, the 23rd ESV Conference, Seoul, Korea.,2013

11. **K. Yang**, E. Y. Du, E. J. Delp, P. Jiang, F. Jiang, Y. Chen, R. Sherony, and H. Takahashi, "An Extreme Learning Machine-based Pedestrian Detection Method", IEEE Symposium on Intelligent Vehicle(IEEE IV), 2013.

12. **K. Yang**, E. Y. Du, P. Jiang, Y. Chen, R. Sherony, and H. Takahashi, “Automatic Categorization-based Multi-stage Pedestrian Detection,” IEEE International Conference on Intelligent Transportation Systems (IEEE ITSC), 2012

LIST OF REFERENCES

LIST OF REFERENCES

[1] U. S. D. o. T. National Highway Traffic Safety Administration, "TRAFFIC SAFETY FACTS 2010 A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System," *DOT HS 811 659*, 2010.

[2] E. Commission, "Road Safety Vademecum--Road safety trends, statistics and challenges in the EU 2011-2012," 2011.

[3] R. Rosenholtz, Y. Li, and L. Nakano, "Measuring visual clutter," *Journal of vision*, vol. 7, pp. 17-17, 2007.

[4] J. M. Wolfe, "Guided search 4.0," *Integrated models of cognitive systems*, pp. 99-120, 2006.

[5] J. Edquist, "The effects of visual clutter on driving performance. Thesis Monash University, Department of Psychology," 2009.

[6] M. W. Eysenck and M. T. Keane, *Cognitive psychology: A student's handbook*: Taylor & Francis, 2005.

[7] A. D. Milner and M. A. Goodale, "The visual brain in action," *New York: Oxford*, 1995.

[8] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, pp. 97-136, 1980.

[9] M. I. Posner, C. R. Snyder, and B. J. Davidson, "Attention and the detection of signals," *Journal of experimental psychology: General*, vol. 109, pp. 160, 1980.

[10] C. W. Eriksen and J. E. Hoffman, "The extent of processing of noise elements during selective encoding from visual displays," *Perception & psychophysics*, vol. 14, pp. 155-160, 1973.

[11] C. W. Eriksen and J. D. S. James, "Visual attention within and around the field of focal attention: A zoom lens model," *Perception & psychophysics*, vol. 40, pp. 225-240, 1986.

[12] J. M. Wolfe, "Guided search 2.0 A revised model of visual search," *Psychonomic bulletin & review*, vol. 1, pp. 202-238, 1994.

[13] J. M. Wolfe, "Visual search," *Attention*, vol. 1, pp. 13-73, 1998.

[14] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience*, vol. 18, pp. 193-222, 1995.

[15] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Matters of Intelligence*, ed: Springer, pp. 115-141, 1987.

[16] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 20, pp. 1254-1259, 1998.

[17] J. K. Tsotsos, S. M. Culhane, W. Y. Kei Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial intelligence,* vol. 78, pp. 507-545, 1995.

[18] K. R. Cave, "The FeatureGate model of visual selection," *Psychological research,* vol. 62, pp. 182-194, 1999.

[19] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search," *Psychological review,* vol. 113, p. 766, 2006.

[20] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *Journal of vision,* vol. 8, pp. 32-32, 2008.

[21] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, "SUN: Top-down saliency using natural statistics," *Visual Cognition,* vol. 17, pp. 979-1003, 2009.

[22] R. P. Rao, "Bayesian inference and attentional modulation in the visual cortex," *Neuroreport,* vol. 16, pp. 1843-1848, 2005.

[23] J. A. Yu and P. Dayan, "Inference, attention, and decision in a Bayesian neural architecture," *Advances in neural information processing systems,* pp. 1577-1584, 2004.

[24] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: A Bayesian inference theory of attention," *Vision research,* vol. 50, pp. 2233-2247, 2010.

[25] J. M. Wolfe, A. Oliva, T. S. Horowitz, S. J. Butcher, and A. Bompas, "Segmentation of objects from backgrounds in visual search tasks," *Vision research,* vol. 42, pp. 2985-3004, 2002.

[26] M. J. Bravo and H. Farid, "Search for a category target in clutter," *-PERCEPTION-LONDON-,* vol. 33, pp. 643-652, 2004.

[27] J. Duncan and G. W. Humphreys, "Visual search and stimulus similarity," *Psychological review,* vol. 96, pp. 433-458, 1989.

[28] M. R. Beck, M. C. Lohrenz, and J. Gregory Trafton, "Measuring search efficiency in complex visual search tasks: Global and local clutter," *Journal of experimental psychology,* vol. 16, p. 238, 2010.

[29] M. B. Neider and G. J. Zelinsky, "Exploring set size effects in scenes: Identifying the objects of search," *Visual Cognition,* vol. 16, pp. 1-10, 2008.

[30] G. Deco and D. Heinke, "Attention and spatial resolution: A theoretical and experimental study of visual search in hierarchical patterns," *Perception,* vol. 36, p. 335, 2007.

[31] L. Reddy and R. VanRullen, "Spacing affects some but not all visual searches: Implications for theories of attention and crowding," *Journal of Vision,* vol. 7, pp.3-3, 2007.

[32] L. I. Voicu, M. Uddin, H. R. Myler, A. Gallagher, and J. Schuler, "Clutter modeling in infrared images using genetic programming," *Optical Engineering,* vol. 39, pp. 2359-2371, 2000.

[33] A. Oliva, M. L. Mack, M. Shrestha, and A. Peeper, "Identifying the perceptual dimensions of visual complexity of scenes," *Proceedings of the 26th Annual Meeting of the Cogn. Sci. Soc*, pp. 101-106, 2004, Chicago, Illinois,USA.

[34] R. Rosenholtz, Y. Li, J. Mansfield, and Z. Jin, "Feature congestion: a measure of display clutter," *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 761-770, 2005, Portland, Oregon, USA.

[35] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," *Proceedings of International Conference on Image Processing*, pp. 444-447, 1995, Washington, DC, USA.

[36] M. C. Lohrenz, J. G. Trafton, M. R. Beck, and M. L. Gendron, "A model of clutter for complex, multivariate geospatial displays," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 51, pp. 90-101, 2009.

[37] S. G. Klauer, T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey, "The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data," National Highway Traffic Safety Administration, 2006.

[38] S. Jenkins, "Consideration of the effects of background on sign conspicuity." *Australian Road Research Board Conference Proc*, vol. 11. 1982.

[39] G. Ho, C. T. Scialfa, J. K. Caird, and T. Graw, "Visual search for traffic signs: The effects of clutter, luminance, and aging," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 43, pp. 194-207, 2001.

[40] J. Edquist, "The effects of visual clutter on driving performance," Monash University Accident Research Centre, 2009.

[41] E.Y. Du, K. Yang, P. Jiang, Y. Chen, R. Sherony, and H. Takahashi, "Pedestrian Behavior Analysis Using Naturalistic Driving Data in USA," *23rd International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, no. 13-029, Seoul, Korea, 2013.

[42] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.*, vol. 1, pp. 886-893, 2005, San Diego, CA, USA.

[43] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," *Computer Vision–ECCV 2006*, pp. 428-441. Springer Berlin Heidelberg, 2006.

[44] K. Yang, E. Y. Du, J. Pingge, C. Yaobin, R. Sherony, and H. Takahashi, "Automatic categorization-based multi-stage pedestrian detection," *15th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 451-456, 2012, Anchorage, Alaska, USA.

[45] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006*, pp. 1491-1498, 2006, New York City, NY, USA.

[46] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *International journal of computer vision*, vol. 73, pp. 41-59, 2007.

[47] S. Nedevschi, S. Bota, and C. Tomiuc, "Stereo-based pedestrian detection for collision-avoidance applications," *IEEE Transactions on Intelligent Transportation Systems,* vol. 10, pp. 380-391, 2009.

[48] J. Ge, Y. Luo, and G. Tei, "Real-time pedestrian detection and tracking at nighttime for driver-assistance systems," *IEEE Transactions on Intelligent Transportation Systems,* vol. 10, pp. 283-298, 2009.

[49] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *IEEE Transactions on Intelligent Transportation Systems,* vol. 6, pp. 63-71, 2005.

[50] D. M. Gavrila, "Sensor-based pedestrian protection," *Intelligent Systems, IEEE,* vol. 16, pp. 77-81, 2001.

[51] S. Milch and M. Behrens, "Pedestrian detection with radar and computer vision," 2001.

[52] C. Premebida, O. Ludwig, and U. Nunes, "Exploiting lidar-based features on pedestrian detection in urban scenarios," *12th International IEEE Conference on Intelligent Transportation Systems, 2009. ITSC'09.*, pp. 1-6, 2009, St. Louis, MI, USA.

[53] S. Sato, M. Hashimoto, M. Takita, K. Takagi, and T. Ogawa, "Multilayer lidar-based pedestrian tracking in urban environments," *Intelligent Vehicles Symposium (IV), 2010 IEEE,* pp. 849-854, 2010, San Diego, CA, USA.

[54] A. Broggi, P. Cerri, S. Ghidoni, P. Grisleri, and H. G. Jung, "A new approach to urban pedestrian detection for automatic braking," *IEEE Transactions on Intelligent Transportation Systems,* vol. 10, pp. 594-605, 2009.

[55] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Transactions on Intelligent Transportation Systems,* vol. 8, pp. 413-430, 2007.

[56] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, pp. 1239-1258, 2010.

[57] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 31, pp. 2179-2195, 2009.

[58] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 743-761, 2012.

[59] C. Gu, J. J. Lim, P. Arbeláez, and J. Malik, "Recognition using regions," *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009.*, pp. 1030-1037, 2009, Miami, FL, USA.

[60] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005.*, pp. 878-885, 2005, San Diego, CA, USA.

[61] Z. Lin and L. S. Davis, "Shape-based human detection and segmentation via hierarchical part-template matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, pp. 604-618, 2010.

[62] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *International journal of computer vision,* vol. 38, pp. 15-33, 2000.

[63] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07.*, pp. 1-8, 2007, Minneapolis, MN, USA.

[64] B. Wu and R. Nevatia, "Simultaneous object detection and segmentation by boosting local shape feature based classifier," *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07.*, pp. 1-8, 2007, Minneapolis, MN, USA.

[65] D. Gerónimo, A. Sappa, A. López, and D. Ponsa, "Adaptive image sampling and windows classification for on-board pedestrian detection," *Proceedings of the International Conference on Computer Vision Systems, Bielefeld, Germany,* vol. 39, 2007.

[66] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," *IEEE 12th International Conference on Computer Vision, 2009*, pp. 32-39, 2009, Kyoto, Japan.

[67] S. U. Hussain and W. Triggs, "Feature sets and dimensionality reduction for visual object detection," *BMVC 2010-British Machine Vision Conference*, pp. 112-1. BMVA Press, 2010, Aberystwyth, UK.

[68] S. J. Krotosky and M. M. Trivedi, "On color-, infrared-, and multimodal-stereo approaches to pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems,* vol. 8, pp. 619-629, 2007.

[69] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1030-1037, 2010, San Francisco, CA, USA.

[70] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on riemannian manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, pp. 1713-1727, 2008.

[71] T. Watanabe, S. Ito, and K. Yokoi, "Co-occurrence histograms of oriented gradients for pedestrian detection," *Advances in Image and Video Technology*, pp. 37-47, 2009.

[72] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *International journal of computer vision,* vol. 63, pp. 153-161, 2005.

[73] C. Wojek and B. Schiele, "A performance evaluation of single and multi-feature people detection," *Pattern Recognition*, pp. 82-91, 2008.

[74] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *British machine vision conference*, pp. 1-11, 2009, London, UK.

[75] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," *BMVC 2010, Aberystwyth, UK*, 2010.

[76] P. Dollár, R. Appel, and W. Kienzle, "Crosstalk Cascades for Frame-Rate Pedestrian Detection," in *ECCV*, Florence, Italy, 2012.

[77] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2903-2910, 2012, Providence, RI, USA.

[78] Y. Xu, D. Xu, S. Lin, T. X. Han, X. Cao, and X. Li, "Detection of Sudden Pedestrian Crossings for Driving Assistance Systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 42, pp. 729-739, 2012.

[79] M. Enzweiler and D. M. Gavrila, "A multilevel mixture-of-experts framework for pedestrian classification," *IEEE Transactions on Image Processing,* vol. 20, pp. 2967-2979, 2011.

[80] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, pp. 349-361, 2001.

[81] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, pp. 1627-1645, 2010.

[82] Z. Lin and L. Davis, "A pose-invariant descriptor for human detection and segmentation," *Computer Vision–ECCV 2008,* pp. 423-436, 2008, Marseille, France.

[83] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, pp. 1-8, 2008, Anchorage, AL, USA.

[84] Y. Freund and R. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," *Computational learning theory,* pp. 23-37, 1995.

[85] G. Overett, L. Petersson, N. Brewer, L. Andersson, and N. Pettersson, "A new pedestrian dataset for supervised learning," *Intelligent Vehicles Symposium, 2008 IEEE*, pp. 373-378, 2008, Eindhoven, Netherlands .

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

[87] D. Gavrila, J. Giebel, and S. Munder, "Vision-based pedestrian detection: The protector system," *Intelligent Vehicles Symposium, 2004 IEEE*, pp. 13-18, 2004, Parma, Italy.

[88] L. Zhao and C. E. Thorpe, "Stereo-and neural network-based pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems,* vol. 1, pp. 148-154, 2000.

[89] D. M. Gavrila and J. Giebel, "Shape-based pedestrian detection and tracking," *Intelligent Vehicle Symposium, 2002. IEEE,* pp. 8-14, 2002, Versailles, France.

[90] P. Geismann and G. Schneider, "A two-staged approach to vision-based pedestrian recognition using Haar and HOG features," *Intelligent Vehicles Symposium, 2008 IEEE*, pp. 554-559, 2008, Eindhoven, Netherlands.

[91] L. Oliveira, U. Nunes, and P. Peixoto, "On exploration of classifier ensemble synergism in pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems,* vol. 11, pp. 16-27, 2010.

[92] G.-B. Huang, L. Chen, and C.-K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Transactions on Neural Networks,* vol. 17, pp. 879-892, 2006.

[93] G.-B. Huang, X. Ding, and H. Zhou, "Optimization method based extreme learning machine for classification," *Neurocomputing,* vol. 74, pp. 155-163, 2010.

[94] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 42, pp. 513-529, 2012.

[95] G. Shafer, "A mathematical theory of evidence", vol. 76: Princeton university press Princeton, 1976.

[96] F. C. Crow, "Summed-area tables for texture mapping," *Computer Graphics,* vol. 18, pp. 207-212, 1984.

[97] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision,* vol. 88, pp. 303-338, 2010.

[98] H. Cho, P. E. Rybski, and W. Zhang, "Vision-based bicyclist detection and tracking for intelligent vehicles," *Intelligent Vehicles Symposium (IV), 2010 IEEE*, pp. 454-461, 2010, San Diego, CA, USA.

[99] H. Cho, P. E. Rybski, and W. Zhang, "Vision-based bicycle detection and tracking using a deformable part model and an EKF algorithm," *2010 13th International IEEE Conference on, Intelligent Transportation Systems (ITSC),* pp. 1875-1880, 2010, Funchal, Madeira Island, Portugal.

[100] B. Bhanu, "Automatic Target Recognition: State of the Art Survey," *IEEE Transactions on Aerospace and Electronic Systems,* vol. AES-22, pp. 364-379, 1986.

[101] C.-P. Yu, D. Samaras, and G. J. Zelinsky, "Modeling visual clutter perception using proto-object segmentation," *Journal of Vision,* vol. 14, p. 4, 2014.

[102] D. E. Schmieder and M. R. Weathersby, "Detection Performance in Clutter with Variable Resolution," *IEEE Transactions on Aerospace and Electronic Systems,* vol. AES-19, pp. 622-630, 1983.

[103] G. Aviram and S. R. Rotman, "Evaluation of human detection performance of targets embedded in natural and enhanced infrared images using image metrics," *Optical Engineering,* vol. 39, pp. 885-896, 2000.

[104] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 187-220, 1972.

[105] B. Luzheng, O. Tsimhoni, and L. Yili, "Using Image-Based Metrics to Model Pedestrian Detection Performance With Night-Vision Systems," *IEEE Transactions on Intelligent Transportation Systems,* vol. 10, pp. 155-164, 2009.

[106] S. S. Stevens, "On the psychophysical law," *Psychological review,* vol. 64, p. 153, 1957.

[107] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters,* vol. 9, pp. 293-300, 1999.

[108] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychological methods,* vol. 1, p. 30, 1996.

[109] G. Tidhar and S. R. Rotman, "New method of target acquisition in the presence of clutter," in *Orlando'91,* pp. 188-199, *Orlando, FL,* 1991.

[110] A. C. Copeland, M. M. Trivedi, and J. R. McManamey, "Evaluation of image metrics for target discrimination using psychophysical experiments," *Optical Engineering,* vol. 35, pp. 1714-1722, 1996.

[111] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision,* vol. 1, pp. 321-331, 1988.

[112] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, pp. 881-892, 2002.

[113] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 194-201, 2012.

[114] D. Sun, R. Stefan and M. J. Black. "Secrets of optical flow estimation and their principles." In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2432-2439. IEEE, 2010.

VITA

VITA

Kai Yang received his B.S degree in Computer Science from Beijing University of Posts and Telecommunications, Beijing, China in 2009, and the M.S degree in Electrical and Computer Engineering from Purdue University. He is a Ph.D. candidate and currently working toward the Ph.D. degree in Electrical and Computer Engineering in Purdue University, West Lafayette, Indiana, USA. He was working as a Graduate Research Assistant in the School of Electrical and Computer Engineering in Purdue University from 2009 to 2015. His research interests include image and video processing, computer vision, intelligent transportation system, pattern recognition and biometrics.

PUBLICATIONS

PUBLICATIONS

Book Chapters

1. **K. Yang** and Y. Du, Trend of biometrics, a chapter in Biometrics: from Fiction to Practice, Pan Stanford Publishing Pte. Ltd.

Journal Publications

2. **K. Yang**, E. J. Delp and E. Y. Du, “A Pedestrian Perception Evaluation Model for the Advanced Driver Assistant System”, IEEE Transaction on Intelligent Transportation System (submitted).
3. **K. Yang**, E. J. Delp and E. Y. Du, “Bicyclist Detection in Large Scale Naturalistic Driving Video Comparing Feature Engineering and Feature Learning”, IEEE Transaction on Intelligent Transportation System (submitted).
4. **K. Yang**, E. J. Delp and E. Y. Du, "Categorization-based two-stage pedestrian detection system for naturalistic driving data", Signal, Image and Video Processing, 2014.
5. Ruijie Huang, Mingyun Li, Meiping Ye, **Kai Yang**, Xin Xu, and Richard Gregory, "Effect of nicotine on Streptococcus gordonii growth, biofilm formation and cell aggregation", Applied and Environmental Microbiology, 2014.
6. **K. Yang**, E. Y. Du, and Z. Zhou, “Consent Biometrics,” Neurocomputing, vol. 100, 153-162, 2013.
7. **K. Yang**, and E. Y. Du, “Speed-up Multi-stage Non-Cooperative Iris Recognition,” International Journal of Biometrics, 4(4), 406-421, 2012.
8. E. Y. Du, **K. Yang**, and Z. Zhou, “Key Incorporation Scheme for Cancelable Biometrics,” Journal of Information Security, 2(4), 185-194, 2011.
9. **K. Yang**, and E.Y. Du**,** “Review of Recent Patents on Cancelable Biometrics,” Recent Patents on Electrical Engineering, 4(2), 125-132, 2011.

International Conference Proceedings

10. Libo Dong, Stanley Chien, **Kai Yang**, Yaobin Chen, David Good, Rini Sherony and Hiroyuki Takahashi, Determination of Pedestrian Mannequin Clothing Color for the Evaluation of Image Recognition Performance of Pedestrian Pre-Collision Systems, ESV 2015 (accepted)
11. **Yang, K**., Liu, C., Zheng, J. Y., Christopher, L., and Chen, Y. Bicyclist detection in large scale naturalistic driving video. IEEE International Conference on Intelligent Transportation Systems (ITSC), 2014

12. R. Tian, L. Li, **K. Yang**, Y. Chen and R. Sherony, Estimation of the Vehicle-Pedestrian Encountering Risk in the Road Based on TASI 110-Car Naturalistic Driving Data Collection, 2014 IEEE Intelligent Vehicles Symposium (IV'14), Dearborn, Michigan, USA, 2014.
13. Tian, R., Li, L., **Yang, K.**, Chien, S., Chen, Y., & Sherony, R. (2014, June). Estimation of the vehicle-pedestrian encounter/conflict risk on the road based on TASI 110-car naturalistic driving data collection. In Intelligent Vehicles Symposium Proceedings, 2014 IEEE (pp. 623-629). IEEE.
14. **K. Yang,** E. Y. Du, E. J. Delp, P. Jiang, Y. Chen, R. Sherony, and H. Takahashi, "A New Approach of Visual Clutter Analysis for Pedestrian Detection", IEEE International Conference on Intelligent Transportation Systems (IEEE ITSC), 2013
15. **K. Yang**, E.Y. Du, P. Jiang, Y. Chen, R. Sherony and H. Takahashi, "In-depth Analysis of HOG based Pedestrian Detection using Naturalistic Driving Data", FAST-zero'13.,2013
16. E.Y. Du, **K. Yang**, P. Jiang, Y. Chen, R. Sherony and H. Takahashi, "Pedestrian Behavior Analysis Using Naturalistic Driving Data in USA", the 23rd ESV Conference, Seoul, Korea.,2013
17. **K. Yang**, E. Y. Du, E. J. Delp, P. Jiang, F. Jiang, Y. Chen, R. Sherony, and H. Takahashi, "An Extreme Learning Machine-based Pedestrian Detection Method", IEEE Symposium on Intelligent Vehicle(IEEE IV), 2013.
18. **K. Yang**, E. Y. Du, P. Jiang, Y. Chen, R. Sherony, and H. Takahashi, "Automatic Categorization-based Multi-stage Pedestrian Detection," IEEE International Conference on Intelligent Transportation Systems (IEEE ITSC), 2012
19. **K. Yang**, and E. Y. Du, "A Multi-stage Approach for Non-cooperative Iris Recognition," IEEE International Conference on Systems, Man, and Cybernetics, 2011.
20. **K. Yang** and E. Y. Du, "Consent Biometrics," IEEE Workshop on Computational Intelligence in Biometrics: Theory, Algorithms, and Applications, pp. 78-83, 2011.
21. **K. Yang**, E. Y. Du, and Z. Zhou, "A New Approach for Willingness Test in Biometric Systems," SPIE Defense, Security, and Sensing, 2011.
22. **K. Yang**, Y. Du, Z. Zhou, C. Belcher, "Gabor Descriptor Based Cancelable Iris Recognition Method", Proceedings of IEEE International Conference on Image Processing (ICIP), pp. 4085-4088, 2010.
23. **K. Yang**, Y. Du, Y. Sui, X. Zou, Z. Zhou, "A new approach for cancelable iris recognition", SPIE Symposium on Defense, Security + Sensing, vol.7708, pp. 77080A, 2010.
24. Y. Sui, **K. Yang**, Y. Du, S. Orr, X. Zou, "A novel key management scheme using biometrics", SPIE Symposium on Defense, Security + Sensing, vol.7708, pp. 77080C, 2010.