Pedestrian Detection and Identification

RAN FEI

Supervised by DR. T. P. STATHAKI

Department of Electrical and Electronic Engineering

Imperial College London

This dissertation is submitted for the degree of

Doctor of Philosophy

2014

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation contains less than 1,000,000 words including footnotes, tables, equations, references and appendices and has less than 150 figures.

RAN FEI
Supervised by DR. T. P. STATHAKI
2014

To be fond of learning is to be near to knowledge.
To practice with vigor is to be near to magnanimity.
To possess the feeling of shame is to be near to courage.

—— ***«Doctrine of the Mean»***

好學近乎智力
行近乎仁知恥近
乎勇　書贈
匯文學校　蔡元培

Acknowlegements

Worldwide, less than 1% adults experienced or are experiencing the life in purchasing a degree of *Doctor of Philosophy*. Thanks for the financial support from my family, I am fortunate enough to be one of them.

I would like to express my gratitude to Dr. T. Stathaki, my supervisor, for her guidance, useful critiques of the research work and her administrative support; to Mrs A. M. Hough, for her advices and assistance in keeping my progress on schedule. I would also like to acknowledge Miss Diao, for her inspiration for the applications of computer vision in material research and production and a generous donation of source images. Furthermore, my grateful appreciation would extend to Ms. Yelland-Sutcliff and others for their tolerance to my ignorance and their encouragement during the struggling period.

Finally, a modest thank to myself.

Abstract

People are the centre of technologies. Understanding, monitoring and tracking the behaviour of people will benefit in various areas including driving assistance, surveillance for safety and caring purposes and applications for machine-people interaction. Particularly, pedestrians attract more attention for two reasons: they restrict the behaviours of people to standing and moving upright; and applications for pedestrian detection and monitoring have positively impact on the quality of life. Pedestrian detection and identification, aims at recognising pedestrians from still images and video frames. Together with pedestrian recognition and tracking, this topic attempts to train computers to recognise a pedestrian.

The problem is challenging. Though frameworks were designed, various algorithms were proposed in recent years, further efforts are needed to improve the accuracy and reliability of the performance. In this thesis, proposing a modifiable framework for pedestrian identification and improving the performances of current pedestrian detection techniques are particularly focused. Based on appearance based pedestrian identification, a modifiable framework is a novel philosophy of developing frameworks which can be easily improved. For pedestrian identification, a novel protocol where layers of algorithms are hierarchically applied to solve the problem. To compare the detected pedestrians, appearance based features are selected, the "bag-of-features" framework is employed to compare the histogram descriptions of pedestrians. To improve the performances of HOG pedestrian detector, the presence of head-shoulder structure is selected as the evidence of the presence of pedestrian. A novel appearance based framework is developed to detect the head-shoulder structure from the detection results of HOG detector. Furthermore, in order to separate multiple pedestrians detected in one bounding box, a novel algorithm is proposed to detect the approximated symmetry axes of pedestrians.

Abbreviations

Abbr.	Phrases
AdaBoost	Adaptive Boosting
API	Application Programming Interface
BG	Background
BoF	Bag Of Features
CBIR	Content Based Image Retrieval
CCTV	Closed-Circuit Television
CoHOG	Co-occurrence Histogram of Oriented Gradients
CSS	Colour Self-Similarity
CV	Open Source Computer Vision Library (OpenCV)
Daimler-DB	Daimler Benz
DET	Detection Error Trade-off
ETH	Eidgenössische Technische Hochschule
FG	Foreground
FOV	Field of View
FPPI	False Positives per Image
FPPW	False Positives per Window
INRIA	Institut National de Recherche en Informatique et en Automatique (National Institute for Research in Computer Science and Control)
IR	Infrared
HOG	Histogram of Oriented Gradients
HSL	Hue Saturation Lightness
HSV	Hue Saturation Value
LPB	Local Binary Patterns
MIT	Massachusetts Institute of Technology
MSCR	Maximally Stable Colour Regions
MSER	Maximally Stable Extremal Regions
OpenCV	Open Source Computer Vision Library
PDF	Probability Density Function
RADAR	Radio Detection and Ranging
RGB	Red Green Blue
ROC	Receiver Operating Characteristic
ROI	Region of Interest
TUD	Technische Universität Darmstadt
SIFT	Scale-Invariant Feature Transform
S-M	Symmetry Measurement
SURF	Speeded Up Robust Features
SVM	Support Vector Machine
V-J	Viola Jones

Contents

List of Figures

List of Tables

Chapter 1

Computer Vision is Incomplete without Pedestrians

Χωρίς εικόνες, η σκέψη είναι αδύνατη.[1]
—— Αριστοτέλης

People as a target object in computer vision has been attracting more and more attentions. Detecting, tracking and identifying people is benefited in various research fields and applications including driving assistance, patients and disabled monitoring, crime and migration surveillance, robotics and machine interactions etcetera [1–3]. To detect and recognise a person is difficult. As a species, people have intra-class variances, biologically or visually; as an individual, they may appear in crowds, may be occluded by other people and objects; as a non-rigid object, the same person may be visually different due to changed viewing angle, poses, clothes and accessories. People detection focuses on people as a species; people tracking and identification focuses on people as an individual [4, 5]. Image files applicable in people detection and identification can be either still images or frames of video clips. People tracking is always applied to frames of video clips. Techniques used in these areas are usually overlapped: identifying the same people detected from every frame in video clips results in people tracking; motion captured by people tracking algorithms would benefit the detection and identification of people [6, 7]. The majority people or human used in the state of art literature are particularly refer to pedestrians, which restricts the target objects to upright people who are standing, walking or running without carrying complex accessories intervening the appearance of human. This simplification not just reduces the complexity of computation,

[1] Thought is impossible without an image. —— *Aristotle*

it specifies its application area. The majority of datasets for the state of art training and testing were collected from streets, traffic areas where human and auto-mobiles interact [1, 8–11]. In this thesis, without particular indication, people or human refer to pedestrians.

Research in this area started decades ago. Studies of learning human motion can date back to the period before digital video systems and the internet were globally commercialised. Frameworks were designed to teach computers to see a walking person [12–15] or to recognise the gestures of a person [16–18]. Early research was restricted by the computation system. Pedestrians were usually simply modelled using connected cylinders and sticks representing the topology of body parts. The symmetry of the human body was employed to separate humans from background. During this period, the study of human detection and motion concerned television imaging devises (CCTV[2]) mainly. Early attempts inspired later research though results of most of the early frameworks were not satisfied. With the development of imaging / videoing system, further research covered wider applications using sensors including conventional camera, wide FOV[3] camera, near-IR[4], thermal-IR, RADAR[5], laser scanner, etcetera [19]. Developments in computational techniques boosted sophisticated frameworks, non-linear categorization algorithms and rich descriptions of pedestrians. Well developed frameworks and algorithms together with the growing datasets (in both quantity and variations) contributed to the state of art people detection and identification. Accelerated processing speed realised the on-line and real-time applications. More and more real services appeared in market especially in driving-aid systems [1, 3, 8–11, 20, 21].

However, even algorithms with commonly recognised high robustness and accuracy tested through various databases may fail in certain pedestrian detection and identification occurrences. In this thesis, two problems are mainly addressed: pedestrian re-identification[6] after a period of total occlusion and the reduction of false alarms and inaccurate pedestrian detections introduced by the Histogram of Oriented Gradients (HOG) pedestrian detector which was originally proposed in [22]

[2] CCTV: Closed-Circuit Television

[3] FOV: Field of View

[4] IR: Infrared

[5] RADAR: Radio Detection and Ranging

[6] Identification is an ambiguous word with different meanings in different registries. In computer science, human identification referred to the process of identifying a human as individual by their measurable characteristics or traits (biometric is an example of human identification). In this thesis the measurable characteristics particularly refer to visually perceptive features used in pedestrian identification.

and trained by CV (2.4.3) [7] (version 2.4.3). The contributions of the thesis lie in following aspects:

- A novel protocol for pedestrian re-identification containing layers of algorithms is proposed: the appearances of pedestrians are quantified according to the complexity of their descriptions and the identification process (parameter settings) can be adjusted according to the complexity of the appearance of the prototype pedestrian. By learning the appearance as well as the complexity of the prototype pedestrian who has been detected in the video / camera system, a bag-of-appearance-feature is constructed for the prototype pedestrian, which is further adapted to reduce repeated false alarms in pedestrian detection in video streams. The novelties of the protocol are:
    - It spends different processing efforts on identifying pedestrians with different appearances.
    - The strategies of pedestrian identification can easily be adapted by adding new algorithms into the protocol.
- To reduce the false alarms introduced by HOG pedestrian detector, the presence of head-shoulder is selected as the evidences of correct detections. Viola-Jones frontal face and upper-body detectors are combined with a novel appearance based head-shoulder detection. The combined strategy reduces the false alarms without much affecting the detection rate. This strategy can also be applied to the reduction of false alarms introduced by other pedestrian detectors which select shape as the discriminative feature of pedestrians.
- Inaccurate detection results of HOG pedestrian detector are discussed: limitations of HOG pedestrian detector caused by parameter settings of the training procedure are analysed; cases that more than one pedestrians recognised as one are particularly focused; a novel framework for separating pedestrians detected in one bounding box is proposed.

The thesis is structured in three parts: as an introductory, this chapter and the next one overview the pedestrian detection and identification problem and review the state of art literatures. Followed by three technical chapters, the main body of the thesis: the first one, (*Chapter 3*), based on appearance related pedestrian description, a novel protocol is presented to identify the reappeared pedestrian in video

[7] Open Source Computer Vision Library

stream after a period of absolute occlusion, which could retain the tracking of the prototype pedestrian. After that, in (*Chapter 4*), algorithms are promoted to reduce the false alarm rate of the pedestrian detection applied to video streams. The appearances of the head-shoulder structures of pedestrians are used as the evidence for the presence of pedestrians. Viola-Jones body-parts detection algorithms are discussed and a novel means of appearance based head-shoulder detection technique is introduced. In the third chapter, (*Chapter 5*), issues related to the HOG-SVM[8] detector trained by CV (2.4.3) are discussed. Particularly, cases that multiple pedestrians are recognised as one are considered and a novel means of symmetry axes detection is applied to separate individuals bounded in one box. Finally, future works in the development of pedestrian detection strategies and the expansion of vision related applications in other subjects will be summarised. A thought on the outlook of object recognition in language of computer vision and machine intelligence will be presented as the end of the thesis (*Chapter 6* and *7*).

[8] SVM: Support Vector Machine.

Chapter 2

The State of the Art

> Όρος εστίν, ο τινός εστι περας· Σξήμά εστι το υπό τινος ή τινων όρων περιεξόμενον.[1]
>
> ——Ευκλείδης

Literatures related to pedestrian detection and recognition will be reviewed in two levels: the *Strategic Level* concerns the philosophy of solving the detection / identification related problems and the *Algorithmic Level* provides computational approaches to model and compare target objects to be detected / identified. Strategies for detecting and identifying pedestrians are further categorised into supervision based strategies and logic based strategies. The former trains detectors to separate pedestrians from non-pedestrians and identify the prototype pedestrian from other pedestrians; the latter defines logics and rules to detect / identify pedestrians. Rules may include the topology of detected body parts / interest points. In *Algorithmic Level*, emphasis on different purposes of applications, detection or identification, algorithms are reviewed in two categories: cognitive shape based algorithm (*Section 2.2.1*) to distinguish pedestrians from background objects and texture / colour based algorithm to identify the individuals. OpenCV (version 2.4.3) [2] is employed for the majority of experiments in this thesis, a short introduction of the tool kit is included in *Appendix A*.

For disambiguation purposes, terms used in this thesis are declared. The prototype pedestrians have different meanings in pedestrian detection and identification: used in pedestrian detection, the prototype pedestrian is related to the model of pedestrians and in pedestrian identification, the prototype pedestrian mean the

[1] A boundary is that which is an extremity of something. A figure is that which is contained by any boundary or boundaries. —— *Euclid*

[2] In later paragraph, OpenCV version 2.4.3 will be abbreviated as CV(2.4.3)

individual who has been detected in the system. Identification in this thesis is related to the problem of the recognition of pedestrians as individuals (using their visual characteristics) and re-identification focuses on the verification of pedestrians appearing in different segments of videos or in different camera systems. The previously detected pedestrians are recognised as the prototypes[3]. *Fig. 2.1* illustrates the relationship of algorithms in the strategic and algorithmic levels in the basic structure of commonly applied pedestrian detection and identification frameworks: the process starts from positive & negative (not always required) examples in the Image Level and outputs the decision of detection and identification; the entire procedure is demonstrated in but not restricted to three levels of algorithms. The level of strategies concerns the means of detection / identification and the level of algorithms supports the the level strategies; in each level, results from the previous levels are processed in selected function block(s) which generate inputs for the next level of processing.

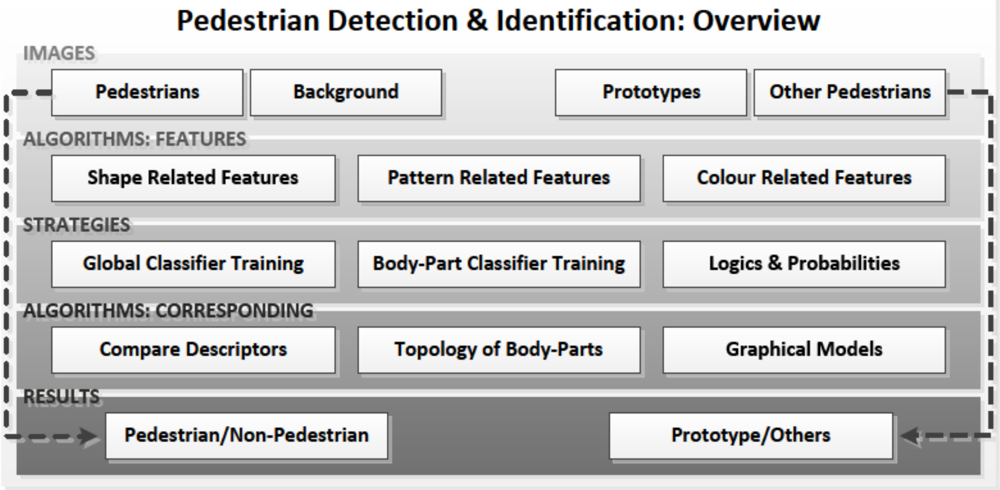


Fig. 2.1 **A General Structure of Pedestrian Detection and Identification: In the image level and the result level, the left hand side represents the Pedestrian Detection and the right hand side represents the Pedestrian Identification. While in other levels, the blocks of algorithms can be applied to either problem, though some blocks may be preferred in one algorithm than another. For example, in the algorithms level, shape related features are more suitable in Pedestrian Detection than in Pedestrian Identification.**

According to the categories in *Fig. 2.1*: in frameworks of Pedestrian Detection, algorithms used in the levels of "Algorithms: Feature", "Strategies", and "Algorithms: Corresponding" are: "Shape Related Features", "Global Classifier Training" and "Compare Descriptors" respectively; in frameworks of Pedestrian Identification, "Colour

[3] The computer is required to re-identify the present of the prototype who has been identified by human in video / camera systems

Related Features", "Logics & Probabilities" and "Compare Descriptors" are used accordingly.

2.1 Strategic Level

The strategic level of algorithms answers the questions:

1. How to obtain the template of the prototype pedestrians?

2. How to correspond the regions of interest (ROI) areas of input images with the template?

The first question concerns the training of pedestrian detector / identifier and the logics existed in the cognition of pedestrians; the second question aims at analysing the input images and corresponding the ROIs of input images with the templates obtained from the the solution to the first question. Strategies used in pedestrian detection and identification are usually overlapped as both problems can be treated as a binary classification problem: pedestrian detection classifies objects into pedestrians and non-pedestrians, pedestrian identification classifies detected pedestrians into prototype and other pedestrians. As shown in *Fig. 2.2*, algorithms in strategic level are reviewed in discriminative and generative two groups. Discriminative learning (*Section 2.1.1*) focuses on training the classifier which response differently to varied input and generative learning (*Section 2.1.2*) model pedestrians according to the observations. Discriminative training strategies are employed in HOG and Viola-Jones pedestrian detection frameworks. In HOG, linear / non-linear SVM training are normally used to discriminate pedestrians from non-pedestrians. And the training strategy introduced by Viola-Jones[23] can be treated as a tree based classification where stages of weak classifiers are cascaded to achieve accurate classification of the prototype pedestrian and other pedestrians.

2.1.1 Training a Global Detector / Identifier (Discriminative)

A global detector / identifier recognises pedestrian as an entire object. The binary classification problem for pedestrian detection is to separate pedestrians from background objects. For pedestrian identification it is to distinguish the prototype pedestrian from other pedestrians. Mathematically, the observations $\mathbf{x} \in \mathbb{R}^N$ are mapped to the decision space $y \in \{0, 1\}$ or $\{-1, +1\}$ through a classifier $y = H(\mathbf{x}, \phi)$, where ϕ represents the set of parameters used in the classifier $H(\mathbf{x}, \phi)$. During training, the

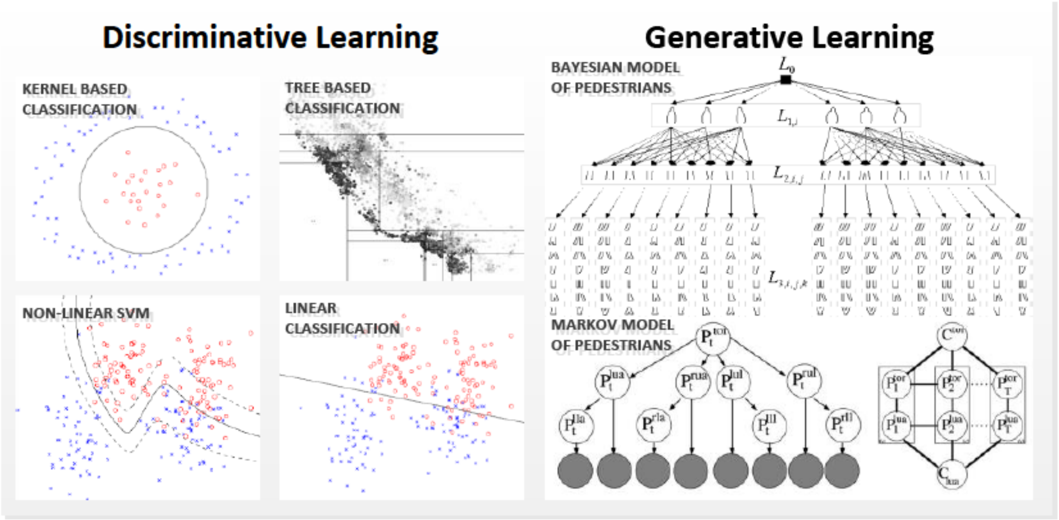

Discriminative Learning
KERNEL BASED CLASSIFICATION
TREE BASED CLASSIFICATION
NON-LINEAR SVM
LINEAR CLASSIFICATION
Generative Learning
BAYESIAN MODEL OF PEDESTRIANS
MARKOV MODEL OF PEDESTRIANS

Fig. 2.2 **Examples of the *Discriminative* and *Generative* models used in the detection and classification[4]. Images one the left-hand side demonstrate several means of binary classification of input observations; in Bayesian model of pedestrians, body-parts in lower levels are dependent on the body-parts in upper levels and in Markov model of pedestrians, image on the right shows the dependencies of the position of body-parts and image on the right show the node map used in judging the present of pedestrians.**

parameters used in the classifier are optimised to maximise the margin between the two classes. Commonly used detectors include Boosting Classification (especially AdaBoost), Support Vector Machine (SVM) and their variations [24, 25]. SVM returns the hyperplane separating the observations and Boosting algorithms return the combination of weak classifiers [26]. Ada (Adaptive) in AdaBoost indicates that the weights applied to weak classifiers are adjusted to their influences on classification results [27].

Weak binary classifiers are required when AdaBoost is selected to train a classifier [27]. Developed by Viola and Jones, Haar-like local features calculated on integral images are one of the apparent choices for weak classifiers as they output binary values in fast calculation [23, 30]. Demonstrated in *Section 4.2.1*, *Appendix A*, the performances of the AdaBoost classifier trained using Haar-like local descriptors in pedestrian detection may be less stable especially when the sizes of dominant Haar-like features are relatively small. Training low level features into binary classifiers using strategies like SVM is another way to obtain weak classifiers. [31] calculated the

[4] Images referenced from [24]:linear classification, non-linear SVM and kernel based classification, [28]:tree based classification; [29]:Bayesian Model, [6]:Markov Model (P represents the position, C represents the appearance; the superscripts indicate the body-parts).

gradient of an image as the low level features which capture the shape features of the object within the region of interest (ROI). The combination of gradients (shapelet) within sliding windows of ROI are trained as the weak classifiers to build the output pedestrian detector. Local HOG descriptors trained using SVM were cascaded in [32] as weak classifiers. The weak classifier training could be complicated. Multiple features can be combined in training the local weak classifiers. For instance, [10] combines Haar-like features, shapelet, HOG and shape context in selections of weak classifiers. In [8], integrated features from different channels of images were selected in the training of Boosting classifier. Integral features refer the different types of features, such as histogram, shape signatures, etc.. The pre-processed images for calculating those features are recognised as channels of images, for instance, shape signatures are calculated on the edge map channel, histogram is calculated on the original image channel, etc. The introduction of channels of images add another dimension of the weak classifiers. As a result, the influences of selected features calculated in one channel will not affect the influences of features calculated in other channels in the trained classifier.

The output classifier of AdaBoost captures both local and global features, which depend on the size and position of the involved weak classifiers. In most literatures, applying AdaBoost training, weak classifiers are calculated in blocks sized from several square of pixels to the size of the entire ROI. AdaBoost trained classifier has a tree-like structure, of which the branches contain information on the position, size and the way to calculate the local feature related weak classifiers. Boosting algorithms and Haar-like features will be further reviewed in *Chapter 4* where the Haar-like feature based body-part detectors are used in false alarm reduction of HOG pedestrian detector.

Support Vector Machine (SVM) is an extended discriminative model of linear classification, which outputs linear combinations of characteristics of inputs $\mathbf{x} \in \mathbb{R}^d$, d is the dimension of the input space:

$$y_i = f(\mathbf{x}_i) = \sum_i \mathbf{w} \cdot \mathbf{x}_i + b, \qquad i = 1, 2, \ldots, N \tag{2.1}$$

$\mathbf{w} \in \mathbb{R}^d$ is the weight applied to $\mathbf{x}$ to classifying the inputs $(\mathbf{x}_i, y_i), y_i = \{+1, -1\}$. b[5] is the offset variable which means the distance between $\mathbf{w} \cdot \mathbf{x}_i$ and the desired labels,

[5] In certain circumstances, b is treated as error or variances. In classification of pedestrian and other objects, b is less meaningful as the contrast positions of two classes are more focused than the absolute ones.

$y_i = \{+1, -1\}$. During supervision, the nearest distance (the margin) between elements in each class $\frac{1}{\|\mathbf{w}\|}$ is maximised. If there exists a hyperplane between the two classes of $(\mathbf{x}_i, y_i)$, the data is separable. In the linear SVM introduced by Vapnik in [26], the hyperplane is located where:

$$\min_{\mathbf{w},b} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \tag{2.2}$$

In reality, the condition is usually relaxed to a non-separable data case of which classes of observations (training datasets) would overlap in decision space as described in *Section 4.1.2*. Furthermore, kernels transforming the inputs to other spaces are applied to obtaining better classification results [24, 25]:

$$\Phi(\mathbf{x}_i) = k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j \tag{2.3}$$

Global descriptors used in SVM training are usually concatenated from local features of one or more types. Single feature used in pedestrian detection ranges from parameters of Haar wavelet transform [33] to HOG descriptors calculated from ROIs [22]. Literatures demonstrated that more than one type of features trained in SVM show promising results: [34] combined HOG with LBP (Local Binary Pattern), [35] combined HOG with CSS (Colour Self-Similarity features)[6] and in [36] 12 different low level features including moment, contrast, colour entropy etc. are involved in calculating the descriptors of pedestrians. Concatenating different features increase the dimension of descriptors at the same time increase the complexity of training an SVM classifier. [36] concluded that SVM may fail in training adequate amount of samples when the descriptors are too long[7]. Dimension reduction algorithms may be required when complex descriptors are desired.

The procedure of pedestrian detection is simple after a global pedestrian detector has been trained. Regions of interest are obtained by sliding a window through input images. The trained global detectors perform as a filter which should positively response to the descriptors of sliding windows where pedestrians present. To achieve scaling invariant, descriptors are calculated from different scaled input images. The performances of trained detector depend on both selected features and the choice of datasets. When more than one type of features are used in describing pedestrians, less information is provided on the significances of features involved in the training.

[6] A low level feature based on the local colour / texture similarity. See *Section 2.2.1*.

[7] Results in [36] show that SVM fail in training more than 2000 samples when the descriptor is longer than 17,000 bit. The length of HOG descriptor using default settings of CV(2.4.3) is 1152 bit per ROI.

Once trained the classifier can hardly be modified.

2.1.2 Defining the Logics of Detection / Identification (Generative)

Logics based pedestrian detection and identification normally apply local classifiers to verify the local structures of perspective pedestrians. The topology of local classifiers are learned using graphical models. Decisions made by local classifiers perform as the nodes of graphical models. The logics among local classifiers and the detection / identification results are modelled using joint probability among the nodes. Two popular strategies employing logic based detection / identification are:

1. Applying several local classifiers to the input image. Output maps of decisions made by local detectors. Explore the map of decisions to detect the best matched topology of local classifiers as described in [37];

2. Using sliding windows to obtain ROIs. Within each window apply the local classifiers and verify the topology of the detected local structures. Then output the final decision on whether a pedestrian is detected / the prototype pedestrian is reappeared.

Graphical models are applied to modelling the topology of the local classifiers. Nodes of local classifiers are normally chosen to catch features of body parts (head, torso and legs for example) [6, 29]. Indirected graphical models like Markov Networks are applied to modelling the topology of local classifiers when the nodes of models have no obvious causal relationship: the appearance of a head may not indicate the appearance of the torso below while the present of both would indicate the presence of a pedestrian. Examples of Markov indirect map applied to pedestrian detection include: [38], which chose segments of edge-map as nodes and [6], which chose appearance based body-parts as nodes. Directed graph would be used to model the conditional dependence between the nodes within the model. In [29, 39]: head-torso-leg parts are modelled from general to specific. The general models of body parts are searched through the images and the suspicious areas are matched with branched specific models which are dependent on the detection of its general version. Bayesian rules are applied to nodes in different levels of models of body parts.

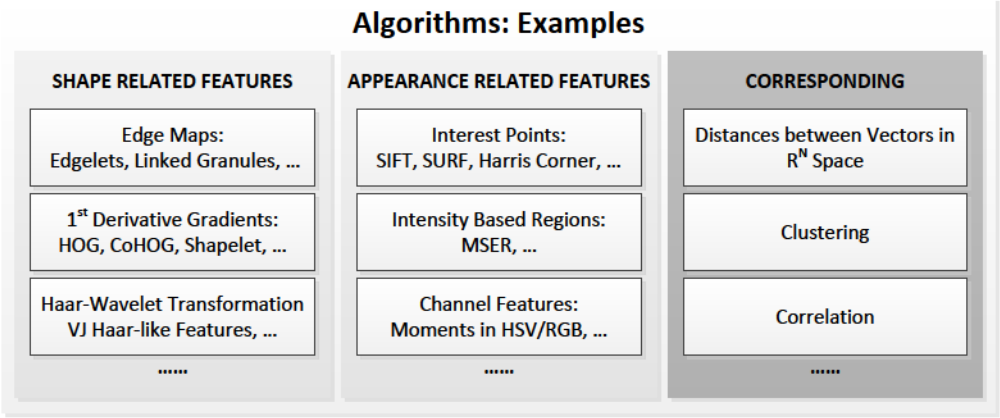
Logics can also be applied to layers of judgements made by different global pedestrian classifiers which are the nodes of graphical models. The conditional dependence / independence between the nodes are modelled using Bayesian / Markov method [40]. Using logics and rules to describe a pedestrian from using predefined features is not an easy task as it is still not clear how we human recognise objects visually. In the current stage of research, rules for recognising a pedestrian are related to the context of image, the detection of symmetrical structures, the topology of body parts and the motion of shape when pedestrian detection is applied to video streams. Here, the context of image refers to either the context between frames or the context between pedestrian and other objects in one frame. Using the context of image as an application relative fact aims at filtering the low suspicious regions which are less probable to contain pedestrians. For instance, in the driving aid system, the shooting angle of images is parallel to the horizon. As a result, the up-left / up-right corner areas of a scene would be less interested. While these areas would be considered important for CCTV monitoring applications when the camera is equipped in the ceiling.

In topics related to pedestrian detection, the discriminative model is popular in the separation of pedestrians and non-pedestrians due to the nature of binary classifiers and the generative model is more widely applied to recognising the actions of pedestrians and the tracking of pedestrians. In object recognition / image understanding, generative models, such as the ”bag of words” [41] and constellation models [42], are widely employed. Compared to the constellation models, which require the topology of parts of interest, the "bag of words" is simple in computation as it only requires the joint probability of parts of interest. When the topology of parts of interest is simple, the constellation models introduce redundant computation procedures. For pedestrian identification in this thesis, when parts of interest are restricted to upper-body, lower-body and background, the framework inspired by the "bag of words" is proposed in *Chapter 3*.

2.2 Algorithmic Level

Algorithmic level relates to the calculation procedure of the strategic level: how to capture features of different types, how to calculate the descriptors of features and how to compare features when relevant. In pedestrian detection, shape is normally modelled to distinguish pedestrians from background objects. Appearance related features introduced by clothes and skin of pedestrians are usually modelled using colour, texture based descriptions. In this thesis: HOG descriptions are calculated on 1^{st} derivative gradients images; Viola-Jones features are calculated on integral images; the histograms of colour images are selected as the appearance based features; and the correlation distances are employed to compare the histogram based descriptions. *Section 2.2.1* reviews the shape / rough shape related feature descriptions and

Section 2.2.2 discusses the appearance based features.



Algorithms: Examples

SHAPE RELATED FEATURES	APPEARANCE RELATED FEATURES	CORRESPONDING
Edge Maps: Edgelets, Linked Granules, ...	Interest Points: SIFT, SURF, Harris Corner, ...	Distances between Vectors in R^N Space
1^{st} Derivative Gradients: HOG, CoHOG, Shapelet, ...	Intensity Based Regions: MSER, ...	Clustering
Haar-Wavelet Transformation VJ Haar-like Features, ...	Channel Features: Moments in HSV/RGB, ...	Correlation
......	......	......

Fig. 2.3 **Relevant algorithms for feature detection, description and correspondence.**

2.2.1 Distinguish Pedestrians from Non-Pedestrians (Detection)

In computer vision, the shape of an object is defined as the area occupied by the object in images, which is a reflection of the space occupied by that object in reality [43]. Shape is an essential feature in vision related tasks. Evidences show that in human vision system the shape information of objects is preliminary processed before other sophisticated cognitive analysis in brain [43, 44]. The shapes of pedestrians are difficult to model. People are non-rigid objects and may carry / wear accessories. Instead of calculating a concrete expression, efforts were paid to achieve the robust descriptions for various occurrences of pedestrians. Rough shape is used in later paragraph referring to the cognitive meaning of the robust shape descriptions. Generally, there are two ways of shape / rough shape modelling applied in pedestrian detection to work with the strategies reviewed in previous sections:

1. Investigating descriptions capturing the cognitive shape feature at the same time tolerating the local variations of boundaries / edges. This is designed especially for applications using classifiers trained by discriminative learning strategy. The shape features captured by HOG descriptions belong to this category;

2. Enumerating variations of the boundaries of pedestrians. This is commonly served in the generatively trained classifiers.

Edge-map is usually selected as an obvious evidence of shape. Edge-map of an input scene is a binary image catching the sharp intensity change which would be cognitively recognised as the boundaries of objects and sketches of patterns / textures. In most frameworks, edge maps are generated using *sobel* [44] or *canny* [45] edge detection algorithms. Detected edge-map is a cognitive feature demonstrating the shape of objects to human but a pile of scattered pixels to computer [44]. Descriptions of edge-map aims in building context within edge pixels. The context between edge pixels of objects can be discovered globally or locally. Applied in pedestrian detection, local edge-map descriptions are normally used to compute the descriptors of edge pixels in a small area. Structures like *Edgelet* in [46] or *Linked Granules (LG)* in [47], segments of line / curve, are usually introduced as vocabularies in edge-map description. In [47], representative edge segments of pedestrians are learnt by combining connect granules which can be seen as short pieces of straight lines. Preprocessing is usually required to reduce the noise effect.

Histograms of edge orientation are also popular to describe the edge-map. The descriptor of an local area is usually the concatenated histograms of each edge pixels / segments. The descriptors of local edge segments can be either trained into weak classifiers which will be selected by Boosting algorithms or performing as nodes in graphical models indicating the topology of pedestrians. Due to the sensitivity of the edge-map to noise and background objects, descriptors of the entire edge-map within ROI are seldom applied to the training of a global pedestrian detector. Normally, the descriptions of rough shapes are calculated to represent pedestrians instead.

Rough shapes do not have to be the exact boundaries of pedestrians at least they could be cognitively recognised as pedestrians. The rough shapes of pedestrians are commonly calculated from 1^{st} derivative gradient image of the input images [22, 31]. Other algorithms to calculate the rough shapes include the Haar-wavelet transform as introduced in [33] and Colour Self-Similarity features (CSS) selected in [35] (by calculating the local similarity, rough areas of body-parts could be perceived after the transformation). To describe the rough shape, histogram based descriptors capturing the local orientations are popular in pedestrian detection. HOG is a SIFT like descriptor calculated on gradient image. It is further modified to Co-occurrence HOG (CoHOG) by [48]. CoHOG extends the HOG descriptors of cells of pixels into anther dimension by introducing the combination of HOG descriptions in 8-connected neighbourhood cells as the basis of descriptors. Training using CoHOG based descriptor is computational expensive due to the increased dimension of descriptors. A detailed review on HOG descriptor and a brief comparison will be demonstrated in *Section 4.1.2*

In pedestrian detection in video stream, the motion of shape is employed. Two common ways of extracting moving shapes include: the computation of optical flow and the stereo images analysis. The optical flow can be obtained either by following the motion constraint equation stating that the intensities of transported pixels between frames will be the same during the moving interval [49] or by calculating the corresponding segments from both frames. Subtraction between frames is commonly employed to capture the optical flow when the background objects are assumed still. Stereo images capturing the scene in different shooting angles are popular tools to analyse the movement of pedestrians. [50] provided a detailed review of motion capturing and developed a combined algorithm to consider both edge map and colour based image segmentation in motion detection. [2] also reviewed the algorithms to capture the shape features related to pedestrian verification and tracking in video streams. It should be clarified, when motion detection is used with body-parts recognition, the motion of a pedestrian is usually considered as the motion of human for the motions of body-parts (arms and hands for instance) may have different moving directions. Besides, [47, 51] introduced an active contour technique to adjust detected edges based on the initial edge detection through video streams.

2.2.2 Recognise Different Individuals (Identification)

Appearance based descriptions play important role in recognising different detected individuals. Differences caused by clothes, skin and accessories are popular selected elements to represent the appearances of pedestrians. In the state of art pedestrian identification, appearance based features, including colour, pattern and textures, are adopted from object classification and image understanding as reviewed in [20, 21]. Detecting and corresponding the descriptions of *Interest Regions* and *Interest Points* are commonly employed techniques to recognise pedestrian as individuals. Strategies applied to building the correspondences between detected features are reviewed in *Section 2.1*.

In pedestrian identification, the prototype pedestrian is one of the detection results, selected manually. Other detected pedestrians, either in other frames or in other camera systems, are defined as target pedestrians which will be compared with the prototype pedestrian. Comparing to the pedestrian identification in general images containing pedestrians, the choice of using the results of pedestrian detection as target images simplifies the identification processing in locating the perspective pedestrian. To build the correspondence between the images of the prototype and target pedestrians, interest regions are usually selected from the areas of body-parts, especially the upper body and lower body which influence the appearance of the pedestrian [19]. The position and size of interested regions can be detected using Boosting algorithms or pre-defined assumptions: symmetrical areas, specially located regions within the bounding box. The problem to identify the prototype pedestrians who are presented in pedestrian detection results will be discussed in *Chapter 3*. A detailed review on how to locate interest regions and how to correspond regions between images of target pedestrian the prototype pedestrian will be provided as well.

When target pedestrians are presented in images with majority areas of backgrounds, the detection of interest points is usually performed to locate the suspicious pedestrians within the image. Interest Points are detected in local areas which are salient to the change of scales and image transformations. Used in pedestrian identification, the saliency requirements are normally less strict than it is in the applications of image understanding and object classification. The pedestrians are upright for most of the time and the scaling effects on pedestrian datasets are less significant especially when shooting cameras are calibrated before the identification. These points can be either pixels located in corner detected using Harris Corner / Harris Laplacian algorithms or pixels located in areas where the gradients are changing significantly, for example, the points detected by SIFT[8] [52] and SURF [9] [53] algorithms. SIFT interest points are detected from the pile of difference of Gaussian smoothed images (DoG). SIFT interest points detection is an approximation of Laplacian corner detection applied in Gaussian smoothed images. The detected interested points that would appear on edge-maps are neglected to avoid the sensitivity of the identification performances to noise. SURF apply a similar strategy to SIFT, rather than calculate a pile of Gaussian smoothed images, SURF apply Haar-like box filter to integrated image[10] to approximate the effect of the difference of Gaussian transform. The relationship between SIFT and SURF procedure is similar to the relationship between Haar-wavelet transform and the Viola-Jones Haar-like feature detection [30]. Histogram based local area description, SIFT / SURF descriptors are popular in pedestrian identification to describe the regions of interest and local areas surrounding the detected interested pixels.

Using the appearance based features to distinguish pedestrians has limitations especially when the appearances captured in the state of the art algorithms are introduced mainly by clothes. This means, two individuals wearing the same clothes

[8] SIFT: Scale-Invariant Feature Transform

[9] SURF: Speeded Up Robust Features

[10] The same definition to the integrated image used in Viola Jones Haar-like features [30]

will be recognised as the same while one person wearing different clothes will be identified as different persons.

2.2.3 Relevant Image Processing

Image normalisation and equalisation are occasionally required to reduce the illuminant effect of input images. Gamma correction, histogram normalisation are commonly employed algorithms for image normalisation. When the default colour spaces of input images are RGB [11]which are default settings applied in OpenCV and Matlab, the colour space transformation would be required to perform the appearance based detection / identification in other colour spaces like HSV, HSL [12]where colours along one dimension (Value in HSV and Lightness in HSL) is cognitively similar to human. This transformation reduces the dimension of colour image intensities and at the same time retains the perceptive meanings of input images to human.

2.3 The Databases of Pedestrians

Databases of pedestrians were established by various research bodies. Frequently cited databases include Caltech [3], ETH [9], INRIA [22], MIT [54], Daimler [55] and TUD [56]. Due to the interest of pedestrians in this thesis, all five listed databases are pedestrian databases. The sources of images (pedestrians and backgrounds) in the databases are obtained according to the research purposes and relevant applications. MIT and INRIA gathered images of pedestrians from photos and images of backgrounds from a random selection in the background area of source images. Images in ETH, TUD and Daimler databases are frames from mobile recordings shot on streets. The training databases may affect the performances of the detectors. Even for the same application, driving assistant for example, different databases would be used to train the detectors: Caltech databases are gathered from both USA and Japan[3, 56] and TUD databases are gathered from both Darmstadt and Brussels [56]. Among the five, Caltech contains the largest number of images. Daimler is the only database containing gray-scaled images.

In this thesis, for pedestrian identification, dataset is gathered from the results of HOG pedestrian detector, and website images where pedestrians are wearing patterned clothes. The dataset contains around 500 images in which ~150 of them are wearing patterned clothes. For pedestrian detection, focusing on the false alarm reduction, the testing dataset is a self-established one where high proportional false

[12] RGB: Red, Green, Blue; HSV: Hue, Saturation, Value; HSL: Hue, Saturation, Lightness

alarms exist in the detection results of HOG. The dataset contains a random selections of 1,000 images of the searching results of "Google pedestrian" and ~3,000 frames of two videos shot in a noisy car park containing sources of false alarms. Popular databases are not employed in this thesis for two reasons:

- Pedestrian Identification in this thesis focuses on how to quantify the complexity of the appearances of pedestrian and identify pedestrians with different appearances. The pedestrians in popular datasets are normally wearing mono-coloured clothes;
- The false alarm rates of HOG pedestrian detector applying to the popular databases are low.

2.4 Conclusions

As reviewed in previous sections, pedestrian detection and identification can be treated as classification problems. Shape is the commonly selected feature to distinguish pedestrian from background objects. Appearance is mainly employed to compare the detected individuals. Generative classification algorithms, such as SVM and Boosting classification, have the advantage of simple processing procedure in applications. But the classifiers trained using such strategy can hardly be changed. Furthermore, complex strategics, which are expensive in computation, may only be compulsory for limited difficult circumstances. In this thesis, attention will be focused on developing frameworks which are editable, capable of quantifying the complexities of problems and improving the performances of trained classifiers.

Chapter 3

Histograms in Pedestrian Re-identification

You cannot build an organ which tells you whether it can be done.

—— John von Neumann

People Re-identification solves problems concerning the recognition pedestrians, especially the reappeared pedestrians, in video streams and camera networks. In video streams, people may be blocked during movement before reappearing in the scene. In camera networks, different views of people may be shot by different surveillance cameras. To recognise the same person in scenes shot in different time and space regardless their poses, shooting angles and background objects is the problem addressed by people re-identification. One important application of people re-identification is the tracking of people, especially the tracking of multiple people for surveillance and monitoring. Due to the probable changes of poses and background of target images with people, the shapes of people are no longer a discriminate feature. The appearances of people, especially the colours and patterns of clothes, play an important role in re-identification problems. Techniques used in content based image retrieval (CBIR), object classification and recognition are usually transferable in this area. Additionally, people re-identification may also consider the information of video clips including the context between frames and the consistency of light conditions.

In people re-identification, a detection of the prototype pedestrian is required and further detection results are compared to this prototype to examine whether they are the same person. This problem can be dealt in either supervised mode or non-supervised mode. The former requires manually labelled prototype people where the latter requires pre-processing steps to separate the prototype people from background so that strategies used in supervised mode can be applied. Even in the supervised mode, people re-identification is a difficult problem. For one thing, there is limited experiences from human cognition and recognition. Until now, how people recognise people as individual or species regardless their appearances is not well understood. For another, at the current stage, only appearances based descriptors are used in people re-identification, which means people in the same piece of clothes (refer to the same colour / pattern) would be recognised the same and one person would hardly be re-detected if his / her clothes has been changed during his / her absence.

To simplify the problem, as in other chapters, only pedestrians are considered, which restrict the agility of human actions and poses. Both supervised (*Section 3.2.1*, *3.2.2*) and non-supervised (*Section 3.2.3*) pedestrian re-identification will be addressed in this chapter. Non-supervised re-identification benefits the applications when it is difficult to select a prototype pedestrian, for example, the tracking of multiple pedestrians or to update the tracked pedestrian automatically during a relatively long period of video streams. The contributions of this chapter lie in three areas:

- A novel pyramid protocol is proposed to describe and recognise the reappeared pedestrian. The appearances of pedestrians are quantified to levels of complexities according to the richness of required descriptors. The lowest level of descriptions are colour related histograms, which is the fundamental level of description. Followed by the descriptions of regular patterns and then the other sophisticated signatures. The more levels of descriptions are applied in descriptions, the more complex the appearances of pedestrians. The aim of the protocol is to spend different computation efforts on the identification of prototype pedestrians with different appearances. The structure of the protocol is modifiable as new levels of descriptions can be added to improve the performance of the existing system;

- In the fundamental level of pedestrian description and re-identification, taking advantages of the "bag-of-features", a novel histogram based strategy is developed to fast identify the reappeared pedestrians. Histogram based descriptions guarantee the speedy processing procedure which is important especially for real-time / online applications. The body parts based codebooks compromise the sensitivity of histogram descriptions to the changes of the size and position of target pedestrians. As a part of the pyramid protocol, fewer histogram features are used to identify pedestrians with relatively simple appearances than the ones with a complex looking;

- A novel attempt of fast image analysis is investigated as an assistant step in building the dictionary of PROTOTYPE and determining the levels of complexities of the appearances. The algorithm is based on the knowledge of pedestrian detection and pixel clustering. The result of the analysis provides the colour and position information of interested cognitive areas within the PROTOTYPES. This analysis is currently only valid for PROTOTYPES with simple appearances: mono coloured clothes, bi-coloured simple patterned clothes (dots, strips, lattices).

3.1 Review of Pedestrian Re-identification

Pedestrian Re-identification requires two sub-tasks:

1. Building correspondences between images containing pedestrians (one performs as the prototype, the other as a target);

2. Generating invariant signature(s) to compare the correspondent parts. Signatures used in object recognition and classification are usually transferable in pedestrian re-identification.

Commonly selected signatures related to colour information include histogram and Maximally Stable Extremal Regions (MSER) [57]; to describe a selected region, SIFT [52], SURF [53], HOG and Hessian affine operator are popular algorithms; Shape Context [58], histogram of accumulated responses of edge detectors are used to provide the shape information of an area.

There are ways of building correspondences between images. Below lists a couple of approaches:

- Constructing the "bag-of-features (BoF)" of the prototype. The features can be a group of interest points detected from images, such as SIFT / SURF interest points, Harris Corners, which are originally utilised in object classification and recognition [41, 59–61].

- Locating the body parts of the pedestrians and corresponding the same body parts in images. In [62], triangular graph for modelling the shape of objects and pictorial structures for modelling the deformable shapes were used to model the body parts of pedestrians. In [63], the symmetry feature of pedestrian was examined to detect the corresponding areas between images.

- Using global signatures and treating the pedestrian as a whole. Biometric features, gaits [64, 65], were selected as signatures for instance. [66] introduced the brightness transfer functions between cameras to adjust colour histograms between images.

- The discriminative Learning approaches were sometimes applied to train the models that employ combinations of local descriptors to represent the pedestrian. Boosting algorithms and SVM are popular training algorithms [67, 68].

In *Table 3.1*, frameworks used in pedestrian re-identification are summarised according to their means of building correspondences between images and the signatures selected to describe pedestrians.[1]

3.2 Code Book Matching

During the tracking of pedestrians, to identify a reappeared pedestrian after a period of complete occlusion can be modelled as a prototype based recognition problem. The pedestrian being tracked is the prototype pedestrian (use PROTOTYPE(s) in later paragraph) and future detections, especially the ones after a period where the PROTOTYPE is blocked and disappeared from the scene, are the target pedestrians (use TARGET(s) in later paragraph). To retain the tracking activity and update the information of the PROTOTYPE, the detected TARGETs are compared with the PROTOTYPE to examine if they are the reappeared PROTOTYPE. If the judgement is true, the information of the PROTOTYPE will be updated. When there is no clear clue showing which biometric features play crucial roles in human recognition activity, features used in re-identification relate to the appearances of pedestrians. It means, a TARGET will be judged as a reappeared PROTOTYPE if the TARGET is wearing the similar patterned clothes and having similar skin-hair colours, if any exposure, with the PROTOTYPE. The difficulties of re-identification problems vary according to the complexities of the appearances of PROTOTYPEs. For example, if the clothes (top / bottom) of a PROTOTYPE are mono-coloured without patterns and are covering the majority of the body, the re-identification of the PROTOTYPE is simpler than the re-identification of PROTOTYPES who are wearing patterned multiple coloured clothes (a multi-coloured top inside an open jackets for instance). The problem is even harder if the exposed body parts (limbs, hairs, etc.) of PROTOTYPES are involved in the the re-identification.

[1] Some are left blank as less information was provided in the literature. This may be because the literature focus only one aspect of re-identification, or the algorithms presented in the literature were assumed versatile with various accompanied algorithms in generating signature or building correspondences.

Literature	Building Correspondence	Signature(s)
[62]2006, Gheissari.	Body parts modelling by triangulated graph	Hue histogram Hessian affine operator
[69]2005, Bird.	10-horizontal stripes	accumulated HSV value
[70]2007, Wang.	Densely computed local descriptors spatio relation matrix	Hue histogram
[63]2010, Farenzena.	Axes of symmetry and asymmetry defined body parts require FG/BG separation	Weighted colour histogram MSCR Recurrent structured patches
[71]2007, Gandhi.		Panorama appearance map
[67]2008, Gray.		Ensemble of local descriptors
[72]2008, Hamdoun.	SURF Interest points	Hessian affine operator
[36]2009, Schwartz.		Texture, Gradient, Colour to low-dimensional space using Partial Least Square
[73]2010, Bak.	Body part detector: HOG and modified face detector using edge information	Colour histogram Pyramid Matching: Body Body parts, 1/4-body parts
[74]2011, Cheng.	Pictorial structure: Single/Multi Shot(s)	Colour histogram
[75]2012, Satta.	Multiple component matching	Appearance of body parts
[76]2013, Zheng.	Relative distances between images	Colour histograms Texture Features using Gabor Filter

Table 3.1 **Frameworks used in Pedestrian Re-identification**

Using an appearance based re-identification framework, if the selected features represent simple appearances only, the processing speed of the re-identification is relatively fast but the adaptability of the framework is narrowed thereafter. Complex frameworks employing multiple features and dedicated judgements are normally computational expensive. And they are overwhelmed for a simple case especially when fast reaction is crucial for pedestrian tracking and other real-time applications.

In this section, using a pyramid structured protocol to solve the re-identification problem, the number of levels of judgements are selected according to the complexities of the appearances of pedestrians:

- Level 0: mono-coloured clothes without much skin-hair exposure;
- Level 1: bi-coloured clothes or mono-colour clothes with significant areas of skin-hair exposure;
- Level 2: clothes with regular patterns containing two colours;
- Level 3: clothes with more than one type of regular patterns or containing two colours or more;
- Level 4 (above): clothes with irregular / random patterns

Fig. 3.1 shows a basic structure to compare the PROTOTYPE with the TARGETS using levels of descriptions in a dictionary. In this protocol, the colour related signatures (histograms calculated in the HSV colour space in this chapter) are selected as the fundamental descriptions to exempt the TARGETS wearing different clothes with the PROTOTYPE, upper-body or lower-body. Enough levels of descriptions should be calculated on the PROTOTYPES for varied purposes while not all levels of those are required during the identification. The selection of descriptions is related to the consideration of speed, accuracy and the stability of performances in applications. Experiments demonstrate when the PROTOTYPE is wearing mono-coloured clothes or simple patterned clothes, only colour related histogram description is enough to make decisions with detection rate > 90% and false alarm < 10%. Observations also demonstrate if pattern information could have been considered in re-identification, the results would be better.

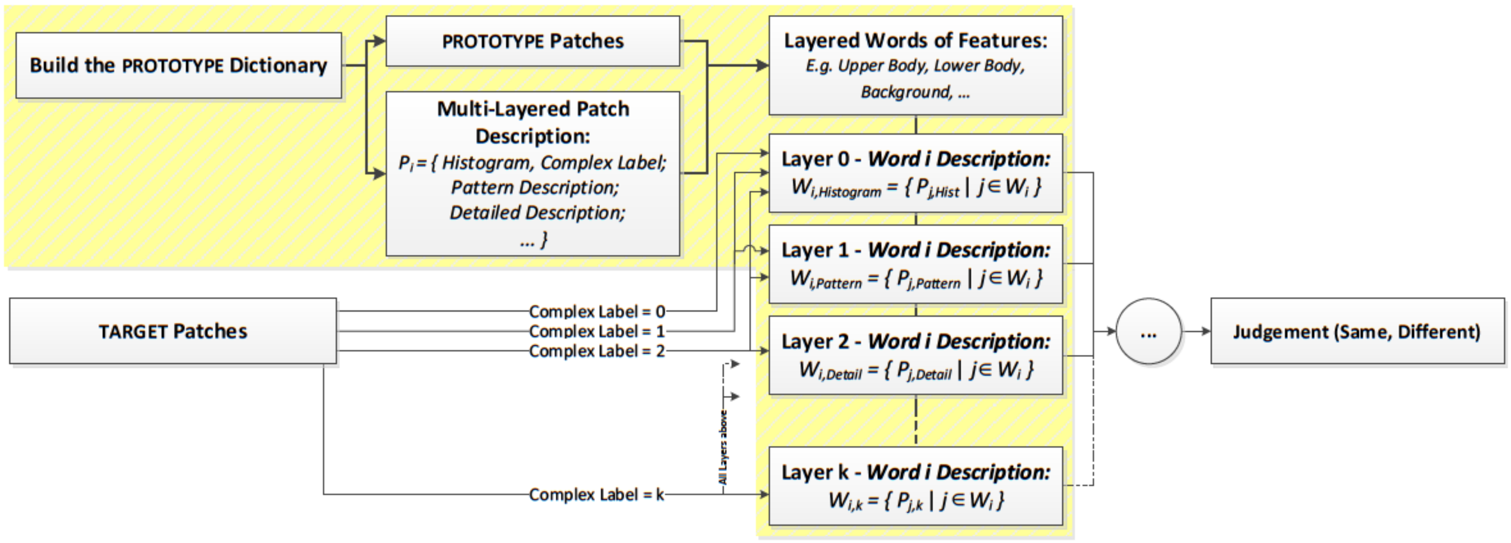

Build the PROTOTYPE Dictionary
PROTOTYPE Patches
Multi-Layered Patch Description:
P_i = { Histogram, Complex Label; Pattern Description; Detailed Description; ... }
Layered Words of Features:
E.g. Upper Body, Lower Body, Background, ...
Layer 0 - Word i Description:
$W_{i,Histogram} = \{ P_{j,Hist} \mid j \in W_i \}$
Layer 1 - Word i Description:
$W_{i,Pattern} = \{ P_{j,Pattern} \mid j \in W_i \}$
Layer 2 - Word i Description:
$W_{i,Detail} = \{ P_{j,Detail} \mid j \in W_i \}$
Layer k - Word i Description:
$W_{i,k} = \{ P_{j,k} \mid j \in W_i \}$
TARGET Patches
Complex Label = 0
Complex Label = 1
Complex Label = 2
Complex Label = k
All Layers above
...
Judgement (Same, Different)

Fig. 3.1 **The dictionaries of prototype pedestrians are built using vocabulary related patches. Words selected in the dictionary include upper-body, lower-body and background, etc.. The patches are described using multiple layers of descriptions. The fundamental level of description is histogram followed by pattern descriptions and other descriptions on upper levels when required. By comparing the TARGET to the dictionary of PROTOTYPE, patches of TARGET are coded by words. This code map is further processed to make a final judgement.**

The TARGETs are desired, but not compulsory, to have the same number of levels of descriptions as the PROTOTYPE. The levels of descriptions of PROTOTYPEs and TARGETs are compared simultaneously or hierarchically. When descriptions on different levels focus on different dimensions of the appearances of pedestrian, shapes and patterns for instance, the descriptions on different levels can be simultaneously compared. The comparisons of descriptions should be done in a hierarchical way under following circumstances: one level of descriptions is more distinctive than the other levels of descriptions or one level of descriptions is dependant on the other levels of descriptions. Using a hierarchical ways of comparing the levels of descriptions, higher levels of descriptions are usually less compulsory than the lower ones.

The content of this section is as follows. *Section 3.2.1* succeeds the "bag of features" strategy and builds a dictionary to describe the PROTOTYPE using histogram based descriptions. *Section 3.2.2* examines the colour histogram based fundamental descriptions of the TARGET and the PROTOTYPE. Considering the character of pedestrian detection: the pedestrian is usually located in the centre and would occupy at least a quarter of the bounding box area. *Section 3.2.3* generates the dictionary of PROTOTYPE automatically. The level of pattern judgement (using hough transform based regular pattern descriptions) will be discussed in the future work (*Section 6.2.1*). For complicated cases, more features could be introduced and more levels of judgements should be considered.

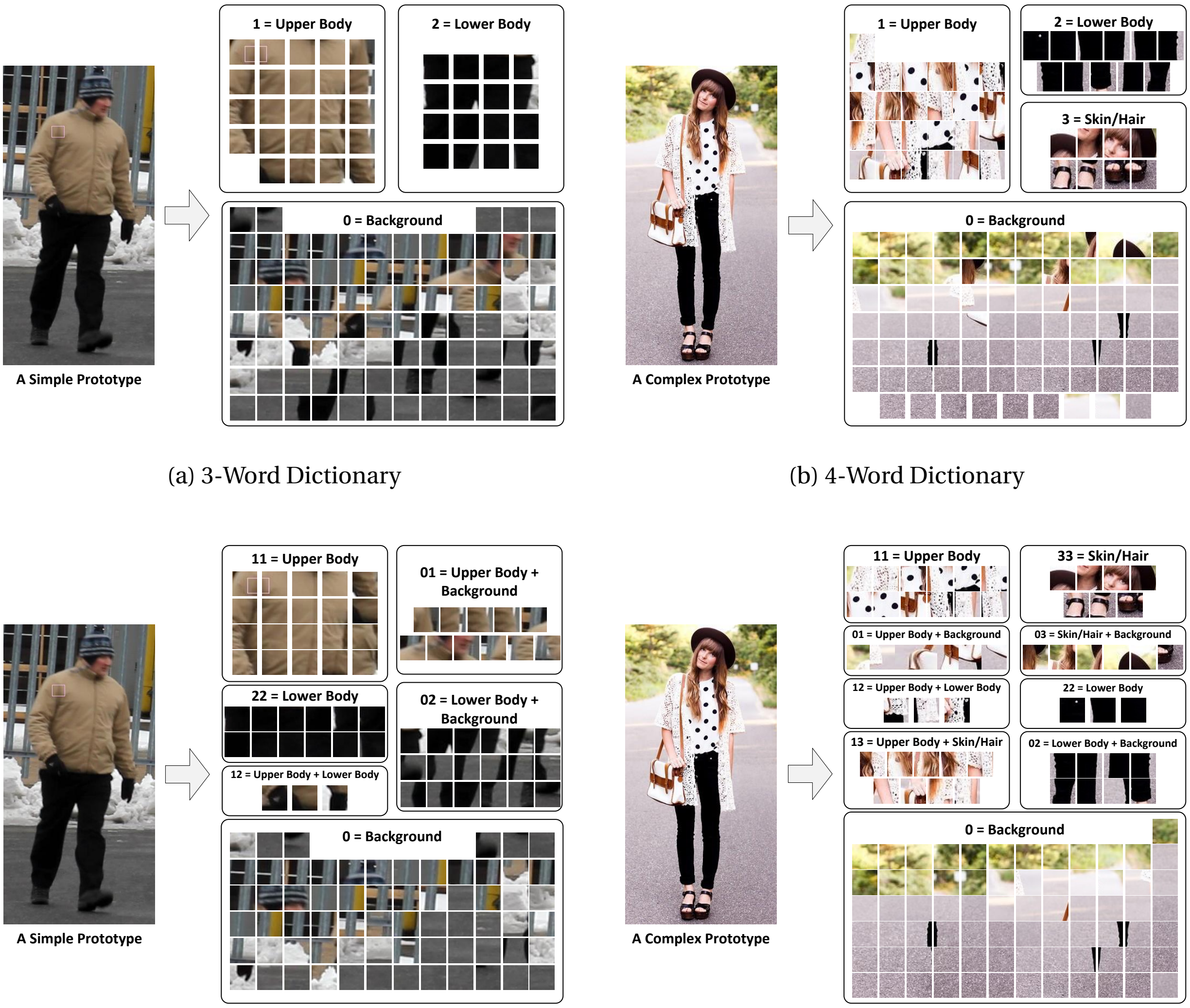
3.2.1 PROTOTYPE Description: build a dictionary

To describe a PROTOTYPE using a dictionary, words are selected to represent the cognitive meanings of parts of the PROTOTYPE. Such parts of interest include body parts and background. The definitions of words are summarised from the descriptions of relevant areas of image. During the correspondences, the sub-image areas of TARGET are given the same word with its matched areas of the PROTOTYPE. The judgement of re-identification is made according to the word map of TARGET. For example, when the upper-body area is labelled above the lower-body area and both of whom are surrounded by the background area in the word map, the pedestrian will be judged as a reappeared PROTOTYPE. Practically, the PROTOTYPE and the TARGETs are uniformly separated into rectangular patches to efficiently locate the matched pairs of sub-image areas. Patches with the same cognitive meanings in the PROTOTYPE are grouped and coded with relevant words in the dictionary. The patches of TARGETs are coded by the word representing their best matched patches of PROTOTYPE. As a reference to the calculation of the word maps of TARGETs, the word map of the PRO-

TOTYPE can be manually generated or use an automatic way (*Section 3.2.3*). When the pyramid judgement structure is applied to identification, the levels of descriptions are calculated on patches of images, the PROTOTYPE and the TARGETS.

The number of words used in the dictionary affect the performance of the re-identification framework. To describe a detected pedestrian, at least three words are required: *upper-body*, *lower-body* and *background*. Sometimes, when the pedestrian occupy a relatively small area of the bounding box, a large number of patches may contain equally areas of upper-body / lower-body / background. Such patches will be categorised into new intersection words, such as *upper-body with lower-body*, *upper-body with background*, *lower-body with background* etcetera. If the areas of skin, hair and accessories are considered crucial in re-identification, words of descriptions can be added to the dictionary. The vocabulary of the dictionary should be large enough to distinguish all interested areas of pedestrians, especially the areas of upper-body and lower-body. Lack of words may introduce errors in re-identification. However, experiments have shown that an over complete vocabulary may also result in less stable performances. *Fig. 3.2* demonstrates the manually generated dictionaries containing three, four, six and ten words, which are used in the experiments of re-identification in later sections.

In the prototype based pedestrian re-identification, building the dictionary manually is not usually a problem as the prototype should be selected before the identification process. Manually built dictionary is accurate according to the human cognition. However, when large number of PROTOTYPEs are desired and when large number of patches are separated from the PROTOTYPE, efficient means of PROTOTYPE analysis are required. Based on the clustering of pixels and the Probability Density Functions (PDFs) of body parts, *Section 3.2.3* attempts to learn PROTOTYPEs with simple appearances using an automatic way.



(a) 3-Word Dictionary

(b) 4-Word Dictionary

(c) 6-Word Dictionary

(d) 10-Word Dictionary*

Fig. 3.2 **The results of the dictionaries of the PROTOTYPE using three, four, six and ten words, built manually. Some of the time, not all words are related with patches in images. As shown in (d), no patch is categorised as "23 = *lower-body with background*". Above are dictionaries of the PROTOTYPE separated into** 8×16 **patches. New dictionaries are required when different numbers of patches are divided in *prototype*.**

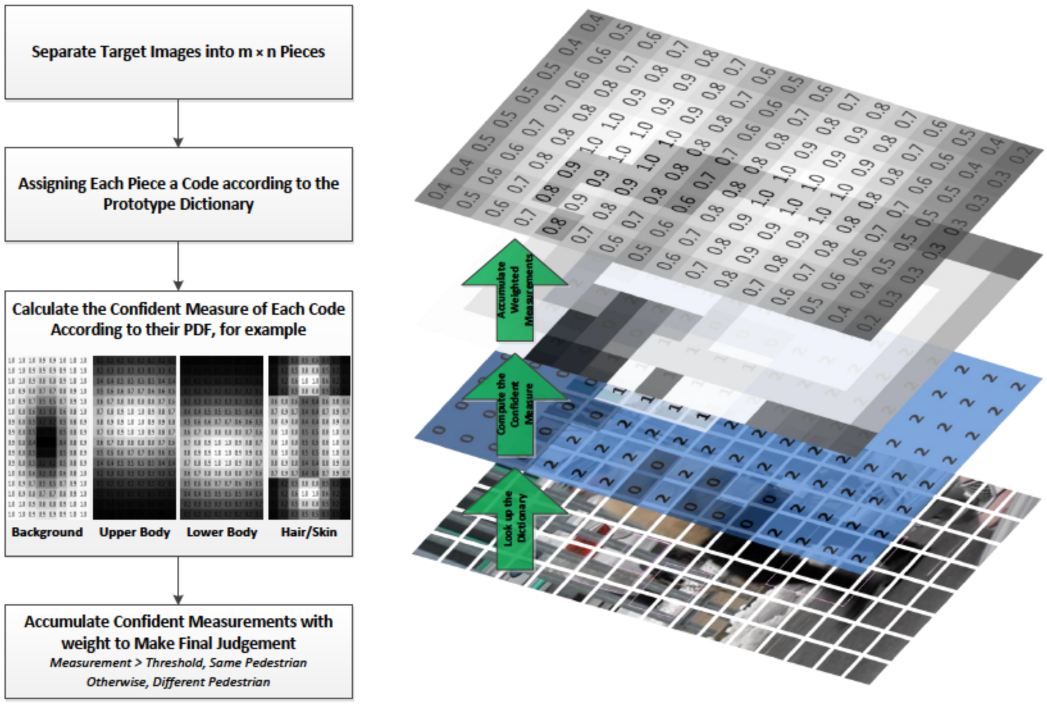

Separate Target Images into m × n Pieces
Assigning Each Piece a Code according to the Prototype Dictionary
Calculate the Confident Measure of Each Code According to their PDF, for example
Background
Upper Body
Lower Body
Hair/Skin
Accumulate Confident Measurements with weight to Make Final Judgement
Measurement > Threshold, Same Pedestrian
Otherwise, Different Pedestrian
Accumulate Weighted Measurements
Compute the Confident Measure
Look up the Dictionary

Fig. 3.3 **Making Judgement of a Target Image**

3.2.2 The Fundamental Level: Check the Histogram

To calculate the similarity of the TARGETs to the PROTOTYPE ($\text{Similarity}_{\text{Total}}$), the word maps of TARGETs should be calculated first. In the fundamental level, this is done by comparing the HSV histogram descriptions of every patch in TARGET to the patches of every word in the dictionary of PROTOTYPE. The correlation distances are employed to compare histograms. The similarity of a patch in TARGET to a word in the dictionary ($\text{Similarity}_{\text{word}}$) is measured using the averaged correlation distances between the patch of TARGET and patches coded with the word in the dictionary. The patches of TARGETs are corresponded with the best matched word which has the largest validated similarity measurement ($\text{Similarity}_{\text{word}} \geq T_c$) among the vocabulary. If there is no word in the dictionary could match the patch ($\text{Similarity}_{\text{word}} < T_c$), the patch is categorised as background.

Considering patches from background may have similar histogram descriptions with patches from body parts, to verify the reliance of the code map of TARGET, confident measurements of elements in the code maps are calculated. The confident measurements indicate how probable the patch in position with meanings of $word_i$ would be a part of the PROTOTYPE or background. The confident measurements of patches in areas of parts of interest are averaged as the similarity between pairs of corresponded parts of interest ($\text{Similarity}_{\text{parts}}$) in TARGET and PROTOTYPE. The $\text{Similarity}_{\text{parts}}$ are weighted and summed as the $\text{Similarity}_{\text{Total}}$. If the $\text{Similarity}_{\text{Total}}$ is over threshold T_j, the TARGET will be judged as a reappeared PROTOTYPE.

When $M \times N$ patches are separated in TARGETs. A word map of TARGET (C, size(C) = $M \times N$) is a $M \times N$ matrix. The confident measurements map (F, size = $M \times N$) of word map C ($F(x, y) = P(\text{part} \mid word_i, \text{position})$) is calculated using Bayesian rules:

$$\begin{aligned} P(\text{part} \mid word_i, \text{position}) &= \frac{P(word_i \mid \text{part}, \text{position}) P(\text{part} \mid \text{position})}{P(word_i)} \\ &= \frac{P(word_i \mid \text{part}, \text{position})}{P(word_i)} \cdot \frac{P(\text{position} \mid \text{part}) P(\text{part})}{P(\text{position})} \end{aligned} \tag{3.1}$$

Where "part" indicates the areas that have the same cognitive meanings like upper-body, lower-body or background; "*word*"s are vocabularies to describe the cognitive parts, for instance a white top, a black cardigan, which are represented by the descriptions of elements in the code map ($C(x, y) = word_i$); "position" relates to the coordinates of patches in the code map ($(x, y), x = 1, \ldots, N, y = 1, \ldots, M$). *Equation 3.1* is a general form of the calculation of confident measurements when the vocabularies and the cognitive parts are not one-to-one mapped. It means a body part can be related to more than one words and one word may represents areas of different body parts. $P(word_i \mid \text{part})$ shows the dependence of $word_i$ to the cognitive parts. Manually selected vocabulary means the probabilities of words are independent with the positions of patches, and *Equation 3.1* becomes:

$$P(\text{part} \mid word_i, \text{position}) = \frac{P(word_i \mid \text{part})}{P(word_i)} \cdot \frac{P(\text{position} \mid \text{part}) P(\text{part})}{P(\text{position})} \tag{3.2}$$

$P(word_i), P(\text{part}), P(\text{position})$ are all constants as "$word_i$", "part" and "position" are manually defined terms. When vocabulary and cognitive parts are bijectively mapped, $P(word_i \mid \text{part}) = 1$. The confident measurement of $F(x, y)$ is proportional to the probability of the patch at (x, y) represented by a cognitive part ($word_i$), which is related to the PDF of the "part". When vocabularies for describing intersection areas are introduced, the confident measurements of patches coded with intersection areas are as follows:

$$P(\text{intersection} \mid word_i, \text{position}) = \frac{1}{2}\Big(P(\text{part}_1 \mid word_i, \text{position}) + P(\text{part}_2 \mid word_i, \text{position})\Big) \tag{3.3}$$

The similarity measurements of parts of interest ($\text{Similarity}_{\text{parts}}$) are the averaged confident measurements $\overline{F(x,y)}$ multiply the percentage of correctly corresponded patches in the areas of parts of interest. The areas of parts of interest are the areas in which the PDFs of patches are over threshold (0.7 in experiment). For example, to calculate the similarity of upper-body, given the word map (C) and its confident measurements map (F) of TARGET, the $\text{Similarity}_{\text{parts}}$ is:

$$\text{Similarity}_{\text{upper}} = \frac{\#C(x,y)_{\text{represents upper-body}}}{\mathit{Area}(\text{upper-body})}\,\overline{F(x,y)},\ (x,y) \in \mathit{upper\text{-}body}$$
$$\text{where} \qquad \mathit{upper\text{-}body} = \{(x,y)|\mathit{PDF}_{\mathit{upper\text{-}body}}(x,y) > 0.7\} \quad (3.4)$$

If parts of interest are not considered equally important in identify if the TARGET is similar to the PROTOTYPE. The $\text{Similarity}_{\text{parts}}$ is the weighted in calculating the similarity of TARGET to the PROTOTYPE. In a general case:

$$\text{Similarity}_{\text{Total}} = w_{\text{upper}}\text{Similarity}_{\text{upper}} + w_{\text{lower}}\text{Similarity}_{\text{lower}} + w_{\text{skin}}\text{Similarity}_{\text{skin}} \quad (3.5)$$

Weights (w_{part}) applied to the similarity of corresponded parts ($\text{Similarity}_{\text{parts}}$) should be chosen according to the complexity of appearances of the PROTOTYPE. In the experiment in *Section 3.3*, for PROTOTYPE wearing mono-coloured top and bottom, w_{upper}, $w_{\text{lower}} = 0.5$ are selected. For PROTOTYPE wearing patterned top and mono-coloured bottom, $w_{\text{upper}} = 0.6$, $w_{\text{lower}} = 0.4$ are selected. Without the comparison of patterns of parts of interest, different weights applied to the parts of interest with different appearances have less effect on the identification performance (the identification rate is increased by ~1% in experiment). When patterns of parts of interest are considered, the identification rate will be increased by around 5%. The calculation of PDFs of parts of interest (body parts and background) will be discussed in *Section 3.2.3.4*.

3.2.2.1 A Fast Comparison of Fundamental Layer

In the re-identification of pedestrians, background is usually less attractive than body parts, upper-body, lower-body or skin-hair for example. In detection results of HOG, the area of background is larger than areas of parts of interest. To fast identify the reappeared PROTOTYPE, patches of TARGETs can be compared with patches of words related to parts of interest only. If no words related to parts of interest are selected, the patch will be coded as background. Discarding the word of background remains the identification rate of this modification may introduce false alarms. As a compensation, if patches located in the areas of parts of interest are coded with words representing the parts of interest, these patches are compared with the patches coded by background in dictionary to verify its cognitive meaning. This improvement reduces the half of the computation time comparing to the former strategy on average. It is significant when large number of patches are separated the PROTOTYPE or TARGETs.

Applying the above structure to the comparison of the histogram descriptions of PROTOTYPE and TARGET wearing mono-coloured clothes, experiments (*Section 3.3*) will show that even the dictionary is built with few words and only a small number of patches separated from the PROTOTYPE and TARGETs, the identification result is promising: in the dataset containing ~500 images, more than 90% images are correctly recognised. Increasing the number of patches separated from the PROTOTYPE and TARGETs, the performances of re-identification can be improved. If the body-parts of PROTOTYPE appear in combined patterns, for instance a mono-coloured top with exposed arms, a top with different coloured vest / cardigan, increasing the number of words used in the dictionary of PROTOTYPE can improve the identification performance: to identify the PROTOTYPE wearing different coloured top and cardigan, using two words in the dictionary of upper-body can increase the identification rate by ~2% than using one word in the dictionary. When the PROTOTYPE is wearing patterned clothes (dots, stripes, lattices, etc.), due to the loss of geometry information in the histogram descriptions, the patterns of two patches having similar histograms cannot be compared. Therefore, further layers for pattern judgement should be included to verify the appearance of TARGETs who have similar histogram descriptions to the PROTOTYPE.

3.2.3 Analyse the PROTOTYPE with Simple Appearance

The previous section introduced the strategy of pedestrian re-identification: how to describe the PROTOTYPEs using a dictionary of which the vocabulary is explained by levels of descriptions of sub-image patches (*Section 3.2.1*); in the fundamental level of comparison using histogram based descriptor, how to generate the code map of TARGETs and its map of confident measurements (*Section 3.2.2*). In this section, how to automatically analyse the PROTOTYPE with simple appearance will be discussed. This analysis is under the assumptions that upright PROTOTYPEs are detected using HOG. The detected pedestrians are roughly located in the the centre of the bounding boxes. Using parts of interest (*abbr. part(s)*), upper-body / lower-body for example, the analysis provides information on both the colour(s) of the parts and the approximate positions and areas of these parts. The colour information is described in HSV space and the positions and areas of parts are determined by the positions of sub-image patches assembling the parts of interest.

3.2.3.1 Colour Information Analysis

The analysis of colour information of PROTOTYPE presented in this section is similar to the strategy of "the colour names description" introduced in [77], both of whom require pixels clustering in RGB colour space. In this section, the number of clusters is restricted without knowing the exact colour of the body parts. [77] modelled the pre-defined colour palettes before clustering. Pixels in the input PROTOTYPE are clustered to k categories using k-means clustering. In the k-means clustering, the RGB values of pixels are observations, the similarity of clusters are measured in Euclidean distance. To avoid mis-clustering, a relatively larger number of clusters are chosen comparing to the number of parts of interest. Due to the lighting conditions and the probable changes in view point, cognitively similar colours may be clustered into different groups. To detect the main colours of body parts, centres of clusters are transformed from RGB space to HSV space as the H-S values are more similar to human cognition. Clusters are combined if their centres are closely projected in "H-S" coordination plane. The new cluster centre is chosen as the one of the clusters with the largest amount of elements. After that, the intensity of each pixel is given the value of its cluster center. The combined clusters are sorted by their size of elements. PROTOTYPE usually occupy more than 50% area of the image. After combination, the clusters containing more than 50% of the total pixels of image are recognised as the dominant clusters. The colours of dominant clusters are recognised as the main colours of the PROTOTYPE. To further decide the colour information of each interested cognitive part, the areas of parts of interest should be located, which is addressed in following contents.

3.2.3.2 Locate the Range of Interested Cognitive Parts

To locate the parts of interest: upper-body, lower-body, and background, the PDF matrices of these parts are initially calculated. If the PROTOTYPE is separated into $M \times N$ patches, the PDF matrix (P_{part}) has $M \times N$ elements. $P_{part}(x, y)$ describes the probability of the patch at (x, y) representing the part of interest. In following paragraph, $P_{upper\text{-}body}$, $P_{lower\text{-}body}$, $P_{skin\text{-}hair}$, $P_{background}$ are used to represent the PDF matrices of "upper-body", "lower-body", "skin-hair", "background" respectively. A patch at (x, y) is labelled as a part of interest if the value of PDF of the part at this patch is the maximum one comparing to the values of PDF of other parts at the patch. For example, if a patch at (x, y) is assigned to the rough upper-body area,

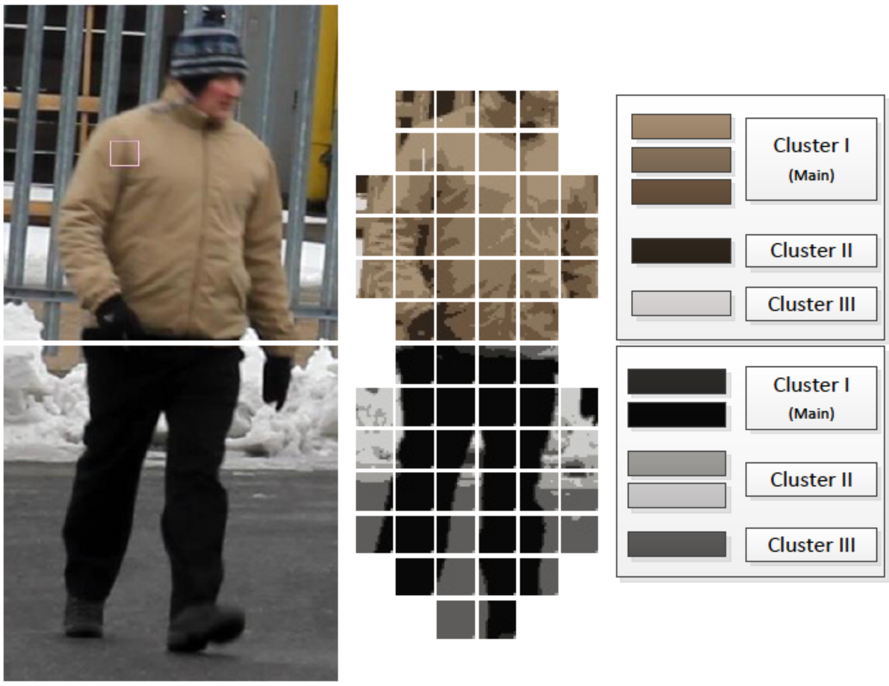


Fig. 3.4 **The result of the clustering of pixels within the rough area of upper-body and lower-body of a prototype wearing mono-coloured top and bottom.**

then the probability of upper-body at that patch has the maximum value among the probabilities of all interested cognitive parts:

$$\text{Label}_{x,y} = \text{arg}_{\max(P_{parts}(x,y))}\,\text{Label}_{x,y}(parts)$$
$$\text{where, } parts = \{upper\text{-}body,\ lower\text{-}body,\ background\} \quad (3.6)$$

If the values of two PDF matrices of cognitive parts are the same, the patch is labelled as the intersection area between the two. This strategy locates the rough area of parts of interest. *Fig. 3.4* demonstrates the rough upper-body and lower-body areas of a simple prototype and the results of clustering of pixels of the PROTOTYPE image (5 clusters are selected for both upper-body and lower-body). Detailed calculations of PDF Matrices of upper-body, lower-body, skin-hair and background are provided in *Section 3.2.3.4*

3.2.3.3 Determine the Sub-image Patches of Interested Cognitive Parts

To locate the parts of interest of the PROTOTYPE, the following assumptions are made:

- The PROTOTYPE is located in the centre area of the bounding box;
- The probability of a patch belonging to a part of interest is measured by the PDF of the

part;

- Clothes with simple appearances means the patterns and colours of clothes can be recognised the same for most of the time and the changes of appearances happen at the edge of two pieces of clothes (top inside a cardigan for example);

- The patterns of a large area are more stable than patterns occupying a small area in re-identification as they may be occluded in folds and drapes of clothes. Unstable patterns may not be visible when the viewing angle of the pedestrian has been changed.

Based on the above assumptions, a learning strategy from central patches to surrounding ones is proposed: several centre patches are selected as the prototype patches of which colours and patterns are recognised as elements of the appearance of the parts of interest (upper-body or lower-body); then patches surrounding the prototype patches are compared with the prototype patches to decide whether they are from the same piece of clothes; surrounding patches that have different patterns with the prototype ones are further compared with its 4-connected neighbourhood patches to check the pattern consistency; if this new pattern is spread over several neighbourhood pieces, these patterns are also recognised as elements of the appearance of the part of interest; otherwise, the patch is recognised as either noise or patterns from other parts of interest according to the relative position of the patches within the bounding box. Practically, the strategy is adjusted with upper-body and lower-body. The consistency of patterns over patches of clothes are examined in two levels, colour and boundary:

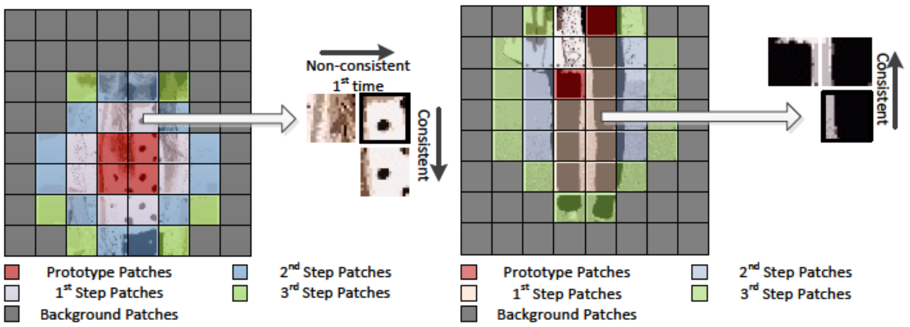
- *Colour consistency*: each patch is given k'-bit histogram description, where k' is the number of clusters after the combination of cognitive similar clusters (*Section 3.2.3.1*). The value of each bit of the description is the percentage of pixels in each cluster in the total number of pixels of the patch. To compare two neighbourhood patches, the euclidean distances between the descriptions of two patches are calculated;

- *Boundary consistency*: edge between two patches are examined, a perspective area around the boundary of two neighbourhood patches are selected. The first and the second derivative of the perspective area are calculated. This step is crucial when a patch contains the main colours of a part of interest but have different descriptions with its neighbourhood patches belonging to the part of interest. This step also determines the boundary of two pieces of clothes as well as the boundary of two parts of interest.

Learning the appearance of upper-body: the prototype patches of upper-body are selected as patches that have PDF values over 0.95. The PDF matrices of parts of interest are modelled using Gaussian distribution (see *Section 3.2.3.4*). These patches usually occupy approximately 15% area of the perspective upper-body area. Colour information obtained from prototype patches are the percentage of each major colour clusters. In experiment, in a prototype patch, clusters containing more than 30% of the total pixels of the patch are recognised as one of the main colour clusters, clusters containing less than 30% but more than 10% of the total pixels of the patch are recognised as the complimentary colour clusters, other colour clusters are ignored. Patches surrounding the prototype patches are examined from the inner ones to the outer ones according to their distance to their nearest prototype patch. Patches are first compared to the prototype patches, the ones have similar descriptions with the prototype patches of upper-body are recognised as patches of upper-body. If the description of a patch is different from the prototype patch, the colour consistency and boundary consistency of the patch are then examined with its 4-connected neighbourhood patches. Demonstrated in *Fig. 3.5*, these patches are compared with two neighbourhood patches inside. To examine these patches, following rules are applied:

- Vertical appearance consistency: a patch will be exempt if it lose its colour consistency to its vertical inner side prototype patches of upper-body. The ones above the prototype patches are recognised as head or background. The ones under are recognised as lower-body.
- Horizontal appearance consistency: if a patch loses the colour consistency to its horizontal inner side neighbour patch of upper-body, the patch will be tolerated as upper-body for there may exist another piece of clothes on the top (cardigan over a top for example). Examining patches in the outside circle, if another patch on the row loses its colour consistency to its horizontal inner side neighbour patch, this patch will be coded as background.
- Patches keep colour consistency with its inner neighbourhood patches which have been assigned as background will be recognised as background patches as well.

Learning the appearance of lower-body: the prototype patches are recognised as the ones below upper-body. To locate the prototype patches, boundaries between upper-body and lower-body are detected by comparing the patches in the perspective lower-body area with their neighbourhood patches above. Patches below the last rows of patches of upper-body are recognised as the prototype patches of lower-body. Patches below and surrounding the prototype patches are compared with the prototype patches. Patches that have different descriptions with prototype patches are further compared with its neighbourhood patches. The appearance of lower-

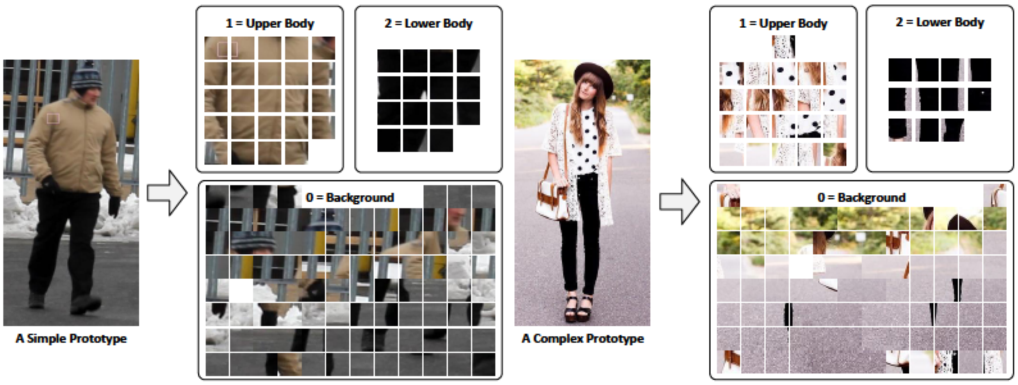


(a) Upper-Body (b) Lower-Body

Fig. 3.5 **Illustration of judgement from inner to outer part of a rough interested parts.**

body is affected by the movement of legs. To decide whether a patch in the perspective lower-body area belongs to lower-body, the patch is compared to its neighbourhood patches inside and above. Similarly, lower-body may have two parts of appearances (e.g., shorts, skirts with stocks or bare legs). If a patch loses its colour consistency to its neighbourhood patches which have been judged as lower-body, it will be tolerated. On the same column of the patch, other patches lose its colour consistency to its neighbourhood patches will be exempt as background. *Fig.* 3.5 demonstrates the rules used in learning the appearances of upper-body and lower-body.

Using the above strategy, the result 3-word dictionaries of the previous simple and complex PROTOTYPE are demonstrated in *Fig.* 3.6. Due to the complexity of appearances introduced by skin-hair and intersection area, dictionaries with 4 words or more have not been considered at this stage.



(a) Simple Prototype (b) Complex Prototype

Fig. 3.6 **3-Word Dictionaries built using the automatic way.**

3.2.3.4 The PDFs of Interested Cognitive Parts

The PDF Matrices of parts of interest reflect the observations of pedestrians detected using HOG or similar algorithm using sliding windows to obtain ROIs. The matrices are used to calculate the confident measurements of patches in TARGET. When the TARGET are separated in patches, the probabilities of pixels in one patch are recognised as the same. Then the PDFs of parts of interest are calculated in $M \times N$ matrices. $M \times N$ is the number of patches separated in TARGET. In experiment, M is usually the twice of N as the length of the bounding boxes is usually twice the width of the bounding boxes. 2D Gaussian distribution is chosen to model the PDFs of upper-body ($P_{upper\text{-}body}$), lower-body ($P_{lower\text{-}body}$), skin-hair ($P_{skin\text{-}hair}$) and background ($P_{background}$). The parameters of 2D normal distribution:

$$P_{parts}(x, y) = \exp\left\{-\frac{x - x_c}{\sigma_x^2} - \frac{y - y_c}{\sigma_y^2}\right\} \tag{3.7}$$

(x_c, y_c) and σ_x, σ_y are parameters to be determined in the following steps:

- Selecting the centre (x_c, y_c) of the interested cognitive part. For upper-body, lower-body and background, only one centre is used while for skin-hair, multi centres are used as shown in *Fig. 3.8*. The coordinates of the centres within $M \times N$ matrix are the means of normal distribution in each dimension;

- Selecting the range of confidence. Patches within the range of confidence have larger confident values than patches outside. This is determined by the variations (σ_x and σ_y) of normal distribution in each dimension. In experiments, assumptions are made that all patches within the areas of a part of interest should have confident measurements over 0.7. It means, all patches within the areas of interest should be located with in the 0.8σ area of the normal distribution as shown in *Fig. 3.7*. This assumption also guarantees the probability of patches of upper-body / lower-body will not attenuated too fast.

Practically, the centres of upper-body and lower-body are selected according to the ratio = $\frac{\text{Length of } \textit{upper-body}}{\text{Length of } \textit{lower-body}}$; the centre of background is located in the centre of the detection results; and the centres of skin-hair are located in the centres of the areas bounded by dashed lines in *Fig. 3.8(c)*.

The ranges of confidence of each interested parts are selected according to observations. And the variations of relevant PDFs in either dimensions are determined accordingly. *Fig. 3.8* demonstrates the predefined area of confidence of each part and the relevant 3σ confident range the PDF. The σs are selected to guarantee the confident measurements of patches within the range of confidence are over 0.7. Calculate

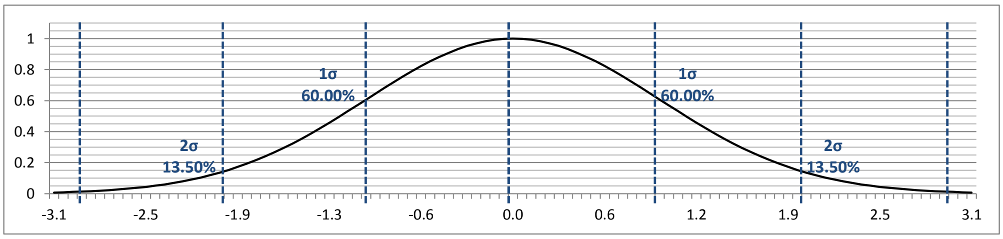


Fig. 3.7 **2D Normal Distribution is used to model the PDFs of parts of interest. The coordinates of the centres of the parts of interest are selected as the means of the Normal Distributions in each dimension. The range of the area of interest is bounded within the 0.8σ range of the Normal Distribution in each dimension. As a result, the confident measurements of patches within the area are over 0.7.**

on $M \times N$ matrix, the PDFs of *upper-body, lower-body, background* and *skin-hair* are:

$$P_{\textit{upper-body}}(x,y) = \exp\left\{-\frac{(x-N/2)^2}{(2M/3)^2} - \frac{(y-(1-\text{ratio})\cdot M/2)^2}{(\text{ratio}\cdot 2M/3)^2}\right\} \tag{3.8}$$

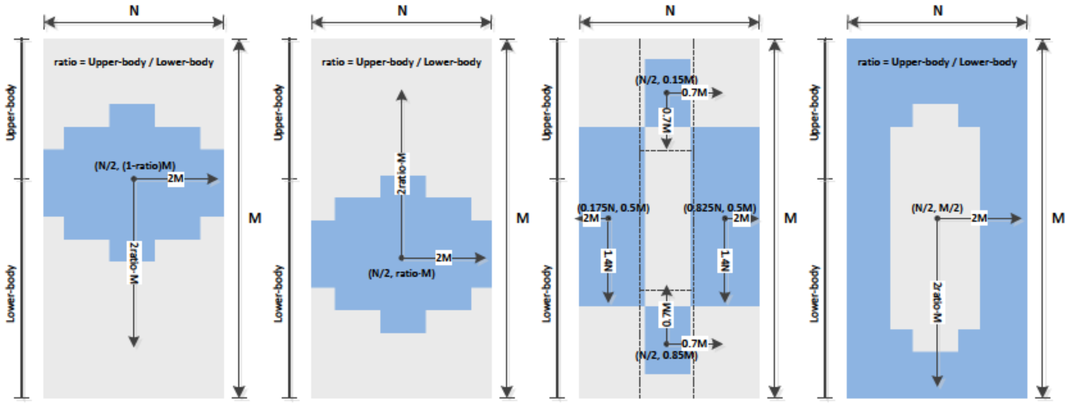
$$P_{\textit{lower-body}}(x,y) = \exp\left\{-\frac{(x-N/2)^2}{(2M/3)^2} - \frac{(y-\text{ratio}\cdot M/2)^2}{(\text{ratio}\cdot 2M/3)^2}\right\} \tag{3.9}$$

$$P_{\textit{background}}(x,y) = 1-\exp\left\{-\frac{(x-N/2)^2}{(2M/3)^2} - \frac{(y-M/2)^2}{(\text{ratio}\cdot 2M/3)^2}\right\} \tag{3.10}$$

$$P_{\textit{skin-hair}}(x,y) = \begin{cases} \exp_{\substack{x=0,\dots,N-1\\ y=0,\dots,0.25M-1}} \left\{-\frac{(x-0.5N)^2}{(0.7M/3)^2} - \frac{(y-0.15M)^2}{(0.7M/3)^2}\right\} & \textit{Head} \\ \exp_{\substack{x=0,\dots,N/2-1\\ y=0.25M,\dots,0.7M}} \left\{-\frac{(x-0.175N)^2}{(1.4N/3)^2} - \frac{(y-0.5M)^2}{(2M/3)^2}\right\} & \textit{Left Arm} \\ \exp_{\substack{x=0.5N,\dots,N-1\\ y=0.25M,\dots,0.7M}} \left\{-\frac{(x-0.825N)^2}{(1.4N/3)^2} - \frac{(y-0.5M)^2}{(2M/3)^2}\right\} & \textit{Right Arm} \\ \exp_{\substack{x=0,\dots,N-1\\ y=0.7M+1,\dots,M-1}} \left\{-\frac{(x-0.5N)^2}{(0.7M/3)^2} - \frac{(y-0.85M)^2}{(0.7M/3)^2}\right\} & \textit{Legs} \end{cases} \tag{3.11}$$

where, $x = 1,2,\dots,N-1$ and $y = 1,2,\dots,M-1$

Above PDFs basically match the observations of these parts in detected pedestrian using popular pedestrian detectors including HOG. Only the PDFs of upper-body, lower-body and background are used in experiments in this thesis, the PDF of skin-hair is provided for generic cases.



(a) Upper-Body (b) Lower-Body (c) Skin-Hair (d) Background

Fig. 3.8 **Using 2D Gaussian distribution to model the PDF of upper-body, lower-body, skin-hair and background. Black dots are the centres of Gaussian distributions and arrows demonstrate the 3σ confidence range of the Gaussian distribution. The areas in blue are the ranges of confidence of interested parts.**

3.2.3.5 Appearance based Pedestrian Identification: Summary

As a summary, the following steps show the routine of building and utilising this framework to identify the reappeared prototype pedestrian:

Preprocessing the (PROTOTYPE):

- Calculate the Probability Density Matrices of parts of interest $P_{parts}(x, y)$: upper-body, lower-body, background, skin-hair using *Equations* 3.8, 3.9, 3.10, 3.11;
- Build the Dictionary:
  1. Choose words and their representing codes: 0 for background, 1 for upper-body and 2 for lower-body for example;
  2. Separate the PROTOTYPE into $M \times N$ equal patches and give each patch a code of word manually or using the algorithm in *Section* 3.2.3 (As illustrated in *Fig.* 3.2 and *Fig.* 3.6);
  3. Record the histogram descriptions of every patch in HSV colour space using 15 bins in the hue channel and 10 bins in the saturation channel. (The number of bins used in descriptors can be changed, see *Section* 3.3.1 *for details.*)

Compare TARGETs **with the** PROTOTYPE:

- Calculate the word map of TARGET $C(x, y)$:

1. Separate the TARGETs into $M' \times N'$ equal patches and calculate the histogram description of each patch in HSV colour space;
2. Calculate the correlation distances between the histogram descriptions of patches in TARGET and patches of words in the dictionary. The similarity of a patch to a word ($\text{Similarity}_{\text{word}}$) is measured by the averaged distances between the patch and patches labelled with the word in dictionary;
3. Select the largest similarity measurement, if the measurement is over the pre-defined threshold ($T_c = 0.2$ in experiment), the patch is labelled using the word that achieves this similarity measurement. Otherwise, label the patch as background.

- Making a judgement:

1. Calculate the map of confident measurements of TARGET ($F(x, y)$) using the PDFs of parts of interest: $Similarity_{(x,y)} = PDF_{part}(x, y)$, a general case is shown in *Equation 3.2*;
2. Calculate the similarities of parts of interest ($\text{Similarity}_{\text{parts}}$) using *Equation 3.4*;
3. Calculate the total similarity ($\text{Similarity}_{\text{Total}}$) by combining the $\text{Similarity}_{\text{parts}}$ using *Equation 3.5*. If $\text{Similarity}_{\text{Total}}$ is over the pre-defined threshold ($T_j = 0.6$ in experiment), the TARGET is judged as the reappeared PROTOTYPE or passed to the next layers of judgemental algorithms when necessary.

3.3 Experiment

In this section, re-identification experiments are carried on the layer of histogram descriptions. The following will be demonstrated: parameters used in the fundamental histogram layer including the means of comparing two histogram, the number of words used in constructing the dictionary and the number of patches used in the PROTOTYPE and TARGETS.

3.3.1 Parameters: The Number of Words and Patches

In our experiment, two prototype pedestrians are chosen as one simple and one complex model. The simple pedestrian wears mono-coloured clothes, top and bottom (as shown in *Fig. 3.2(a),(c)*) and the complex model wears a white based black dotted top partially covered by a white cardigan and a pair of black trousers. The pedestrian also has a significant amount of hair / skin exposure (as shown in *Fig.*

3.2(b),(d)). For both PROTOTYPEs. Versions of dictionaries are built using all four kinds of vocabularies (3 / 6 / 4 / 10-word, refer to *Fig. 3.2*). The number of patches separated in the PROTOTYPE is from 3 × 6 to 16 × 32 pieces. As demonstrated in *Fig. 3.9*, for the identification of the simple prototype example, 3-word dictionary results in the best performances, followed by the 4-word one and the 6 / 10-word dictionaries have less stable performances than the other two especially when the number of patches are limited (< 4) or overwhelmed (>12). This is also observed from the experiment results of using the complex prototype as shown in *Fig. 3.10*. Increasing the number of words used in the dictionary only improve the accuracy by 2 - 3% in either case.

Effectively increasing the number of patches used in separating the PROTOTYPE and TARGETs results in more stable and accurate performances, while overwhelmed separation also introduces problems. Observed from *Fig. 3.9*, the peak performance reaches at when the number of patches separated in TARGETs are between 7×14to13×26 pieces no matter how many patches are used to separating the PROTOTYPE. Too less or too more may result in unstable performances and high error rates in re-identification. For the complex prototype example, as demonstrated in *Fig. 3.10*, the performance of re-identification is improved with increased number of patches separated in both PROTOTYPE and TARGETs. After these images are all separated into more than 7 × 14 patches, on average, 80% of TARGET images can be correctly identified. Errors are mainly introduced by the TARGETs who are wearing clothes with a similar hue but different pattern comparing to the PROTOTYPE (see Table 3.2). The peak performances comes at when the number of patches used in PROTOTYPE and TARGETs are 14 × 28 or 15 × 30 pieces.

Insufficient pieces separated from the PROTOTYPE or TARGETS means the word may not be accurately represented and labelled to patches while finely divided patches may introduce noise pieces situated within drapes / shades of clothes. According to current stage of experiments, to identify a pedestrian wearing mono-coloured clothes, a separation of 8×16 pieces of patches in the PROTOTYPE and TARGETs and labelling them using a 3-word dictionary achieves promising result: in the testing sets containing 500 images, more than 95% are correctly identified. To identify a PROTOTYPE with complex appearances, more pieces of patches should be divided in both PROTOTYPE and TARGETs to achieve accurate and stable performances. When 9 × 18 patches are separated from both PROTOTYPE and TARGETs, ~85% images can be correctly judged. More patches separated in PROTOTYPE and TARGETs only improve the identification rate by ~5%. Over separation, when the images of detected pedestrians are divided into 16×32 pieces or more, the performances of re-identification will be unstable. This is because the sizes of each separated pieces are too small to be cognitively described and categorised.[2] Furthermore, when the images are finely divided, the re-identification require more processing time, the processing time spent on comparing the PROTOTYPE and TARGETs separated into 16 × 32 patches will be 16 times more than the case when 8 × 16 patches are separated in those images and 256 times more than the case when only 4 × 8 are divided.

To compare the patches in TARGETs and in the dictionary, the correlation distance performs the best among popular selected histogram comparison methods, such as χ^2 distance, intersection distance. The Bhattacharyya distance has similar performance with the correlation distance. The threshold ($T_c = 0.2$) used in labelling the patches in TARGET is selected small to guarantee the detection rate. Using 15 bins in Hue and 10 bins in Saturation, $T_c > 0.3$ will result in ~40% reduction in the detection rate and $T_c < 0.15$ will increase the false alarm rate by ~20%. More bins used in histogram will not affect the identification result, while less bins will reduce the identification rate of pedestrians (wearing either mono-coloured clothes or patterned clothes). Using 8 bins in Hue and 5 bins in Saturation reduce the identification rate by around 15%. Threshold (T_j), which is to determine if the TARGET the same as the PROTOTYPE, is selected 0.6 when 0.7 is chosen as the minimum confident measurements of parts of interest as described in *Section 3.2.3.4*.

3.4 Summary

Pedestrian Re-identification applied to surveillance means to retain the identity of pedestrians being tracked / monitored. The applications of pedestrian re-identification associate pedestrians in video frames and multi-viewed camera system. Instead of identifying a unique person, re-identification in this chapter focused on the recognition of different occurrences of one pedestrian. The differences of appearances of people are referred to the differences of their clothes rather than their sizes and shapes of figures or other traits which may be important for human to recognise a person. It means, within a short interval the same person is assumed to appear in clothes with similar appearances: tops and bottoms.

In this chapter, dictionary representing the prototype is constructed manually or automatically. The automatic ways rely on the PDFs of body parts. Manually constructed dictionary may be more accurate comparing to the automatic one. There is no restriction to the size of the vocabularies. Experiment results show that the size

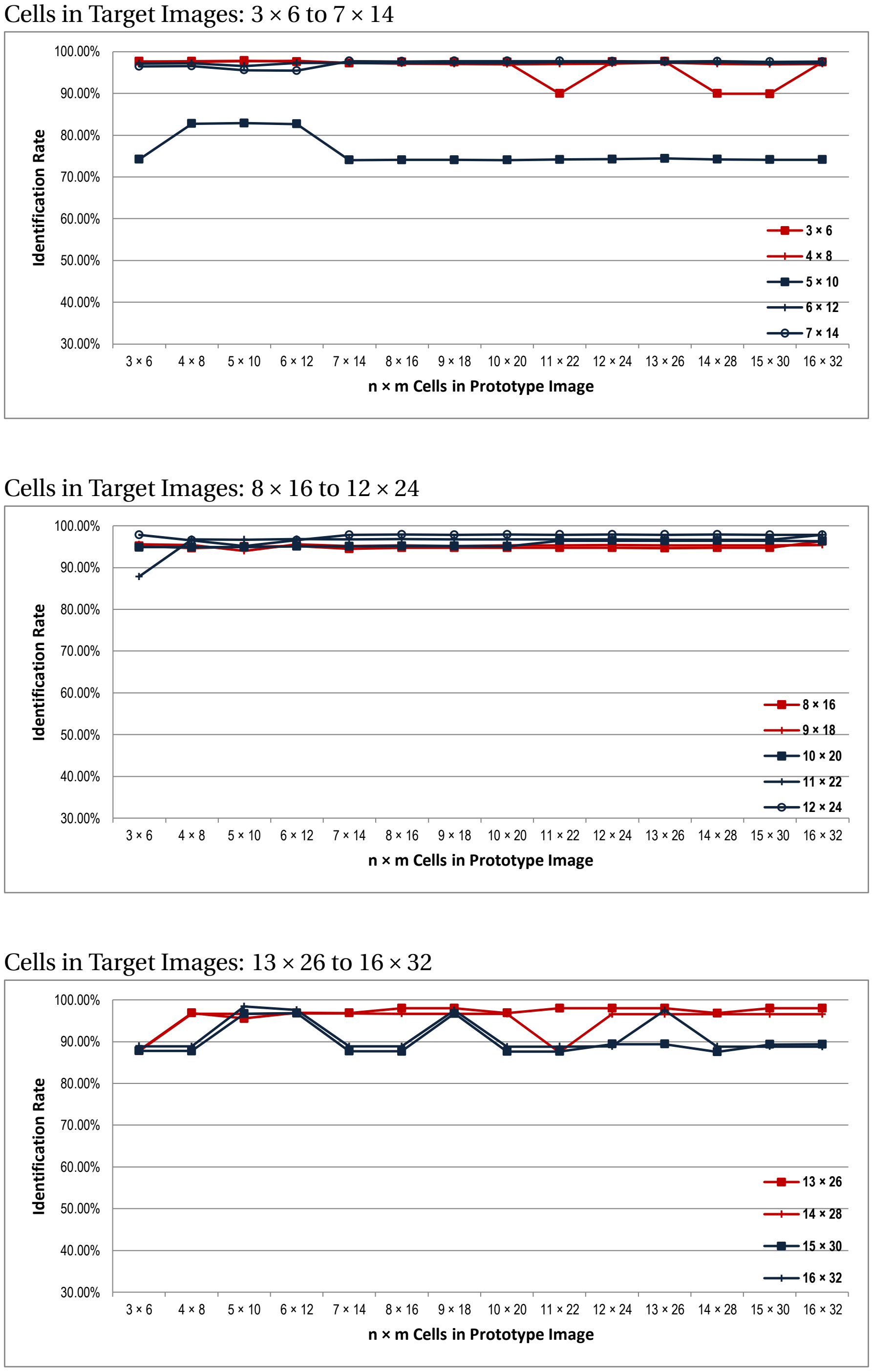
[2] For pedestrians detected by HOG, the size of bounding box is usually around 100 × 200pixel^2. 16 × 32 patches means that each patch contains ~30 pixels only.

of vocabulary may not be a high influenced parameter compare to the number of patches used to divide the prototype and target images. On average, 3-words (upper-body, lower-body and background) dictionary in identification achieve best results. This is because, extra words representing either intersections of parts of interest or exposed skin / hair occupy limited area within the images. The descriptions of such words containing few patches are less reliable comparing with the descriptions of words containing a bigger set of patches.

Local descriptions of patches guarantee the scaling and transportation invariant of the framework. This is important as pedestrians detected in bounding box may not always tightly bounded in the centre of the bounding box. Applying the pyramid protocol, reduced calculations will be spent on the identification of pedestrians wearing mono-coloured clothes than the identification of the ones wearing patterned clothes. Using HSV histogram (especially the hue channel) for colour information, the algorithm achieve short processing time. For the identification of pedestrian wearing mono-coloured top and bottom, the separation of 4×8 patches in the PROTOTYPE and TARGETs is recommended. For pedestrian wearing patterned clothes, 10×20 patches should be separated in images to obtain stable identification (error rate $< 10\%$). The number of patches used in separating the TARGET can be different with it is used in building the PROTOTYPE dictionary.

Current experiments do not consider pedestrians with complex appearances beyond regular patterns including stripes, dots and lattices. The lack of geometry information in histogram based description means that the current re-identification algorithm may fail in separating different patterns with the same combination of colours. As observed in *Table* 3.2, pattern judgement is crucial to improve the performances. Lighting condition and the interference of appearances introduced by accessories of pedestrians will be addressed in future work.

Simple PROTOTYPE: mono-colour top & bottom

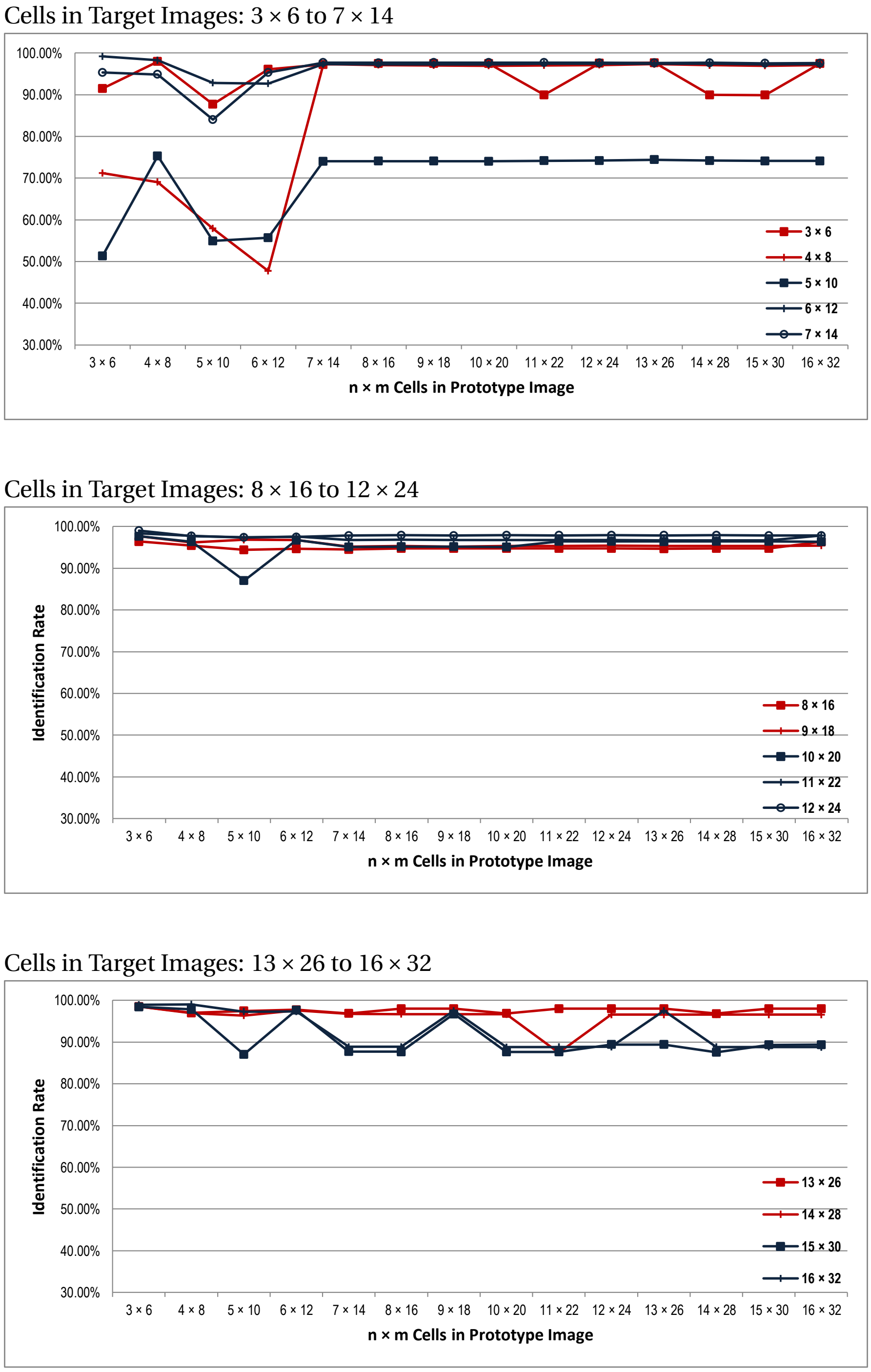

Cells in Target Images: 3 × 6 to 7 × 14

Cells in Target Images: 8 × 16 to 12 × 24

Cells in Target Images: 13 × 26 to 16 × 32

(a) Using 3-code Dictionary: 0 1 2

Fig. 3.9 **The effects of using different number of words in Dictionary, different numbers of patches to separating the PROTOTYPE and TARGETs in a simple prototype re-identification example. 0 1 2 represent background, upper-body and lower-body respectively.** 5 × 10 **patches have worse performances than others. This also happen with other PROTOTYPE examples. It maybe because the odd number 5 disturbs the appearances of patches within the PROTOTYPE when pedestrian is an approximated symmetry object.**



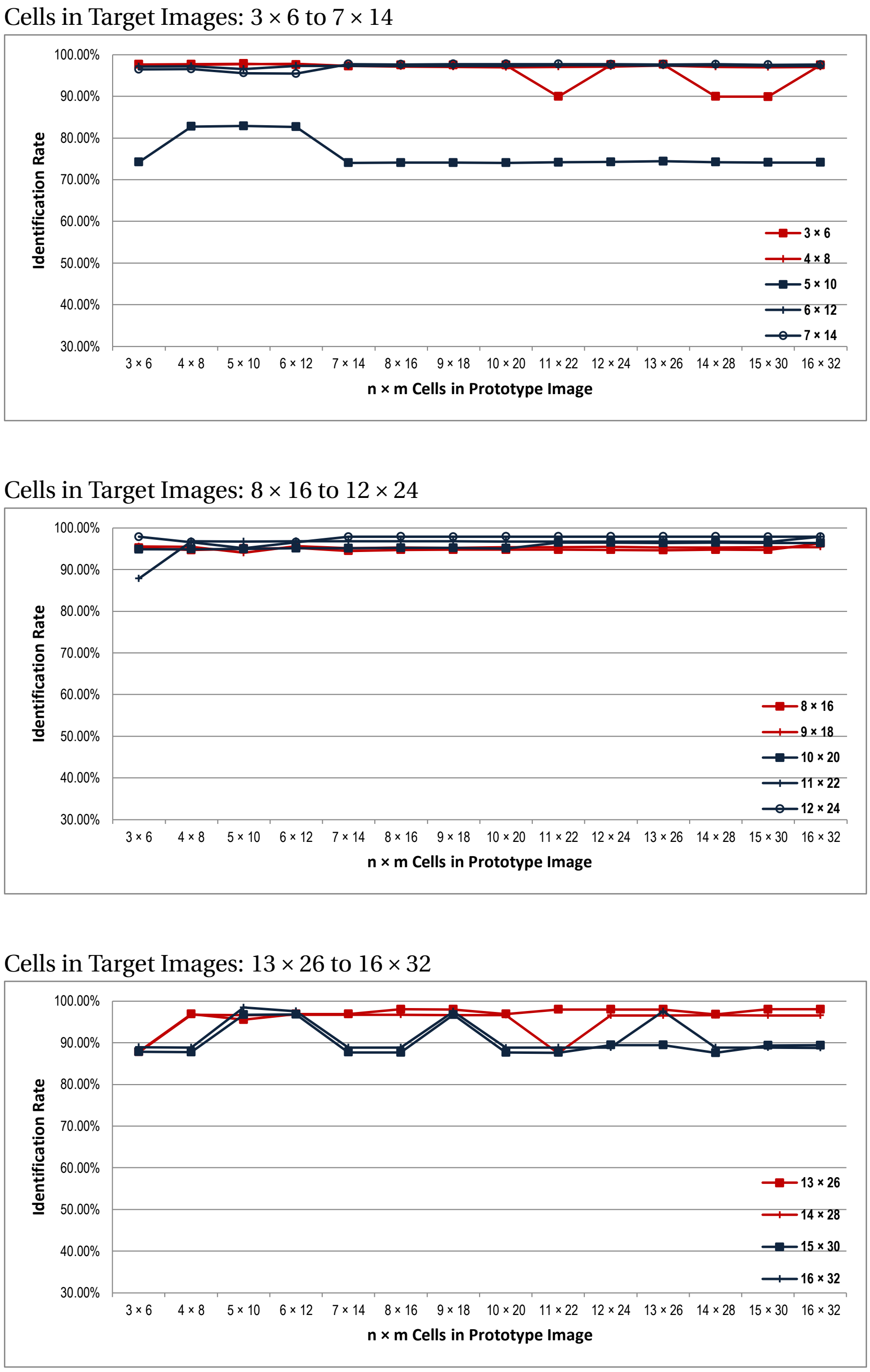
Cells in Target Images: 3 × 6 to 7 × 14

Cells in Target Images: 8 × 16 to 12 × 24

Cells in Target Images: 13 × 26 to 16 × 32

(b) Using 6-code Dictionary: 0 1 2 11 12 22

Fig. 3.9 **(Cont.) Simple** PROTOTYPE: **mono-colour top & bottom, where the codes are represented by: 0: background, 1: upper-body& background, 2: lower-body& background, 11: upper-body, 12: upper-body& lower-body, 22: lower-body.**



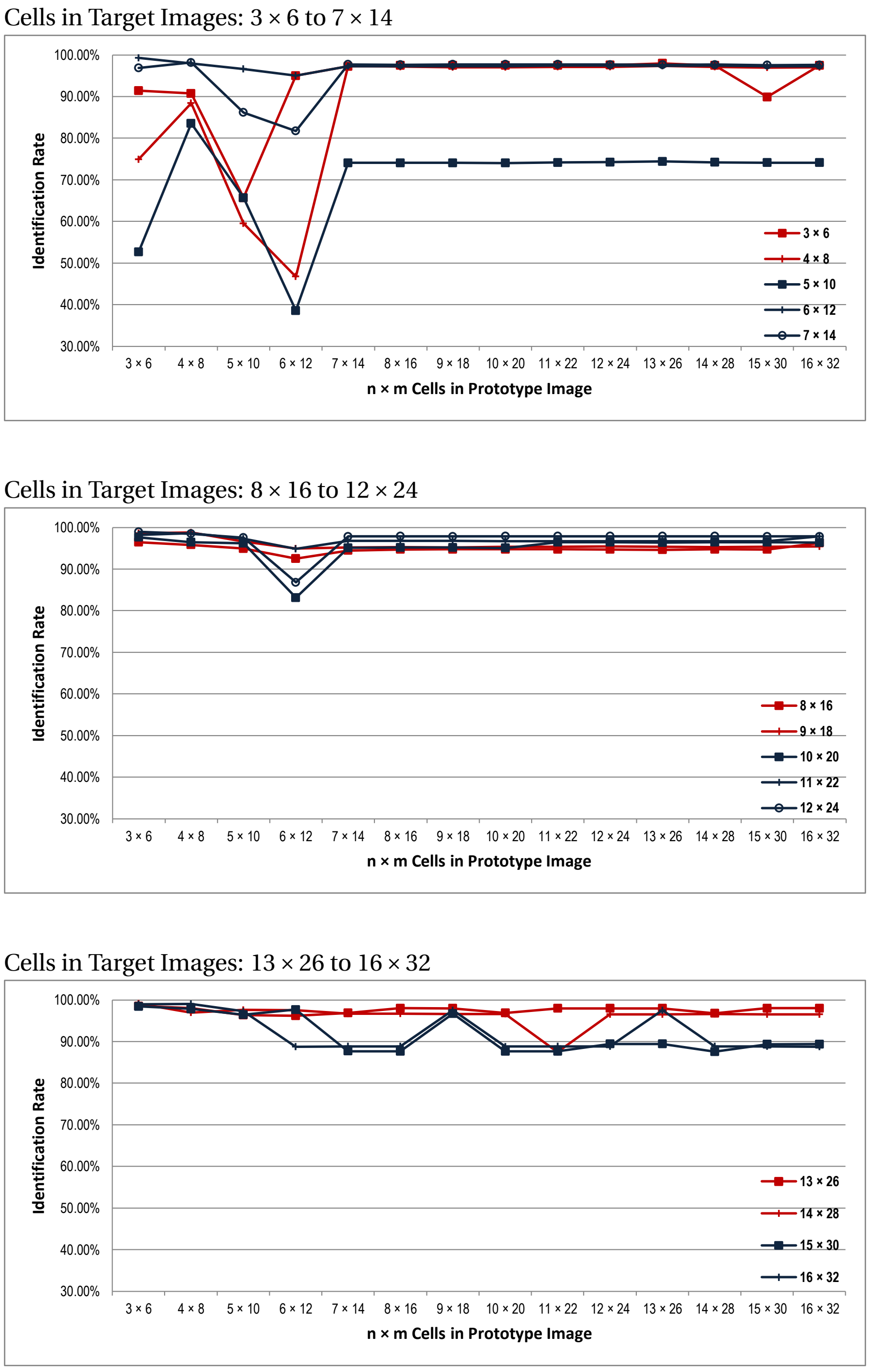
Cells in Target Images: 3 × 6 to 7 × 14

Cells in Target Images: 8 × 16 to 12 × 24

Cells in Target Images: 13 × 26 to 16 × 32

(c) Using 4-code Dictionary: 0 1 2 3

Fig. 3.9 **(Cont.) Simple PROTOTYPE: mono-colour top & bottom, where the codes are represented by: 0: background, 1: upper-body, 2: lower-body, 3: skin-hair.**



Cells in Target Images: 3 × 6 to 7 × 14

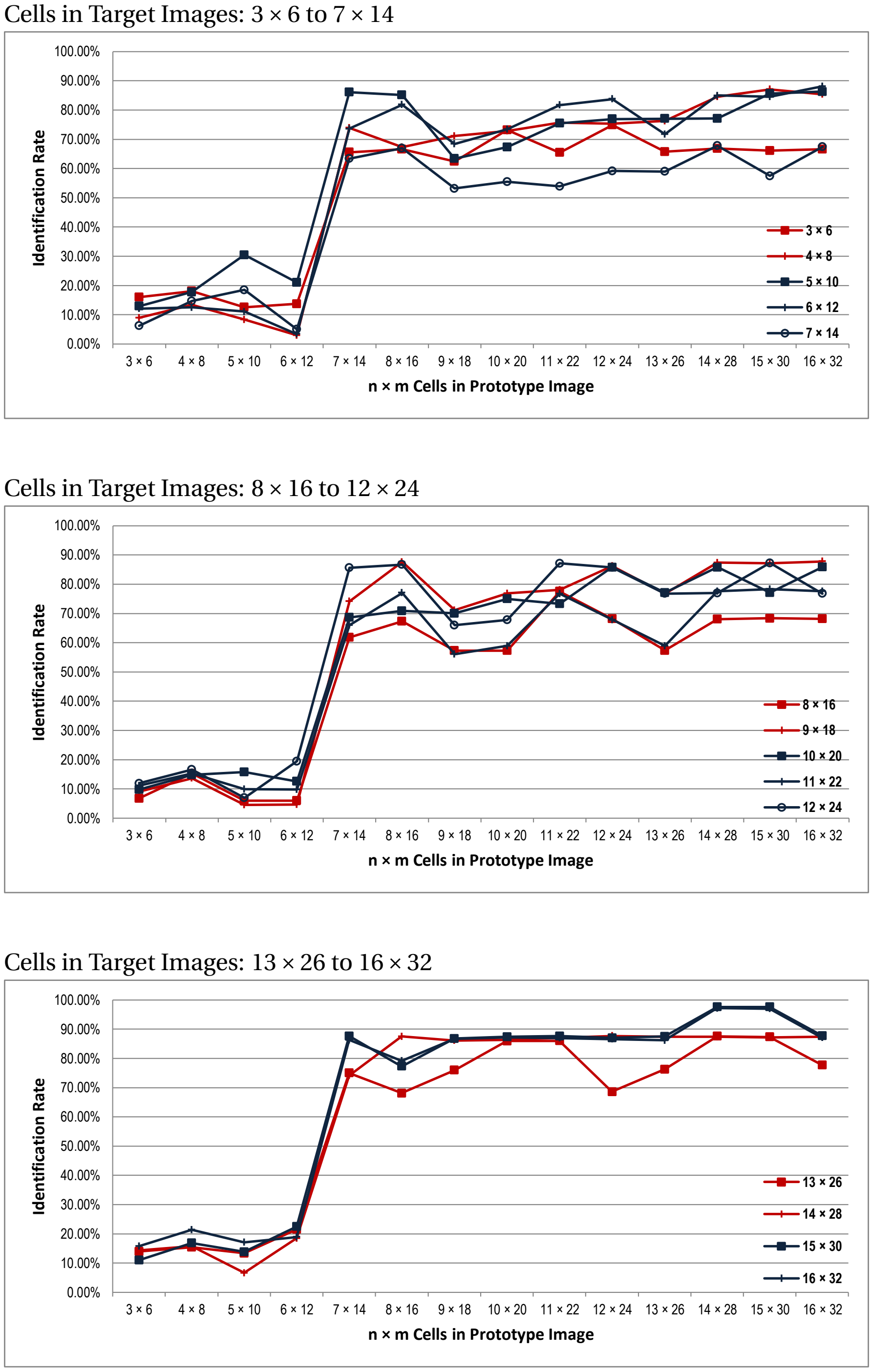
Cells in Target Images: 8 × 16 to 12 × 24

Cells in Target Images: 13 × 26 to 16 × 32

(d) Using 10-code Dictionary: 0 1 2 3 11 12 13 22 23 33

Fig. 3.9 **(Cont.) Simple** PROTOTYPE: **mono-colour top & bottom, where the codes are represented by: 0: background, 1: upper-body& background, 2: lower-body& background, 3: skin-hair& background 11: upper-body, 12: upper-body& lower-body, 13: upper-body& skin-hair, 22: lower-body, 23: lower-body& skin-hair, 33: skin-hair.**

Complex PROTOTYPE: patterned top with cardigan

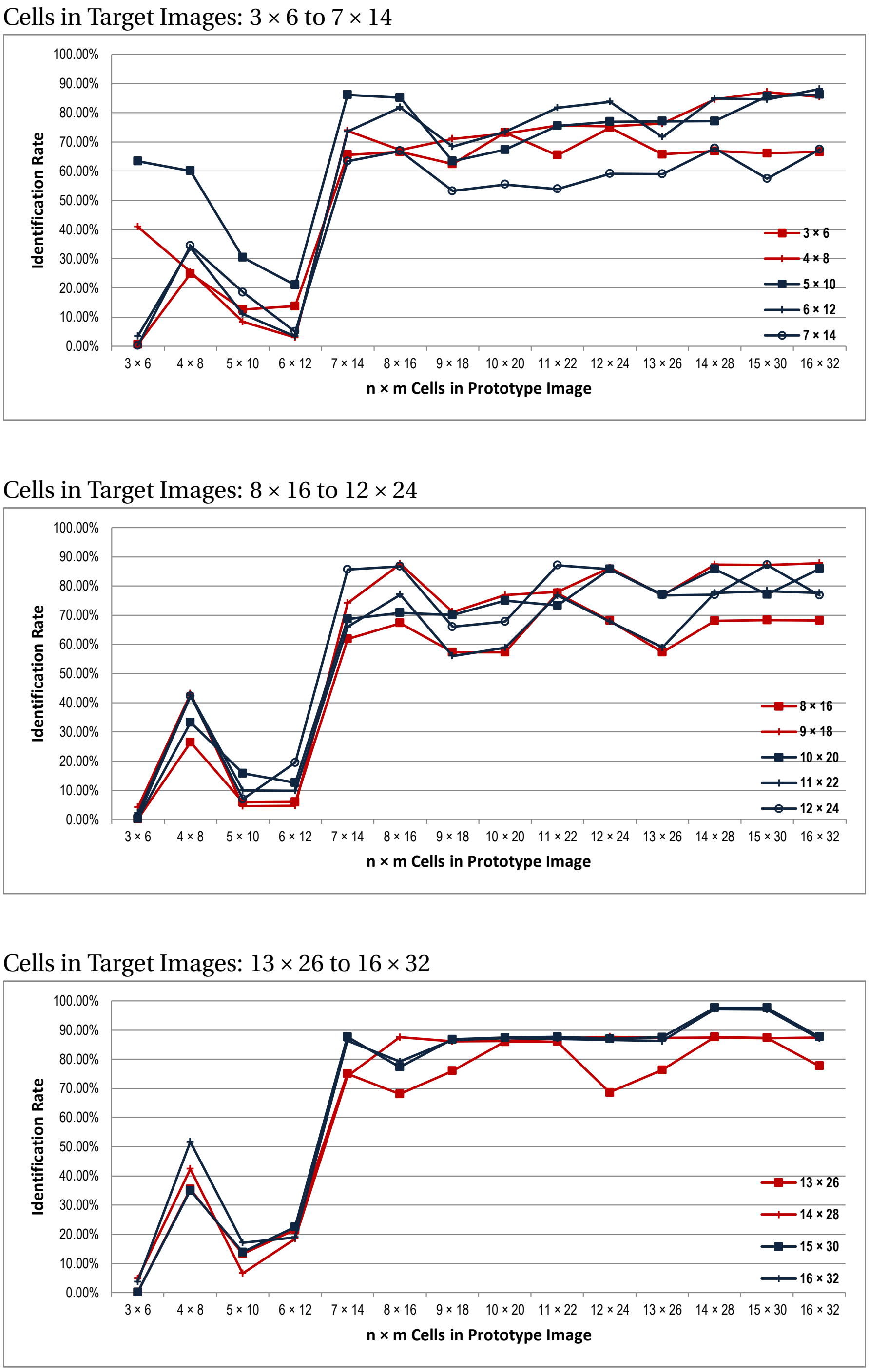

Cells in Target Images: 3 × 6 to 7 × 14

Cells in Target Images: 8 × 16 to 12 × 24

Cells in Target Images: 13 × 26 to 16 × 32

(a) Using 3-code Dictionary: 0 1 2

Fig. 3.10 **Effects of using different number of words in Dictionary, different numbers of patches to separating the PROTOTYPE and TARGETs in a complex prototype re-identification example. 0 1 2 represent background, upper-body and lower-body respectively.**

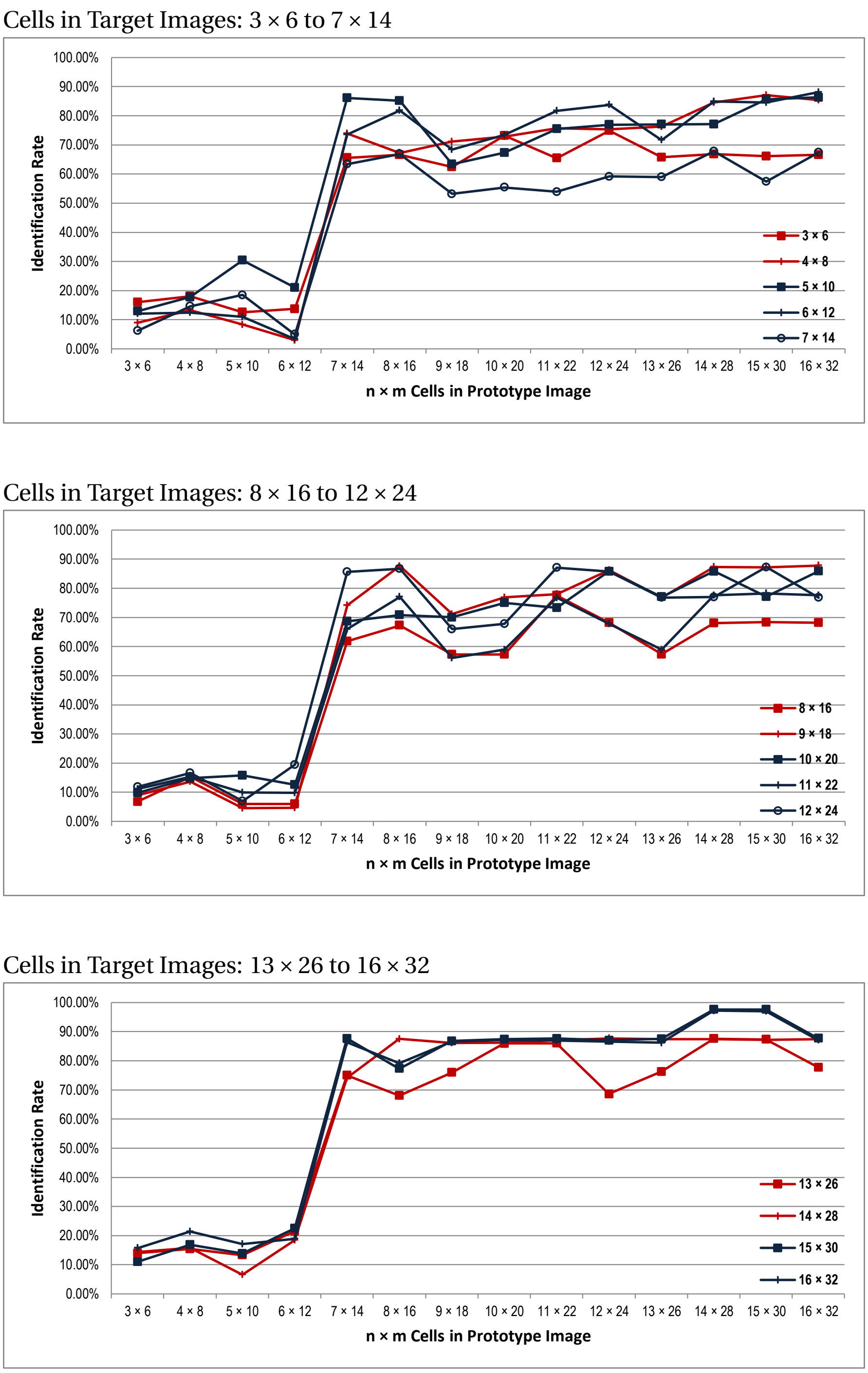

Cells in Target Images: 3 × 6 to 7 × 14

Cells in Target Images: 8 × 16 to 12 × 24

Cells in Target Images: 13 × 26 to 16 × 32

(b) Using 6-code Dictionary: 0 1 2 11 12 22

Fig. 3.10 **(Cont.) Complex PROTOTYPE: patterned top with cardigan, where the codes are represented by: 0: background, 1: upper-body& background, 2: lower-body& background, 11: upper-body, 12: upper-body& lower-body, 22: lower-body. The peak appear at the bottom figure when** 4 × 8 **patches are separated in PROTOTYPE not usually happen with other prototype. This is due to the frequency of the repeated pattern. Such peaks may locate in other place when not enough patches are separated in PROTOTYPE with different pattern repetition rates.**

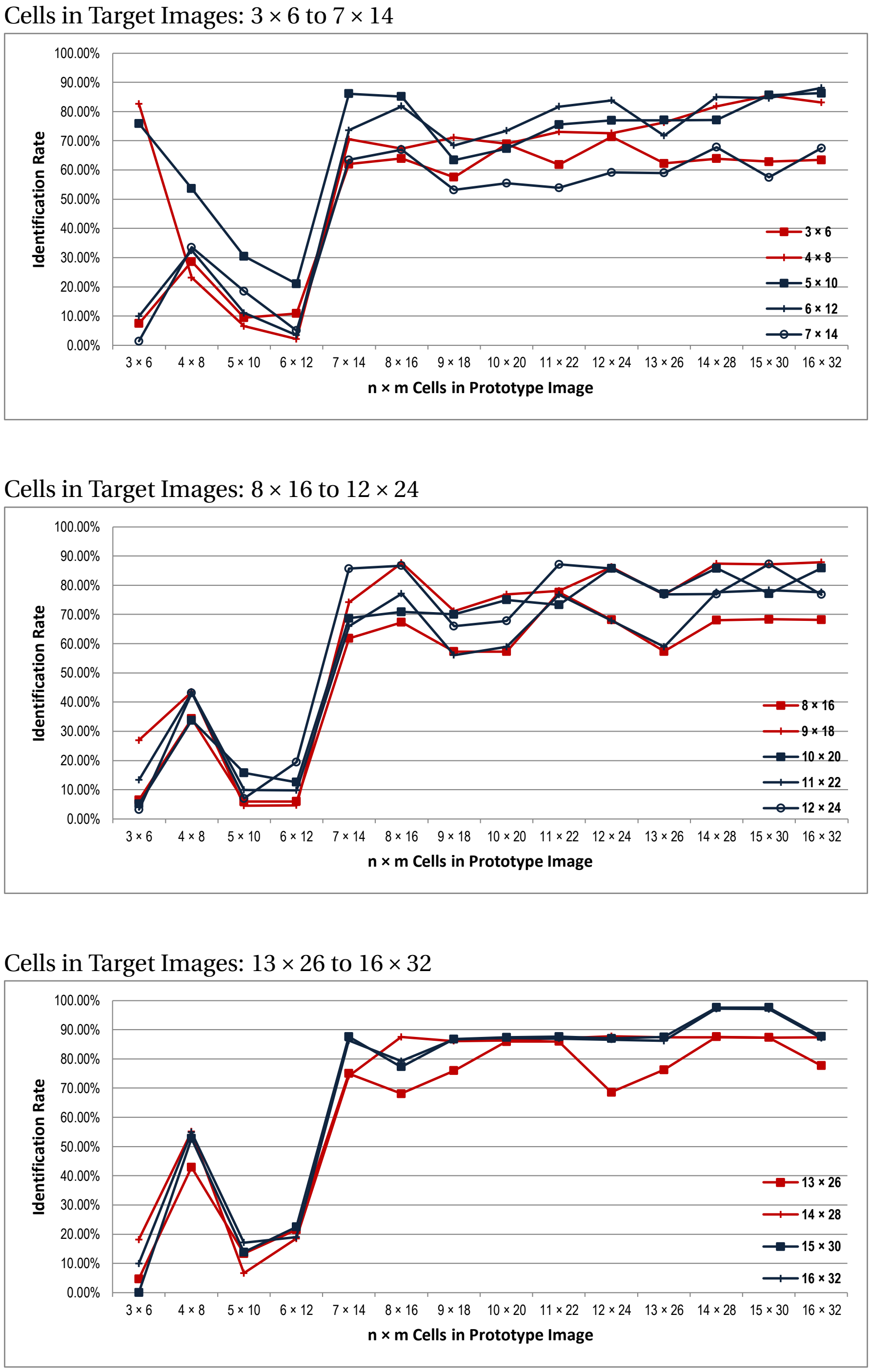

Cells in Target Images: 3 × 6 to 7 × 14

Cells in Target Images: 8 × 16 to 12 × 24

Cells in Target Images: 13 × 26 to 16 × 32

(c) Using 4-code Dictionary: 0 1 2 3

Fig. 3.10 **(Cont.) Complex** PROTOTYPE: **patterned top with cardigan, where the codes are represented by: 0: background, 1: upper-body, 2: lower-body, 3: skin-hair.**


Cells in Target Images: 3 × 6 to 7 × 14

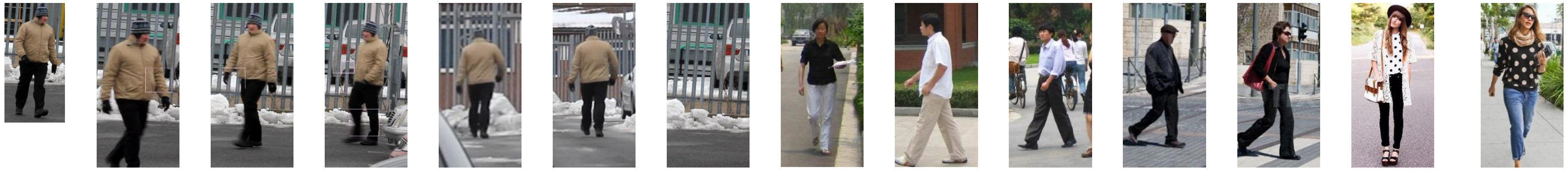
Cells in Target Images: 8 × 16 to 12 × 24

Cells in Target Images: 13 × 26 to 16 × 32

(d) Using 10-code Dictionary: 0 1 2 3 11 12 13 22 23 33

Fig. 3.10 **(Cont.) Complex** PROTOTYPE: **patterned top with cardigan, where the codes are represented by: 0: background, 1: upper-body& background, 2: lower-body& background, 3: skin-hair& background 11: upper-body, 12: upper-body& lower-body, 13: upper-body& skin-hair, 22: lower-body, 23: lower-body& skin-hair, 33: skin-hair.**

(a) Simple PROTOTYPE, 3-Word Dictionary



315 × 630	315 × 630	325 × 650	351 × 702	105 × 210	174 × 348	325 × 650	170 × 339	157 × 313	175 × 350	109 × 218	184 × 368	279 × 560	221 × 511
3 × 6	0.60	0.57	0.56	0.53	0.50	0.26	0.29	0.00	0.25	0.22	0.29	0.23	0.26
4 × 8	0.61	0.59	0.55	0.53	0.55	0.28	0.21	0.06	0.25	0.30	0.29	0.26	0.19
5 × 10	0.67	0.69	0.65	0.27	0.30	0.34	0.14	0.00	0.29	0.00	0.28	0.29	0.27
6 × 12	0.72	0.72	0.71	0.68	0.69	0.31	0.40	0.04	0.30	0.35	0.30	0.29	0.40
7 × 14	0.72	0.73	0.70	0.73	0.71	0.32	0.21	0.04	0.31	0.35	0.31	0.32	0.45
8 × 16	0.73	0.73	0.73	0.76	0.72	0.33	0.53	0.37	0.32	0.34	0.30	0.33	0.49
9 × 18	0.73	0.77	0.75	0.75	0.74	0.33	0.53	0.43	0.32	0.31	0.28	0.31	0.49
10 × 20	0.71	0.77	0.76	0.76	0.75	0.32	0.54	0.43	0.33	0.32	0.28	0.33	0.48
11 × 22	0.74	0.77	0.77	0.78	0.75	0.34	0.26	0.35	0.33	0.32	0.29	0.33	0.51
12 × 24	0.74	0.75	0.77	0.80	0.76	0.34	0.26	0.00	0.33	0.29	0.30	0.32	0.54
13 × 26	0.74	0.77	0.76	0.82	0.76	0.32	0.23	0.00	0.33	0.32	0.28	0.33	0.57
14 × 28	0.72	0.77	0.77	0.79	0.76	0.33	0.25	0.36	0.35	0.32	0.28	0.30	0.56
15 × 30	0.73	0.77	0.77	0.35	0.76	0.35	0.26	0.00	0.34	0.32	0.29	0.33	0.54
16 × 32	0.73	0.76	0.78	0.34	0.76	0.33	0.27	0.00	0.35	0.35	0.30	0.34	0.59

Table 3.2 **The examples of re-identification of both simple and complex prototypes. Prototype and target images are separated using the same number of patches during comparison. The value in the table are the similarity between the TARGET on the 1st row and the PROTOTYPE in the top-left corner.**

(a) Complex PROTOTYPE, 10-Word Dictionary

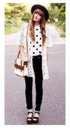
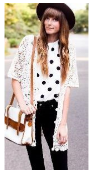
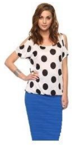
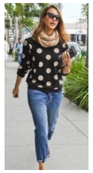
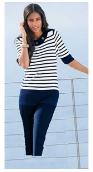
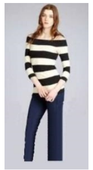
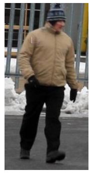
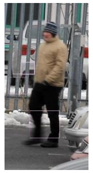
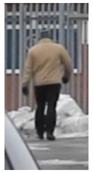
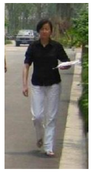
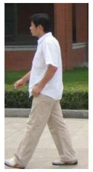
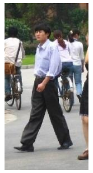
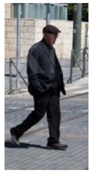
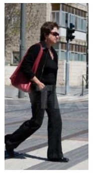
													
279 × 560	282 × 617	203 × 380	221 × 511	720 × 1468	205 × 366	315 × 630	351 × 702	105 × 210	170 × 339	157 × 313	175 × 350	109 × 218	184 × 368
3 × 6	0.66	0.31	0.18	0.29	0.22	0.31	0.49	0.29	0.29	0.25	0.56	0.45	0.49
4 × 8	0.52	0.53	0.52	0.49	0.13	0.44	0.54	0.45	0.53	0.09	0.51	0.57	0.54
5 × 10	0.20	0.18	0.32	0.28	0.00	0.12	0.16	0.10	0.33	0.25	0.12	0.00	0.22
6 × 12	0.41	0.18	0.29	0.28	0.00	0.12	0.16	0.10	0.33	0.28	0.12	0.19	0.20
7 × 14	0.56	0.54	0.35	0.49	0.22	0.26	0.26	0.45	0.19	0.30	0.62	0.27	0.28
8 × 16	0.63	0.61	0.25	0.60	0.00	0.30	0.31	0.61	0.25	0.00	0.67	0.55	0.29
9 × 18	0.62	0.58	0.19	0.57	0.28	0.32	0.49	0.33	0.23	0.34	0.70	0.51	0.29
10 × 20	0.64	0.60	0.23	0.53	0.24	0.31	0.30	0.49	0.26	0.00	0.67	0.55	0.28
11 × 22	0.66	0.62	0.24	0.67	0.00	0.34	0.32	0.55	0.27	0.00	0.69	0.33	0.29
12 × 24	0.67	0.64	0.25	0.40	0.00	0.36	0.32	0.56	0.27	0.00	0.68	0.56	0.30
13 × 26	0.66	0.62	0.23	0.40	0.30	0.34	0.31	0.61	0.24	0.00	0.71	0.34	0.28
14 × 28	0.66	0.62	0.23	0.41	0.00	0.35	0.33	0.33	0.24	0.00	0.70	0.32	0.30
15 × 30	0.68	0.65	0.23	0.41	0.00	0.36	0.33	0.35	0.26	0.00	0.35	0.33	0.31
16 × 32	0.68	0.64	0.27	0.43	0.00	0.37	0.33	0.34	0.27	0.00	0.71	0.35	0.31

Table 3.3 **(Cont.) Examples of re-identification of both simple and complex prototypes. Prototype and target images are separated using the same number of patches during comparison. It can be observed, false alarms appear when no pattern judgement is performed. The value in the table are the similarities between the TARGET on the 1st row and the PROTOTYPE in the top-left corner.**

Chapter 4

False Alarm Rate Reduction in Videos

Jedes hinreichend mächtige, rekursiv aufzählbare formale System ist entweder widersprüchlich oder unvollständig.[1]

—— *Kurt F. Gödel*

In the last chapter, the re-identification of pedestrians after a period of complete occlusion is discussed. Strategies are proposed based on HOG pedestrian detection or similar algorithms where the sliding windows are applied to obtaining ROIs. In pedestrian re-identification, the prototype and the target pedestrians are results of pedestrian detections. It means, the performances of pedestrian detections may affect the performance of pedestrian re-identification. In this chapter, a novel approach is applied to reduce the false alarms introduced by HOG pedestrian detector. After that, the framework of pedestrian re-identification introduced in *Chapter 3* will be adopted to reduce the reappeared false alarms in the results of pedestrian detection applied to video streams. All detectors used in the experiments of this chapter are trained by CV (v. 2.4.3). Before presenting the algorithms, HOG pedestrian algorithm and its OpenCV trained detector will be reviewed.

4.1 HOG Pedestrian Detector in CV (2.4.3)

HOG, is a supervision based framework for pedestrian detection. Generally, it follows the structure demonstrated in *Fig. 4.1*: train a detector using labelled descriptors of positive and negative examples (pedestrians and background) and then apply the detector to ROIs of target images to deciding the presences of pedestrians within

[1] Any effectively generated theory capable of expressing elementary arithmetic cannot be both consistent and complete. —— *Kurt F. Gödel*

the ROIs. The detected pedestrians are bounded in bounding boxes. In *Fig. 4.1*: the positive and negative training images are labelled using +1 and −1; ROIs are obtained using sliding windows; by repeating the trained detection procedures on the down sized input images the scale invariant of the detector is achieved. The trained HOG detectors are capable to detect pedestrians in real-time application though the training procedure is time consuming. As reviewed in *Chapter 2*, the performance of a supervision based detector is affected by the training process: the strategy and the dataset. The training strategy, which translates numerical features to binary judgement is essential in designing the frameworks of pedestrian detection. It is believed by researchers [3] that a larger, more complete training dataset may result a better detector. However there is no universal definition to quantify the completeness of a dataset and no observations show the relationship between the performance of the trained detector and the scale of its training dataset. A trained detector may perform differently when it is applied to different testing images. Before presenting the approach to improve the performances of HOG detector provided by CV (2.4.3), the false alarms occurred in HOG detection will be analysed in this section.

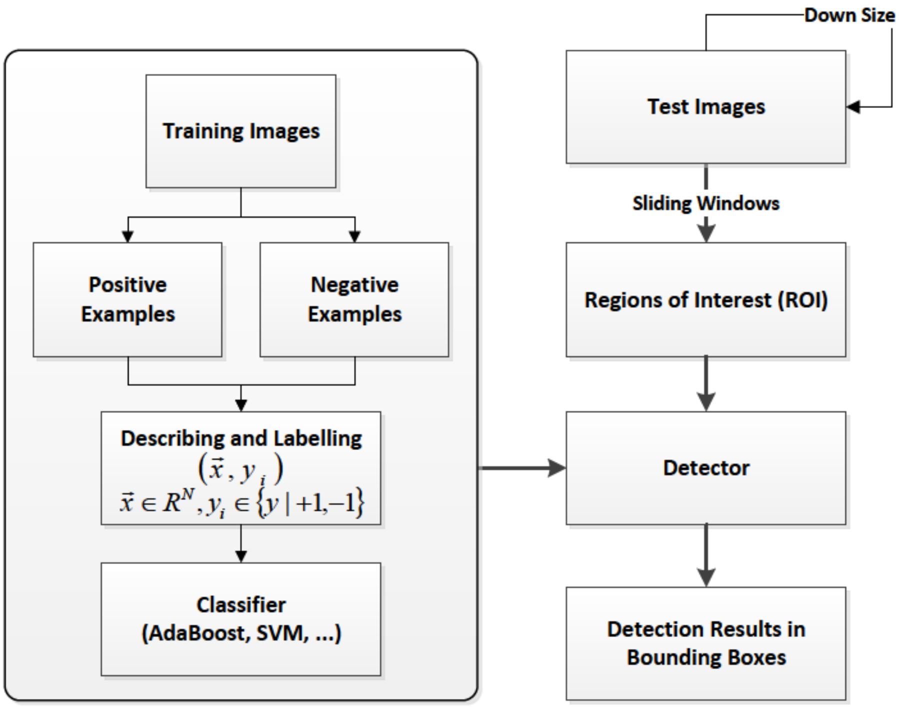

Training Images
Positive Examples
Negative Examples
Describing and Labelling
$(\vec{x}, y_i)$
$\vec{x} \in R^N, y_i \in \{y \mid +1, -1\}$
Classifier (AdaBoost, SVM, ...)
Down Size
Test Images
Sliding Windows
Regions of Interest (ROI)
Detector
Detection Results in Bounding Boxes

Fig. 4.1 **Structure of Training and Testing a HOG/Haar-like Feature based Detector**

HOG descriptions capture the cognitive shapes of pedestrians. Cognitive shape tolerates the intra-class variation of the shapes of pedestrians, which may vary due to the different poses of different pedestrians. It means the shapes of false alarms are usually similar to the shapes of the correctly detected pedestrians. To improve the performances of HOG pedestrian detection, features other than cognitive shapes of pedestrian should be considered, for instances, local features introduced by body parts. In this chapter, assumptions are made that the appearances of the head-shoulder parts of pedestrians play more important roles in deciding the presence of pedestrians than other body-parts.

4.1.1 Why HOG?

As reviewed in *Chapter 2*, HOG description is not the only method to capture the features of pedestrians and SVM is one of the algorithms for training a detector. For various purposes, pedestrian detectors were trained by various research institutes using a selections of datasets. The popular ones, such as HOG-SVM[2] and Viola-Jones[3], have been integrated in the libraries of various programming languages (C++ / C, etc.). But not many discussions were made on how to quantify the detection results. This is may be because:

- The judgement of the detection results is usually done manually. It means, the righteousness of a detection may vary from one research to another, especially when the detection results contain a part of a pedestrian.
- The detection rate is proportional to the false alarm rate. Though the desired pedestrian detector is high in detection rate and low in false alarm rate, normally, the improved detection rate may result in the increased false alarm rate and the reduced false alarm rate may sacrifice the detection rate.

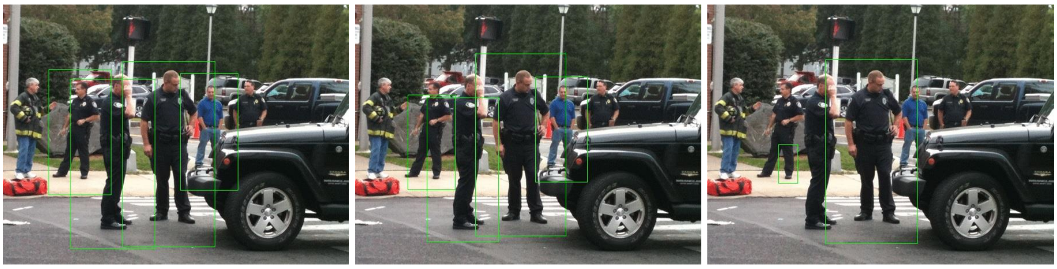
In recent studies on pedestrian detection, the performances of frameworks are evaluated using the curves of the detection rates verses the false rates, which include the DET (Detection Error Trade-off) and ROC (Receiver Operating Characteristic) curves [22, 78]. These curves are usually plotted during training. Experimentally, the prepared datasets are randomly divided in two groups: one for training and the other for testing. To calculate the DET curve, the parameters / thresholds used in the training procedure are tuned to achieve different performances of the trained

[2] HOG-SVM: the abbreviation of HOG-SVM pedestrian detector. In this chapter, HOG-SVM refer to the pedestrian detector trained by CV (2.4.3)

[3] Viola-Jones refer to the Viola-Jones pedestrian / body-part detectors trained by CV (2.4.3)

detectors applying to the testing images. Similarly, ROC curve introduced in [22] measured the decreasing rate of the percentage of detection rates with regard to the reduction of false alarm rate, which is usually measured by FPPW (False Positives per Window[4]) or FPPI (False Positive per Image)[5]. For example, to measure the performances of HOG-SVM pedestrian detector in [22], the DET or ROC curve is calculated by changing the margins between two classes of observations during SVM training. When using cascaded decision tree to train the detector, this is done by changing the number of levels and stages of weak classifiers cascaded in decision trees.

The training datasets affect the performance of pedestrian detection as well. A dataset may never complete and one well behaved detector may fail in other circumstances. In [78], a new term, "confidence", was defined to evaluate the consistencies of frameworks when apply to different datasets. [3] compared several large pedestrian databases and concluded that among these databases, INRIA (followed by Daimler-DB) outperform others in both image quality and result stability.



(a) HOG-SVM Detector (b) HOG Cascaded Detector (c) Haar-like Cascaded Detector

Fig. 4.2 **The detections of three pedestrian detectors provided in CV** (2.4.3)

HOG is chosen as the pedestrian detection algorithm in this chapter due to the following reasons:

1. Its performance is tested stable and promising in a number of literatures;
2. The principle of HOG descriptions is cognitively perceivable. The shape feature detected by HOG have cognitive meanings to human. In other words, the possible false alarms can be retrieved or predicted;
3. HOG detector has been well trained in CV (2.4.3).

As a comparison, *Fig. 4.2* demonstrates an example pedestrian detection results using three trained detectors provided by CV (2.4.3):

[4] FPPW is the averaged false alarms out of the number of ROIs on the testing images;
[5] This is introduced in this [22] only, not seen in other literatures.

- HOG-SVM: HOG descriptor, training using SVM;
- HOG Cascade: HOG descriptor calculated on local areas are trained using SVM to weak classifiers. They are cascaded using AdaBoost to decision trees; [6]
- Haar-Pedestrian: Haar-like features trained using AdaBoost to cascaded decision trees (known as Viola-Jones method). [7]

	False Alarms	Detection Rate	Dataset (Testing)	Dataset (Training)
HOG-SVM	~ 1×10^{-4}FPPW	~90%†	INRIA[11]	INRIA[11]
	~7%	~80%	Website	CV (2.4.3)
	~74%	~94%	Video Clips - 1	CV (2.4.3)
	~93%	~90%	Video Clips - 2	CV (2.4.3)
HOG Cascading	~ 1×10^{-4}FPPW	~90%†	INRIA[32]	INRIA[32]
	~7%	~70%	Website	CV (2.4.3)
	~75%	~85%	Video Clips - 1	CV (2.4.3)
	~93%	~80%	Video Clips - 2	CV (2.4.3)
Viola-Jones	~40 FPPI	~90%‡	MIT-CMU[79]	unknown[1] [79]
	~30%	~60%	Website	CV (2.4.3)
	~80%	~70%	Video Clips - 1	CV (2.4.3)
	~95%	~63%	Video Clips - 2	CV (2.4.3)

Table 4.1 **The performances of frameworks are measured by the ROC curves [3, 11, 32]. The images from "Website" are a selection of the results of Google "pedestrians". Frames from video clips 1 and 2 are shot in a car park area with fences.**

† The false alarm rate refers to the approximate FPPI value when the detection rate is 90% or the miss rate is 10%.

‡ The performance of the frontal face detector trained using the Viola-Jones frameworks.

In the table, the first row of each detector is referenced from its original literature followed by the performances of the detectors applied to the testing dataset used in this chapter. The dataset includes randomly selected 1,000 images using the results of Google "pedestrian" and ~3,000 frames of two video clips shot in a noisy ground as introduced in *Section 2.3*. Images from the website contain several difficult cases including images containing crowds. Video clips are shot using steady camera placed in a car park with striped fences, where high amount of false alarms introduced by cars and constructions surrounding. In CV (2.4.3) manual, no information

[6] Cascaded HOG detector in CV (2.4.3) is trained following the instructions introduced [32]. In each region of interest with size 48 × 96 pixel^2, blocks from size 12 × 12 pixel^2 to size 64 × 128 pixel^2 are described using HOG descriptors calculated on integral images. The 36-bit Descriptors of cells are concatenated and normalized within 2 × 2 pixel^2-cell block, 9 bin per cell (as shown in *Fig. 4.8*). Cascaded HOG detector have similar performances with HOG-SVM for using the same feature descriptor and the same sized regions of interest.

[7] See *Section 4.2.1*

is provided on the training dataset of the trained detectors. Applied to testing images selected from databases including INRIA, these detectors can basically repeat the experiment results provided in their original literatures as summarised in *Table 4.1*. Apply the listed detectors to the dataset of this chapter, the results are quantified as follows: the detection rate is the number of detected pedestrian using listed detectors over the number of pedestrians recognised by human; the false alarm rate is the percentage of detected false alarms of the total number of detections using selected detectors. Applied listed detectors to the testing dataset of this chapter: the detection rate of HOG-SVM and HOG Cascade are similar. The false alarm rate of HOG Cascade is higher than HOG-SVM. Haar-Pedestrian performs the worst among the three. This may be because the sizes of a number of ROIs used in training are relatively small (14×28 pixel^2).

4.1.2 HOG Pedestrian Detector and Its Limitations

The descriptors used in HOG succeed the intensity based descriptor of SIFT [52] with some variations (R-HOG, C-HOG and Single-Centred C-HOG). Using Support Vector Machine (SVM), HOG descriptions of positive and negative examples are trained to a detector which transforms HOG descriptions into the decision space (pedestrian, non-pedestrian in this case). To detect pedestrians in target images, sliding windows are used to obtain the regions of interests (ROI). In each window of input images, the trained detector is applied to the HOG descriptions of the window. Windows with positive responses to the detector are recognised as pedestrians. In multi-scale pedestrian detection, the sliding windows are applied to down-scaled input images to achieve the scale invariant of the framework.

Rather than calculating descriptions on the edge maps like Shape Context [58], cognitive shapes are chosen as features to increase the the intra-class compatibility and the noise sustainability of pedestrian detector. The attempt was presented in [33] using Haar-wavelet block transform to capture the cognitive shape feature. The length of wavelet bases determines the roughness of the captured shape, the longer the descriptors, the more detailed shapes are captured. The coefficients of the wavelet transform are used as the input observations of SVM to train the detector. Based on the similar idea, HOG descriptor in [22] uses the 1st order derivative image to capture the cognitive shape feature. Histogram oriented descriptors are locally normalised with the descriptors of neighbourhood cells.

R-HOG and SIFT descriptor have many commons in the calculation procedure. SIFT description captures the local appearances of interest points (detected using

SIFT interest point detection) while HOG captures the rough shapes of objects. The differences between HOG and SIFT descriptor are as follows. A brief comparison of HOG and similar descriptors are summarised in *Table 4.2*.

- SIFT descriptor is calculated on Gaussian smoothed images while HOG is calculated on the 1^{st} derivative images;

- There are three ways of calculating the HOG descriptors of cells as introduced in [22]: R(Rectangular)-HOG, C(Circular)-HOG and Single centre C-HOG. R-HOG is calculated using the SIFT way. C-HOG and single centre C-HOG are calculated on the log-polar coordination mapped 1^{st} derivative image.

- SIFT descriptor is normalised within the region of area to calculate the descriptions (16×16pixel^2). The aim of the normalization is to reduce the effects of luminance and contrast. HOG descriptor is normalised within local neighbourhood cells (block), the normalization strengthens the common gradient orientations of the descriptions of cells in the block and weakens the others. After the normalization, the descriptions of cells participated in constructing the cognitive shapes of pedestrians are emphasized.

- The scale invariant of SIFT is achieved by selecting the scale salient interest points of input images. In HOG this is achieved by repeatedly applying the detection procedure to different scaled images.

- The rotation invariant is considered in SIFT. The histograms of gradient orientations are normalised to the major orientation within the 16×16pixel^2 cell. The rotation of pedestrians is not considered in HOG (at least not in its original article) as pedestrians usually appear upright in images. In some literatures, slight angle variations were adjusted in camera calibration stage or by applying the trained detector to sliding windows that are slightly tilted of the vertical axis.

Kernel based SVM is applied to train the HOG pedestrian detector. During training, the sizes of positive and negative training images are fixed. Image pre-processing, such as gamma correction, is normally required to reduce the influence of illumination effects. SVM training transforms the descriptions of positive and negative observations to the decision space where the margins between the two classes of input data can be maximised. Given training vectors $\mathbf{x}_i \in \mathbf{R}^d$, $i = 1, \ldots, l$, labelled by $y_i \in \{1, -1\}$ into two classes, a general form of Support Vector Machine solves the following optimization problem:

$$\begin{aligned}
\min_{w,b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\textstyle\sum_{i=1}^{l}\xi_i && (4.1)\\
\text{subject to} \quad & y_i\mathbf{w}^T\phi(\mathbf{x}_i + b) \geq 1 - \xi_i && (4.2)\\
& \xi_i \geq 0, i = 1, \ldots, l && (4.3)
\end{aligned}$$

where $\phi(\mathbf{x}_i)$ is the kernel transforming the descriptor $\mathbf{x}_i \in \mathbf{R}^d$ to the decisions space. $\mathbf{w}$ and b are the parameters to be optimised in linear support vector machine. $C \geq 0$ is the parameter corresponding to errors. When two classes of input data are overlapped in the decision planes, the input data are non-separable. In this case, ξ is introduced as a slack variable to relax the classification constraints. The summation of the slack variables ($\sum_i \xi$) is the upper bound of training errors.

$$\begin{aligned}
\mathbf{w}^T\phi(\mathbf{x}_i + b) \quad & \geq +1 - \xi \quad \text{for } y = +1 && (4.4)\\
\mathbf{w}^T\phi(\mathbf{x}_i + b) \quad & \leq -1 + \xi \quad \text{for } y = -1 && (4.5)
\end{aligned}$$

In some literatures, HOG is trained using Latent SVM which adds latent factors to the kernel functions. Different images are trained using different kernels in SVM. When the latent factor is universal for all training images, the latent SVM becomes the above linear SVM. The latent SVM attempts to differ the contributions of different descriptions in training. Difficult examples which may easily be miss classified should have more significant effect in training than the simple ones.

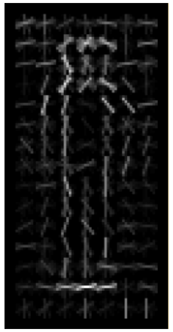
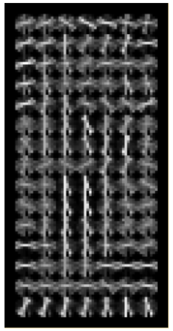
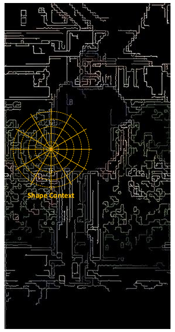
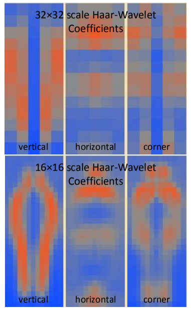
	HOG[22]	SIFT[52]	Shape Context[58]	Haar-Wavelet[33]
Calculation Area	ROI of 4×4 pixel^2 cells	4×4 cells, size: 4×4 pixel^2 surround the Interest Points	Radii of n pixel log-polar area surround edge pixels	whole ROI
Image Pre-processing	1st Gradient of Smoothed Image	Smoothed Image	Edge Map	n/a
Feature Illustration†	Cognitive Shape† 	Local Texture† 	Edge Map† 	Cognitive Shape†  32×32 scale Haar-Wavelet Coefficients vertical horizontal corner 16×16 scale Haar-Wavelet Coefficients vertical horizontal corner
Length‡	$4 \times 4 \times 9$	$4 \times 4 \times 8$	6×12	5×13 or 13×34
Sensitivity	Moderate	Moderate	Sensitive	Moderate

Table 4.2 **R-HOG, SIFT, shape context and Haar-wavelet descriptors are all used to capture local features. Used in different circumstances, these descriptors work with different features. R-HOG is calculated over the entire ROI to build the descriptor; SIFT is served to describe a local area around the interest point which is used as the features of objects; shape context provides the descriptions of the edge maps of objects; Haar-wavelet uses the coefficients of Haar-wavelet transform of ROIs to describe the rough shape feature.**

† Image is quoted from [22]; HOG and SIFT Illustrations are referenced from [22];
Edge map is detected by Canny Edge detection; Wavelet Transform image is referenced from [33]

‡ Parameters are referenced from [22, 33, 52, 58] respectively.

4.1.2.1 False Alarms introduced by HOG

False positives are usually introduced by objects that have similarly shapes to pedestrians, objects with approximated rectangular shape for example. In [3], errors and inaccurate detected results were categorised into groups as shown in *Figure 4.3* [8]. Categorised in four, they are:

- Rectangular objects. Examples include windows, doors, (part of) logos, telephone boxes, trash bins, etcetera;
- Objects in rough rectangular shape having structures like the shoulders of pedestrians. Such objects may be a wheel of vehicle with its fender, church windows with arch top, street lamp with shoulder like decorations.
- Overlapped or less-accurately detected pedestrians: pedestrians and part of pedestrians may be detected several times in different scaled input images. The detected bounding boxes may be overlapped; furthermore, the bounding boxes may not be accurately centred at the centroids of the pedestrians.
- More than one pedestrians are detected as one.

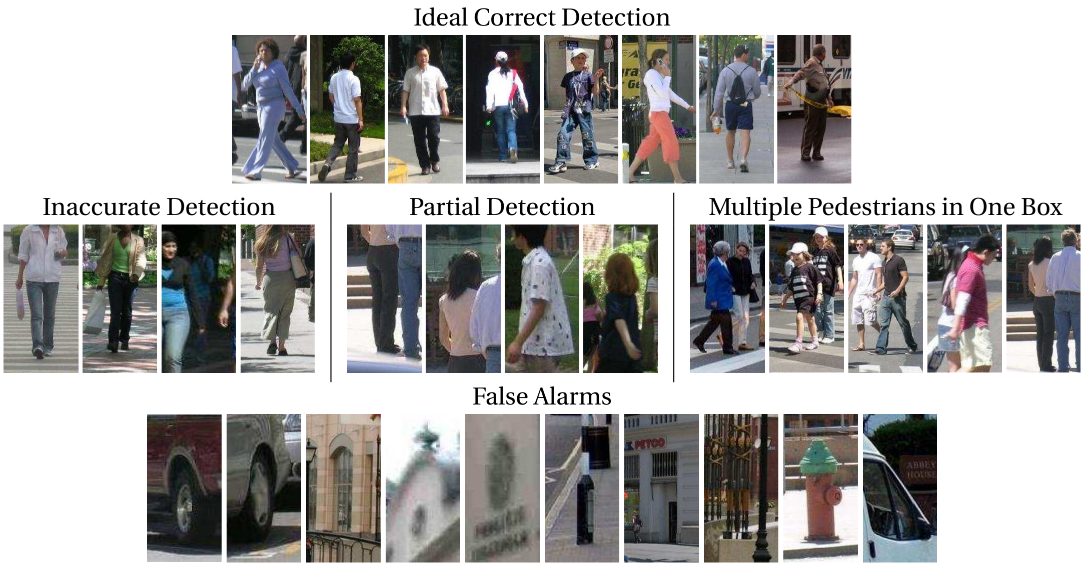


Fig. 4.3 **Sample Results of HOG-SVM Pedestrian Detector**

The "Inaccurate Detection" and "Partial Detection" are usually related to the parameters used in training the HOG detector as discussed in the following section. Reducing the the "False Alarms" according to the appearance of head-shoulder part

[8] Images demonstrated in this Table follow the ideas in [3] but they are not exactly the same.

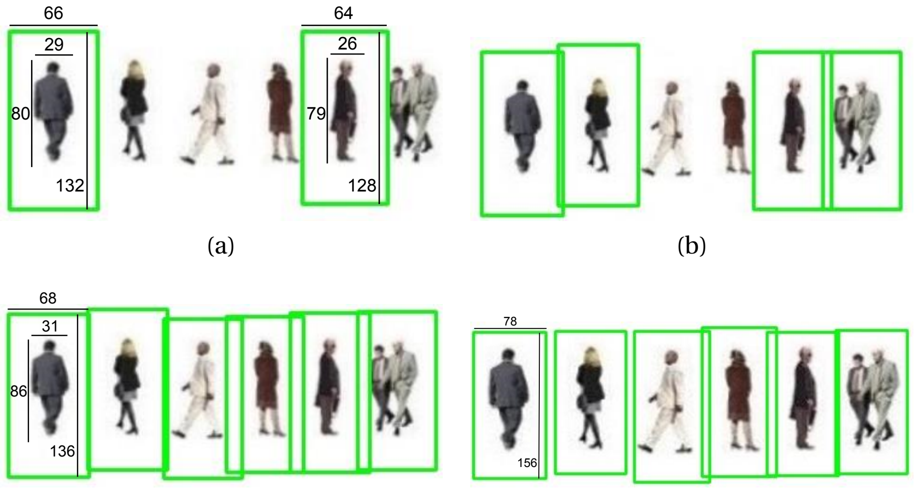
will be discussed in *Section 4.2* and separating the "Multiple Pedestrians in One Box" will be discussed in the next chapter (*Chapter 5*).

4.1.2.2 Parameters used in CV (2.4.3) HOG-SVM Pedestrian Detector

Parameters used in training and obtaining ROIs affect the performances of the trained HOG-SVM detector: missed detections and inaccurate detection results. In [22], the effects of tuning the parameters were demonstrated using the DET curve when applying the HOG detector to INRIA and MIT pedestrian databases. Suggestions were provided on the selection of the values of these parameters. HOG-SVM detector provided by CV (2.4.3) basically follow the recommendations in [22]. Key parameters concerned in [22] includes (values in brackets are settings of HOG-SVM detector trained by CV (2.4.3)):

- The size of a cell (8×8 pixel^2) and the number of bins (9-bin per cell) used in calculating the descriptor: The size of a cell should not be too large to capture the rough shape feature. Finely divided cells may fragment the appearances of local areas.
- The size of a block (2×2 cell^2 which is equivalent to 16×16 pixel^2): the descriptors of cells within a block are normalised.
- The sizes (128×64 pixel^2) and the stride (8 pixel in each direction) of sliding windows to obtain the regions of interest: the descriptor of a window is concatenated by the normalised descriptors of all cells in the window (total length of the descriptor will be $9\times16\times8=1152$ bit).
- The parameters and kernels used in SVM training: linear SVM is the default setting for training a SVM classifier. A selection of kernels are provided in CV (2.4.3), including Polynomial kernel, Radial basis function (RBF) and Sigmoid kernel.
- The down scaling factor used in multi-scale pedestrian detection: after each round of complete search, the original images are down-scaled by 1.05 to achieve the scaling invariant.

The smallest bounding box detected by HOG is 128×64 pixel^2. It determines the lower limit of the sizes of detectable pedestrians in images. Usually, the detected pedestrians occupy around 30% to 60% area of the bounding boxes. This means, within the smallest bounding box, the pedestrian can hardly be detected if its size is smaller than 80×30 pixel^2. As shown in *Fig. 4.4*, no pedestrians are detected in images containing pedestrians smaller than the ones shown in *Fig. 4.4 (a)*. *Fig. 4.4* *(b),(c),(d)* show that the performance of HOG-SVM is better when the input images are enlarged.



(a) (b)

(c) (d)

Fig. 4.4 **The Smallest Detectable Pedestrians of HOG Detector: in experiment, the original image containing pedestrians with height** 50 **pixels and width around** 20 **pixels. In this image, no pedestrians are detected using HOG detector trained by CV** (2.4.3)**. Proportionally expanded the image by 5% in each step, (a) is the first image where there are detected pedestrian. The average size of pedestrians in (a) is** $\sim 70 \times 25$ **pixel**^{2}**, in (b) it is** $\sim 72 \times 27$ **pixel**^{2}**, in (c) it is** $\sim 75 \times 30$ **pixel**^{2} **and in (d) it is** $\sim 100 \times 40$ **pixel**^{2}

Pedestrians in positive training examples are centred in the bounding box. If 16×8 cells are used in HOG description, the width of the background surrounding the pedestrian in positive training images is usually around the width of 1 to 2 cells. Including sufficient amount of background in positive training images is crucial to obtain the cognitive shape of the pedestrian when the descriptions of cells are normalised with its neighbourhoods. Pedestrians who are close to the boundaries of images would be ignored. This means the HOG detector which only responses positively to the bounding boxes containing the rough shapes of pedestrians in the centre. As illustrated in *Fig. 4.5*, not all pedestrians near the boundary are detected except when they can be centred in bounding boxes. The effect is even obvious when pedestrians are placed in the background of streets. An extension around 10pixels to all boundaries of an input image will solve the problem (under current CV (2.4.3) HOG detector setting).

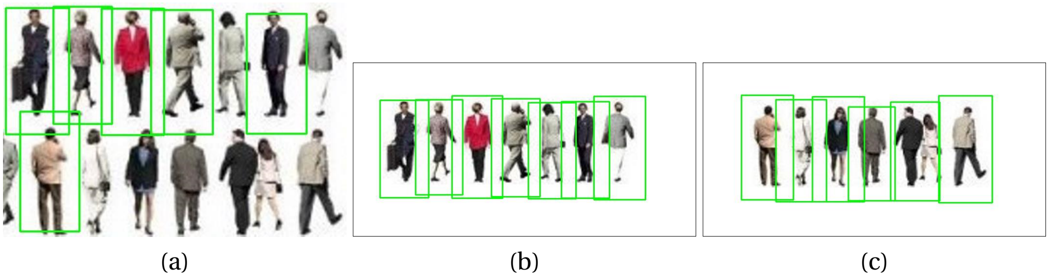


Fig. 4.5 **HOG Performance with Pedestrian near Boundary**

4.2 Upper-body Detection

In the last section, pedestrian detectors provided in CV (2.4.3) are discussed. Conclusions were made that HOG-SVM detector provided in CV (2.4.3) outperforms other built-in pedestrian detectors. HOG-SVM detector is competent for the majority of the pedestrian detection cases but may fail in certain circumstances. Considering false alarms usually have less typical head-shoulder appearance, algorithms for upper-body detection will be employed to eliminate false-alarmed background objects. As an assistant algorithm, the upper-body detector will be applied to detections results of HOG-SVM pedestrian detector. Furthermore, this algorithm should be fast in calculation. Two algorithms are considered: Viola-Jones frontal face and upper-body detector and a novel way of head-shoulder verification.

Inspired by Haar-wavelet based pedestrian detection introduced in [33] and further developed in [80], Viola-Jones body (part) detection [23] was initially designed for the frontal face detection. This framework has two main contributions in computation reduction of its predecessor in [33]: Haar-like features are calculated in integrated images as weak classifiers and a fast cascaded AdaBoost training procedure is employed to combine selected weak classifiers into a strong one. Viola-Jones have drawbacks: features selected in detectors rely on training dataset; once trained the detector cannot be modified. In following sections, the algorithms will be reviewed; factors affect the performance of the detectors provided in CV (2.4.3) will be unveiled.

After that, an appearance based algorithm for the verification of the head-shoulder structure of detection results is proposed to separate false alarms from the results of pedestrian detection. *Section 4.4* will demonstrate that the above two algorithms significantly improve the performances of HOG-SVM detector.

4.2.1 Viola-Jones Algorithm

Haar-like features are the bases of Viola-Jones approach for pedestrian / body-part detection. Calculated in integrated images, Haar-like features capture the colour changing within a local block. The size of the block can vary from several squares of pixels to the size of ROI. The Haar-like features are performed as weak classifiers which will be selected using boosting algorithms. Weights applied to the weak classifiers will be adjusted during training. Heavy weight will be applied to important features cascaded in the final classifier. The performances of Viola-Jones algorithm are affected by the Haar-like features selected in the output detectors, especially the ones with heavy weights.

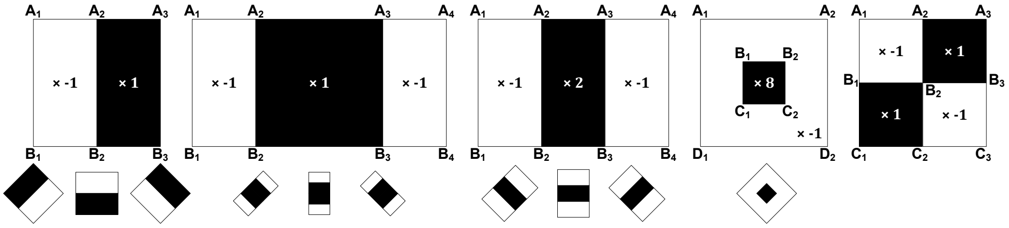
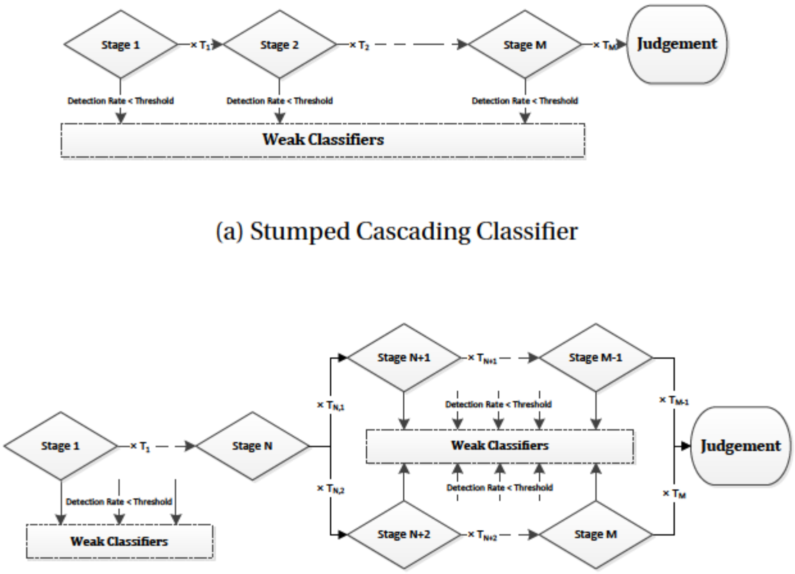


Fig. 4.6 **Haar-like Feature bases (2-rectangle, 3-rectangle(wide line), 3-rectangle(narrow line), centre surround, 4-rectangle diagonal). To calculate each feature using above bases, the size of black and white area are normalised. For example, the black area in 2-rectangle, 3-rectangle (wide line) and 4-rectangle bases have the same size with the rest white area so equal weights are assigned to black and white area** $(+1, -1$ **respectively). While in the 3-rectangle(narrow line) and the centre surround bases, the weight assigned to the black area is equal to the ratio of the size of white area to black. Detailed calculation is attached in *Appendix. A*.**

4.2.1.1 Haar-like Features

Features shown in *Fig. 4.6* are calculated in rectangular blocks from 1×2 pixel^2 to the size of ROI. Each feature compare the accumulated value of pixels in black area to it is in the white area and returns a positive or negative sign to indicate which part is darker. For training purposes, the sign of each feature is translated to meanings of object / non-object. A Haar-like feature is a binary classifier $f(x_i)$ with low accuracy. To reduce calculation, integrated images were introduced , where the value of pixel (x, y) is the summed values of pixels in the area above and to the left of (x, y). *Table. A.1* show the calculation of Haar-like features in integral image. In CV (2.4.3), both original Haar-like features introduced in [23] and extended ones introduced in [30] are considered. Working together, these features capture the colour changing near boundaries, corners, crossing points and lines.

The number of Haar-like features within ROI is normally huge no matter it is for pedestrian detection or body-part detection. In a $n \times m$ pixel^2 area, there are $\binom{n}{2}\binom{m}{2} - mn$ rectangles to be evaluated as weak classifiers. The number is proportional to $(mn)^2$. There are usually hundreds of thousands features in one ROI even only 2 feature bases are calculated in one rectangle block. Majority Haar-like features will be eliminated during AdaBoost training. In [23], AdaBoost training was applied to remaining Haar-like features for several times, stage detectors obtained from each round are cascaded to improve the performance. The cascading procedure increases the effect of strong features to the final result as they may reappear in cascading stage of classifiers. Cascaded decision stump combined by weighted stage classifiers is a specific form of cascading trees which have more than one branches of combined stage classifiers as shown in *Figure 4.7*.



(a) Stumped Cascading Classifier

(b) A Cascading Tree with Two Branches

Fig. 4.7 **The Structures of Cascaded Classifiers**

4.2.1.2 AdaBoost Cascaded Decision Trees

The philosophy behind boosting algorithm is to achieve a strong classifier by combining several weak ones. Statistically, boosting algorithms optimise the binary classifier by building an additive logistic regression model. Additive indicates the ways of combining of weak classifiers in training rounds to obtain the final classifier [27]. The performance of boosting is improved in [81], known as Discrete AdaBoost. Various AdaBoost strategies are supported by CV (2.4.3): Discrete AdaBoost, Real AdaBoost, LogitBoost and Gentle AdaBoost all of which aims to minimise the training error. AdaBoost follow the procedures below:

AdaBoost Training (N Images)

1. Assign weights $\omega_i = 1/N, i \in 1, \dots, N$ to training images $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots,$ $(\mathbf{x}_N, y_N)$. $\mathbf{x}_i \in \mathbf{R}^d$ is the image description; $y_i \in \{+1, -1\}$ is the judgement;
2. Repeat for $m = 1, \dots, M$:
   (a) Fit the classifier function $f_m(\mathbf{x})$ that minimise $E_\omega(exp\{-yF(\mathbf{x})\})$ with weights ω_i. y is the response of classifier $F(\mathbf{x})$; E_ω represents a weighted expectation;
   (b) Update $F_m(\mathbf{x}) \leftarrow F_{m-1}(\mathbf{x}) + f_m(\mathbf{x})$
   (c) Update ω_i and renormalise them for the next round so that $\sum_{i=1}^{N} \omega_i = 1$;
3. Output classifier sign$[F(\mathbf{x})]$ = sign$[\sum_{m=1}^{M} f_m(\mathbf{x})]$

The differences between these strategies lie in the measurements of training errors and the means of updating the classifier function. Discrete AdaBoost minimises the population expectation via Newton-like optimization. Real AdaBoost estimates the probability of miss classifications and update the final classifier by minimizing the squared error loss which is an approximate of weighted expectation. Gentle AdaBoost, changes the ways of estimating the weighted class probabilities in Real AdaBoost to achieve more stable performances. AdaBoost algorithms utilise a stage-wise estimation procedures that coefficients already engaged in the classifier will not be modified in further training round. Detailed analysis can be found in [27]. Discrete AdaBoost was chosen in [79] to train the cascaded Viola-Jones classifiers. Gentle AdaBoost is the default Boosting training mode in CV (2.4.3) and is applied to the training of built-in Viola-Jones detectors. Experiments in [30] showed that Gentle AdaBoost outperformed other AdaBoost Strategies.

Adaboost excels in real-time applications using linear programming. The performance of the strategy has limited reliance on the tuning of parameters used in training. Only three factors may affect the AdaBoost trained classifiers: the threshold used to reject the weak classifiers (t), the round of training (T) and the number stages (M) / branches (M') used in cascaded decision trees. T is dependent on t as t is related to the average error rate (p_e) of weak classifiers. The final detection rate of the additive weak classifiers after T round of training can be predicated as $1 - p_e^T$. This means, increasing the round of training (T) will reduce the averaged error rate (p_e). [82] also pointed that too many rounds of training may result in over fit which lacks observations. [82] concluded that increasing the number of boosting stages M and M' may improve the performances of the final classifier. Theoretically, more cascading levels (M and M') would improve the detection performances, while no statistical analysis shown this relationship. In current stage of research, cascaded decision trees used in pedestrian / body-part detection contain a maximum of two branches of cascaded stage classifiers.

4.2.1.3 **Viola-Jones Body-part Detectors in CV** (2.4.3)

In CV (2.4.3), the Haar-like cascaded detector use the tree structure where different coefficients are applied to positive and negative training images. Cascading decision trees containing two branches were trained for the frontal-face, upper-body and full body detector. The number of stages in a branch are usually flexible. If additional branches / stages cannot improve the performance (converged classifier) of the output detector, training stops. [30] pointed that when 20 × 20 pixel^2 ROIs are selected, in the training of frontal face detector using Gentle AdaBoost, an average of 30 weak classifiers will be cascaded in one stage of the output classifier. The improvement of performance of the output classifier stagnated until new stages of weak classifiers are cascaded in. A disadvantage of Viola-Jones training strategy is the features selected in the output classifier is unknown until the end of the training procedure. The training result is less perceivable to human: in a cascading decision tree with two branches containing hundreds of weak classifiers in cascaded in 50-60 stages, it is difficult to unveil which are the most influential local features. Normally, key weak classifiers have heavier weight than other weak classifiers and are normally employed in multiple stages of the trained classifier. *Fig. 4.9* demonstrates several frequently appeared weak classifiers that have relatively heavier weights than others in the final output classifier. The characteristics of Viola-Jones pedestrian / body-part detectors in CV (2.4.3) are summarised in *Appendix A*.

Applied in different applications, the sized of ROIs are different. In selected detectors: the sizes of ROIs of Viola-Jones frontal face detector are 20 × 20 pixel^2, 22 × 20 pixel^2 for Haar-like upper-body detector, 19 × 23 pixel^2 for Haar-like lower-body detector and 14 × 28 pixel^2 for Haar-like pedestrian detector. Above parameters are referenced from the CV (2.4.3) library.

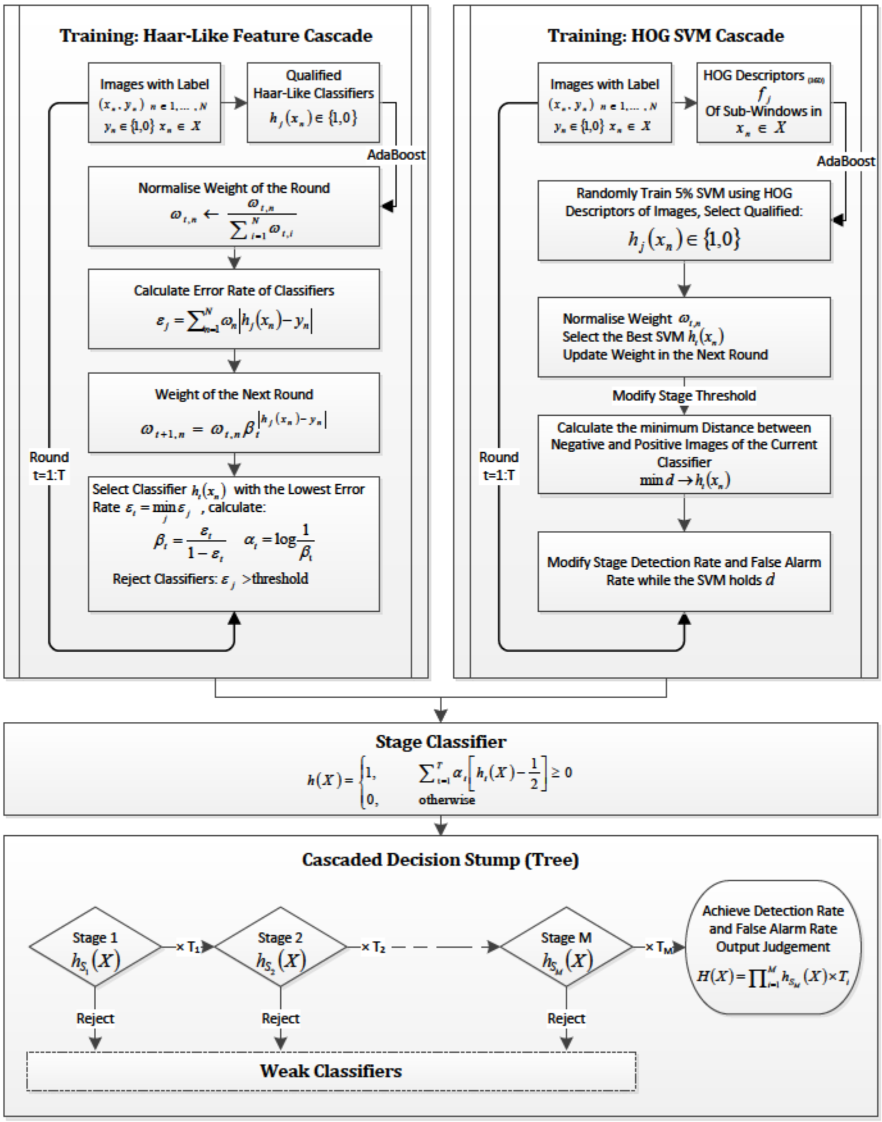
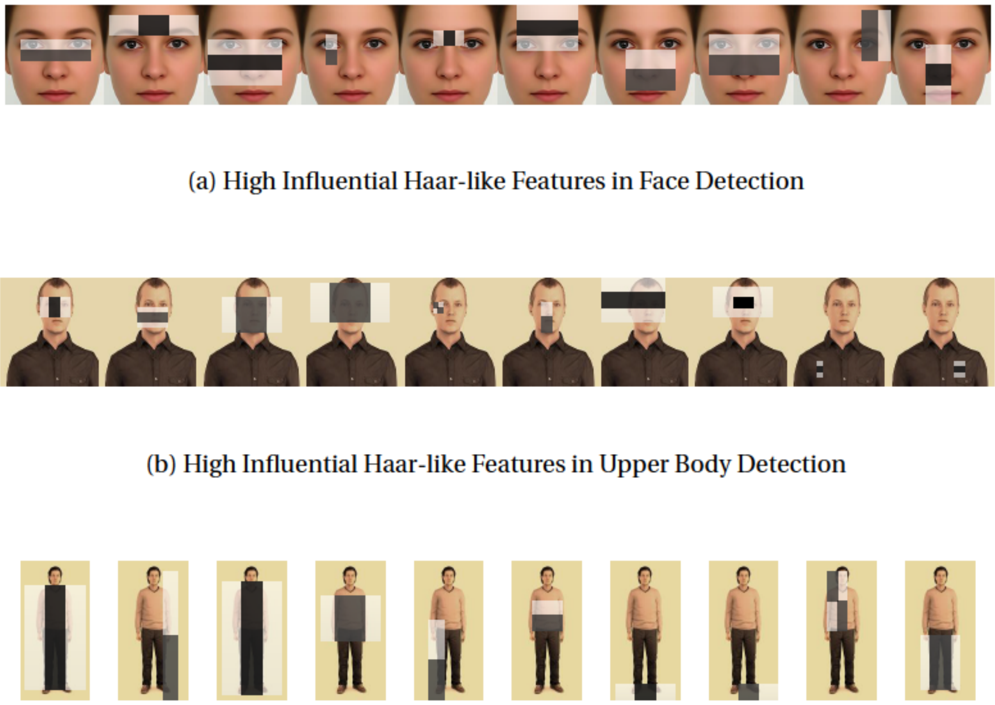

Training: Haar-Like Feature Cascade
Images with Label
(x_n, y_n) $n \in 1, \ldots, N$
$y_n \in \{1,0\}$ $x_n \in X$
Qualified
Haar-Like Classifiers
$h_j(x_n) \in \{1,0\}$
AdaBoost
Normalise Weight of the Round
$\omega_{t,n} \leftarrow \frac{\omega_{t,n}}{\sum_{i=1}^{N} \omega_{t,i}}$
Calculate Error Rate of Classifiers
$\varepsilon_j = \sum_{n=1}^{N} \omega_n \left| h_j(x_n) - y_n \right|$
Weight of the Next Round
$\omega_{t+1,n} = \omega_{t,n} \beta_t^{\left| h_j(x_n) - y_n \right|}$
Round
t=1:T
Select Classifier $h_t(x_n)$ with the Lowest Error
Rate $\varepsilon_t = \min_j \varepsilon_j$, calculate:
$\beta_t = \frac{\varepsilon_t}{1 - \varepsilon_t}$ $\alpha_t = \log \frac{1}{\beta_t}$
Reject Classifiers: ε_j >threshold
Training: HOG SVM Cascade
Images with Label
(x_n, y_n) $n \in 1, \ldots, N$
$y_n \in \{1,0\}$ $x_n \in X$
HOG Descriptors $_{(360)}$
f_j
Of Sub-Windows in
$x_n \in X$
AdaBoost
Randomly Train 5% SVM using HOG
Descriptors of Images, Select Qualified:
$h_j(x_n) \in \{1,0\}$
Normalise Weight $\omega_{t,n}$
Select the Best SVM $h_t(x_n)$
Update Weight in the Next Round
Modify Stage Threshold
Calculate the minimum Distance between
Negative and Positive Images of the Current
Classifier
$\min d \rightarrow h_t(x_n)$
Round
t=1:T
Modify Stage Detection Rate and False Alarm
Rate while the SVM holds d
Stage Classifier
$h(X) = \begin{cases} 1, & \sum_{t=1}^{T} \alpha_t \left[h_t(X) - \frac{1}{2} \right] \geq 0 \\ 0, & \text{otherwise} \end{cases}$
Cascaded Decision Stump (Tree)
Stage 1
$h_{S_1}(X)$
× T₁
Stage 2
$h_{S_2}(X)$
× T₂
Stage M
$h_{S_M}(X)$
× T_M
Achieve Detection Rate
and False Alarm Rate
Output Judgement
$H(X) = \prod_{i=1}^{M} h_{S_M}(X) \times T_i$
Reject
Reject
Reject
Weak Classifiers

Fig. 4.8 **The training procedure of a cascaded classifier: Viola-Jones [79] and Cascaded HOG-SVM [32]**



(a) High Influential Haar-like Features in Face Detection

(b) High Influential Haar-like Features in Upper Body Detection

(c) High Influential Haar-like Features in Full Body Detection

Fig. 4.9 **High influential Haar-like blocks in trained cascading decision trees: These blocks have large coefficients in heavy weighted stages. Some of the blocks appear in more than one stages.**

4.2.2 Appearance Based Upper-body "Detection"

By verifying the upper-body and frontal face of a pedestrian using Voila-Jones method can greatly reduce the false alarms introduced by HOG pedestrian detector. But the frontal face detector fails in the verification of backward facing pedestrians and the upper-body detector is less stable in performances. The upper-body detector may response to any shoulder like structures, such as the fender-wheel part of vehicles, arch or arch structured windows/doors. Furthermore, due to the large amount of small sized Haar-like features cascaded in the Viola-Jones upper-body detector, the detector is sensitive to noise and details in the appearances of pedestrians. In this section, a novel means of false alarm reduction based on the detection of the head-shoulder structure will be introduced. This algorithm is to verify the presence of the head-should structure inside the perspective head area of the detected "pedestrians". Experiments will show that the algorithm reduces more false alarms than using Viola-Jones upper-body and frontal face detectors. After applying this algorithm to the HOG detection results, the total false alarms can be reduced to less than 5%

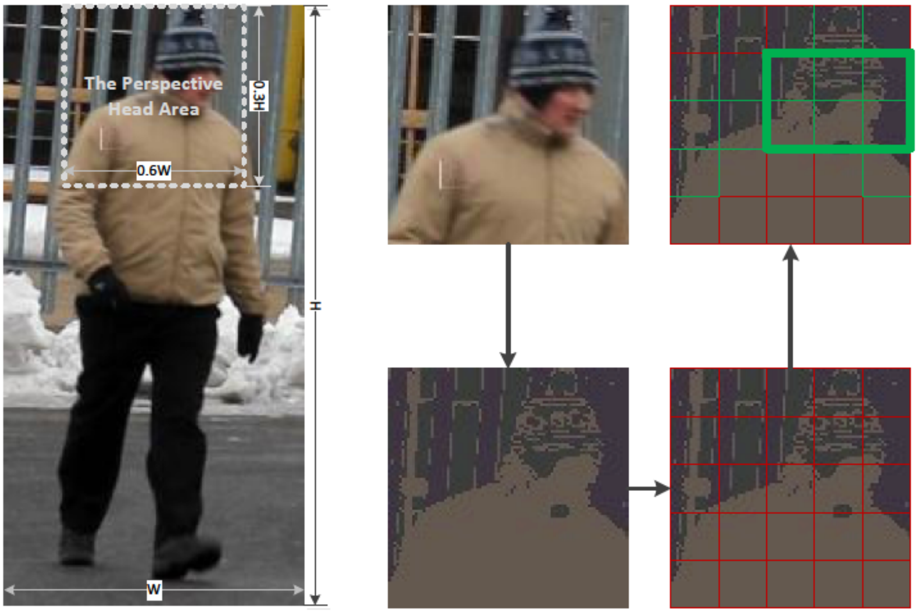
without significantly sacrifice the detection rate. The algorithm is designed based on the following assumptions:

- The detected "pedestrian", either pedestrians, inaccurate detected pedestrians or false alarms, should have a rough rectangular appearances placed the centre of the detected bounding box;
- The head-shoulder structure is recognised as the feature to distinguish between pedestrians and false alarms;
- The head-shoulders of the ground truth pedestrians should be clearly visible;

The first assumption is observed from the detection results of a HOG pedestrian detector as reviewed in *Section 4.1*. For most of the time, the second assumption is true: the head-shoulder structure is hardly seen from false alarms and inaccurate detections including artificial manufactured objects or a part of the pedestrian. The third assumption restricts the application of this algorithm to images without crowds or noisy background for it is even difficult for human to recognise pedestrians in such images. Still, as an assistant algorithm to the HOG pedestrian detector, the algorithm should be simple without complex computation. Based on these assumptions and requirements, the algorithm first locates the perspective head area within the bounding box and then transfer the perspective head area into appearance related descriptions. If the descriptions of a blob with reasonable size (head) are different from its neighbourhood area, it will be recognised as a head. If the area below the blob also has different appearances with its upper neighbourhood, then the shoulder is recognised. Bounding boxes containing such head-shoulder structure would be recognised as pedestrians.

4.2.2.1 Within the Perspective Head Area

The Perspective Head Area: observed from results of HOG detector, the head would appear in the top 1/3 part of the bounding box. In experiment, the perspective head area is selected as shown in *Fig. 4.10 (a)*. This area within the bounding box is independent with the training datasets as the training images for HOG pedestrian detection all have the height to width ratio of 2. Pedestrians in training images are all placed in the centre. Sometimes, the pedestrian detected by HOG may not appear in the centre, this usually happen when multi-scaled detection is applied to images containing pedestrian who is relatively small with regard to the size of the ROIs. In this chapter, such detections are considered as inaccurate detections or even false alarms.



(a) The Perspective Head Area

(b) Head Detection

Fig. 4.10 **Processing to detect the probable head within a bounding box of pedestrians.**

Clustering: to describe the sub-areas within the perspective head region, the pixels within this area are clustered using algorithms introduced in *Section 3.2.3.1*:

1. Pixels are clustered in RGB spaces into K clusters;
2. The centres of these clusters are then transformed to HSV space;
3. The clusters with their centres closely mapped in the H-S coordination plane are combined as they are cognitively similar to each other;
4. After the combination of clusters, the intensities of pixels in one cluster are the values of their cluster centre.

Unlike the clustering algorithm introduced in the last chapter *Section 3.2.3.1*, the number of K is not necessarily to be large to discover the colour information of the part. The larger the K, the slower the clustering speed. The colour clustering aims at simplifying the appearance of the perspective head area. In experiment, $K = 3$.

Consistent Matrix: the perspective head area is evenly divided into $M \times N$ rectangular cells. The description of each cell is calculated as the histogram of pixels after clustering. To detect an area which has different descriptions with its surrounding areas, cells are compared to its 4-connected neighbourhood cells (cells upward, downward, on the left and on the right). The appearances of two cells are considered inconsistent (Matrix value = 1) if the correlation distance between the histograms of two cells is larger than a pre-defined threshold ($T_c = 0.6$). Otherwise the appearance of the two cells are consistent (Matrix value = 0). The consistencies of appearances of cells are recorded in the consistent matrix as shown in *Fig. 4.11*. If $M \times N$ cells are divided vertically and horizontally within the perspective head area, the size of the consistent matrix is $(2M-1) \times (2N-1)$. A cell on the y^{th} row and the x^{th} column ($x = 1,\dots,N, y = 1,\dots,M$) of the perspective head area will be mapped to the $(2x,2y)$ element in the consistent matrix. The consistent measurements between the cell $(2x,2y)$ and its 4-connected neighbourhood cells are calculated if both of them are within the perspective head area:

- $(2x-1,2y)$ is the consistency between the current cell and the cell on the left (in matrix: $(2x-2,2y)$);
- $(2x+1,2y)$ is the consistency between the current cell and the cell on the right (in matrix: $(2x+2,2y)$);
- $(2x,2y-1)$ is the consistency between the current cell and the cell above (in matrix: $(2x,2y-2)$);
- $(2x,2y+1)$ is the consistency between the current cell and the cell below (in matrix: $(2x-2,2y+2)$)

Cells located at the boundary of the perspective head area ($x = 1,\ x = N$ or $y = 1,\ y = M$) will only be compared with its neighbourhood cells which are inside the perspective head area. In the consistent matrix, $(2i,2j)$ are meaning less elements showing the position of the cells divided within the perspective head area, $(2i+1,2j+1)$ are meaning less elements to fill gaps in the consistent matrix and the rest of the elements are the consistency measurements between cells. $(2i+1,2j+1)$ elements will be coded when the consistent measurements are calculated between the current cell to its 8-connected neighbourhood cells (the current four plus the cells on up-left, up-right, down-left and down-right directions). The consistent matrix is to determine if there is a probable head within the area. A head-shoulder structure is considered as existing if a reasonable sized rectangle is detected from the perspective head area.

The size of the Head-Shoulder Structure: the reasonable size is approximated from observations of detected "pedestrians" using HOG. If the bounding box has height H and width W, the area of the detected "pedestrians" is normally larger than $1/2W \times 2/3H$. Then the size of the head area would be around $1/6W \times 1/10H$. In experiment, the size of the perspective head area is $3/5W \times 1/3H$. If 5 cells are separated in both dimensions, the valid rectangles to be recognised as a head-shoulder structure

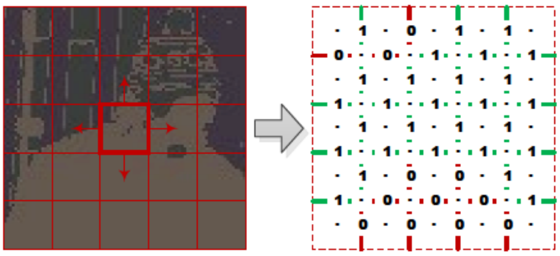


Fig. 4.11 **The Consistent Matrix of the perspective head area: the meaning less dots hold the positions of cells as the consistent measurements are calculated in the pairs of 4-connected cells. The dots will be replaced by 0s for the search of sub-matrices representing Head-Shoulder structures demonstrated in *Fig. 4.12*.**

should occupy an area of 2 × 1 cells or larger. As shown in *Fig. 4.10 (b)*, the borders between cells with inconsistent appearances are highlighted. Rectangles can be detected from the highlighted borders. Using the consistent matrix in detection, the models of the head-shoulder structures are transferred to sub-matrices as shown in *Fig. 4.12*. To locate the head-shoulder structure, the sub-matrices are searched within the appearance consistent matrix.

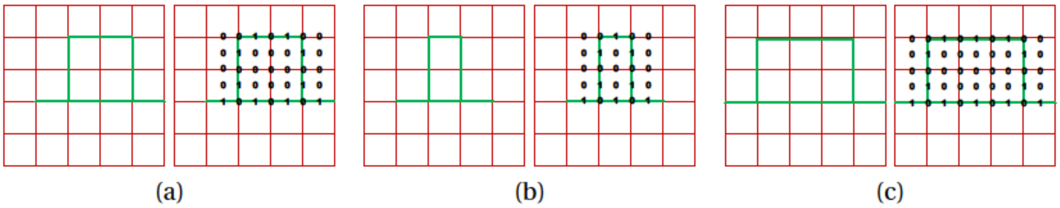


Fig. 4.12 **Sub-matrices searched with the judgement** 9×9 **consistent matrix when** 5×5 **cells are separated within the perspective head area.**

4.2.2.2 Not a Proper Upper-body Detection

The above algorithm is an assistant process to reduce false alarms introduced by HOG pedestrian detector. Though a perspective head-shoulder structure is eventually detected, this is not a head (blob) detection as the inconsistent appearances between the blob and its neighbourhood cells can hardly guarantee there the existence of a blob in general cases. However, when the target images of the algorithm are the results of HOG pedestrian detection. Such images usually contain a cylinder shaped object within the box. If pedestrians are presented in the bounding boxes, the head-shoulder structure has inconsistent histogram representations comparing to its neighbourhood area. This kind of inconsistencies can hardly be observed from false alarms like trash bins, cylinder logos, arch windows, car fender-wheels or parts of body (legs for example). The algorithm may not be accurate enough for pedestrian recognition, but it is a fast way to reduce false alarms of HOG pedestrian detector.

4.3 Re-appeared False Alarms in Video Streams

When pedestrian detection is applied to video streams false alarms may reappear in frames especially when the shooting position of the camera is fixed. Observations show that ~60% of false alarms will reappear in subsequent frames. In previous sections, algorithms are introduced to reduce the false alarms by verifying the head-shoulder structure within the bounding box. To avoid calculations in clustering, generating the consistent matrix and searching the sub-matrices, the pedestrian re-identification algorithm described in the last chapter *Chapter 3* is adopted with necessary modifications:

1. Verify the detections in several frames using algorithms introduced in *Section 4.2*. Separate the pedestrians from false alarms;

2. Build the Pedestrian-False Alarm dictionary using the verified results: divide both pedestrians and false alarms into $M \times N$ equal patches and calculate the HSV colour histogram descriptions for each patch;

3. For the HOG detection results of a new frame, separate the detected images into $M' \times N'$ patches, each patch is compared to patches in the Pedestrian-False Alarm Dictionary. The corresponding patch is selected as the one which has the largest correlation distance from the current one. The current patch is recognised as pedestrian / false alarm if its corresponding patch belongs to the words of pedestrian / false alarm. Otherwise, the patch is recognised as unfamiliar. To avoid calculations, the current detection is priorly compared to the detections which are geometrically located near them in the dictionary. The "geometrical near" is measured according to the movement of shooting camera.

4. If the majority of patches (>80%) within the bounding box are coded with pedestrian / false alarm, the detection is recognised as pedestrian / false alarm. If patches within a bounding box are equally coded with pedestrian or false alarm or unfamiliar, this detection is probably a new appeared pedestrian or false alarm. The perspective upper-body area of this detection will be verified using algorithms presented in previous sections.

In experiment, 6×3 pieces of patches are required in either images of dictionaries or reappeared false alarms. Comparing to the appearance based false alarm reduction introduced in *Section 4.2.2*, the calculations spent on clustering the perspective head area and searching the sub-matrices are reduced. For videos shot using still cameras, the "geometry near" in experiment means the circle area centred at the centroid of the detected image with the radius twice the length of the diagonal of the image. For videos shot using moving cameras, the circle area is calculated according to the movements of the cameras using formula introduced in [83].

4.4 Experiments

In this section, the effect of using upper-body verification strategies to reduce false alarms of HOG pedestrian detection will be demonstrated. HOG-SVM is selected as pedestrian detector as it outperforms other detectors provided by CV (2.4.3). All following syntheses are based on pedestrian detection experiments using default settings of detectors provided in CV (2.4.3). Two algorithms for the verification of upper-body are applied to detection results of HOG-SVM detector. Testing images are frames from two video clips shot using still cameras and a selection of internet images. As shown in *Table 4.3*, appearance based upper-body verification perform better than Viola-Jones frontal face with upper-body detectors. The detection rate is calculated as the number of detected pedestrians using algorithms over the number of manually detected pedestrians. The false alarm rate is calculated as the number of false alarms over the number of total detected occurrences:

$$\text{Detection Rate} = \frac{\text{\#Detected Pedestrians}}{\text{\#Manually Detected Pedestrians}} \tag{4.6}$$

$$\text{False Alarm Rate} = \frac{\text{\#False Alarms}}{\text{\#Total Detections}} \tag{4.7}$$

In above equations, inaccurate detections where pedestrian may be occluded but still occupy the majority area of the bounding box are recognised as correct detection result while detected body parts (legs, arms, upper-body only for example) are recognised as false alarms.

False alarms are greatly decreased when the algorithms are applied to the testing datasets. The detection rate is not greatly affected (decreased by ~7% on average). The appearance based algorithm is affected by the complexity of background and may not always perform better than Voila-Jones upper-body detector. The appearance based algorithm cannot accurately locate the upper-body as the Voila-Jones methods. The way that it eliminates the false alarms relies on the changing of appearances around the head-shoulder structure which is hardly seen from false alarm objects. It means, the appearance based algorithm may have problems with noisy

	Viola-Jones Detector		Appearance based Algorithm	
Video Clip 1	**Before**	**After**	**Before**	**After**
Detection Rate	~ 94%	~ 92%	~ 94%	~ 92%
False Alarm	~ 74%	~ 3%	~ 74%	~ 3%
Video Clip 2	**Before**	**After**	**Before**	**After**
Detection Rate	~ 90%	~ 73%	~ 90%	~ 82%
False Alarm	~ 93%	~ 21%	~ 93%	~ 10%
Internet Images	**Before**	**After**	**Before**	**After**
Detection Rate	~ 80%	~ 70%	~ 80%	~ 75%
False Alarm	~ 7%	~ 7%	~ 7%	~ 5%

Table 4.3 **Reducing the false alarms introduced by HOG-SVM pedestrian detector in two ways.**

background and may be interfered by skin colours. *Table 4.4* demonstrates examples of upper-body verification using the two algorithms based on the results of HOG-SVM detector provided by CV (2.4.3). 5 × 5 is the minimum pieces separated in the perspective head area to achieve performances shown in *Table 4.3*. More pieces separated in the perspective head area slightly improve the performance: if 9 × 9 pieces are separated in the perspective head area, another 1 - 2% of the total false alarms will be reduced on average. If 5 × 5 cells are separated in the rough upper-body area, according to the calculations in *Section 4.2.2.1*, the size of the consistent matrix is 9 × 9. Then the sub-matrices representing the head-shoulder structures are searched in the consistent matrix, as shown in *Fig. 4.12*. During the searching, few miss matched digits in the sub-matrices can be tolerated (2 in this chapter). As shown in *Table 4.3*, above settings achieve promising results with testing images selected in this chapter. The accuracy of the Viola-Jones frontal face detector is high while its detection rate is moderate; the performance of Viola-Jones upper-body detector is moderate. The novel appearance based head-shoulder verification algorithm outperforms the Viola-Jones upper-body detector in false alarm reduction but this algorithm is less accurate than the Viola-Jones frontal frontal face detector. When accuracy is desired over detection rate, Viola-Jones frontal face detectors should be considered with priority. But the Viola-Jones frontal face detector is less capable to recognise the backward viewed head.

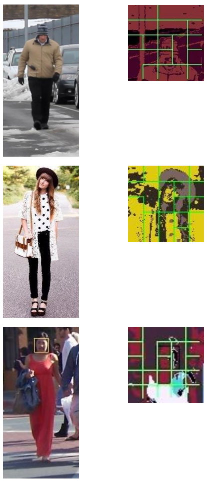
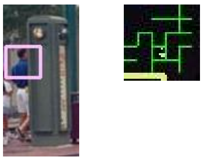
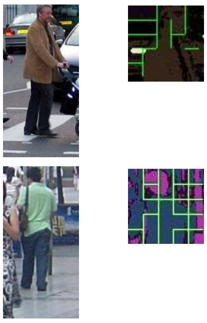
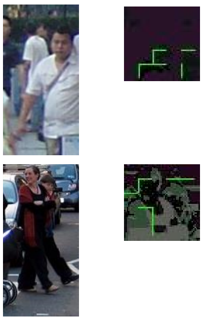
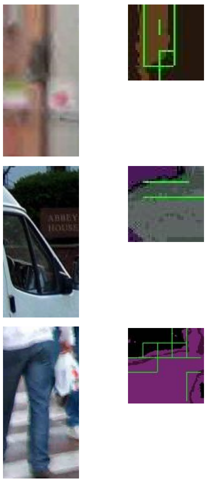
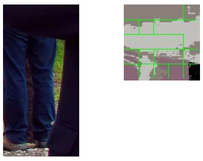
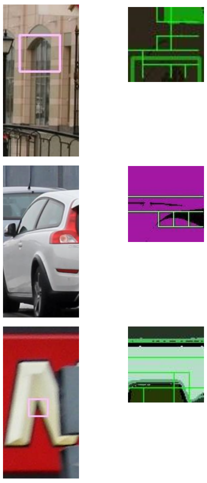
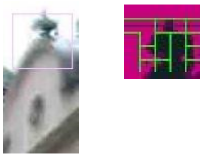
	Voila Jones Correct	Voila Jones Correct	Voila Jones Wrong	Voila Jones Wrong
	Appearance Correct	Appearance Wrong	Appearance Correct	Appearance Wrong
Pedestrian				
False Alarm				

Table 4.4 **Upper-body verification using Viola-Jones upper-body detector and the appearance based algorithm. The first row of image show the reactions of two algorithms on the ground truth pedestrians and the second row show the example results of two algorithms on ground truth false alarms.**

4.5 Summary

In this chapter, two issues of pedestrian detection are considered: false alarm reduction in the results of HOG-SVM pedestrian detection and the identification of reappeared false alarms when apply pedestrian detection to video streams. HOG and Viola-Jones approaches for pedestrian detection are reviewed. The performances of trained pedestrian detectors in CV (2.4.3) are compared.

Cited from the originated literatures, recent review papers and experiments performed in this chapter, HOG-SVM pedestrian detector outperforms other detectors provided by CV (2.4.3): HOG cascaded detector and Haar-like cascaded detector for example. For the majority cases, they produce promising results as shown in *Table 4.1*. When applied to certain images, they may perform less satisfactory and introduce high false alarms rate in detection results. According to observations in this chapter, large number of false alarms come from fences and vehicle parts. To reduce the false alarms that have similar shapes with pedestrians, features from body-parts are considered. In this chapter, head-shoulder structure is believed as an important evidence in separating false alarms from correct detections. Viola-Jones is a commonly used algorithm in body-parts detection. A wide range of body parts Viola-Jones detectors are provided by CV (2.4.3), including the ones for detecting "frontal face", "profile face", "upper body" and "lower body" using Haar-like features[9]. These Haar-like detectors can be combined with HOG-SVM detector as the sizes of ROIs used in these detectors match the sizes of ROIs of HOG-SVM detector. However, the performances of Voila-Jones detectors are difficult to control and may fail with images containing backward / side viewed pedestrians. Another means of appearance based upper-body verification is provided. Applying the algorithm to pedestrian detection results of HOG detector, false alarm rate is effectively reduced.

When applying pedestrian detection to video streams, it is complicated to verify every detected occurrences especially when the same false alarms will reappear in hundreds of frames in a video. The re-identification algorithms introduced in the last chapter are adopted to simplify the procedure.

[9] LPB cascading face detector is also provided in CV (2.4.3). Due to the lack of LPB upper-body detector, the detector is not considered in our experiments.

Chapter 5

Multiple Pedestrians Detected in One Box

Reality is merely an illusion, albeit a very persistent one.

——Albert Einstein

In [3], inaccurately detected pedestrians are categorised into inaccurate detection, partial detection, multi-pedestrians and non-pedestrians. In *Section 4.1.2*, relevant parameters used in HOG pedestrian detector originating the inaccurate detections were analysed. The reduction of some kinds of inaccurate detections and false alarms were discussed. In this chapter, a novel means of separating multiple pedestrians detected in one bounding box is proposed. The whole chapter is presented in two parts: the reason to multiple pedestrians detected in one bounding box will be briefly reviewed; after that, the novel framework of separating closely standing pedestrians detected as one will be introduced.

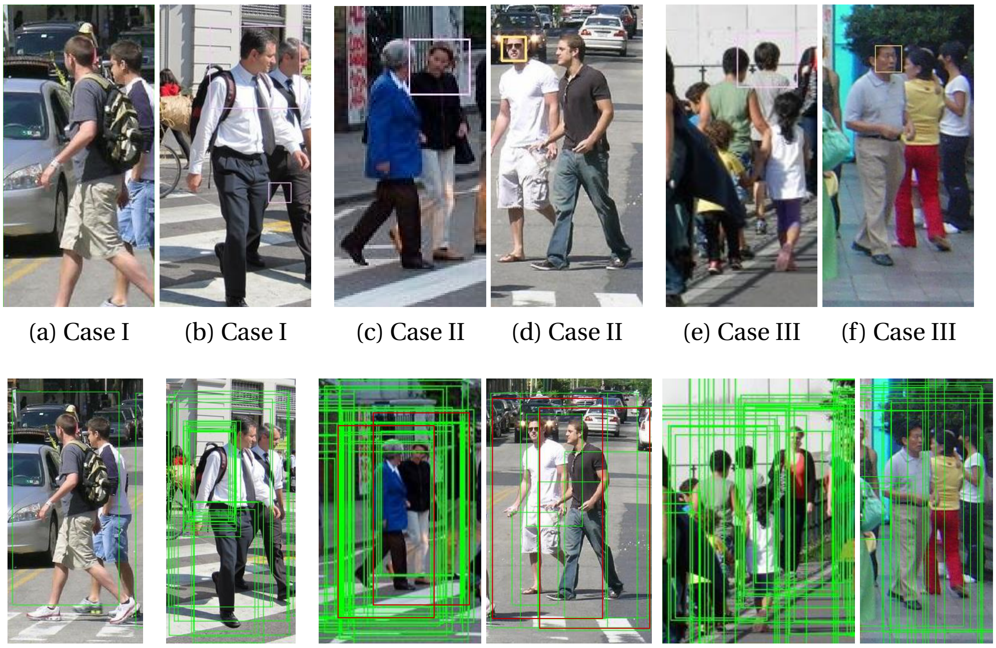
5.1 Multiple Pedestrians Detected in Single Box

When two pedestrians are standing closely to each other, they may be detected in one bounding box. Before promoting strategies to detect the number of pedestrians in the bounding box and separate multiple individuals, the reasons to this observations will be discovered. In this chapter, observations of multiple pedestrians detected as one are categorised in three groups as shown in *Fig. 5.1*:

1. A pedestrian is partially occluded by the other pedestrian and the occluded one is miss detected;

2. Both pedestrians are detected. But their bounding boxes are too close to be clustered into one;

3. Detected pedestrians were in a crowd.



(a) Case I (b) Case I (c) Case II (d) Case II (e) Case III (f) Case III

(g) Result I (h) Result I (i) Result II (j) Result II (k) Result III (l) Result III

Fig. 5.1 **Commonly observed multiple pedestrians detected as one, the first row (a) - (f); and their detection results when no bounding box clustering is performed in the second row (g) - (l). In Case I, the occluded pedestrian may be neglected by the HOG pedestrian detector; In Case II, both pedestrians have positive responses to the HOG detector. The red bounding boxes in (i) and (j) show the average position of bounding boxes which have positive responses. But they are too close to be clustered as one; In Case III, pedestrians in crowds are difficult to be separated even for human. HOG may have positive responses to individuals in the crowds.**

In type I cases, the occluded pedestrian may not be detected even when no bounding box clustering is performed. The pedestrian in the centre of the bounding box is the main detection, the occluded one would appear occluded next to the main detection as a part of background. In type II cases, all appeared pedestrians are detected but their bounding boxes are merged to one in clustering. Pedestrians detected in case II are usually occupying equally amount of areas within the bounding box. Multiple pedestrians detected in the two types of cases are surrounded by other background objects. In the third type of cases, the pedestrian detected in the centre of the bounding box is surrounded by a crowd of other pedestrians, full size or occluded. In following sections, type II cases of a pair pedestrians detected in one box will be mainly addressed using the novel means of rough symmetry axis detection. To separate the type I and III cases of multiple pedestrians detected in single box, the appearance based head area detection algorithm introduced in *section 4.2.2* should be considered to obtain other evidences of the presence of pedestrian.

5.1.1 How Close the Pedestrians Who are Detected as One?

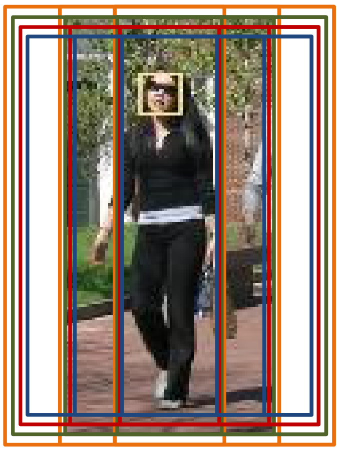
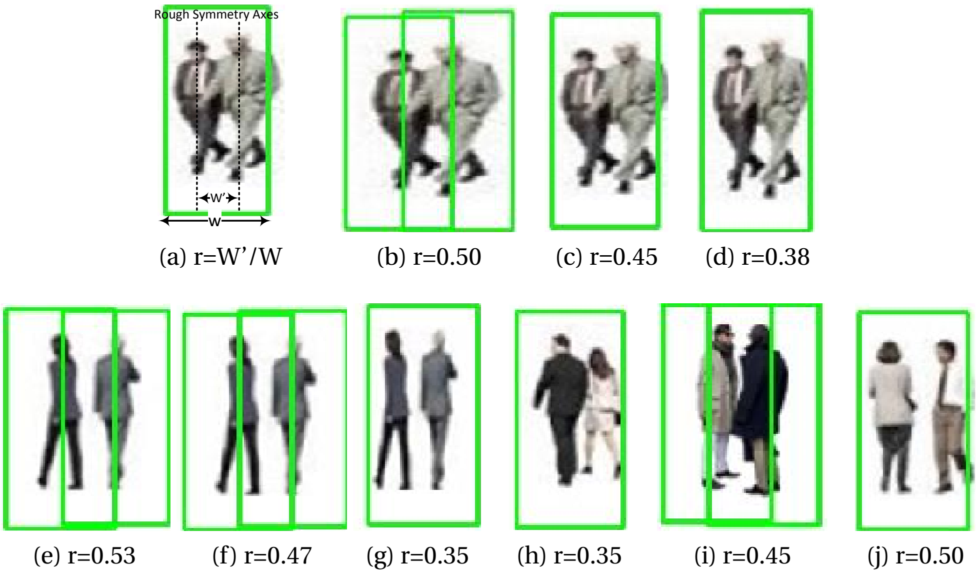


Fig. 5.2 **A pedestrian may have multiple positive responses to sliding windows in different scales. Bounding boxes in different colours are sliding windows from different scales of images. The sliding stride is a quarter of the width of the window and the scalding factor is 1.05**

Default values of parameters selected in CV (2.4.3) HOG-SVM detector are an optimised settings considering both detection rate and false alarm rate as illustrated in [22]. Without clustering the detected bounding boxes, it is difficult to locate the detected pedestrians due to over detection. After clustering, bounding boxes with similar sizes and positions are clustered. The averaged box of the cluster is calculated as the new bounding box for detected pedestrians. As shown in *Fig. 5.2*, using the scaling factor 1.05 and sliding window stride equals to a quarter of the width of the window, a pedestrian may have positive responses upto three neighbouring sliding windows in more than five scales. When two pedestrians are closely standing in an image, positive responses of both pedestrians may overlap. After clustering the bounding boxes, the output bounding box will contain two or more pedestrians.

Due to the stride of sliding windows is a quarter of the width of sliding windows, assumptions are made that when the distance between centroids of pedestrians is smaller than a half of the width of sliding windows, the two pedestrians would be detected as one. In observations, the ratio of the distance between centroids of pedestrians to the width of the bounding boxes are calculated. The centroid of a pedestrian



(a) r=W'/W (b) r=0.50 (c) r=0.45 (d) r=0.38

(e) r=0.53 (f) r=0.47 (g) r=0.35 (h) r=0.35 (i) r=0.45 (j) r=0.50

Fig. 5.3 **HOG performance in separating closely appeared pedestrians with Different Ratio**

is on the rough symmetry axis of the pedestrian as shown in *Fig. 5.3*. Experiments shown when the ratio is smaller than 0.5 (ratio < 0.4), pedestrians may always be detected in one box; when ratio > 0.6, pedestrians may have separate responses to sliding windows; when ratio ~ 0.5, the results may vary from circumstances to circumstances. *Fig. 5.3* illustrates the performances of HOG in separating two closely appeared pedestrians (with limited occlusion) in blank background.

5.1.2 The Rough Symmetry Axes Detection

In this section pedestrians bounded in one box will be separated according to their symmetry axes. Unlike manufactured objects, pedestrians are not mathematically symmetrical though symmetry axes of pedestrians could be located cognitively. The term rough symmetry is used to describe this situation where the two symmetrical counterparts are similar in appearance. The word similar tolerates the differences introduced not only by the affine transformation or light conditions but also by slightly changed shapes or patterns of the pedestrians. Clothes on pedestrians may have drapes or asymmetrical patterns, pedestrians may not pose in a symmetrical gesture etcetera.

Not many literatures considered the symmetry axes detection of objects or pedestrians [84, 85]. The state-of-the-art algorithms used to detect symmetry axes in objects can be categorised in two groups: one detect the symmetry axes by matching the interest points as introduced in [86] and the other correspond regions of interests to locate the symmetry axes [87, 88]. Clustering and Hough transform are commonly used techniques to detect the correspondent pixels / regions. These sophisticated algorithms are less capable in detecting human like quasi-symmetrical structure. [88] illustrated that only few symmetry pairs can be detected from a human face, which result in less reliable symmetry axes. Still, as a support algorithm to identify and separate multiple pedestrians detected in single bounding box, the processing procedure should be fast. None of the above algorithms are efficient enough for real-time applications.

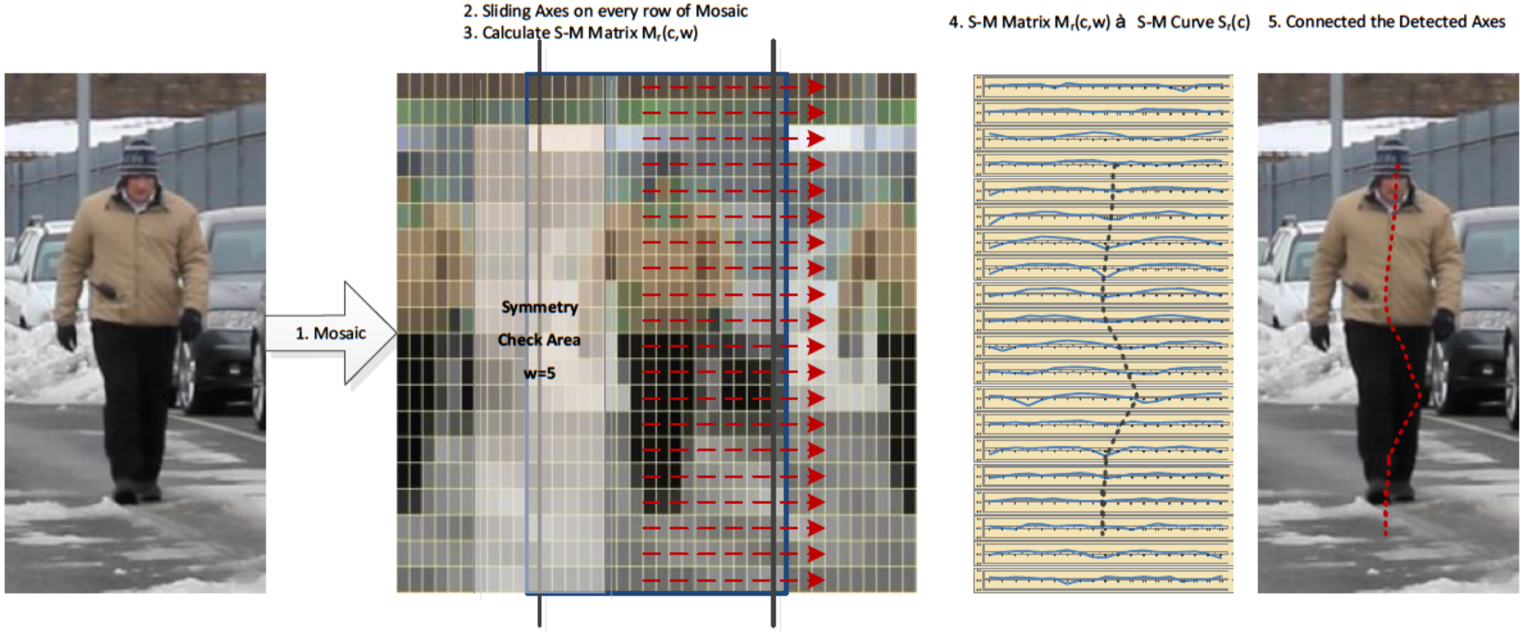

1. Mosaic
2. Sliding Axes on every row of Mosaic
3. Calculate S-M Matrix M$_r$(c,w)
Symmetry
Check Area
w=5
4. S-M Matrix M$_r$(c,w) à S-M Curve S$_r$(c)
5. Connected the Detected Axes

Fig. 5.4 **Demonstration of the rough symmetry axes detection in *Section 5.1.2* : 1. HOG detections are transferred to mosaic images; 2. Sliding axes on every row of the mosaic image and calculate the S-M Matrix; 3. Transfer the S-M Matrix to S-M Curve and normalise the S-M Curve; 4. Detect the peak value of every S-M Curve; 5. Connect the detected symmetry axes on each row. Note: the values in S-M Curve in this image are inverse proportional to the symmetrical condition, the lower the symmetrical measurement, the more probable the symmetry structure.**

The detection of the rough symmetry axes within the bounding box containing probable more than one pedestrians follows below steps as demonstrated in *Fig. 5.4*:

1. *Mosaic Image*: Transfer the input detections into mosaic image where the intensities of pixels in tiles of mosaic images are the same. In experiment, input images (detected bounding boxes using HOG-SVM) are separated into $n \times m$ cells (n rows and m columns). To maintain the cognitive appearance of the mosaic image with its origin, the median HSV value of pixels is calculated as the intensity of the piece of tile.

2. *Symmetrical Measurements of the r^{th} row*: The approximated symmetry axis of a pedestrian may not a straight line. To detected such symmetry axes, in a $n \times m$ tile mosaic image, the symmetry axes are detected on every row and then connected as the symmetry axis of the pedestrian. To determine the symmetry axes of the r^{th} row, the Symmetrical Measurement Matrix (S-M[1] Matrix), $\mathbf{M}_r$ will be calculated. The matrix is calculated by sliding an axis from the common border of the first two cells to the common border of the last two cells with sliding stride equals to one cell. In each sliding position, the right and left areas within width of w cells ($w = 1, \ldots, m$) to the sliding axis are compared. The element on the c^{th} row and w^{th} column of S-M Matrix $\mathbf{M}_r(c, w)$ is the $2w$ wide symmetrical measurement of sliding axis on the c^{th}-column and the r^{th}-row of the mosaic image. This measurement is the median of w colour differences $\mathbf{d}$ between pairs of reflected tiles to the sliding axis, $\mathbf{M}_r(c, w)$ is (I_i is the intensity of the i^{th} tile on the r^{th} row of the mosaic image):

$$\mathbf{d}_i = \| I_{c+i} - I_{c-i} \| \quad i = 1, 2, \ldots, m \tag{5.1}$$

$$\mathbf{S}_r(c, w) = \tilde{\mathbf{d}} \tag{5.2}$$

$$c + i \equiv c + i(\text{module } m) \quad \text{when } c + i > m \tag{5.3}$$

$$c - i \equiv c - i(\text{module } m) \quad \text{when } c - i < 0 \tag{5.4}$$

3. *Verify the Symmetrical Measurements*: The symmetrical measurements calculated on a range of width aim at filtering the weak symmetrical structures. As shown in *Fig. 5.5*, even when the width w is over the actual range of the symmetrical structure, the measurements of strong symmetrical structure are still over the setting threshold ($T = 0.15$). Further more, this can be treated as the reference of the range of the symmetrical structure: the w that reaches the

[1] S-M: Symmetrical Measurement

peak can be believed as the width of the symmetrical structure. To decide if the sliding axis on the c_{th} column and the r^{th} row of the mosaic image is a strong symmetrical axis, the median value of every row of S-M Matrix $\mathbf{M}_r(c, w)$ is selected as the symmetrical measurement of the row. This reduce the S-M Matrix to S-M Curve $\mathbf{S}_r(c)$, where:

$$\mathbf{S}_r(c) = \tilde{\mathbf{M}}_r(c, w) \; w = 1, 2, \ldots, m \tag{5.5}$$

See *section 5.1.2.1* for details.

4. *Normalise the S-M Curve*: Symmetrical measurements over threshold at position c calculated in last step is the sufficient but not necessary condition to a rough cognitive symmetry axis. A solid colour block may have valid symmetrical measurements everywhere according to the rules above. The symmetry axes of a rough symmetrical object should meet two conditions:

   (a) It should be in the valley position of the symmetrical measurement curve;

   (b) Its symmetrical measurement should over threshold.

   To detect the validated symmetry axes of rows in mosaic image, the S-M Curve of r^{th} row $\mathbf{S}_r$ should be normalised with the median value ($\tilde{\mathbf{S}}_r$) of the curve. The normalised S-M Curve is projected into negative logarithm coordination so that peaks over median value are amplified. The normalised S-M Curve is:

$$\bar{\mathbf{S}}_r = -\log \frac{\mathbf{S}_r}{\tilde{\mathbf{S}}_r} \tag{5.6}$$

5. *Detect the Rough Symmetry Axes of r^{th} Row*: Calculate the first derivative of the normalised S-M Curves:

$$\bar{\mathbf{S}}'_r = \frac{d\bar{\mathbf{S}}_r}{dc} = \bar{\mathbf{S}}_r(c) - \bar{\mathbf{S}}_r(c-1) \tag{5.7}$$

   Detect the peak points where the first derivative value change from positive to negative. If its symmetrical measurement is over threshold, the position is recognised as a symmetry axis of the row. This axis of the row at position c should meet following criterion:

   (a) $\bar{\mathbf{S}}'_r(c-1) > 0$;

   (b) $\bar{\mathbf{S}}'_r(c) < 0$;

   (c) $\bar{\mathbf{S}}_r(c) > T$ ($T = 0.15$ in experiment);

6. *Connect Symmetry Axes of Rows*: Multiple axes may be detected in each row of the mosaic images. Map the positions of detected symmetry axes of each row to $n \times (m-1)$ matrix P where n and m are the numbers of rows and columns in mosaic image respectively. If a symmetry axis on row j is determined between the common border of i^{th} and $(i+1)^{th}$ cell, $P(j,i) = 1$, otherwise $P(j,i) = 0$. In matrix P, 4-connected value 1 elements are connected. The more value 1 elements connected, the longer the the axes. The variance of the column indices of connected elements of the perspective axes with length over threshold (T) are calculated to verify the perpendicularity of the rough symmetry axis as pedestrians are normally appeared up-right in images:

$$\mathrm{Var}(I) = \mathrm{E}[(I - \bar{i})^2] \leq C, i \in I, i = 1, \ldots, m-1, P(j,i) = 1 \tag{5.8}$$

   Consider the type I and type II cases of multi-pedestrians detected as one, where pedestrians would occupy at least 50% of the bounding box area, axes longer than 30% the height of the image (T = $0.3n$) are transformed to the input image as the rough symmetry axes. In type III cases, when pedestrians in the bounding boxes are significantly occluded, shorter symmetry axes can be tolerated. This will not be considered in this thesis.

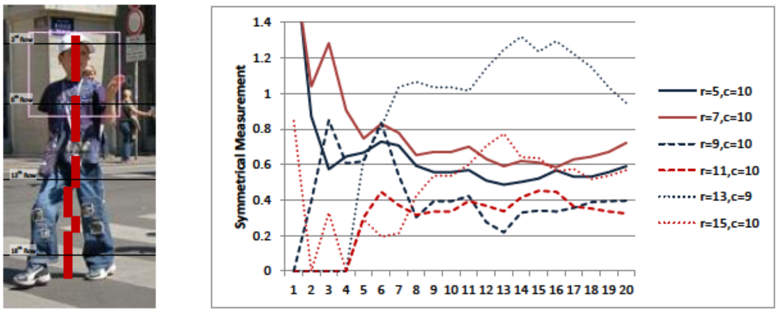
Using mosaic image in symmetry axes detection greatly improve the processing speed as pixels in cells are identical. Image intensity clustering also improve the symmetrical appearance of a pedestrian, but applying clustering in an image containing $\sim 10,000$ pixels[2] or more is computational expensive.

5.1.2.1 The Width w in Calculating the Symmetrical Measurement

The width w is an important parameter used in calculating the symmetrical measurements. Symmetrical measurement of sliding axis at one position is obtained as the median value of the distances between mirror reflected cells to the axis. Using median value in detecting the rough symmetry axes, the symmetrical measurement will be mainly affected by the number of pairs of similar and different mirror reflected cells to the axis rather than the similarity values between those pairs of cells. This is crucial when the actual width of the symmetrical structure is unknown. Increasing the width w from 1 to the actual width of the rough symmetrical structure, the symmetrical measurements will increase as more symmetrical pairs of cells are

[2] The smallest detected bounding box using HOG pedestrian detector trained by CV (2.4.3) has size 128 × 64 pixels. This means, the results of HOG pedestrian detection usually contain more than 8000 pixels.

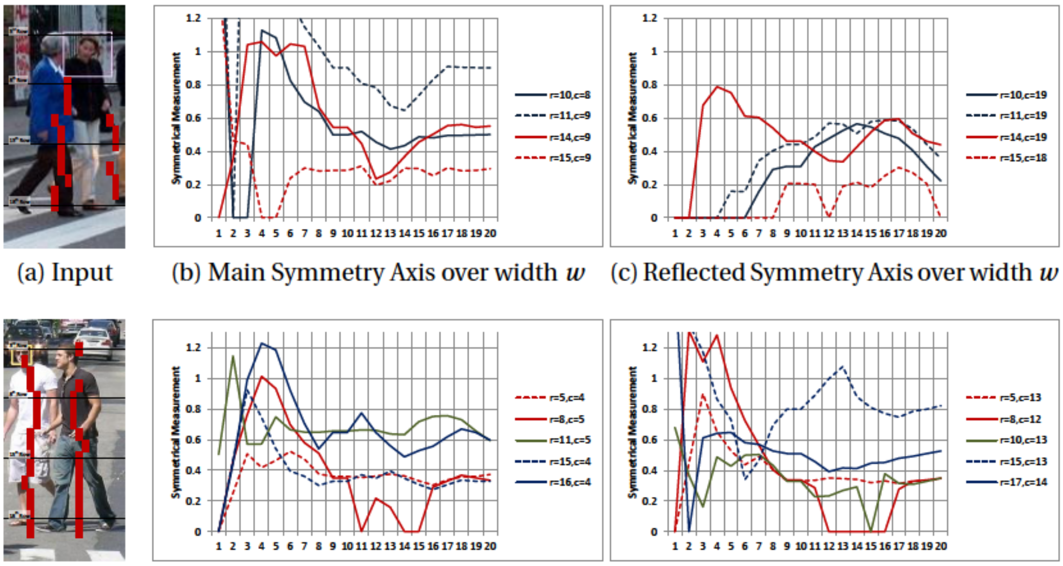
added to the series of symmetrical measurements; further increasing the width w beyond the actual width of the rough symmetry, the median value may not be significantly affected by non-identical pairs of cells with different colours. As shown in *Fig.* 5.5, the symmetrical measurements of symmetry axes on each row are all over threshold ($\forall\, w = 1, \ldots, 20 : \mathbf{M}_r(c, w) > 0.15$ in experiment). By selecting the median value of the symmetrical measurements to series of w, strong symmetry axes of rows could be separated from the weak ones. More examples are shown in *Table 5.1*.



(a) Input (b) Symmetrical Measurement over width w

Fig. 5.5 **The values of** 9^{th}, 10^{10} **rows of S-M Matrix** ($c = 9, 10$, **the position of strong symmetry axes) calculated on the** $5^{th}, 7^{th}, 9^{th}, 11^{th}, 13^{th}, 15^{th}$ **row of mosaic image. The symmetrical measurements are calculated over a range of** w **from 1 cell to 20 cells (the width of the row). The 4 horizontal lines on the input image indicate the position of the** 3^{rd}, 8^{th}, 13^{th} **and** 17^{th} **row of its mosaic image.**

The curve of symmetrical measurements of a symmetry axis to the width w provides information on the width of the actual symmetry structure. In experiment, 20 cells are separated in the mosaic image, this curve is calculated by using w ranges from 3 to 20 cells. One cell is used as the unit of increment. Symmetrical measurements of axes when $w = 1$ or 2 are ignored as the measurements is less convinced for the symmetry axes of pedestrians when only a small area is considered. When w over half the width of the row, due to the usage of circular indexing, pairs of mirror reflected cells are compared repeatedly. Sometime, the peaks of the symmetrical measurement to w curve may appear twice, once near the actual width of the symmetry structure and the other echoed at the half width of image away from the first peak. A detected symmetry axis may have a reflected symmetry axis around half the image width away as shown in *Fig.* 5.6 *(a)*. Twin symmetry axes happen usually when two pedestrians bounded in one box have similar appearance. Normally, the symmetrical measurement of the reflected symmetry axis is smaller than the symmetrical measurement of the main one and the peak position of $\mathbf{M}_{r,w=1,\ldots,20}(c, w)$ curve of the reflected axis is usually beyond the half image width. *Fig.* 5.6 *(a) to (c)* demonstrate a typical twin symmetry responses to the changes of width w and *Fig.* 5.6 *(d) to (f)* are the responses of two correctly located symmetry axes to width w. But this cannot be used as the judgement of twin symmetries. To determine if a detected symmetry axes indicate a pedestrian and to reallocate the twin symmetry axes introduced by similar appeared pedestrians, HOG should be reapplied to the area around the detect symmetry axes.



(a) Input (b) Main Symmetry Axis over width w (c) Reflected Symmetry Axis over width w

(d) Input (e) First Symmetry Axis over width w (f) Second Symmetry Axis over width w

Fig. 5.6 **The values of rows of S-M Matrix (c, the position of strong symmetry axes) calculated on the r^{th} rows of mosaic image. The symmetrical measurements are calculated over a range of w from 1 cell to 20 cells (width of the row). When two pedestrians having similar appearance are bounded in one box (a) - (c), the main symmetry axis would be detected between the two pedestrian and a reflected symmetry axis would be detected half the image width away from the main symmetry axis. (d) to (f) show an example of two well detected symmetry axes. The horizontal axes of the charts are the indices of the common border of cells in mosaic image: i is the common border between the i^{th} cell and $(i+1)^{th}$ cell.**

5.2 Experiment

In this experiment, input images are the detected pedestrian(s) of HOG. False alarms are not considered in this section as symmetry is not a feature to separate pedestrian from the artificial manufactured objects which have more symmetry appearance than human. Input images are transferred into mosaic image of 20 × 20 cells. The colour of a cell is the median HSV intensity value of pixels in the cell. In the mosaic image, following steps are performed to detect the symmetry axes:

1. In each row of the mosaic image, slide an axis at each common border of neighbourhood cells, calculate the 18 symmetrical measurements at each position for width $w = 3$ to 20 cells:

   (a) For each w, calculate the series of differences between mirror reflected pairs of cells to the sliding axis. The differences are calculated using Euclidean distance between the HSV intensities of two mirror reflected cells to the sliding axis. Circular index is used when cells are out of range;

   (b) Normalise the series of differences using the median of the series and transform the value of series to the negative log coordination plane (the more symmetrical the structure, the larger the value);

   (c) Calculate the first order derivative of the series of normalised distances and record the peak position of which the symmetrical measurement is over threshold = 0.15;

2. Select the median value from the 18 symmetrical measurements calculated using w from 3 to 20 cells. If the value is over threshold (= 0.15), the position of the current sliding axes is recognised as a strong symmetry axes;

3. Connect 4-connected symmetry axes in rows of mosaic image;

4. Re-apply HOG pedestrian detector to the local area of the detected symmetry axes. If the size of the input image is $M \times N$, choose an area sizing $1.5M \times N$ centred at the axis to re-apply HOG detector. Average the position of positive responses around the detected axes as an updated pedestrian.

Twin symmetry axes are rarely seen in a box of pedestrians with different appearances of single correctly detected pedestrian. This is because the width of one pedestrian is relatively small and the symmetry responses to width w is less strong for a reflected symmetry axis. For type II cases of multiple pedestrians bounded in one box where limited occlusion is seen from the detection and all pedestrians have positive response to HOG pedestrian detector, above algorithm show promising result:

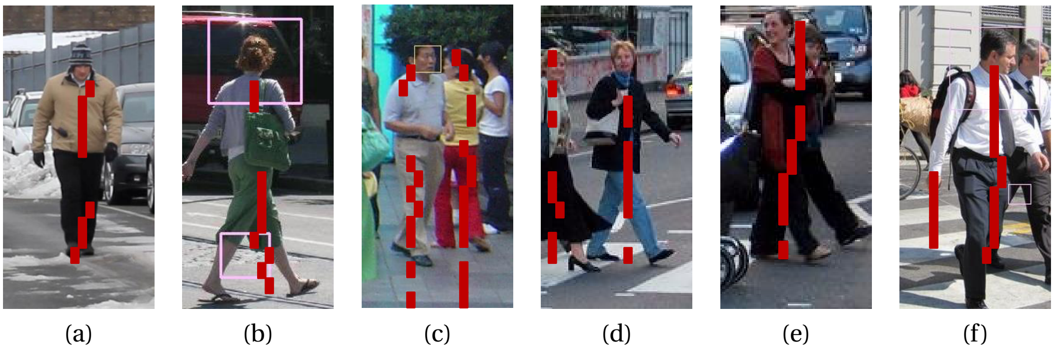


Fig. 5.7 **Examples of Symmetry Detection in Multi-pedestrian Separation**

in testing dataset containing ~ 100 HOG detections of multiple pedestrians, around 90% of cases, multiple symmetry axes are detected. For type I cases where one pedestrian is occluded, the performance of the symmetry axes detection may vary due to the condition of occlusion. For type III cases, severe occluded pedestrians would be neglected while the symmetry axis of pedestrians with limited occlusion would still be detected. The background is object or other pedestrians in a HOG detection have less effect in the performance of this symmetry detection algorithm. *Fig. 5.7* demonstrate more symmetry axes detection on HOG detected pedestrians. The symmetry appearance will be affected by accessories.

For occluded pedestrians with visible head-shoulder, applying the appearance based head part verification algorithm introduced in *section 4.2.2* to the relevant head part, positive responses would indicate a presented head. While that algorithm is not a proper head detection algorithm, repeated performances can be retrieved. This is a promising attempt when an occluded pedestrian who is neglected by HOG pedestrian detector. Still, only the positive responses detected at the top of the symmetry axes area within the perspective pedestrian images should be considered as this verification algorithm can hardly decide if the object is pedestrian or background object in general environment.

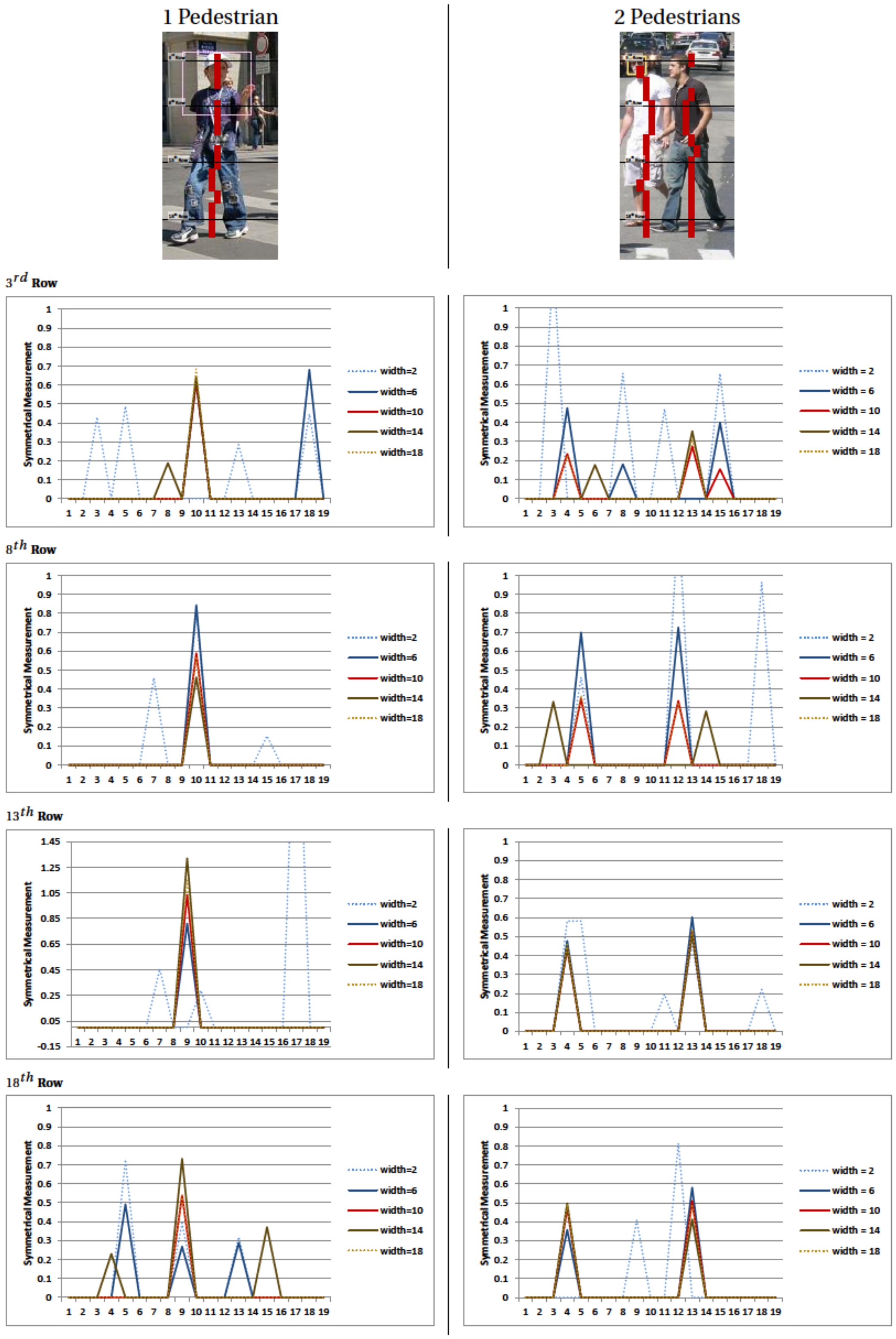
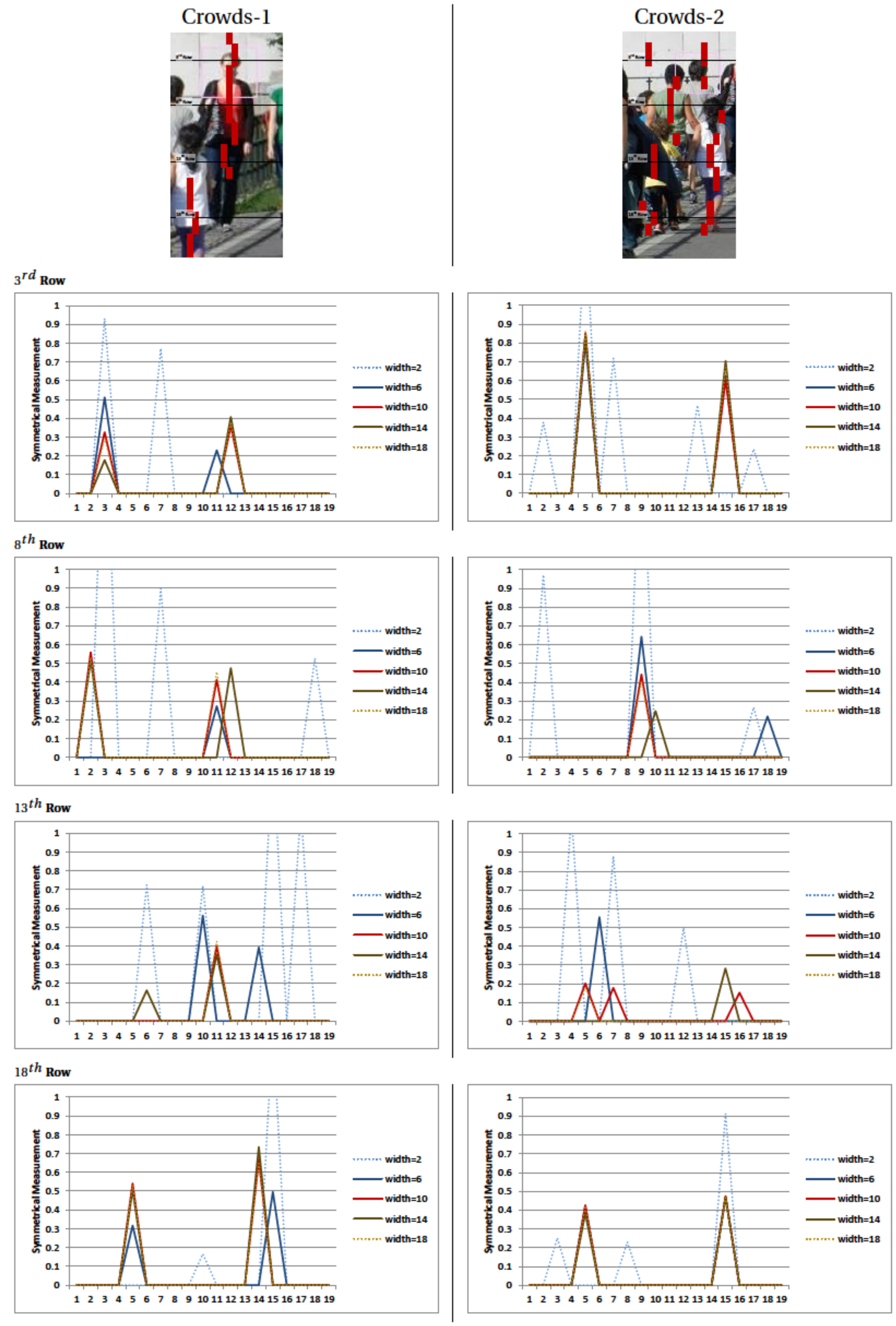


Table 5.1 **The $2^{nd}, 6^{th}, 10^{th}, 14^{th}, 18^{th}$ column of S-M Matrix (w) over range of the position of sliding axes (c). The horizontal axes of the charts are the indices of the common border of cells in mosaic image: i is the common border between the i^{th} cell and $(i+1)^{th}$ cell.**



3^{rd} Row

8^{th} Row

13^{th} Row

18^{th} Row

Table 5.1 **(Cont.) The $2^{nd}, 6^{th}, 10^{th}, 14^{th}, 18^{th}$ column of S-M Matrix (w) over range of the position of sliding axes (c). The horizontal axes of the charts are the indices of the common border of cells in mosaic image: i is the common border between the i^{th} cell and $(i+1)^{th}$ cell. When pedestrians are occluded, more than one symmetry axes will be detected, in the "Crowds-2" image, symmetry axes detected are normally short and their symmetrical measurements are usually low.**

5.3 Summary

In this chapter, more than one pedestrians detected as one are considered and a novel symmetry axes detection algorithm is proposed to separate the pedestrians in one bounding box. Human is not perfectly symmetrical but they have quasi-symmetry appearance. When sliding an axis horizontally and compare the mirror reflected pixels to the sliding axis in an area, the detection of symmetry axes would be sensitive to the appearances of pedestrians: patterns on clothes, limb gestures, clothes drapes may affect the symmetrical measurement and cause the pedestrian mathematically asymmetry. To reduce the sensitivity of the detected symmetry axes:

- The cell based mosaic images are introduced to replace the pixel based input images;
- The detection of an integral symmetry axis is replaced by the connected symmetry axes detected on each row of the mosaic image.

The substitution of the pixel images for mosaic images greatly reduces the processing speed: the complexity of the calculation spent on the verification of symmetrical structures is reduced to $\mathcal{O}(n)$ from $\mathcal{O}(n^2)$ using the original pixel based image, where n is the number of pixels in the image. This is especially important when the algorithm is designed as an assistant step to reduce the less-ideal detections of HOG pedestrian detector. To select the validated symmetry axes, a wide range of width w (from 3 to 20) are used in calculating the symmetrical measurements of the sliding axis, only the axes whose symmetrical measurements are less affected by the change of w are considered as candidate axes. In area of 4-connected neighbourhood of a detected candidate axis, this current axis is connect to the candidate axes detected on adjacent rows. After connection, axes have considerable length [3] will be recognised as the symmetry axes of pedestrians. Sometimes, the position of the detected symmetry axes may deviated from the centres of pedestrians especially when two closely appeared pedestrians are similar in appearances. In this case, HOG pedestrian detector will be reapplied to the surrounding areas of detected symmetry axes to determine the positions of pedestrians.

In cases of multiple pedestrians detected as one, two pedestrians in one detection is more commonly seen than other cases where a crowd is detected. A pair of pedestrians with limited occlusion [4] detected as one is due to the clustering of sliding windows in HOG detection. In this case, the symmetry axes detection algorithm

[3] In experiment, 30% of the total height of the image is considered valid;

[4] Less than 30% of the total area is occluded.

has promising performances: 90% of ~100 HOG detection results containing multiple pedestrians can be separated. When pedestrians in the bounding boxes have negative responses to the HOG pedestrian detector due to occlusion, though two symmetry axes may be detected in the detection, the one of the occluded pedestrian will hardly be verified using current technologies. As an attempt, the result of the head-shoulder verification algorithm can be considered as an evidence of presence. Further research is required to discover the performances of the algorithm. Similar situation happen when a group of pedestrians are detected in one box. Though the symmetry axes detection may not be much affected by the background (objects or pedestrian), only strong symmetry shape should be recognised. The symmetry axes detection algorithm is not a proper algorithm for pedestrian detection as symmetry is not a feature to separate pedestrians from background objects which are more mathematically symmetrical.

Chapter 6

Conclusions and Future Work

We can only see a short distance ahead, but we can see plenty there that needs to be done.

—— *Alan M. Turing*

Several issues in pedestrian detection and recognition have been discussed in this thesis. Appearance related features based algorithms for pedestrian re-identification were introduced. A protocol was developed to quantify the complexity of the appearances of pedestrians so that different effort would be spent on identifying different prototype pedestrians. After that, strategies were proposed to reduce the false alarms in HOG pedestrian detection and to separate multiple pedestrians detected in one bounding box using the HOG-SVM pedestrian detector trained by CV(2.4.3). Based on discussions in last several chapters, conclusions will be summarised, future work will be presented in this chapter. Moreover, a thought on vision and perception related tasks in machine learning will be narrated in the next chapter.

6.1 Conclusions

Pedestrian Identification and Detection are two challenging topics in computer vision. It is difficult to decide which features play a key role in detecting and identifying a pedestrian from images, though human are expert in this area. Seen from the state of the art literatures, shape related features are usually selected to distinguish pedestrians from other background objects. Furthermore, appearance based features, such as colours, patterns and textures, are popular choices for recognising the identity of a prototype pedestrian. Taking advantage of improved computation technology, employing combined features and algorithms is a recent trend in pedestrian detection and identification. However, problems emerged with the development of techniques for pedestrian detection and identification:

- Applications for pedestrian detection and identification, such as surveillance and monitoring, normally demand real-time processing. Rich descriptions and sophisticated frameworks increase the difficulties in the development of relevant industrial applications. Furthermore, complex algorithms are not necessary for every circumstances;
- Classification algorithms are normally based on statistical modelling and human experiences based logics, both of which have been proved unreliable to imitate the human vision system. This means, the state of the art frameworks may have limitations with certain identification and detection cases. As presented in *Chapter 4*, large amount of false alarms exist in the detection results of the famous HOG SVM pedestrian detector applied to the video frames shot in the car park.

Inspired by above problems, a novel protocol is proposed in *Chapter 3*. The protocol contains layers of algorithms, the more layers applied, the more capable the framework to solve the complex identification and detection problems. The hierarchical layers introduced in the protocol increase the flexibility of the framework. Whenever required, algorithms can be added into the protocol. This overcomes the limitations of the popular classification strategies, especially the generic ones including SVM and Boosting.

Discussions in pedestrian detection and identification can never avoid pedestrian tracking. As discussed in [7], this is similar to the chicken-egg problem. The continuous detection is an equalised tracking and the tracking of a pedestrian involves the initial detection and the identifications of later appeared individuals. The pedestrian detection applied to video frames in *Chapter 4* can be treated as the tracking of pedestrians. Using pedestrian-detection-by-tracking false alarms may reappear in later frames. It is expensive in calculations to apply false alarm reduction algorithms to every frame. Simple appearance based algorithms for pedestrian identification are adopted to reduce the reappeared false alarms.

Algorithms are always desired to improve the performances of the existing frameworks. To improve the performance of HOG SVM pedestrian detector: in *Chapter 4*, the presences of head-shoulders of the pedestrians are selected as the evidence of the presences of pedestrians. The appearance based algorithm and Viola-Jones body part detectors are applied to verify the head-shoulder structures of pedestrians. In *Chapter 5*, the approximated symmetrical feature is selected as the evidences of the numbers of pedestrians. Both algorithms are fast in calculation. In false alarm reduction, only 10% of the total area of the detection results are examined. To separate the multiple pedestrians in one box, the usage of mosaic images reduces the calculations spent on the verification of symmetrical structures in scale of $\mathcal{O}(n)$, where n is the total number of pixels of detection results. However, nothing is perfect, both the protocol and the algorithms proposed in this thesis demand future works.

6.2 In the Recent Future

6.2.1 Clothes Pattern Analysis

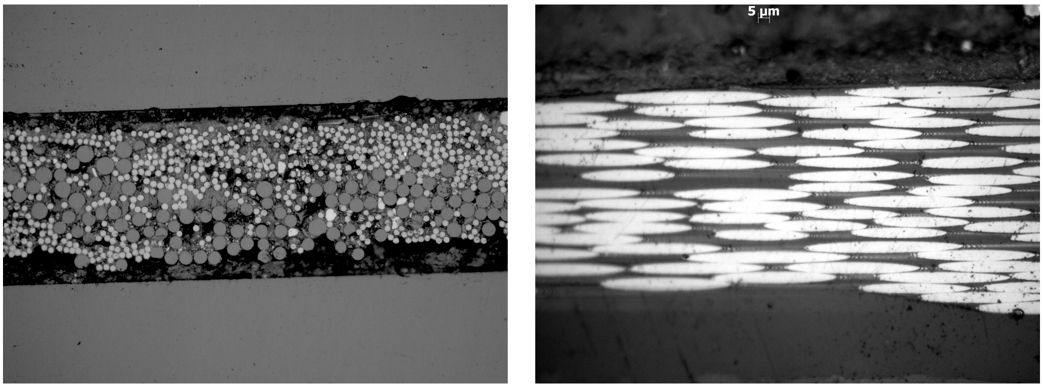
In *chapter 3*, histogram based descriptions were applied to synthesis the prototype and target pedestrians. To obtain more reliable results, descriptors for patterned clothes should be introduced. The complexity of patterns should be quantified by their colours and regularity when applied to the proposed pedestrian-re-identification protocol: the computation effort should be proportional to the complexity of the problem which is normally related the appearance of the pedestrian. Common patterns including dots, stripes and lattices in two or three colours may have lower levelled complicity comparing to variegated floral and irregular patterns. In current stage of research, stripes, simple lattices and dots pattern with limited occlusion can be detected and modelled using Hough transform. Less research has been done in description of other patterns.

6.2.2 Symmetrical Appearance based Pedestrian Detection

In *chapter 4 and 5*, algorithms were developed to detect the head-shoulder structures and symmetrical axes in the detection results of HOG-SVM pedestrian detector (trained by CV(2.4.3)). Despite the fast processing speed of the two algorithms, both of them introduce a necessary but less sufficient condition of the presence of a pedestrian. The application of head-shoulder detection is currently restricted to the detection results of HOG-SVM detectors. Future work will be done to improve the adaptation of the appearance based head-shoulder detection. Research will investigate the combination of the two algorithms into a novel protocol of pedestrian detection: to detect a pedestrian according to its rough symmetrical shape with a fixed range of geometry ratio and a head-shoulder appearance on the top.

6.2.3 Pattern Recognition in Chemistry and Biology

Pattern description and recognition will not only support the modelling of pedestrians. The recognition of circles and ellipses from a noisy background may have direct applications in fibre material production: the desired product may have different geometry measurements (radius for circles and eccentricities for ellipses). Manually counting the desired shapes of fibre materials out of total fabricated fibre materials from microscopic image is the current ways of evaluating the quality of the productions. For example, carbon fibre and glass fibres are usually mixed to obtain desired strength and flexibility. The quality of the combined fibre materials depends on the divergence and uniformity of the two mixtures. Recognising the number of different sized fibres (carbon fibres are thinner than the glass fibre) and determining their distribution through hundreds of sampled microscopic profile images will improve the efficiency of quality control which is currently processed manually during manufacture. Furthermore, recognition could be spanned to cells and emulsion droplets where the number of these objects are vastly required in relevant research and manufacture.As shown in *Fig. 6.1*[1]



(a) Mixture of Fibre Materials

(b) Single Kind of Fibre Materials

Fig. 6.1 **Two Examples of Pattern Recognition applied in Chemical Engineering: (a) is the cross-section image of carbon fibres (small white dots) mixed with glass fibres (big gray dots), to model the distribution of fibres in mixture, fibres of different materials should be recognised and located; (b) is the cross-section image of a kind fibre materials in production, the eccentricities of ellipses (the fibre material) in the images are required to analyse the quality of fibre materials. Above are lab images in excellent condition: boundaries of fibres can be clearly detected without overlaps or occlusion, fibres of different materials appear in both different colour and dimension.**

[1] Images from Polymer and Composite Engineering Group, Department of Chemical Engineering, Imperial College London

Chapter 7

An Evolving Recognition System: A Thought

Έν οίδα ότι ουδέν οίδα.[1]

—— Εωκράτης

Recognition is a basic function of human and some animals but it is a difficult problem for computers. There is even no precise definition for recognition rather than "to get something recognised". To simplify the problem, research in this area focus on activities of verification, detection and identification like problems addressed in this thesis. The human activity would be the desired model for the counterpart research in computer vision. If compare the recognition activities as the inputs and outputs with regard to the black box of the recognition system, recognition research aims in constructing the rules to process the input visual signals to the output cognition meanings which should as similar as the responses produced by the human black box. However, even the research of activities are difficult: the definition of the activity is based on the observations of human activity while the mathematical modelling is based on the conjecture and reasoning of the activity; the database of the activity cannot be enumerated; the evaluation of the activity is subjective and insufficient, the algorithms surviving from certain databases may fail in others.

Rather than developing an organ which could see and percept using machines and computers, it would be more practical to develop several genes relating to the function. Furthermore, with the evolution of computer vision, new genes will be collaborated with the existing ones, ancient genes would be mutated when updated replacements are developed. The work of the thesis would perform some of these

[1] I know nothing except the fact of my ignorance.—— *Socreates*

genes in the evolution of the recognition system. In the future, the recognition system may be developed in following routes:

- Take advantage of brains: similar to the invention of neural controlled arms, by discovering the brain activity in neural signal level may assistant in developing an interpreter of visual signal to cognition;

- Take advantage of computation: assembling computer vision algorithms for various vision related activities.

The first route will detect the ports where computer vision can be collaborated with human vision and the second one will create a parallel vision system which intimate the human vision system based on the observations of human vision activities. Human is the expert in recognition and identification which are difficult to be modelled or imitated by computer. There is not enough information on the procedure how human process and translate visual signals received by eyes to cognitive meanings in brain. It is foreseen that more and more complicated computer vision system will be developed. People said that computer vision may never be fully developed until the truth of human vision and perceptions are discovered like there was no aeroplanes until the mechanics of flying of birds could be modelled. Thesis like this may not contribute on human self-understanding in the ways of thinking nor perception. But it is an attempt in providing means of perception for computers to learn, though different to human vision, its application will benefit in certain area.

References

[1] M. Enzwiler and D. M. Gavrila, “Monocular Pdestrian Detection: Survey and Experiments,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.

[2] D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, “Survey of Pdestrian Detection for Advanced Driver Assistance Systems),” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1239–1258, 2010.

[3] 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, no. 4, pp. 743–761, 2012.

[4] D. Gray, S. Brennan, and H. Tao, “Evaluating Appearance Models for Recognition, Reacquisition, and Tracking,” in *IEEE International workshop on performance evaluation of tracking and surveillance*, 2007.

[5] S. Paisitkriangkrai, C. Shen, and J. Zhang, “Performance Evaluation of Local Features in Human Classification and Detection,” *Computer Vision, IET*, vol. 2, no. 4, pp. 236–246, 2008.

[6] D. Ramanan, D. A. Forsyth, and A. Zisserman, “Tracking People by Learning Their Appearance,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 1, pp. 65–81, 2007.

[7] M. Andriluka, S. Roth, and B. Schiele, “People-Tracking-by-Detection and People-Detection-by-Tracking,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[8] P. Dollár, Z. Tu, P. Perona, and S. Belongie, “Integral Channel Features,” in *BMVC*, vol. 2, no. 3, 2009, p. 5.

[9] A. Ess, B. Leibe, and L. van Gool, “Depth and Appearance for Mobile Scene Analysis,” in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[10] C. Wojek and B. Schiele, “A Performance Evaluation of Single and Multi-Feature People Detection,” in *Pattern Recognition*. Springer, 2008, pp. 82–91.

[11] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

[12] D. Hogg, “Model-based Vision: a Program to See a Walking Person,” *Image and vision computing*, vol. 1, no. 1, pp. 5–20, 1983.

[13] T. Tsukiyama and Y. Shirai, “Detection of the Movements of Persons from a Sparse Sequence of TV Images,” *Pattern Recognition*, vol. 18, no. 3, pp. 207–213, 1985.

[14] S. Riter, A. Bernat, and D. Schroder, “Computer Detection and Tracking of Moving People in Television Images,” in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, vol. 2, Aug. 1988, pp. 1013–1016.

[15] K. Rohr, “Towards Model-based Recognition of Human Movements in Image Sequences,” *CVGIP: Image Understanding*, vol. 59, no. 1, pp. 94–115, 1994.

[16] J. O’Rourke and N. I. Badler, “Model-based Image Analysis of Human Motion using Constraint Propagation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 6, pp. 522–536, 1980.

[17] M. K. Leung and Y. H. Yang, “First Sight: A Human Body Outline Labeling System,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 4, pp. 359–377, 1995.

[18] A. P. Pentland, “Machine Understanding of Human Action,” in *Proceedings of 7th International Forum Frontier of Telecommunication Tech.*, Nov. 1995.

[19] G. Doretto, T. Sebastian, T. Tu, and J. Rittscher, “Appearance-based Person Reidentification in Camera Networks: Problem Overview and Current Approaches,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 2, no. 2, pp. 127–151, 2011.

[20] M. Bauml and R. Stiefelhagen, “Evaluation of Local Features for Person Re-identification in Image Sequences,” in *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on.* IEEE, 2011, pp. 291–296.

[21] T. D’Orazio and G. Cicirelli, “People Re-identification and Tracking from Multiple Cameras: A Review,” in *Image Processing (ICIP), 2012 19th IEEE International Conference on.* IEEE, 2012, pp. 1601–1604.

[22] N. Dalal, “Finding People in Images and Videos,” Ph.D. dissertation, Institut National Polytechnique de Grenoble, 2006.

[23] P. Viola and M. Jones, “Robust Real-time Object Detection,” in *Proceedings of the Second International Workshop on Statistical and Computational Theories of Vision - Modeling, Learning, Computing, and Sampling*, Jul. 2001.

[24] V. Franc and V. Hlavác, “Statistical Pattern Recognition Toolbox for Matlab,” *Prague, Czech: Center for Machine Perception, Czech Technical University*, 2004.

[25] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2008.

[26] O. Chapelle, P. Haffner, and V. N. Vapnik, “Support Vector Machines for Histogram-based Image Classification,” *Neural Networks, IEEE Transactions on*, vol. 10, no. 5, pp. 1055–1064, 1999.

[27] J. Friedman, T. Hastie, and R. Tibshirani, “Additive Logistic Regression: A Statistical View of Boosting,” *The Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[28] W.-Y. Loh, “Classification and Regression Trees,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[29] Z. Lin and L. S. Davis, “A Pose-invariant Descriptor for Human Detection and Segmentation,” in *Proceedings of the European Conference on Computer Vision*, 2008, pp. 423–436.

[30] R. Lienhart and J. Maydt, “An Extended Set of Haar-like Features for Rapid Object Detection,” in *Proceedings of International Conference on Image Processing*, vol. 1, 2002.

[31] P. Sabzmeydani and G. Mori, “Detecting Pedestrians by Learning Shapelet Features,” in *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on.* IEEE, 2007, pp. 1–8.

[32] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, “Fast Human Detection Using a Cascade of Histograms of Oriented Gradients,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2006, pp. 1491–1498.

[33] M. Oren, T. Poggio, E. Osuna, P. Sinha, and C. Papageorgiou, “Pedestrian Detection using Wavelet Templates,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 1997, pp. 193–199.

[34] X. Wang, T. X. Han, and S. Yan, “An HOG-LBP Human Detector with Partial Occlusion Handling,” in *Computer Vision, 2009 IEEE 12th International Conference on.* IEEE, 2009, pp. 32–39.

[35] S. Walk, N. Majer, K. Schindler, and B. Schiele, “New Features and Insights for Pedestrian Detection,” in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on.* IEEE, 2010, pp. 1030–1037.

[36] W. R. Schwartz and L. S. Davis, “Learning Discriminative Appearance-based Models using Partial Least Squares,” in *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on.* IEEE, 2009, pp. 322–329.

[37] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object Detection with Discriminatively Trained Part-based Models,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[38] Y. Wu, T. Yu, and G. Hua, “A Statistical Field Model for Pedestrian Detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 1023–1030.

[39] D. M. Gavrila, “A Bayesian, Exemplar-based Approach to Hierarchical Shape Matching,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1–14, 2007.

[40] M. Enzweiler and D. M. Gavrila, "A Multilevel Mixture-of-Experts Framework for Pedestrian Classification," *Image Processing, IEEE Transactions on*, vol. 20, no. 10, pp. 2967–2979, 2011.

[41] F.-F. Li and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, 2005, pp. 524–531.

[42] R. Fergus, "Visual Object Category Recognition," Ph.D. dissertation, Department of Engineering Science, University of Oxford, Oxford, UK, Dec. 2005.

[43] J. N. Wilson and G. X. Ritter, *Handbook of Computer Vision Algorithms in Image Algebra*. CRC press, 2010.

[44] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. New Jersey: Pearson Education Inc., 2008.

[45] J. Canny, "A Computational Approach to Edge Detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 6, pp. 679–698, 1986.

[46] B. Wu and R. Nevatia, "Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 90–97.

[47] W. Gao, H.-Z. Ai, and S. Lao, "Adaptive Contour Features in Oriented Granular Space for Human Detection and Segmentation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1786–1793.

[48] T. Watanabe, S. Ito, and K. Yokoi, "Co-occurrence Histograms of Oriented Gradients for Pedestrian Detection," in *Advances in Image and Video Technology*. Springer, 2009, pp. 37–47.

[49] B. K. P. Horn and B. G. Schunck, "Determining Optical Flow," *Artificial intelligence*, vol. 17, no. 1, pp. 185–203, 1981.

[50] M.-P. Dubuisson and A. K. Jain, "Contour Extraction of Moving Objects in Complex Outdoor Scenes," *International Journal of Computer Vision*, vol. 14, no. 1, pp. 83–105, 1995.

[51] C. Hilario, J. M. Collado, J. M. Armingol, and A. de la Escalera, "Pedestrian Detection for Intelligent Vehicles based on Active Contour Models and Stereo Vision," in *Computer Aided Systems Theory–EUROCAST 2005*. Springer, 2005, pp. 537–542.

[52] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[53] H. Bay, A. E. ana T. Tuytelaars, and L. van Gool, "Speed-up Robust Features (SURF)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 110, pp. 346–359, 2008.

[54] C. Papageorgiou and T. Poggio, "A Trainable System for Object Detection," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[55] M. Enzweiler and D. M. Gavrila, "Monocular Pedestrian Detection: Survey and Experiments," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 12, pp. 2179–2195, 2009.

[56] C. Wojek, S. Walk, and B. Schiele, "Multi-Cue Onboard Pedestrian Detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, 2009, pp. 794–801.

[57] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide-baseline Stereo from Maximally Stable Extremal Regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.

[58] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition using Shape Contexts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 4, pp. 509–522, 2002.

[59] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation," in *Computer Vision–ECCV 2006.* Springer, 2006, pp. 1–15.

[60] M. Varma and A. Zisserman, "A Statistical Approach to Texture Classification from Single Images," *International Journal of Computer Vision*, vol. 62, no. 1-2, pp. 61–81, 2005.

[61] J. Winn, A. Criminisi, and T. Minka, "Object Categorization by Learned Universal Visual Dictionary," in *Computer Vision, 2005. ICCV 2005. 10th IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 1800–1807.

[62] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person Reidentification using Spatiotemporal Appearance," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1528–1535.

[63] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person Re-identification by Symmetry-Driven Accumulation of Local Features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on.* IEEE, 2010, pp. 2360–2367.

[64] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette Analysis-based Gait Recognition for Human Identification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 12, pp. 1505–1518, 2003.

[65] A. Bissacco and S. Soatto, "Hybrid Dynamical Models of Human Motion for the Recognition of Human Gaits," *International journal of computer vision*, vol. 85, no. 1, pp. 101–114, 2009.

[66] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling Inter-camera Space-time and Appearance Relationships for Tracking across Non-overlapping Views," *Computer Vision and Image Understanding*, vol. 109, no. 2, pp. 146–162, 2008.

[67] D. Gray and H. Tao, “Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features,” in *Computer Vision–ECCV 2008*. Springer, 2008, pp. 262–275.

[68] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, “Person Re-identification by Support Vector Ranking.” in *BMVC*, vol. 2, no. 5, 2010, p. 6.

[69] N. D. Bird, O. Masoud, N. P. Papanikolopoulos, and A. Isaacs, “Detection of Loitering Individuals in Public Transportation Areas,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 6, no. 2, pp. 167–177, 2005.

[70] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, “Shape and Appearance Context Modeling,” in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[71] T. Gandhi and M. M. Trivedi, “Person Tracking and Reidentification: Introducing Panoramic Appearance Map PAM for Feature Representation,” *Machine Vision and Applications*, vol. 18, no. 3-4, pp. 207–220, 2007.

[72] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux, “Person Re-identification in Multi-camera System by Signature based on Interest Point Descriptors Collected on Short Video Sequences,” in *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on*. IEEE, 2008, pp. 1–6.

[73] S. Bak, E. Corvee, F. Brémond, and M. Thonnat, “Person Re-identification using Spatial Covariance Regions of Human Body Parts,” in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*. IEEE, 2010, pp. 435–440.

[74] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, “Custom Pictorial Structures for Re-Identification.” in *BMVC*, vol. 2, no. 5, 2011, p. 6.

[75] R. Satta, G. Fumera, and F. Roli, “Fast Person Re-identification Based on Dissimilarity Representations,” *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1838–1848, 2012.

[76] W.-S. Zheng, S. Gong, and T. Xiang, “Reidentification by Relative Distance Comparison,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 3, pp. 653–668, 2013.

[77] C.-H. Kuo, S. Khamis, and V. Shet, “Person Re-identification using Semantic Color Names and RankBoost,” in *WACV*, 2013, pp. 281–287.

[78] M. Hussein, F. Porikli, and L. Davis, “A Comprehensive Evaluation Framework and a Comparative Study for Human Detectors,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 10, no. 3, pp. 417–427, 2009.

[79] P. Viola and M. Jones, “Rapid Object Detection Using a Boosted Cascade of Simple Features,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[80] C. Papageorgiou, M. Oren, and T. Poggio, "A General Framework for Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1998.

[81] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," in *Machine Learning: Proceedings of the Thirteenth International Conference*, 1996, pp. 148–156.

[82] ——, "A Short Introduction to Boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, 1999, in Japanese, translation by Naoki Abe.

[83] T. Hashiyama, D. Mochizuki, Y. Yano, and S. Okuma, "Active Frame Subtraction for Pedestrian Detection from Images of Moving Camera," in *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, vol. 1, Oct 2003, pp. 480–485.

[84] X. Chen, P. J. Rynn, and K. W. Bowyer, "Fully Automated Facial Symmetry Axis Detection in Frontal Color Images," in *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on.* IEEE, 2005, pp. 106–111.

[85] G. Ma, A. Kummert, S.-B. Park, S. Müller-Schneiders, and A. Loffe, "A Symmetry Search and Filtering Algorithm for Vision Based Pedestrian Detection System," SAE Technical Paper, Tech. Rep., 2008.

[86] D. P. Mukherjee, A. Zisserman, and M. Brady, "Shape from Symmetry: Detecting and Exploiting Symmetry in Affine Images," *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, vol. 351, no. 1695, pp. 77–106, 1995.

[87] T. Tuytelaars, A. Turina, and L. van Gool, "Noncombinatorial Detection of Regular Repetitions under Perspective Skew," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 4, pp. 418–432, 2003.

[88] H. Cornelius, M. Perd'och, J. Matas, and G. Loy, "Efficient Symmetry Detection using Local Affine Frames," in *Image Analysis.* Springer, 2007, pp. 152–161.

[89] docs.opencv.org, *The OpenCV Reference Manual - Release 2.4.3*, Itseez ltd., Nov. 2012.

Appendix A

Viola Jones Detectors Provided by OpenCV

OpenCV is multi-platform library for computer vision algorithms. In this thesis, its C++ API[1] is used in programming. As described, trained pedestrian and body-parts detectors used in this thesis are provided by OpenCV (version 2.4.3). The trained HOG pedestrian detector is a series kernel parameters of SVM classifier. According to the OpenCV reference manual[89], parameters used in HOG training follow the recommendations in [22], though no information was provided about the training datasets, the training result can be easily retrieved using functions in HOG class.

The trained Viola Jones based detector is a series of Haar-like features with its weights. Even following the exact training routines introduced by [30], the selected rectangular features and their weights may vary especially when the training datasets are unknown. As presented in previous context, the size and the position of the chosen Haar-like features in the detector may affect the accuracy and the stability of the performances. To better understand the trained Haar-like feature based detectors. The files of "Frontal Face Detector", "Upper-body Detector" and "Pedestrian Detector" were examined. The summary of these detectors are provided below, more detailed information can be retrieved in OpenCV documentation[2]. As a supplement, *Table A.1* demonstrates the detailed calculations of Haar-like rectangular features on integral images.

[1] API: Application Programming Interface.

[2] URL: `docs.opencv.org`

Haar-like Face Detector file: "haarcascade_frontalface_alt_tree.xml"[3]

- Face detector detect face area from the middle of the forehead to the middle of the chin;
- Colour changing around eye, nose and mouth area play more important role in face recognition.

Haar-like Upper-Body Detector file: "haarcascade_mcs_upperbody.xml"

- This upper-body detector detect upper-body from head top to middle of chest;
- There are overlaps of features used in upper-body detection and frontal-face detection. This may introduce the miss detection of the upper-body in back view;
- A fairly amount (~25%) of small/slim sized Haar-like feature blocks were observed in upper-body detector. Small means the block is sized under 3×3 pixel^2 and slim indicate the width of the block is less than 3 pixel. This would introduce instability to the performance of the detector.

Haar-like Pedestrian Detector file: "haarcascade_fullbody.xml"

- The fullbody detector detect the upright human from head to feet.
- High influential features capture the colour changing around the body, between head and torso and between upper-body and lower-body;
- Haar-like blocks were used to provide more shape information of human.

[3] In the OpenCV package, the trained Viola Jones detectors are stored in ".xml" file, where the trees / stumps of rectangular features are hierarchically coded using following parameters: the Coordinates of Top-Left Corner within the ROI followed by the width and height of the rectangle feature. The weights of features are accompanied.

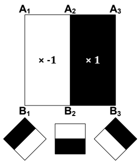
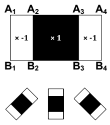
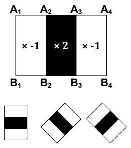
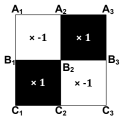
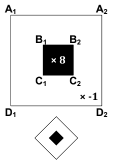
Haar-like Feature	Calculate in Integral Image
(a) 2-Rectangular (Edge) 	$(B_3 - A_3 - B_1 + A_1) - 2(B_2 - A_2 - B_1 + A_1)$ $= B_3 - 2B_2 + B_1 - A_1 + 2A_2 - A_3$
(b) 3-Rectangular (Edge) 	$(B_4 - A_4 - B_1 + A_1) - 2(B_3 - A_3 - B_2 + A_2)$ $= B_4 - 2B_3 + 2B_2 - B_1 - A_4 + 2A_3 - 2A_2 + A_1$
(c) 3-Rectangular (Line) 	$(B_4 - A_4 - B_1 + A_1) - 2(B_3 - A_3 - B_2 + A_2)$ $= B_4 - 2B_3 + 2B_2 - B_1 - A_4 + 2A_3 - 2A_2 + A_1$
(d) 4-Rectangular 	$(C_3 - A_3 - C_1 + A_1) - 2(C_2 - B_2 - C_1 + B_1)$ $-2(B_3 - A_3 - B_2 + A_2)$ $= C_3 - 2C_2 + C_1 - 2B_3 + 4B_2 - 2B_1 + A_3 - 2A_2 + A_1$
(e) Centre 	$(D_2 - D_1 - A_2 + A_1) - 2(C_2 - C_1 - B_2 + B_1)$ $= D_2 - D_1 - 2C_2 + 2C_1 + 2B_2 - 2B_1 - A_2 + A_1$

Table A.1 **Haar-like Feature and Detailed Calculations in Intensity Images**