A Study on Pedestrian Detection and Person Re-identification Based on Different Features

張　国棟

A Thesis submitted to the University of Tokushima in
partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2017

The University of Tokushima
Graduate School of Engineering
Information Science and Systems Engineering

Contents

List of Figures

List of Tables

Acknowledgment

First, I would like to thank my advisor Professor Kenji Kita. With his kind help and support I was able to continue my study on the basic theories of image processing and image analysis. He created pleasant research environment for me on my Ph.D study and research.

I would also thank Professor Fuji Ren for his support and encouragement. He gave me much inspiration and help in my study.

I would thank Mr.kazuyuki Matsumoto, Yoshida Minoru, Prof. Motoyuki Suzuki, Mr. Peilin Jiang and Mr. Xin Kang. They are very kind and helped me a lot when taking my doctoral courses.

Finally, thanks to all the Japanese language teachers in Tokushima university. They helped me work out my Japanese problems during my course of the thesis.

Abstract

Many object detection algorithms assume that the pedestrian scale is fixed during detection, such as the DPM detector. However, detectors often give rise to different detection effects under the circumstance of different scales. If a detector is used to perform pedestrian detection in different scales, the accuracy of pedestrian detection could be improved. A multi-resolution DPM pedestrian detection algorithm is proposed in this paper. During the stage of model training, a resolution factor is added to a set of hidden variables of a latent SVM model. Then, in the stage of detection, a standard DPM model is used for the high resolution objects and a rigid template is adopted in case of the low resolution objects. In our experiments, we find that in case of low resolution objects the detection accuracy of a standard DPM model is lower than that of a rigid template.

Person re-identification, which aims to track people across nonoverlapping cameras, is a fundamental task in automated video processing. Moving people often appear differently when viewed from different nonoverlapping cameras because of differences in illumination, pose and camera properties. The color histogram is a global feature of an object that can be used for identification. This histogram describes the distribution of all colors on the object. However, the use of color histograms has two disadvantages. First, colors change differently under different lighting and at different angles. Second, traditional color histograms lack spatial information. We used a perception-based color space to solve the illumination problem of traditional histograms. We also used the spatial pyramid matching (SPM) model to improve the image spatial information in color histograms. Finally, we used the Gaussian mixture model (GMM) to show features for person re-identification, because the main color feature of GMM are more adaptable for scene changes and improve the stability of the retrieved results for different color spaces in various scenes. Through a series of experiments, we found the relationships of different features that impact person re-identification.

Chapter 1

Introduction

Video monitoring system can be used for different purpose and are currently applied in all sectors[21]. Generally speaking, many monitoring systems only record information. For instance, monitoring system on road only records the license plate number of cars, while indoor monitoring system only records movements of people inside[1]. Such information cannot be processed or analyzed without human, which adds great restriction to the application of video monitoring, since human plays a dominant role in the system and all the decisions need to be made by human. In the future, video monitoring system would be intelligentized and automatic. It can help the user make decision and provide unique stability and precision. In specific situation, it can even make decision on behalf the user. Effective and stable pedestrian detection can greatly improve the quality and effect of monitoring.

1.1 Pedestrian Detection

Pedestrian detection is widely used in multimedia storage and search[60]. If the algorithm efficiency and precision of pedestrian detection and tracking is quite high, we can analyze specific situation from the video recorded in closed circuit television system. If a child is missing, we can find all the possible children from the features of the missing child (e.g. color of clothes and height et. al.). Besides, we can also analyze how many people entered a specific place as well as the movement trend of the crowd.

On a driverless car, the built-in intelligent system should be able to detect how many

people are crossing the road and if there are any people that needs special cares (the elder and the disabled), detect the time it takes for pedestrian to cross the road and analyze the road condition, so as to ensure the safety of everyone when the car is moving. Therefore, pedestrian detection is very necessary. However, if a pedestrian is not crossing the road, but walking toward the edge of the road, analysis on the movement trend of the pedestrian is necessary, which adds difficulty to pedestrian detection. In this case, RFID technology might be an effectively solution.

Pedestrian detection is very useful in the aspect of intelligent robot[10][61]. When providing domestic service, robots not only need to perform several functional movements according to the command of the user, but also need to make response to certain incidents (e.g. a old people in the family fell down). Rescue robots can offer great help when disaster occurs. They can help people survive or help rescue workers find the injured. All these need the help of pedestrian detection.

1.1.1 The Research Problem in Pedestrian Detection

A major challenge in studying pedestrian detection is the different features of the people under detection. Meanwhile, diversified of the detection environment also adds difficulties to the process. The details are as follows:

Illumination Illumination differs at different times. Places with bright light can achieve better detection effect than places with dark light. Imaging quality in places with weak light is always poor, which is shown in Fig 1.1

Figure 1.1: The pedestrians with different illumination condition

occlusion Another difficulty of pedestrian detection lies in the blocking of pedestrian. Generally speaking, under high resolution ratio, detection effect can be better when there is no blocking, but would be bad when there is blocking. Under low resolution ratio, detection effect would be even worse. In the real world, blocking could not be predicted, which makes detection much more difficult, as shown in Fig 1.2. Occlusion occurs in different ways, and is also inevitable.

Figure 1.2: The Pedestrians were obscured in a Sample Picture

Resolution Ratio Image of pedestrian would differ as the distance between pedestrian and camera differs, which is related to the resolution ratio of pedestrian detection. The higher resolution ratio is, the better the pedestrian detection effect would be. Detection of 300 pedestrians with high pixel is quite different from detection of 30 pedestrians with high pixel. The higher the pixel is, the more information is contained and the easier the detection would be. At present, many algorithms are detection under fixed resolution ratio. When pedestrians of different size appear in the image, the effect of detection would be less ideal, as shown in Fig 1.3. The appearance of pedestrians under different resolution ratio adds difficulty to the detection process.

Figure 1.3: The different resolution of Pedestrians in a Sample Picture

Visual Angle Different visual angle makes the detection more difficult. Visual angle in video monitoring is usually a downward vision with certain angle, which differs with the visual angle of automobile detection and video monitoring. Difference visual angle will also bring certain difference in features of pedestrian, as shown in Fig1.4.

Figure 1.4: The Pedestrians with different visual angle in a Sample Picture

Pedestrian Direction Position and orientation of pedestrians in the image are not fixed. The same person would present different features in different orientation, as shown in Fig 1.5.

Figure 1.5: The different pedestrians direction in a Sample Picture

Pedestrian Posture People would present different posture in daily life. Pedestrian in the image would be standing, leaning or sitting. Different posture would present different image features, which would make detection more difficult, as shown in fig 1.6.

Figure 1.6: The Pedestrians with different postures in a Sample Picture

1.1.2 Pedestrian Detection method

Pedestrian detection has been a hotspot in computer vision research[19]. The corresponding detection algorithm has been developed towards high precision and instantaneity[55]. For a driverless automobile, the usage of which has become popular nowadays, its intelligent system should be able to detect the locations and quantities of pedestrians ahead, to analyze the road conditions, and to guarantee the safety of these pedestrians. For such cases, the pedestrian detection is an inevitable procedure. The pedestrian detection problem is difficult because that the target people often have various characteristics and the surrounding environments also change frequently[30]. The pedestrian sizes in real world

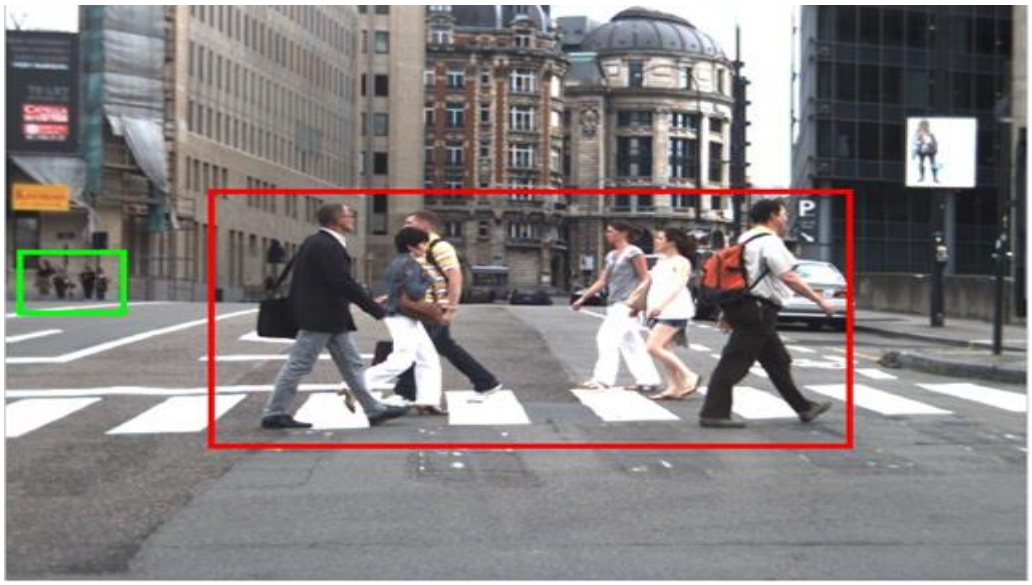

Figure 1.7: The Pedestrians with Multiple Resolution in a Sample Picture

are different from each other. Besides the height diversity of different people, many imaging differences are incurred by the different distances between people and the camera. As shown in Fig 1.7, a high resolution corresponds to the large pedestrian scale and a low resolution corresponds to the small pedestrian scale, in the process of pedestrian detection. Pedestrians contain rich information in the case of high resolution, and it is more likely for them to be detected. Even if they are locally overlapped, many algorithms have the capability to detect these targets[54]. However, in the case low resolution, the pedestrians which contain a small amount of information cannot be detected easily. Meanwhile, low resolution pedestrians are very vulnerable to the interferences of the surrounding environments. In most cases, a detection algorithm has a much better detection result for the high resolution pedestrians than that for the low resolution pedestrians. Dalal and Triggs propose a HOG detector[13]. If the detection window is fixed to pixels during training and detection, this detector can generate good effects at the time of detecting pedestrians with pixels greater than . However, when the target pedestrians are smaller than , the detector almost fails to detect any pedestrian. Although the target can be increased to larger than pixels by means of interpolation, the detection accuracy is still brought down. The DPM pedestrian detector makes use of a root filter and several part filters to describe the pedestrians. Information in the pedestrians of high resolution is sufficient.

Fig 1.8 and 1.9 are results obtained by utilizing a standard DPM detector to detect

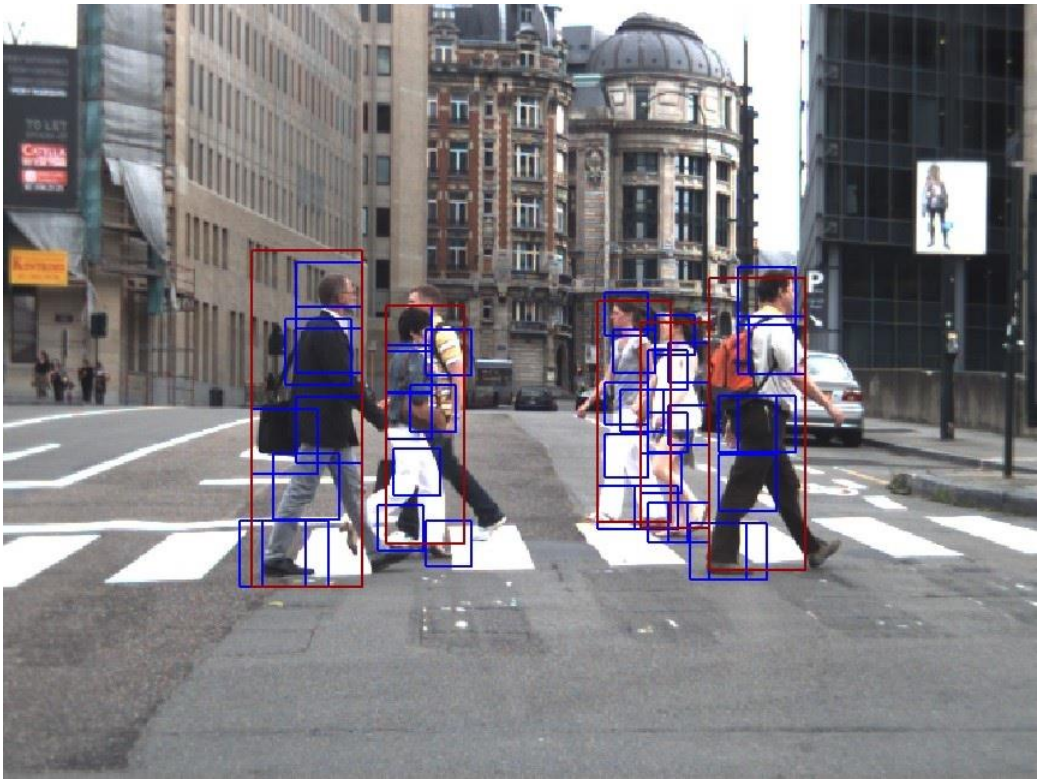

Figure 1.8: The part filter and root filter in DPM.

pedestrians in Fig 1.7. It is obvious that the small-scale pedestrians cannot be detected successfully.

Therefore, the overall detection effect can be improved if we can improve the detection effect for low resolution pedestrians and prevent affecting the detection effect for high resolution pedestrians. In this paper, we propose a multi-resolution DPM pedestrian detection algorithm, which takes advantage of the standard DPM framework in training the pedestrian with the resolution factor as a hidden variable. For the high resolution pedestrians, the response can be figured out in the first place. And its location can be estimated with the combination of this high resolution response and the response under a corresponding low resolution. However, for the low resolution pedestrians, the judgment over possible locations of these targets is carried out by only calculating the responses under the low resolution. High resolution and low resolution are only intuitive concepts in the common sense. In addition, resolution is closely associated with the heights of pedestrian samples. Pillor divide samples in the Caltech Pedetrian Library into distant ($h < 30$), medium ($30 < h < 80$) and near ($h > 80$) types, according to the heights (h)

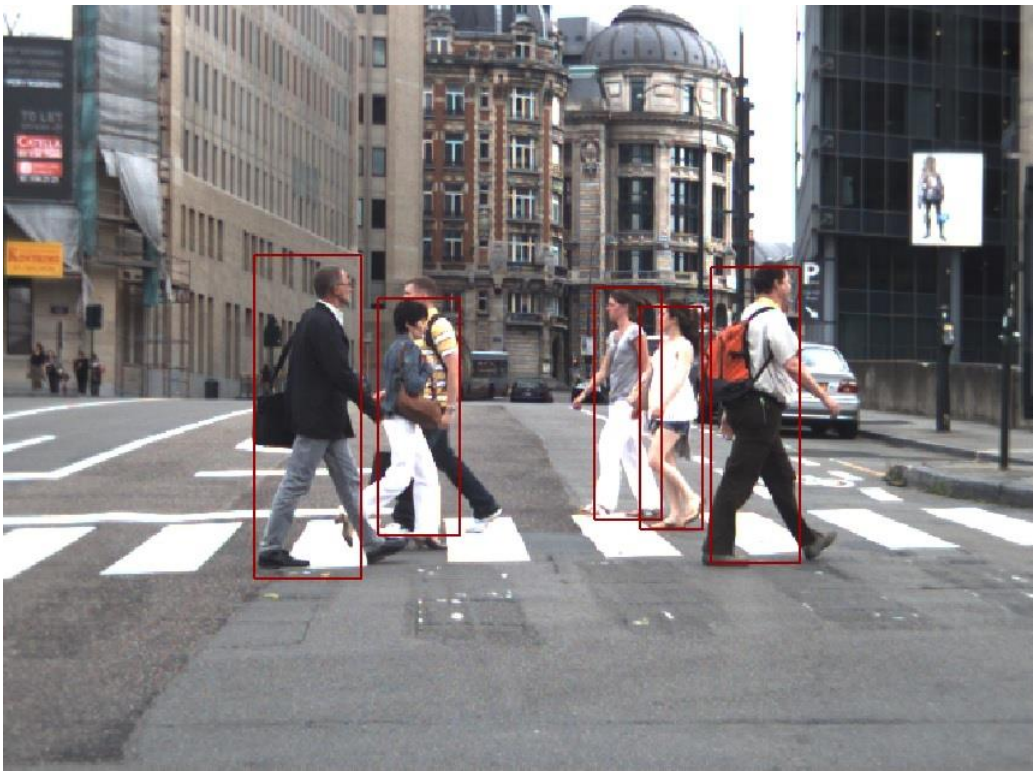

Figure 1.9: The detection result of standard DPM.

of pedestrians. In this work we define targets with height $h > 100$ as the high resolution target, targets with height $h < 50$ as the low resolution targets, and targets with height $50 < h < 100$ as the unknown resolution targets in which case the resolution factor is treated as a hidden variable.

1.2 Person Re-identification

At present, the bottleneck in the field of video application is how to highly effectively extract video information, how to perform standard data exchange with other informational systems, interconnection and semantic inter operation[2][53]. The key technique to solve this problem is video structural description technique[41]. Video structural description is a technique for extracting video content information. According to semantic relationship, it employs processing methods such as spatiotemporal segmentation, feature extraction and object recognition etc. to organize video content into information understandable to human and machines[23][4], and further to implement the transformation from video data to information. Traditional video systems need to be transformed by video structural

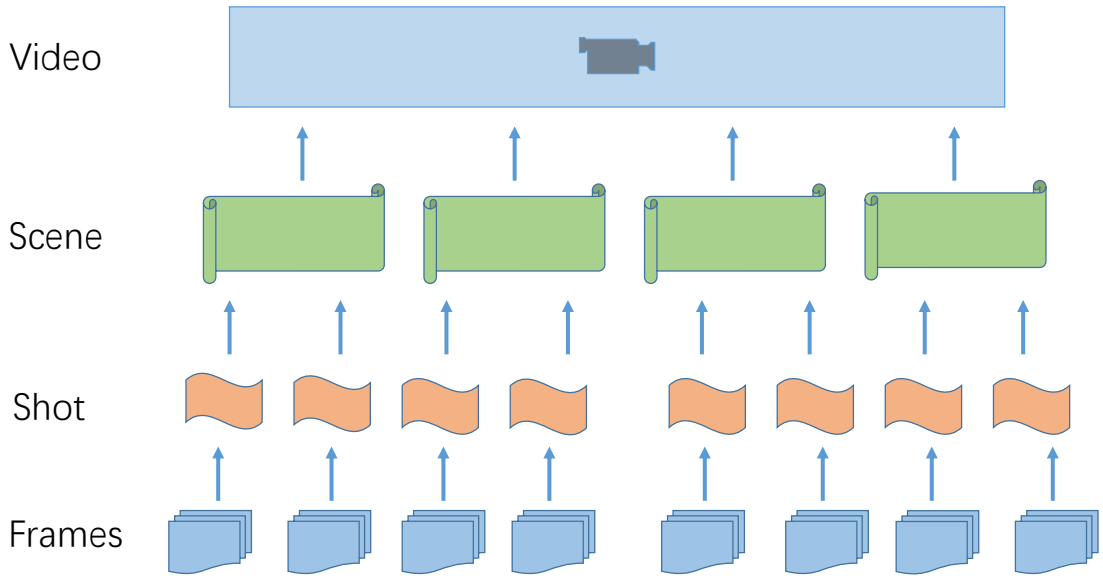

Figure 1.10: The video of structure and hierarchy diagram

description technique into new generation semantic video systems which are smart and semantic.

1.2.1 Video Structure Analyze

Video is non-structural stream data. The basic unit is frame. Video stream is a set of frames which have temporal and spatial relationships. Temporally, features in continuous video frames change gradually. Spatially, some sub-blocks remain the same in continuous video frames[18]. Video frames are not mutually independent.

From top to bottom, the video hierarchical structure could be divided into video, scene, shot and key frame, as shown in fig 1.10, with increasingly fine expression granularity.

Frame

Basic unit in video stream. Video data stream is comprised of image frame. Each frame could be regarded as an independent image.

Shot

An uninterrupted frame sequence captured by the camera. Its the foundation of further restructuration of video data stream[47]. Shot usually is regarded as the smallest structural

unit in a video. In the same group of shots, the image features of video frames basically remain stable. If drastic feature changes are found between adjacent image frames, a shot transition is believed to occur.

Scene

A number of shot groups which are semantically related and temporally adjacent constitute a scene. A scene is the expression of higher abstract concepts and semantic contained in a video.

Key frame

In a shot, the frame that could represent the main content in the shot is the key frame. In a video shot, the number of key frames is way less than the number of image frames in the shot. Use key frame to represent a shot could considerably reduce the computational complexity.

Region

An image could be decomposed into multiple regions. Target in an image normally is not the entire image but only a region within. Thus, this region level could represent some specific concepts more precisely.

Video

Video sequence is made of orderly arranged frame images and the most common data structure used to represent a frame image in computer vision is a matrix. Elements in a matrix are all integer values corresponding luminance or other properties of pixels in the sampling grids. Image information in the matrix could be retrieved by coordinates of pixels[26]. Coordinates correspond to row numbers and column numbers. Matrix is a complete representation of image. It not only contains the data information of the image but also implies the spatial relationship between image components. Normally how accurate the representation of image by matrix is depends on the value range of elements in the matrix. Binary image, i.e. black and white image, can be just represented by the matrix containing only 1 and 0.

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

Figure 1.11: Images showing the same person in different camera views (a) Pose change (b) Illumination change (c) Occlusion (d) Low resolution

1.2.2 The method of Person Re-identificationn

The public security technology has become increasingly intelligent, surveillance cameras have been set up in public places such as airports and supermarkets[59]. These cameras provide huge amounts of nonoverlapping video data. It is often necessary to track an object or person of interest that appears on video from multiple cameras under different illumination conditions. When searching for moving people in surveillance video data, object retrieval systems for intelligent video surveillance experience the following problems.

1. Object retrieval results in video surveillance depend on motion segmentation and video analysis. Digital video is a series of images, constituted by frames that contain rich information. If an image frame contains moving objects, then object retrieval detection can be used to segment a moving target. Object retrieval results depend on the object segmentation[28]. If video analysis cannot separate the foreground and moving objects, the target object cannot be retrieved from the many irrelevant foreground objects. A good object retrieval system should adapt to various levels of video quality for foreground detection, which could eliminate unrelated objects and retrieve the target[17].

2. Specific object retrieval in video surveillance faces technical limitations. The moving objects of interest in surveillance video are often persons and cars. Facial features are the most distinctive elements for person recognition, and relatively mature methods are available for this process. However, low camera resolution often makes it difficult to extract perceivable information about facial expression. The mature tech-

nology of video object retrieval based on facial features should receive more technical exploration.

3. External factors greatly influence an object's appearance under video surveillance. A robust object retrieval system should be able to compensate for the following factors.

    - Person pose variation. A moving person may have arbitrary poses. (Fig 1.11 (a))
    - Varying illumination conditions. Illumination conditions usually differ between camera views. (Fig 1.11 (b))
    - Occlusion. A person's body parts may be occluded by other subjects, such as a carried bag, in one camera view. (Fig 1.11 (c))
    - Low image resolution. Due to surveillance camera performance, images of a moving person often have low resolution. (Fig 1.11 (d))

1.3 Thesis Organization

This thesis contains the theories, methods, results, and discussions about the person detection and person re-identification, which is organized in the rest chapters as follows.

Chapter 2: Background

In this chapter, we begin by reviewing the person detection and person re-identification. We review the person detection theory and overview of person re-identification theory, discuss the technology in the person detection and re-identification. We illustrate the methods and models that we used in the detection and re-identification experiment.

Chapter 3: Related Work

In this chapter, we review the study of person detection and person re-identification.

Chapter 4: Overview of Proposal Methods

In this chapter, first we discuss method that we have developed for person detection. We first introduce the person detection algorithm. Then we introduce the Multi-resolution DPM algorithm for pedestrian detection. Then we provide the methods for re-identification of persons in real-life video. At the end, we discuss the Color histogram

feature extraction from video object, and introduce the histogram of color feature similarity measurement

Chapter 5: Evaluating Pedestrian Detection and Person Re-identification

we introduce the data set and the evaluation method in the pedestrian detection and person re-identification. The data source named INRIA and Caltech which have been widely employed in recent researched. The INRIA consist of 288 positive samples (containing 1,126 pedestrians) and 453 negative samples. Images in the INRIA data set are mainly collected from google, GRAZ-01 and personal photographs The Caltech Pedestrian Database is a large scale database. It consists of videos of 640x480 pixels with 30 frames per second, captured by in-vehicle cameras for about 10 hours. Within these videos, 250,000 frames (around 137 minutes), 350,000 bounding boxes, and 2,300 pedestrians are manually annotated by human experts.

Chapter 6: Experiments of Pedestrian Detection and person Re-identification

In this chapter experiment results will be evaluated by a series of evaluation criteria including Precision. We analyse the result of pedestrian detection and person Re-identification in detail.

Chapter 7: Conclusion

we conlude the merits and demerits of our proposed methods in pedestrian detection and person re-identification.

Chapter 2

Background

2.1 Overview of Person Detection Theory

2.1.1 Current Research of Pedestrian Detection

Pedestrian detection has always been a hot topic in study on computer vision. Its detection algorithm is being developed to provide high-precision and real-time service. At present, the speed of detection algorithm can reach 2.67 frame/second[62]. On large-scale dataset INRIA, detect precision can reach over 89%. According to the number of cameras, pedestrian detection can be divided into single-camera detection and multi-camera detection[52]. Most pedestrian detection is based on monocular camera. Multi-camera detection can make use of the movement information of pedestrian. We usually conduct pedestrian detection according to static image or video. Under video detection, we can use inter-frame information. Detection based on sliding window has become the mainstream. The most critical issue in pedestrian technology is feature extraction and design of classifier. Currently, the HOG (Histogram of Oriented Gradient, HOG) proposed by Dalal is the low-level feature with the best detection effect[13]. Other low-level features include Edgelet, shaplet, LBP, and CSS. SVM and BOOST classifier is the most popular classifier in pedestrian detection.

2.1.2 Feature Extraction in Pedestrian Detection

A crucial technology in pedestrian detection is feature abstraction and design of classifier. Pedestrian feature descriptor can be divided into two categories: low-level feature and integration of multi-features[8]. Low-level features include gradient direction graph (HOG), textural features, CSS and color information. Multi-feature integration conducts detection by making use of combination of various low-level features.

Histogram of Oriented Gradient (HOG)

HOG uses first-order derivative to reflect the changing relations among different pixels[48]. It defines cells to describe local information of images and combines small cell features into the feature of a large image block, and finally obtains the HOG feature of the whole image. HOG feature is insensitive to local deformation and it made insensitive to light through normalization.

Algorithm process description :

(1) Standardized Gamma space or Color space.

Firstly, the input image is normalized, which could reduce the influence of light. In reality, the image is obtained by transforming the gray-scale map.

(2) Calculating Pixel Gradient

Computing every pixel gradient in an image. The pixel gradient (x, y) could be calculated as follow:

$$G_x(x, y) = H(x + 1, y) - H(x - 1, y) \tag{2.1}$$

$$G_y(x, y) = H(x, y + 1) - H(x, y - 1) \tag{2.2}$$

In the formula, $G_x(x, y)$ is the horizontal direction gradient, $G_y(x, y)$ is the vertical direction gradient $H(x, y)$ is the value of pixel in the coordinate point (x, y)

The gradient magnitude and gradient direction in pixel (x, y) are as follow:

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \tag{2.3}$$

$$\alpha(x, y) = \tan^{-1}(\frac{G_y(x, y)}{G_x(x, y)}) \tag{2.4}$$

During actual computation, we use $[-1, 0, 1]$ operator to get the gradient of x direction against image convolution and use $[1, 0 - 1]^T$ operator to get the gradient of y direction against.

(3) Construction of HOG

Add the gradient direction of local cell into the graph and get the HOG vector of one cell.

(4) Combine cell unit into a large block and uniformize HOG

Since the size of a cell is 4×4 or 8×8, combine several cells into one block (4 cells constitutes one block) and cascade the HOG feature of cells in block to obtain the HOG feature of the block and uniformize HOG feature.

(5) Collecting the HOG feature in image

Generally speaking, the size of a block is 16×16, (i.e. the size of cell is8×8; one block contains 4 cells); define a sliding window at the size of one block and can on the image; combine the HOG feature obtained from each scanning operation and get the HOG feature of an overall image.

2.1.3 Textural features

The so-called texture of a certain point usually refers to the relationship between the point and its peripheral points, that is, the relationship between the point and other points around it[37]. In pedestrian detection, Local Binary Pattern (LBP) is usually used to describe the local textural feature of image. The original LBP operator is defined within the window of, with the center pixel of the center being the threshold value. Comparing the grey value of 8 adjacent pixels with the threshold value, if peripheral pixel value is larger than center pixel value, the position of the pixel point should be marked as 1, otherwise it should be marked as 0. In this way, through comparison, the 8 points can generate 8 bits of binary number (usually transformed into decimal number, that is, LBP code, 256 kinds in total), that is, obtain the LBP value of the center pixel of the center and use the value to reflect the textural information of the area[15]. Many improved LBP operators are as follows:

(1) Circular LBP Operator

One problem of the original LBP is that its neighborhood is fixed, which cannot meet the demand of changing size. Therefore, Ojala modifies the original LBP algorithm and ex-

pand the neighborhood scope to random neighborhood and replace square neighborhood with circular neighborhood. The improved LBP operator is allowed to have various pixel points in neighborhood randomly.

(2) LBP rotation invariant model

LBP cannot meet the requirements of rotation invariance. Rotation of image would generate difference LBP value. Maenpaa et. al. has proposed LBP operator with rotation invariance. By rotating circular neighborhood continuously, we can obtain a series of LBP value, and the minimum value should be used as the LBP value of the neighborhood.

(3) LBP Uniform Pattern

A LBP operator has two modes: 0 and 1. As for the circular area with a radius of R, the LBP operators of P sampling points would generate P^2 kinds of modes. As the number of sampling points increases, types of binary modes soar. Too much binary modes would impose negative influence on the abstraction, recognition, expression and classification of texture as well as storage and withdraw of information. Therefore, it is necessary to reduce dimensions of the original LBP, so as to best represent information of image when data quantity is decreasing. Ojala [40]proposed a Uniform Pattern to reduce the dimensions of mode categories of LBP operators. Ojala et. al. think that most LBP modes contain at most two times saltus step from 1 to 0 and from 0 to 1. Therefore, they defined the Uniform Pattern as: when the cyclic binary number corresponding to a LBP has at most two times saltus step from 0 to 1 or from 1 to 0, the binary corresponding to the LBP should be called as a Uniform Pattern type, including 00000000 (0 saltus step), 00000111 (only contain one-time saltus step from 0 to 1), 10001111 (two times saltus step from 1 to 0 and then from 0 to 1). Such improvement greatly reduces the types of binary mode and would not lose any information. Number of modes reduce to $P(P-1)+2$ from $2P$. As for the 8 sampling points in 3×3 neighborhood, types of binary mode reduces to 58 from the original 256, which reduces the dimensions of feature vector and can reduce the influence of high-frequency noise.

Self-similarity of color(CSS)

Self-similarity of color (CSS) is widely used in image classification[38]. Detection effect can be greatly improved by combining it with other low-level feature. In pedestrian detection,

we usually dont use color feature of image directly, because color is highly restricted in usage and greatly influenced by light and background. CSS mainly uses certain features of pedestrian color. For example, the skin colors of people are always similar, so are the colors of the clothes of most people. CSS collects color histogram information on image block at the size of 8×8 and uses trilinear interpolation to minimize aliasing phenomenon. Experiment has shown that it is the best choice to calculate CSS feature effect in HSV space. Histogram intersection is used to calculate the color similarity between two difference blocks. At last, CSS feature vector is uniformized.

Shape feature

Shape feature is also used for pedestrian detection. Gavrila and Philomin[22] use $Hausdorff$ distance and hierarchy template to match image edge rapidly, so as to conduct pedestrian detection. Wu and Nevatia[58] use *edgelet* feature to described local shape feature. Similarly, *shapelets* is shape describer, usually obtained by learning the gradient of local sections. Then, they use Boosting classifier to integrate various *shapelets* into a panorama detector.

Motion feature

Motion feature is also quite useful in pedestrian detection. When the camera is in static status, Viola et. al. calculated Haar-like feature on different images and obtained sound detection effect. However, then the camera is moving, motion feature would impose negative influence on the detection result. Dalal et[13]. al., based on the difference of internal optical flow field, found that we can compensate the motion feature of local image by modeling motion feature.

Multi-feature integration

Up to now, HOG is a single feature that can achieve the best detecting effect. Therefore, it is an effectively solution to make up for the deficiencies of HOG feature with other features. Wojek and Schiele[16] achieved better detection effect by combining Haar-like, shapelets, shape context than they did with any single feature. Walk et. al., based on the study of Wojek, added local color self-similarity and MULTIFTR+MOTION into the

combination process. Wu and Nevatia combined HOG, edgelet and covariance feature to conduct pedestrian detection. Wang et. al[57]. used HOG and LBP feature to conduct pedestrian detection.

2.1.4 Hard Example Mining (SVM)

In the training procedure, there are usually more negative samples than the available positive samples. Taking pedestrian detection for example, the images of pedestrians are positive samples, and the images without pedestrian are negative samples. In this case, 10^5 samples can be generated from an image, most of which are negative samples. It is almost impossible to take all negative samples into consideration. Therefore, we select the positive samples and the hard examples for constructing a training set. The hard examples are referred to those which are incorrectly classified at the first time. The Bootstrapping classification algorithm is employed for training an initial negative sample set[54]. The algorithm collects the incorrectly classified samples at the first time, add these samples to the negative sample set to form the hard samples. The process is repeated for several times until until a good classification result is achieved. We define the hard example and easy sample as follows:

$$H(\beta, D) = \{(x, y) \in D \mid yf_\beta(x) < 1\}, \tag{2.5}$$

$$E(\beta, D) = \{(x, y) \in D \mid yf_\beta(x) > 1\}, \tag{2.6}$$

where, $H(\beta, D)$ denotes the incorrectly classified samples at the first time or the samples located within the classification boundary. $E(\beta, D)$ denote the correctly classified samples. The samples on the classification boundary do not belong to $H(\beta, D)$ or $E(\beta, D)$. $\beta^*(D) = \arg\min_\beta L_D(\beta)$.

Because L_D is strictly convex, $\beta^*(D)$ is the single result in the optimization problem. Given a sample library D, we want to find a small sample set C with $C \subseteq D$ and $\beta^*(C) = \beta^*(D)$. To solve this problem, we firstly define an initial set which contains all the training samples. We train an LSVM model, and renew the previous set by removing the simple samples and by adding the new hard examples.

2.1.5 Hard Example Mining (LSVM)

For the LSVM, mining hard examples is equivalent to optimize $L_{D(Z_p)(\beta)}$ rather than $L_D(\beta)$. This constraint turns the whole optimization problem to a convex optimization problem.

As for the hard example mining with SVM, we define an set with samples in the form of (x, z), in which $z \in Z(x)$. In the real application, the set consists of $\Phi(x, z)$ rather than (x, z). We define a vector set $F = (i.v)$, in which i denotes sample index, and $v = \Phi(x, z)$ with $z \in Z(x_i)$. Because the hidden variable z is not fixed, for each sample x_i, there may be multiple corresponding $(i, v) \in F$. We then define $I(F)$ as the index of vectors in the vector set F, and define the target function for β with the feature vectors in F:

$$L_F(\beta) = \frac{1}{2}\|\beta\|^2 + C \sum_{i \in I(F)} \max(0, 1 - y_i(\max_{(i,v)\in F} \beta \cdot v)). \tag{2.7}$$

L_F can be optimized with the gradient-descent algorithm. We use $V(i)$ as the set of feature factors v. The gradient-descent algorithm is described as follows.

(1) a_t is the learning rate for iteration t.

(2) $i \in I(F)$ is the index for samples in F.

(3) $v_i = \arg\max_{v \in V(i)} \beta \cdot v$.

(4) If $y_i(\beta \cdot) \geq 1$, then $\beta = \beta - a_t\beta$.

(5) Otherwise, $\beta = \beta - a_t(\beta - Cny_iv_i)$.

We set $\beta^*(F) = \arg\min_\beta L_F(\beta)$ and try to find $\beta^*(F) = \beta^*(D(Z_p))$ in a sample set $D(Z_p)$ of small size. As for the hard example mining with standard SVM, we define the feature vectors for hard and simple example in training set D as follows.

We find the hard examples by calculating $\beta^*(D(Z_p))$. F_1 is defined as the initial feature vector set. The LSVM hard example mining algorithm is given below.

1. Train model with $\beta_t := \beta^*(F_t)$.

2. If $H(\beta, D(Z_p)) \subseteq F_t$, stop the iteration and return β_t.

3. Remove the simple samples by $F_t^{'} := F_t \backslash X$, where $X \subseteq E(\beta_t, F_t)$.

4. Add new hard examples by $F_{t+1} := F_t^{'} \cup X$, where $X \cap H(\beta_t, D(Z_p)) \setminus F_t \neq \phi$.

In the 3rd step simple samples are removed from the training set, while in the 4th step new hard examples are added to the training set. The entire iteration procedure terminates when there is no hard examples to add.

2.1.6 Fixed Resolution Model

Let x represent an image window, and $\phi(x)$ represent the image feature. As many slide window detection algorithms, we have

$$f(x) > 0 \; where \; f(x) = w \cdot \phi(x), \tag{2.8}$$

in which x is marked as pedestrian. We train the above model with the positive and negative samples of the training set (x_i, y_j), in which $y_i \in \{-1, 1\}$. The commonly available training algorithms include SVM and boosting, and we employ the linear SVM for training the parameter w:

$$w^* = \arg_w \min \frac{1}{2} w \cdot w + C \sum_i \max(0, 1 - y_i w \cdot \phi(x_i)) \tag{2.9}$$

in which x_i is assumed to be of a fixed size during training and testing. we define a feature vector $\phi(x)$ to deal with the windows of different sizes.

2.1.7 Models with Fixed Resolutions

If an image contain objects of different resolutions at the same time, the detector of a fixed resolution usually cannot detect all different objects simultaneously. Because we can describe the different distances of pedestrians in an image, for each window x we can define a binary variable s to represent the distance of a pedestrian. We use $s = 0$ to represent the distant the target pedestrians, and use $s = 1$ to denote the close target pedestrian. Our classifier is the same as the previous one. $f(x, s) = w \cdot \phi(x, s)$,

$$\phi(x,s) = \begin{bmatrix} \phi_0(x) \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad if \quad s = 0 \qquad and \qquad \phi(x,s) = \begin{bmatrix} 0 \\ 0 \\ \phi_0(x) \\ 1 \end{bmatrix} \quad if \quad s = 1 \tag{2.10}$$

where $\phi_0(x)$ and $\phi_1(x)$ denote the features at different scales, such as a pedestrian of 50 pixels and a pedestrian of 100 pixels.

2.2 Overview of Person re-identification Theory

2.2.1 Segmentation of Video Objects

A single video file is usually a video sequence composed of several scenes and each scene contains one or more shots. These shots could be continuous or with intervals. Each shot contains several continuous image frames. To structurally process video, the video in hand needs to be reduced to many single shots. This process is called segmentation of shots or segmentation of video.

Objects within a video show spatiotemporal continuity of movement[9]. The inter frame differences between shots are larger than inter frame differences within one shot. In segmentation of shots, shot boundary is recognized using differences between continuous frames. Shot boundary could be the final frame of the last shot or the starting frame of the next shot. Video is ready to be segmented once the position or time stamp of boundary frame is located. Therefore, generally segmentation of shots could also be called detection of shot boundary.

2.2.2 Key frame extraction

In a content-based video retrieval system, usually the key frame of each shot needs to be extracted[56]. The usage of key frames reduces the amount of information considerably and provides an organizational frame for retrieving videos. Key frame is the image frame that describes the key content in a shot. It usually reflects the main content in a shot. The extraction of key frame maps a shot onto several image frames of different contents. While extracting key frame, information of bottom level and other available information is to be used and in many cases the conservative principle of better wrong than less. In the situation that representative features are not specific, repetitive and redundant video frames are generally to be removed and several independent frames are selected to include more information.

The extraction of key frame based on inter frame differences

The main idea of this method is to use the first frame in the video as key frame. Then features of video frames such as color features, texture features, shape features, edge features and motion features are extracted and distances between these features are compared in order to measure inter frame differences. Image frames after the current key frame are compared with the key frame in the order of timeline. If the difference between the two frames is smaller than a certain threshold, then the two frames are regarded similar and next frame is to be compared with. Otherwise the frame becomes the new key frame. And the comparison continues until the shot ends.

2.2.3 Features Extraction from video objects

The foreground object isolated from video image frames is essentially still a target image. Multiple target images from these image frames are the video objects we are retrieving. Content-based retrieval of video objects mainly focuses on the contents of video objects and retrieval results are given according to similarity of contents of video objects. How to rapidly and effectively obtain the content features of image, as well as compare similarity between images is the focus of researches on content-based retrieval of video objects[7]. Contents of images are explained and described through image features. The raw pixels in an image have not undergone any processing and explanation and therefore contain the most complete and richest image information. We need a low-level visual feature which could sufficiently reflect the properties of image contents while exhibits relatively small amount of information to explain the image. These visual features may label differences between images through color, texture, contour, spatial relationship and some other features based on local invariance of the image.

2.2.4 Color Feature

The perception and classification of various objects in the world by human all could base on the inherent color features. Color feature is one of the fundamental features in images as well as a most widely used low-level visual feature in content-based image retrieval. Compared with other visual features, color feature is very stable and has the

lowest dependency on rotation, translation, scaling and even all sorts of deformation of images. In other words, color feature has low dependency on the dimension, direction and perspective of images and shows rather strong robustness. And the extraction of color feature is relatively easy. Therefore, while processing and analyzing images we often use color feature as descriptor to simply target extraction and classification[12]. The first step of expressing color feature is to select the color space corresponding to the specific application to describe the color feature of image[49]. Secondly, vectorize the color feature and convert the retrieval of video object into similarity comparison of eigenvectors. Finally, measure the similarity of color features of different images through the pre-defined standard of similarity distance.

2.2.5 Color Histogram

In 1991, Swain and Ballard[51] proposed that color histogram could be used as color feature. Color histogram is most widely used in the research and application of CBIR. It describes the proportions of different colors in the entire image and does not care the spatial positions of each color. Color histogram is a sort of global color statistical feature and the probability of occurrence of each color in the color space is calculated based on color histogram. Image similarity is measured by distance between histogram vectors. Color histogram is the most common representation of the color feature. Color histogram is simple to calculate and effective, as well as insensitive to geometrical transformations such as local minute changes, rotation, scaling, translation and etc. or transformation of image quality. But as the calculation increases with histogram order this method is inefficient regarding large images. However, retrieval based on color histogram does not retain the spatial information in the original image. Color histogram is also sensitive to color changes. Color histograms of some images are basically the same but the images represented are completely different[33].

2.2.6 Color Moments

The color moment is a color feature which is simple and effective. It was proposed by Stricker and Orengo[50]. The mathematical basis of this method is that the color distribution in a image could be represented by its moments. In addition, St and Or believed

that image information was concentrated in the low-order moments. Therefore, using only the first-order moment (mean), second-order moment (variance) and third-order moment (skewness) is sufficient to represent the color information of an image. Compared with color histogram, another advantage of color moments is that vectorization of features is not necessary.

There are a total of nine components in the color moments of a colorful image. Each image has three color components and each color component has three low-order moments. Only a few moments are used. This method is simple and compact compared with other methods. However, the weak recognition capability of this simplicity would lead to too many false alarms. Thus color moments are often used together with other features.

2.2.7 Color Coherence Vector

The color coherence vector (CCV) is another color-based retrieval and matching method proposed by Pass and Zabih et al[43]. to supply the spatial information of image color unavailable using color histogram or color moment. It is a spatial evolution of image color histogram, with each pixel in a given color bucket as either coherent or incoherent. A pixel is coherent if the size of its connected component exceeds a fixed threshold value; otherwise pixel is incoherent. A color coherence vector stores the number of coherent versus incoherent pixels with each color. For a color bin i, let the number of coherent pixels be i, the number of incoherent pixels be i, and the color coherence vector be x. Then the color histogram of the image would be. This color feature not only reflects the ratio of the number of pixels of a color to the total color space of the image, but also augments with the color spatial information which cant be summarized by the color histogram and color moment. Hence the color coherence vector can provide superior results to color histograms for image retrieval. However, as a shortcoming, CCV does not capture the relationship between the shape of the largest-size color component to its background. Then Zachary et al[32]. proposed the CCV-TEV (Threshold Edge Vector) algorithm as an extension.

2.2.8 Color Auto-Correlogram

Huang J[65] defined a new image feature called color correlogram. The traditional color histogram only describes the proportion of the pixels of a certain color to the total pixels,

depicting the color with global statistical relations. In addition to the information provided by color histogram, the color correlogram is able to express how the distribution of pixels changes with distance, reflecting the spatial correlation between color pairs. The color correlogram can be viewed as a table indexed with color pairs $< i, j >$, where the k-th entry for $< i, j >$ specifies the probability of finding a pixel of color $c(i)$ at a distance k from a pixel of color $c(j)$ in the image. If the correlation between any of the colors is taken into account, the color correlogram would be extremely complex and large ($O(N2d)$). Then its simplified version, the color auto-correlogram (CAC), was developed. It only examines the spatial correlation between pixels of the same color, reducing the dimension to $O(Nd)$. Color, as the most widely used feature that is easy to be extracted, is very important in image retrieval. It should be noted, however, that its inherent sensitivity to lighting conditions can lead to a change in its value and thereby affect the matching result. There is a limitation on its application.

2.2.9 Shape feature

Shape is one of the key visual features of image retrieval. Compared with color and texture, it contains more stable information that reflects the inherent attributes and does not change along with the surrounding environment such as illumination[12]. Humans identify objects based more on shape information than color or texture information. Shape provides a higher level of visual features (targets, objects) than simple visual features such as color or texture, which play an important role in our semantic recognition by images semantic information[29]. It should be noted that the extraction of shape parameters relies on image preprocessing and image segmentation. Image segmentation has an unavoidable effect on the accuracy of the parameters, and even can make it impossible to extract shape parameters for too poor processing. Furthermore, the shape of the object in the image would be slightly deformed with the angle of view. So shapes similarity matching must be based on its translation, rotation and scale invariance. Currently, shape is not used as a retrieval feature in low resolution video retrieval.

2.2.10 Spatial feature

The spatial positions of each object in an image and the spatial feature among objects are also very important features in image retrieval[39]. As to the images containing multiple independent objects, understanding the spatial features of objects in the image could facilitate the classification of image contents. The representation methods of spatial features mainly fall into three categories: metric, directional and topological. Cartesian coordinate system is the most direct representation. Spatial features could supplement other object features which could not determine the spatial relationship of objects.

Spatial Relationship

The university of Pittsburgh first proposed to use generalized $2D$ string hierarchical representation to retrieve spatial relationships of images. The $2D$ strings record the spatial characteristics of each object in the image. The position relationship of the whole image is projected along the X and Y axes. The two strings u and v represent the spatial relationships of the X and Y projections, respectively, of the objects in the image. A $2D$ string over V and A is defined by (u, v). It is a simple method and in some cases, some symbol images can be reconstructed with their $2D$ strings. However, the position of the object in the image is only expressed by the centroid and the description of the spatial relationship is too simple to handle the complex situation.

Structural Relation

The utilization of spatial relation and direction relation has the same problem of automatic image segmentation. Structure is another simple method to provide the relationship between basic spatial characteristics of the image. This method pre-divides the image into sub-blocks (which can be overlapped with each other) equally or in a specific ratio, and then extracts the other features of each sub-block. Similarity comparison can start from similarity comparison of sub-blocks of corresponding positions, and then calculate the similarity of the whole image by weighting. The spatial relations are expressed in a simple way, but have high costs of computation and storage. And being sensitive to the rotation, flipping or scale changes of the target object, it is unable to effectively represent the image

information. These shortcomings limit the practical application of these methods.

2.2.11 Traditional Color Space

Color space is defined as a model that represents color intensity values. Typically, an image is mapped to a color space of 1 to 4 dimensions, with each dimension corresponding to a color channel and representing the corresponding color channel gray information. Color spaces such as RGB and HSV are somehow correlated and can be converted into each other through mathematical formulas.

RGB Color Space

RGB color model is composed of three primary colors, red (R), green (G) and blue (B). It is mainly used in color CRT monitors and color raster graphics. As any desired color can be produced by combining the three primary colors in proportion in the RGB color model, red, green and blue are also called additive primary colors. RGB color model is based on a Cartesian coordinate system (CCS). In the figure, the R, G and B values for each color range from 0 to 255, with each representing 256 discrete levels. A 256-level RGB color model can represent 16777216 colors (known as 24-bit colors). The RGB values of main colors are shown in fig 2.1:

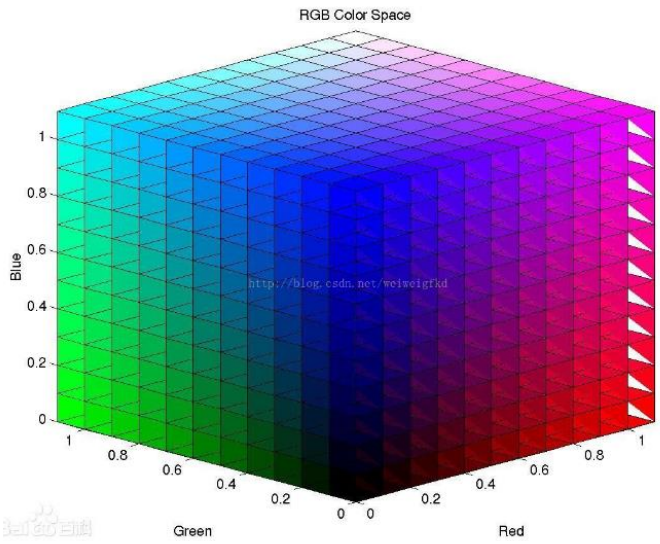

Figure 2.1: The RGB color space

HSV Color Space

HSV color model has three components, hue, saturation and value. Value represents the intensity of color, and has little to do with the color information of an image; hue component and saturation component are closely related with humans visual perception of color, which are the physiological basis for common image processing algorithms. The HSV color model is shown in Figure 2.2:

In the figure, hue change is reflected by an angle ranging from 0 to 360, starting at red, passing through green, blue, pinkish red, and then wrapping back to red. Both 0 and 360 represent red. Saturation value ranges from 0 to 1 and reflects the unsaturated to fully saturated hues (from gray to pure color). Brightness, ranging from 0 to 1, indicates color lightness.

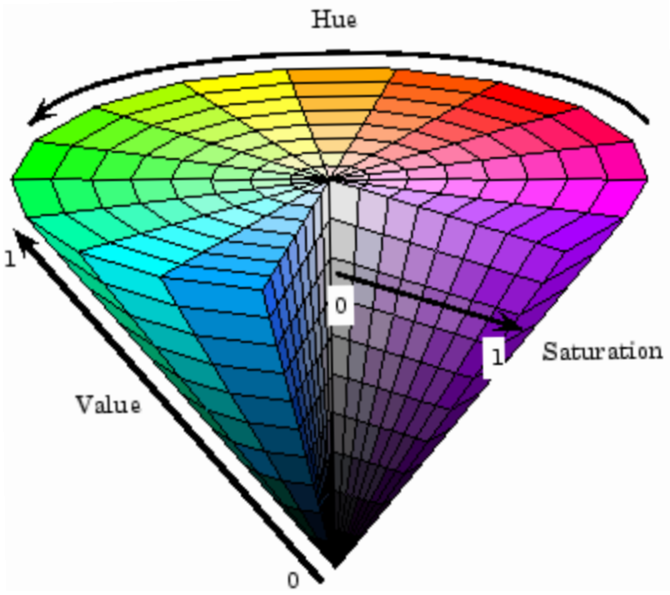

Figure 2.2: The HSV color space

Conversion between RGB color model and HSV color model

In order to select the most suitable color model for each application, we often need to convert the color model. A color model used by a retrieval system should have color perception difference, or enable color difference perceived by human to be reflected approximately by mathematical Euclidean distance. All the HSV model color can be represented by R, G

and B values of the RGB color model.

$$\begin{aligned} H &= \cos^{-1} \frac{\frac{1}{2}[(R-G)+(R+G)]}{\sqrt{(R-B)(R+B)+(R-G)^2}} \\ S &= 1 - \frac{3}{R+B+G}[\min(R,G,B)] \\ V &= \frac{1}{3}(R+B+G) \end{aligned} \tag{2.11}$$

However, conversion from HSV color model to RGB color model is relatively more complex. Different conversion formula is provided for the color of different angle based on hue attribute, where R, G and B are quantized to 0.0 to 1.0.

2.2.12 Perception-based Color Space Histogram Feature

Computations in the RGB and HSV color spaces cannot solve the problem of background illumination sensitivity. The color spaces always affect the computing accuracy of the color histogram. We attempted to use perception-based color space, which exhibits good performance in image processing. As the name suggests, the perception-based color space associated metric approximates perceived distances and color displacements, capturing relationships that are robust to spectral changes in illumination.

RGB color space can be transformed to perception-based color space through the following steps.

1. Transform RGB to XYZ color space using the following formula 2.5:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \frac{1}{0.177} \begin{pmatrix} 0.49 & 0.361 & 0.20 \\ 0.177 & 0.0812 & 0.011 \\ 0.00 & 0.01 & 0.99 \end{pmatrix} \begin{bmatrix} G(R) \\ G(G) \\ G(B) \end{bmatrix} \tag{2.12}$$

where G() is the gamma-correction function. The gamma correction function addresses color distortion and rediscovers the real environment to a certain extent.

2. Transform XYZ to UVW color space.

In UVW color space, the influence of lighting conditions is simulated by the tristimulus multiplication values and scale factor, as shown in the following formula 2.12:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \rightarrow \begin{bmatrix} U \\ V \\ W \end{bmatrix} = B^{-1}DB \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{2.13}$$

Transfer matrix from the current color coordinates to the base coordinates. The non-linear transfer uses the following formula 2.13:

$$\begin{bmatrix} U \\ V \\ W \end{bmatrix} = A\left(\widehat{ln}\left(B \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}\right)\right) \tag{2.14}$$

where A and B are invertible 3×3 matrices and denote the component-wise natural logarithm. Matrix B transforms the color coordinates to the basis in which relighting best corresponds to multiplication by a diagonal matrix, while matrix A provides degrees of freedom that can be used to match perceptual distances. Based on similar color experiments in the database, the A and B matrix-value formulas are shown as formula 2.15 and 2.16:

$$A = \begin{pmatrix} 27.07439 & -22.80783 & -1.806681 \\ -5.646736 & -7.722125 & 12.86503 \\ -4.163133 & -4.579428 & -4.576049 \end{pmatrix} \tag{2.15}$$

$$B = \begin{pmatrix} 0.9465229 & 0.2946927 & -0.1313419 \\ -0.117917 & 0.9929960 & 0.007371554 \\ 0.0923046 & -0.046457 & 0.9946464 \end{pmatrix} \tag{2.16}$$

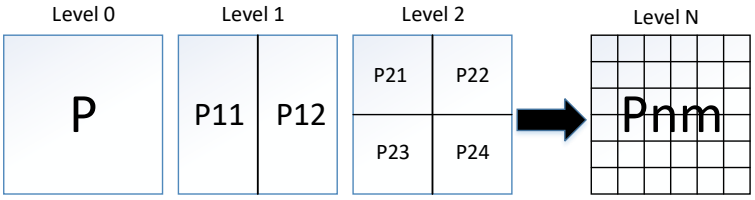

Figure 2.3: The Method of SPM.

2.2.13 SPM Model

Lazebnik et al.[27] proposed the Spatial Pyramid Matching(SPM) in 2006. The SPM model coarse spatial information. Based on spatial consider the information in color histogram in order to restrain the no order of space information. The model divides the image into different levels, which can then be further refined. The SPM model space is shown in Fig 2.3. The level 0 image P based on the original image feature information. But the image feature is based on the global unordered color information, Leve1 show image be separated as space geometry. P11 and P12 are expressed by a spatial order that contains simple space information.

P11 and P12, which also lack internal space information, are in level 1. If internal space information is necessary in P11 and P12, they must be separated using the same process. The level i+1 feature is divided by level i. The levels of division are decided by the actual situation.

The SPM Histogram Feature

Image similarity is computed by the levels corresponding to parts in SPM model. For two images P and Q, the formula is as follows:

$$d(P,Q) = \sum k_{ij} d(p_{ij}, q_{ij}) \tag{2.17}$$

where P_{ij} is the image P histogram feature of the part j in level i; d(p_{ij},q_{ij}) is the feature-similarity degree image P and Q; and K_{ij} is the weight of the similarity calculation. In this case, we focus on part j of level i. The weight of calculation should be set high.

2.2.14 Gaussian Color Model

GMM is constantly used for color image segmentation according to the classification and clustering of image characteristics[31]. The image is divided into different parts based on pixel classification. We considered the main part of person identification to be based on minutia matching and ignored details. The retrieval of similar objects in a video system prioritize the main part of similarity matching and does not emphasize accurate detail matching, so we considered the main colors as the features of the Gaussian color model.

Gaussian Distribution

The Gaussian distribution is a parametric probability density function that is a mean value and variance continuous distribution maximum information entropy[44]. As shown in equation 2.11, when distributing a unit value that fits the normal distribution random variable, the frequency of the variable that follows the Gaussian distribution is entirely determined by the mean value μ and variance σ^2. As x approaches μ, probability increases. σ means the dispersion, and the value of σ is a much greater degree of dispersion.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{2.18}$$

For an image, the Gaussian distribution describes the distribution of specific pixel brightness that reflects the frequency of some gray numerical value. A single-mode Gaussian distribution cannot represent a multicolored image. Therefore, we used a multiplicity of Gaussian models to show different pixel distributions that approximately simulate a multicolored image. Theoretically, we could increase the numbers of models to improve the descriptive ability.

Every pixel of the color image could be represented as a d dimensional vector x_i (color image $d = 3$, gray image $d = 1$). The whole image could be represented as $X = (x_1^T, x_2^T, \cdots, x_N^T)$, where N is the sum of all pixels in a picture, X is represented as M states in GMM, and the value of M is usually restricted from 3 to 5. The linear stacking of the M Gaussian distributions could show the GMM of the probability density function,

as shown in equation 2.12: x is the pixel sampling of a picture.

$$P(x) = \sum_{k=1}^{M} p(k)p(x|k) = \sum_{k=1}^{M} \pi(k)N(x|\mu_k, \sum k) \tag{2.19}$$

$N(x|\mu_k, \sum k)$ is the single Gaussian density function. $k = 1...M$ indicates the Gaussian density function of $No.k$. μ_k is the sample mean vector, $\sum k$ is sample covariance matrix, and $\pi(k)$ is the nonnegative coefficient of weight that describes the proportion of $No.k$ data in the total data, and $\sum_{k=1}^{M} \pi(k) = 1$.

2.2.15 Color Histogram Feature Extraction

The histogram of an image is related to the probability distribution function of the images pixel density. When this concept is extended to a color image, it is necessary to obtain the joint probability distribution value for multiple channels. In general, a color histogram is defined by the following equation 2.13:

$$h_{A,B,C} = N \bullet Prob(A = a, B = b, C = c) \tag{2.20}$$

where A, B and C indicate three color channels (R, G and B or H, S and V) and N is the sum of all pixels in the image. In terms of computing, the first step is to discretize the pixel values of the image, creating statistics for the number of pixels of each color for color histogram.

Chapter 3

Related Work

3.1 Pedestrian Detection

The pedestrian detection could be consider as using the bounding boxes to tag the human by some algorithms or methods in a given image or video. Pedestrian detection had been applied in many fields of image processing. In video surveillance, pedestrian detection is one of important step to identify and search the specific human. The target is to find the pedestrians from image or video. Various features have been used in pedestrian detection. The previous work focused on three aspects, features, classifiers and learning algorithms.

3.1.1 Shape and Appearance Feature

Later studies[14][11] explored the low-level features, such as edge, texture, or colour in an image, to learn the appearance features for the object detection. In the study of[64]citecao2013transfer, the authors used a set of multi-scale orientation (MSO) features to detect the edges of human in the images. The MSO features are contained by coarse and fine features. The coarse feature is the unit orientation, and the fine feature is the pixel orientation histogram. The two features are extracted from the square image blocks of various size. In their study the Adaboost and SVM algorithm are employed to learn the two-stage classifiers for pedestrian detection. The study of [3] also used the Histogram of Oriented Gradient(HoG) algorithm to detected the relevant information by enhancing the human shape. Usually, the HoG algorithm is used as baseline to develop the new human detection algorithms. The authors presented a method that based on the extrac-

tion of features by using Gabor filters. They built a Gabor filter bank by convolution of the images, And the HoG algorithm is used to extract the relevant features from the Gabor filter bank. In the last a classifier is used to judge the original image is or not a person. The study of[46] also used the extended histogram of Gradients for detecting the human in images. ExHoG is a new feature that could improve the weakness of the HoG algorithm. When a bright human in a dark background, the HoG maps gradients of opposite directions in the same histogram bin. It will make the different objects to produce the same features, and could not differentiate some local structures. In their study, the authors proposed a Minimum Mahalanobis Distance classifier(MMDC) that could use the inverse of the covariance matrices estimated from the training samples, and use an Asymmetric Principal Component Analysis(APCA) for improving quadratic classification. The study[45] proposed Discriminative Robust Local Binary Pattern to solve the problem of LBP that could not discriminate the human parts between weak contrast local regions and similar strong contrast ones. Local Binary Pattern (LBP) is always used in the texture classification and face detecting. For human detection, the histogram of LBP codes often has the similar weight, these form of histograms are considered as having the similar features[57]. In order to reduce this problem, the authors used a weighting scheme, the gradient magnitude of each pixel would be computed and the values is used to weigh the LBP code. The DRLBP used the gradient weight and determined the difference of the bins in the LBP codes by their respective complement codes.

3.1.2 Classifiers and Learning Algorithms

The later studies of human detection[55] is formulated as a matter of quadratic classification, which used linear support vector machines to classify the collected features from the background of images. The human bodies in the image background could provide useful information about the view and posture variation. The piecewise discriminative function could construct a classification boundary, that has a good performance in discriminate multiview and multiposture human bodies in a high-dimensional. The authors use a different piecewise SVMs that have a membership degree maximization criterion in training step. Some subspace is separated from the feature space which can be better discriminative for the SVM.

The motion feature[42] is often used in the human detection from video sequences, which is extracted from the two consecutive frames of a video sequence. The State-of-the-art methods for human detection are make use of deformable part model(DPM),In their study, Spatio-temporal Matching algorithm is proposed which has three main components. Firstly, the STM algorithm could learn a bilinear spatio-temporal 3D model from motion information in the video frames. Then, the STM extracts 2D featureand evaluates the pseudo-likelihood of each pixel in the body parts. At last, STM would find a subset of trajectories that correspond body of 3D joints in the STM model. A patch-based method [6] is a general tree representation of a feature sets from an image, which could represent the positive features or negative features. In their study, the labeled ordered tree was used to represent a block based descriptor such as HOG is extracted by partitioning the window into some size blocks. The pedestrian detection method combined multiple local features by a Random Forest ensemble (RF). They used the HOG and LBP to reuse the same features by the multiple local experts. Experiment showed that the method has a good computing cost and speed, and also has a better performance with the state-of-the-art.

3.2 Person re-identification

3.2.1 Color Feature

Color features are one of the low-level feature types that have been widely used in content-based image retrieval (CBIR). Compared with other features, color exhibits little dependence on image rotation, translation, scale change and even the shape change. Color is thus thought of as almost independent of the images dimensions, direction and view angles. Most representations in previous approaches are based on appearance. Gray et al[36]. used a similarity function that trained from a set of data. These authors focused on the problems of unknown viewpoint and pose. The method is robust to viewpoint change because it is based on the ensemble of localized features (ELF). Farenzena et al[20]. presented an appearance-based method based on the localization of perceptually relevant human parts. The information features contain three parts: overall chromatic content, the spatial arrangement of colors into stable regions and the presence of recurrent local motifs with high entropy. The method is robust to pose, viewpoint and illumination variations. Zhao

et al[63]. transformed person re-identification into a distance learning problem. Using the relative distance comparison model to compute the distance of a pair of views, these authors considered a likely true match pair to have a smaller distance than that of a wrong match pair. These authors also used a new relative distance comparison model to measure the distance between pairs of person images and judge the pairs of true matches and wrong matches. Angela et al. proposed a new feature based on the definition of the probabilistic color histogram and trained a fuzzy k-nearest neighbors (KNN) classifier based on an ad-hoc dataset. The method is effective at discriminating and reidentifying people across two different video cameras regardless of viewpoint change. Metternich et al[35]. used a global color histogram and shape information to track people in real-life surveillance data, finding that the appearance of the subject impacted the tracking results. These authors also focused on the performance of matching techniques over cameras with different fields of view.

3.2.2 Metric Learning

Hirzer et al[24]. focused the matching method of metric learning on person re-identification. These authors accomplished metric learning from pairs of samples from different cameras. The method benefits from the advantages of metric learning and reduces the required computational effort. Good performance can be achieved even using less color and texture information. Khedher et al[25]. proposed a new automatic statistical method that could accept and reject SURF correspondence based on the likelihood ratio of two Gaussian mixed models (GMMs) learned on a reference set. The method does not need to select the matching SURF pairs by empirical means. Instead, interest point matching over whole video sequences is used to judge the person identity. Matsukawa et al[34]. focused on the problem of overfitting and proposed a discriminative accumulation method of local histograms for person reidentification. The proposed method jointly learns pairs of a weight map for the accumulations and employs a distance metric that emphasizes discriminative histogram dimensions. This method can achieve better reidentification accuracies than other typical metric learning methods on various sizes of datasets.

Chapter 4

Overview of Proposal Methods

4.1 The Methods for Pedestrian Detection

4.1.1 Overview of the Pedestrian Detection Algorithm

Our proposed multi-resolution DPM is similar to a hybrid deformable model with two target models. But there are also big differences between these methods. Firstly, many parameters are shared in our deformable model, while while all parameters in the hybrid deformable model are independent from each other. Secondly, our multi-resolution deformable model consist a different procedure for the variable s_i. At the training state, s_i is a hidden variable, while at the test state s_i becomes a visible variable. The procedure of pedestrian detection by multi-resolution DPM is shown in Fig 4.1 First, DPM parameters w_0 and w_1 are initialized by the initialization method as illustrated before. w_0 is the model parameter at a low resolution, while w_1 is the model parameter at a high resolution.

At the training state, we set the value of s by considering the height h of a trained sample, for which $s = 1$ if $h > 100$ and $s = 0$ if $h < 50$. For the samples with $s = 1$, LSVM can be used to train a standard DPM model, which renders the model parameter w_1. For the samples with $s = 0$, a linear SVM can be used to train a DPM model (no hidden variable is involved), which renders the model parameter w_0. For the samples with height $50 < h < 100$, we add s to the set of hidden variables, and train the model with the LSVM algorithm.

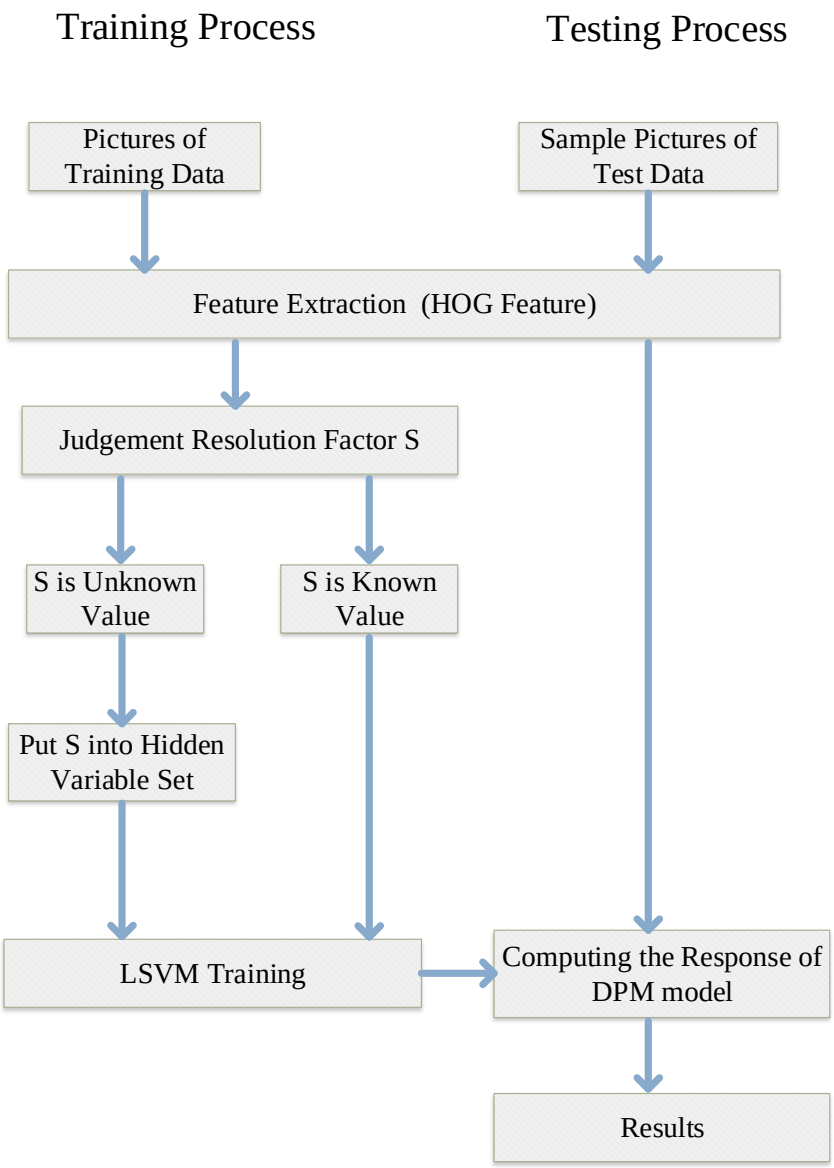

Figure 4.1: The Processing Steps of Pedestrian Detection on Multiple Resolution.

4.1.2 Multi-resolution DPM Algorithm for Pedestrian Detection

A significant feature of the above method is that a rigid template is used for object detection at both large and small scales. The description operators at low levels (e.g., HOG feature) are adaptable to small image deformation. However, such method is not applicable in case of large scales. For example, HOG feature detector is invariant to different postures of a 50 pixel height pedestrian, but not invariant to the 100 pixel height pedestrians. If we are to detect a large-scale target, we can choose a low-resolution template. And if we hope to gain more information, we can select a high-resolution template. For a good

adaptability to the deformation at a large scale, we adopt a DPM model. As a hidden parameter z is defined in the DPM model, we use $\phi_1(x, z)$ as the combination of HOG feature and the deviation.

$$\phi(x,s) = \begin{bmatrix} \phi_0(x) \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad if \quad s=0 \qquad and \qquad \phi(x,s,z) = \begin{bmatrix} \phi_0(x) \\ 1 \\ \phi_1(x,z) \\ 0 \end{bmatrix} \quad if \quad s=1 \tag{4.1}$$

The classifier passes through all the hidden variables at last.

$$f(x,s) = \max_z w \cdot \phi(x,s,z) \tag{4.2}$$

$f(x, s)$ is transformed into a standard linear template for calculating the response at a low resolution. For calculating the response at a high resolution, $f(x)$ would need to search all part models to find the model which makes the maximum response. Suppose the distances between different parts and the root filter are independent from each other, the following formula can be calculated with the QP algorithm:

$$\max_z w_1 \cdot \phi_1(x,z) = \max_z \sum_j w_j \cdot \phi(x, z_j) + \sum_{j,k \in E} w_{jk} \cdot \phi(z_j, z_k), \tag{4.3}$$

in which z_j denotes the location of part j, w_j denotes the template of part j, w_{jk} denotes the deformable model of part j and k, and E denotes the boundary. $\phi(x, z_j)$ represents the HOG feature at location z_j, and $\phi(z_j, z_k)$ represents the deformation difference between part j and part k. For a given training set (x_i, s_i, y_i), we can employ the LSVM algorithm for training the model parameter w.

At the stage of model training, we set $s_i = 1$ if the training sample has a large scale ($h > 100$), in which s_i is no longer a hidden variable. And we set $s_i = 0$ if the training sample has a small scale (h¡50), in which s_i is not a hidden variable, either. If the training sample has a medium scale ($50 < h < 100$), the training sample can be considered as both a high resolution object and a low resolution object. In this case, s_i becomes a

hidden variable, which can be added to the set of hidden variables of the LSVM model for training. The rough procedure consists of the random initialization of variables s_i and z_i, the calculation of model parameter β in model training, and the acquisition of value for the hidden variable in accordance to the maximum response which would be taken into the next iteration.

4.2 Methods for Re-identification of Persons in Real-life Video

4.2.1 Measuring Video Object Similarity

The calculation of image similarity is a key step in image retrieval. And the measurement method has a significant influence on the sorting of the retrieval results. Different from the traditional text-based retrieval methods, which based on point and range queries, content-based video object retrieval relies on similarity matching. The similarity of the image to be calculated is the degree of similarity between the eigenvectors representing the image content. Similarity calculation methods vary with the attributes of image content. Generally, the image features are extracted and quantified and then represented by vectors, which can be viewed as points in a multi-dimensional space. The similarity between images is measured by the distance between the points in the feature space. In addition, relation analysis and correlation coefficient are also utilized.

Distance Measurement

After extracting the image features, the most intuitive method is to measure the similarity of two images directly by the distance between the image feature vectors. Distance measure is to measure the distance between feature vectors in space. The longer the distance, the greater the difference. Image retrieval is to find those points with nearest feature vector distance with the queried image. Distance Mi is defined as follows:

$$d(x, y) = (\sum_{i-1}^{n} |x_i - y_i|^r)^{\frac{1}{r}} \tag{4.4}$$

where, X and Y are images, and x_i and y_i are the $i-th$ feature vector of Image X and Image Y, respectively. when r=1,the distance is called $Manhattandistance$, which is the distance of absolute distance When r=2, the distance is called $Euclideandistance$, which is the distance we commonly referred to as. The Euclidean distance treats feature difference equally, though each attribute may play a different role in distinguishing features. When r goes to infinity, the distance is called $Chebyshev$ distance, which can be expressed as Equation :

$$d(x,y) = \max_{1\leq i\leq n} |x_i - y_i| \tag{4.5}$$

In addition, P. C. Maharanobis, an Indian statistician, introduced the $Mahalanobis$ distance ($Msdistance$), which takes advantage of covariance matrix as follows:

$$d(X,Y) = (X-Y)^T W^{-1}(X-Y) \tag{4.6}$$

where $d(X,Y) = (X-Y)^T W^{-1}(X-Y)$, W is the covariance matrix that is unable to be calculated without priori knowledge about different classes. If each of these axes is rescaled to have unit variance, then $Mahalanobis$ distance corresponds to standard Euclidean distance. Different from Euclidean distance, $Mahalanobis$ distance takes into account the relationships between the various features and is independent of scale.

Correlation Measurement

In the distance measurement, the feature vector is a point in the feature space, where a smaller spatial distance indicates a higher similarity between two images. In the correlation measurement, the correlation between two feature vectors is calculated to indicate similarity. The following is some commonly used methods of calculation. The inner product correlation is a non-normalized correlation.

$$C = \sum_{i=1}^{n} x_i y_i \tag{4.7}$$

The inner product in this calculation $C = \sum_{i=1}^{n} x_i y_i$ method is possibly infinite. If the inner product of the two eigenvectors is 0, they are not correlated. When the inner product is not 0, the infinity of its value brings trouble to the comparison and ordering. It

does not directly reflect the correlation degree. Cosine formula is a normalized correlation calculation which uses eigenvector length to normalize the inner product of vectors, as shown in equation:

$$\cos\theta = \frac{\sum_{i=2}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \tag{4.8}$$

In this calculation method $\cos\theta = \frac{\sum_{i=2}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$, the cosine of an angle between vectors is used to represent the correlation. Cosine similarity ranges from 0 to 1. If $\cos\theta = 1$, the two vectors are scalar multiples of each other and exactly the same; if $\cos\theta = 0$, the two vectors are orthogonal and not correlated. If the denominator is the standard value 1, then the cosine correlation is the same as the inner product correlation. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. It is commonly used in cluster analysis and multimedia retrieval. For color histogram intersection calculating similarity, a normalized histogram intersection method is as follows:

$$r_{XY} = \frac{\sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i=1}^{n} x_i} \tag{4.9}$$

In fact, if the denominator is calculated as Equation (4.9), it is a correlation coefficient:

$$r_{XY} = \frac{\sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i=1}^{n} \max(x_i, y_i)} \tag{4.10}$$

The correlation calculation is mainly used for color similarity measurement, especially for color histograms, for its insensitivity to image size, histogram dimension, partial occlusion and other conditions.

Multi-feature Similarity Calculation

The above similarity measurements are for cases where there is only one single eigenvector. In fact, the description of an image generally requires more than one feature. We make use of a multi-dimensional composite feature integrating features such as color, texture or shape to fully describe the images attributes. The image similarity based on this multidimensional composite feature can be expressed by a linear weighted sum of each single features similarity. Suppose k features are used to describe a given image, if the

similarity of the $i-th$ feature is δ_i, then the overall similarity δ is as follows:

$$\delta = \sum_{i=1}^{k} \lambda_i \delta_i \tag{4.11}$$

where,$\lambda_i \geq 0$, indicating the importance of the $i-th$ feature which is determined by experiments generally.

In many cases, the image also has binary feature, such as whether the image contains text information or not. Here, relational coefficient should be used in similarity calculation. Suppose there are k features; the similarity of the $i-th$ feature is δ_i, with a weight of λ_i; and there are j binary features, whose similarity is binary ρ (1 or 0). The overall similarity is as follows:

$$\delta = \Pi_i \rho_i(x,y) \sum_k \lambda_k \delta_k(x,y) \tag{4.12}$$

The binary feature plays a key role in the overall retrieval results. If any of the binary features does not match, the image is not matching. It is worth mentioning that the inter-image similarity measurement discussed above does not take into account the effects of human visual characteristics. Studies of visual perception show that humans judgment of similarity also depends on the high-level semantic cognition of image content.

4.2.2 An Overview of the Proposed System

The techniques of moving person retrieval information from a video database include shot segmentation, person detection, scene segmentation, feature extraction and similarity calculation. As shown in Fig 4.2, shot segmentation refers to automatically segmenting video clips into shots as the basic unit for indexing. One second of video contains 20 to 30 video frames[17], and neighboring frames are very similar to each other. There is no need to perform retrieval and matching for each frame, and frame differentiation is used to detect and extract the moving person. Frame differentiation relies on the change of pixel value between neighboring key frames. A change value greater than the established threshold value marks the pixel position of the moving person[5]. This step is important in video parsing and directly affects the effectiveness of moving person retrieval.

The measurement method for similarity calculation influences the results ranking of

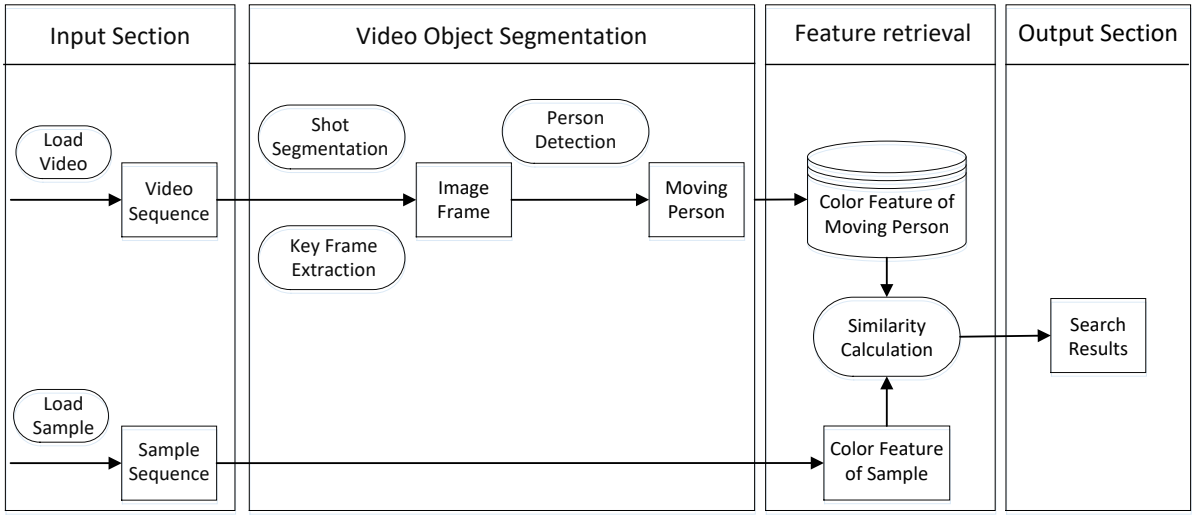

Figure 4.2: Overview of Person Re-identification System.

object retrieval. Essentially, image similarity calculation computes the content of feature vectors from the objects. Each feature attribute selection can employ a different similarity computing method. Frequently, image features are extracted in the form of feature vectors that can be regarded as points in multidimensional space. The most common similarity measure method uses the distance between two spots in feature space. We also use distance measurement and correlativity calculation to scale the comparability between images.

Our proposed method is presented in Fig 4.3 We use traditional histogram and SPM histogram to retrieve the object. The traditional histogram method contains three parts, the color histogram feature extraction, color histogram distance computing and outputting. The difference between SPM histogram and traditional histogram is the histogram distance computing part. The sample image and matching image are segmented into three parts, the upper, middle and lower part. The three parts are then separately computed the color histogram distance, and use average distance to evaluate the results. Then the system uses GMM model to filter the top 20 results, extracts the GMM main color feature and computes the similarity of them. Finally, the system outputs the rank of top 10 results.

4.2.3 Color Histogram Feature Extraction

The histogram of an image is related to the probability distribution function of the images pixel density. When this concept is extended to a color image, it is necessary to obtain the

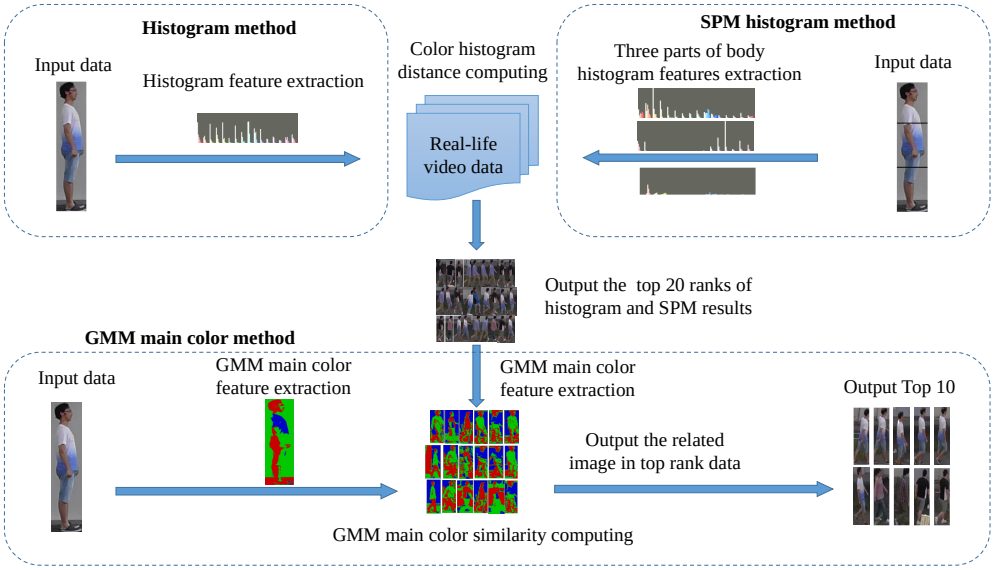

Figure 4.3: Overview of Proposed System.

joint probability distribution value for multiple channels. In general, a color histogram is defined by the following equation (9):

$$h_{A,B,C} = N \bullet Prob(A = a, B = b, C = c) \tag{4.13}$$

where A, B and C indicate three color channels (R, G and B or H, S and V) and N is the sum of all pixels in the image. In terms of computing, the first step is to discretize the pixel values of the image, creating statistics for the number of pixels of each color for color histogram.

4.2.4 Histogram of Color Feature Similarity Measurement

Several methods exist to calculate and weigh the similarity measurement of the histogram. The distance formula of the similarity measure between images is based on the color content. Euclidean distance, histogram intersection and histogram quadratic distance are widely used in image retrieval.

The Euclidean distance of the histogram between two images is given by the following equations:

$$d^2(h, g) = \sum_A \sum_B \sum_C (h(a, b, c) - g(a, b, c))^2 \tag{4.14}$$

where h and g are two histograms and a, b and c are the color channels. The formula subtracts the pixel value in same bin of histogram h and g.

The formula for histogram intersection distance is as follows:

$$d(h,g) = \frac{\sum_A \sum_B \sum_C min(h(a,b,c) - g(a,b,c))}{min(|h|,|g|)} \tag{4.15}$$

where $|h|$ and $|g|$ stand for the pixel values of image sampling in histogram h and g, respectively. The method image sampling measure which could reduce nonmeaningful distractors element in a complex background.

Chapter 5

Evaluating the Pedestrian Detection and Person Re-identification

5.1 Evaluation of Pedestrian Detection

5.1.1 Data set of Person Detection

1.INRIA

The INRIA data set is a static pedestrian detection database which has been widely employed in recent researches. The training set consists of 614 positive samples (containing 2,416 pedestrians) and 1,218 negative samples.The test set consist of 288 positive samples (containing 1,126 pedestrians) and 453 negative samples. Images in the INRIA data set are mainly collected from google, GRAZ-01 and personal photographs.

2.Caltech Pedestrian Database

The Caltech Pedestrian Database is a large scale database. It consists of videos of 640x480 pixels with 30 frames per second, captured by in-vehicle cameras for about 10 hours. Within these videos, 250,000 frames (around 137 minutes), 350,000 bounding boxes, and 2,300 pedestrians are manually annotated by human experts. The data set is divided into 10 sets, among which sets 00-05 are used for training, and sets 06-10 are used for testing. In our experiment, we also employ sets 00-05 for training and sets 06-10 for testing.

5.1.2 Evaluation of Person Detection

Whole Image Evaluation

The detection result for an image consist of a set of bounding boxes (BB) and a corresponding confidence score. If the detection result (BB_{dt}) and the standard result (BB_{gt}) has a great extent of overlapping, we consider the detection result matches the standard result. We define that a detection result matches the standard result if they have over 50% parts overlapped:

$$a_0 = \frac{area(BB_{dt} \cap BB_{gt})}{area(BB_{dt} \cup BB_{gt})} > 0.5 \tag{5.1}$$

Each BB_{dt} can match to at most one BB_{gt}, which means that every detected BB_{dt} can only match one BB_{gt} but not multiple BB_{gt}s. Therefore, if a detection result BB_{dt} could match multiple BB_{gt}s, we only select the BB_{gt} with the highest confidence as the final detection result. If a BB_{dt} is not matched with any BB_{gt}, it is labeled as false positive. And if a BB_{gt} is not matched with any BB_{dt}, it is labeled as false negatives as well.

5.2 Evaluation of Person Re-identification

5.2.1 Data Set of Person Re-identification

Figure 5.1: The Location 1 in Person Re-identification.

Figure 5.2: The Location 2 in Person Re-identification.

We evaluated the performance of different color spaces for real-life video data. Uneven illumination distribution should affect person re-identification results in color images. Therefore, we created a video data set to test the validity and robustness of our method. We recorded the video data on a school campus. Six pedestrians walked from left to right in order under a surveillance camera, as shown in Figure 5.1 and Figure 5.2. Our real-life video data consists of two videos that were recorded simultaneously at different locations. Location 1 was bright and location 2 was dark. The videos were recorded at 25 frames per second.

SARC3D dataset

The SARC3D dataset consists short video clips of 50 people which has been captured with a calibrated camera. We employ the SARC3D dataset to effectively evaluate different person re-identification methods. To simplify the image alignment process, we manually select four frames for each clip which correspond to the predefined positions and postures, i.e. back, front, left, and right, of these people. The selected dataset consists of 200 snapshots with four views for each person. For person re-identification, we randomly choose one of the four views for each person, calculate the similarity scores with all other images, and find the most similar images by sorting their similarities with the chosen image. The images of the same person with different positions and postures should be ranked higher

than the other images.

5.2.2 Evaluation Method of Person Re-identification

We focused on the degree of search result accuracy using evaluation parameters for precision. Precision reflects the capability of filtering irrelevant content. These video retrieval system performance criteria reference the evaluation method for information search systems. For a retrieval object, the retrieval system returns a sort of search results. The precision rate expresses the number of correct relevant retrieval results divided by the number of total retrieval results.

$$Precision(\%) = \frac{A}{A+B} \times 100 \tag{5.2}$$

$$AveragePrecision(\%) = \frac{1}{n}\sum_{i=1}^{n} Precision(i) \tag{5.3}$$

in formula (11),(12), A is the number of correct relevant retrieval examples, B is the number of irrelevant video retrieval examples.

Chapter 6

Experiments of Pedestrian Detection and Person re-identification

6.1 Experiment of Pedestrian Detection

6.1.1 The Result in INRIA

We evaluate our algorithm with the INRIA pedestrian database and the Caltech pedestrian database, respectively. There are nearly one thousand pictures in INRIA pedestrian database. Our model is trained on the INRIA training set, and evaluated on the INRIA test set.

The multi-resolution DPM-based pedestrian detection algorithm acquires a precision of 87.2% on INRIA pedestrian database, which is slightly better than 86.9% as the precision of the standard DPM. As shown in fig 6.1 and 6.2, this is mainly because that there are too few picture samples in this data set, and the pedestrian scale is too big in the pictures. Only a few pictures contain the small-scale pedestrians, in which case the multi-resolution DPM detector cannot change the detection method at the high resolution. Therefore, the detection result is almost the same as the standard DPM result.

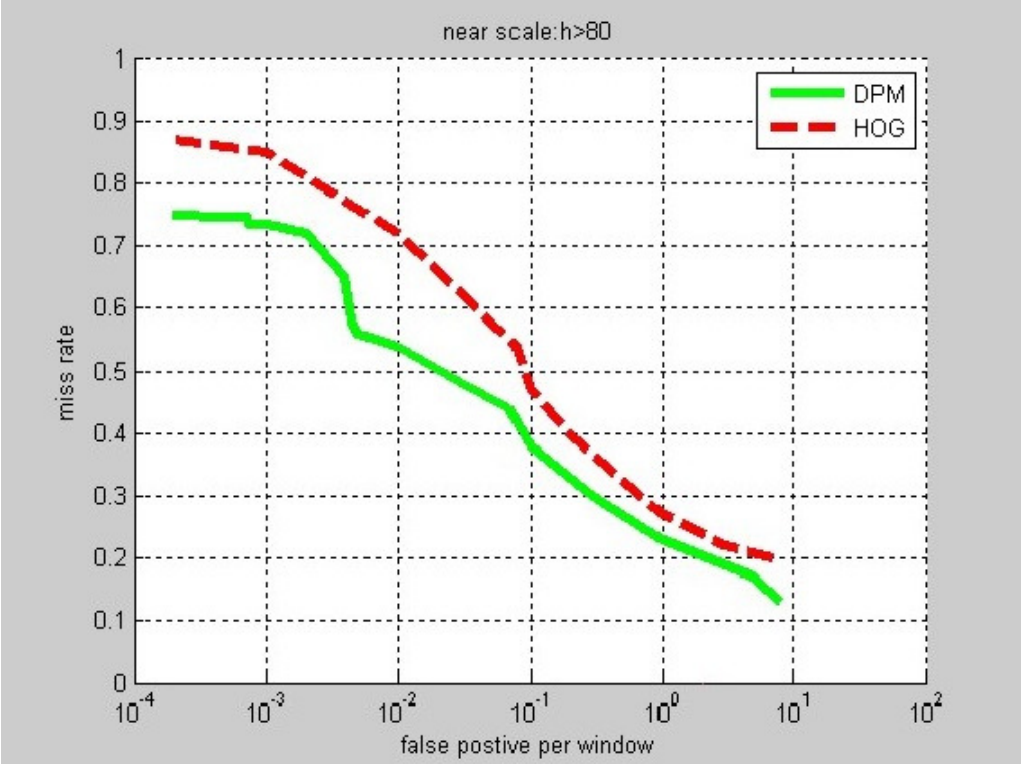

Figure 6.1: The Result on Near-Distance in INRIA.

6.1.2 The Result in Caltech

Training our model on the Caltech pedestrian database is more challenging than on the INRIA dataset. The number of samples in the Caltech pedestrian dataset is large, which is much larger than the number of samples in the INRIA dataset. In addition, samples are mainly acquired from the in-vehicle cameras. The images are close to the real world, which is different from the images collected on the Internet in the INRIA dataset. Finally, the Caltech pedestrian dataset consists samples of 640×480 pixels resolution, in which many pedestrian targets of small scales are included.

In this section, *set*0 to *set*5 in Caltech database are employed for training, and set6 to set11 are chosen for testing. The experiment result suggests that the multi-resolution DPM algorithm renders better detection results than that the standard DPM in terms of all testing sets, figure 6.3 shows the result based on all test samples. The multi-resolution DPM-based pedestrian detection algorithm achieves a missing rate of 52% (1FPPI), which is much better than the missing rate 59% (1FPPI) achieved by the DPM pedestrian detection algorithm. This result suggesting that multi-resolution DPM has better detection

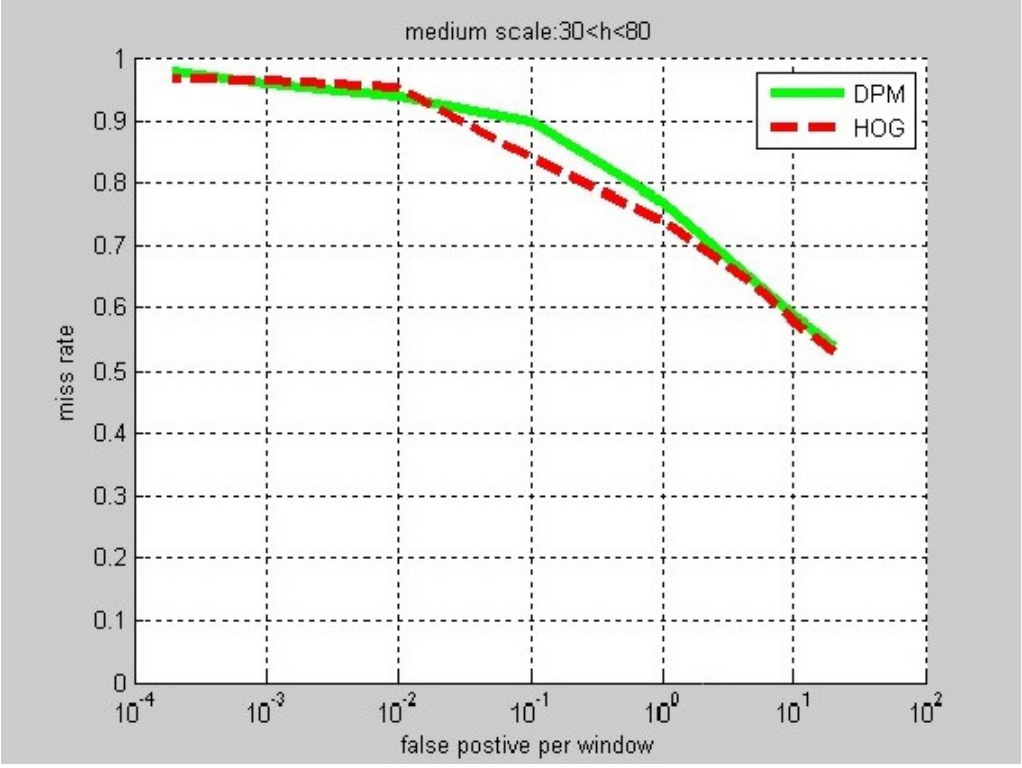

Figure 6.2: The Result on Middle-Distance in INRIA.

effect than the standard DPM. In terms of the large-scale samples (h¿80) in Caltech database, the multi-resolution DPM-based detection algorithm renders a missing rate of 18% (1FPPI), while the standard DPM-based detection algorithm renders a missing rate of 26% (1FPPI), which suggests that the multi-resolution DPM also outperforms the standard DPM for large-scale targets, as shown in fig 6.7 and 6.8. For small-scale samples, the difference in both models is not as obvious, which is mainly because that the useful information at small scales is very limited (the objects at less than 30 pixels are known as small-scale objects). At this point, the detection algorithm cannot acquire enough information for detection.

Figure 6.4, 6.5 and 6.6 show the detection results for near-distance, middle-distance and far-distance samples, respectively. Specifically, the near-distance corresponds to pedestrians with height $h > 80$ pixels, the middle-distance corresponds to pedestrians with height $h > 30$ but $h < 80$, and the far-distance corresponds to pedestrian with height $h < 30$.

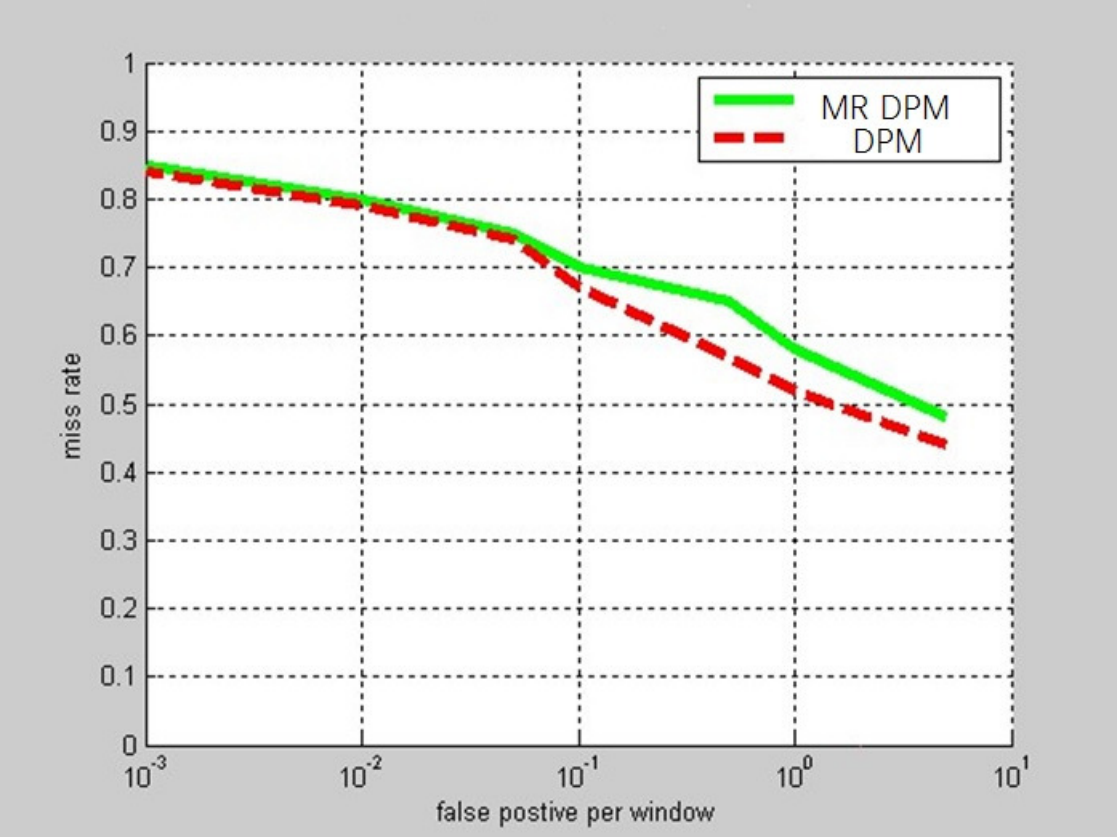

Figure 6.3: The Result on All-distance Samples in Caltech database.

6.1.3 The Result in Part of Caltech

In order to better explain the experiment result, we further select 3,000 pictures from the standard Caltech database for evaluation. This experiment is conducted to compare the detection effects of the detection algorithm at a high resolution (corresponding to the standard DPM algorithm), the detection algorithm at a low resolution (corresponding to merely the root filter-based DPM algorithm), and the proposed multi-resolution DPM algorithm. The experiment result is shown in Figure 6.9, in which LR represents the low-resolution detection algorithm, HR represents the high-resolution detection algorithm, and MR represents the multi-resolution detection algorithm. According to the overall comparison, the multi-resolution DPM algorithm is better than the high-resolution detection algorithm on the testing set. We find that the missing rate on test set is 52% with a rigid template. This is lower than 59%, which is the missing rate with the standard DPM algorithm. This is because most of the pedestrian samples are with height $30 < h < 80$, and even the high-resolution algorithm cannot detect the small-scale targets.

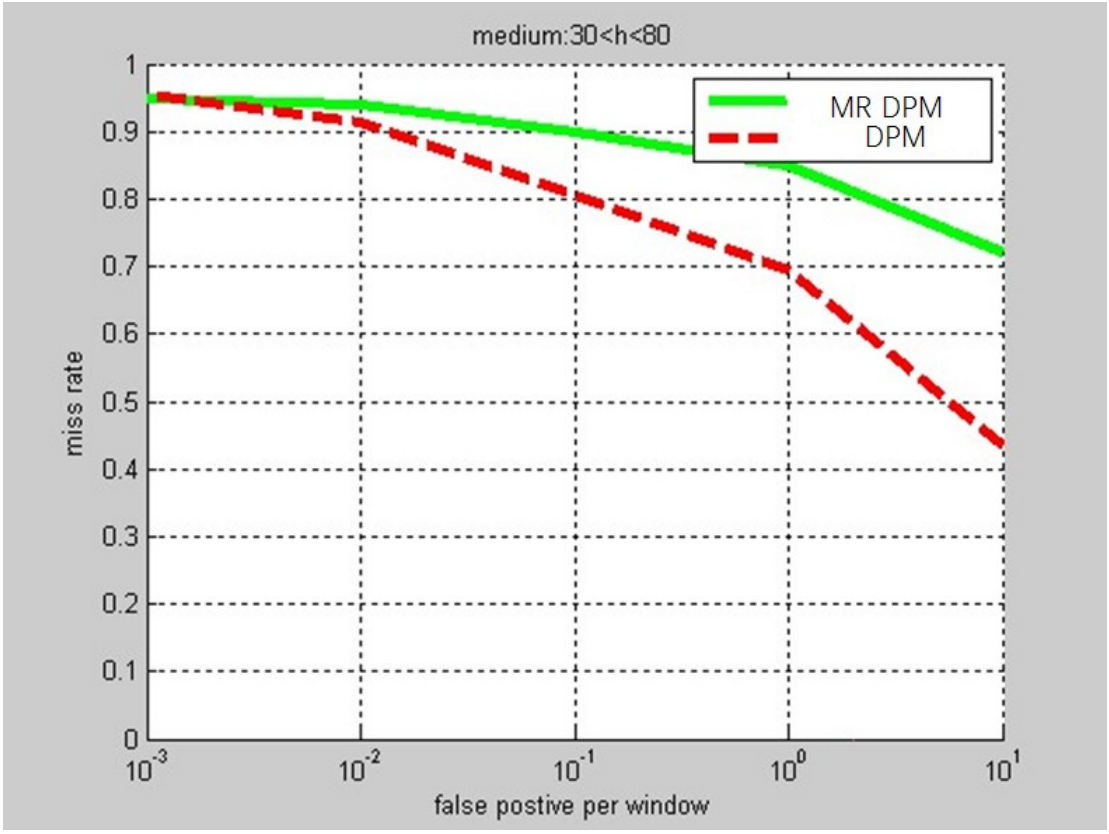

Figure 6.4: The Result on Near-Distance in Caltech database.

The standard DPM algorithm achieves a good detection effect for the large-scale targets. In terms of large-scale target detection, the multi-resolution DPM algorithm achieves similar performance as the standard DPM algorithm. Both algorithms are better than the low-resolution detector. In terms of the small-scale target ($h < 90$) detection, we find that the detection effect of the standard DPM algorithm drops quickly as the target scale grows small. The detection effect of the multi-resolution DPM algorithm is slightly lower than that of the rigid template.

6.2 Experiment of Person Re-identification

We compared the following three methods in our experiments.

Method 1: Traditional color histogram in RGB, HSV, UVW color space.

Method 2: SPM color histogram RGB, HSV, UVW color space.

Method 3: GMM main color feature RGB, HSV, UVW color space.

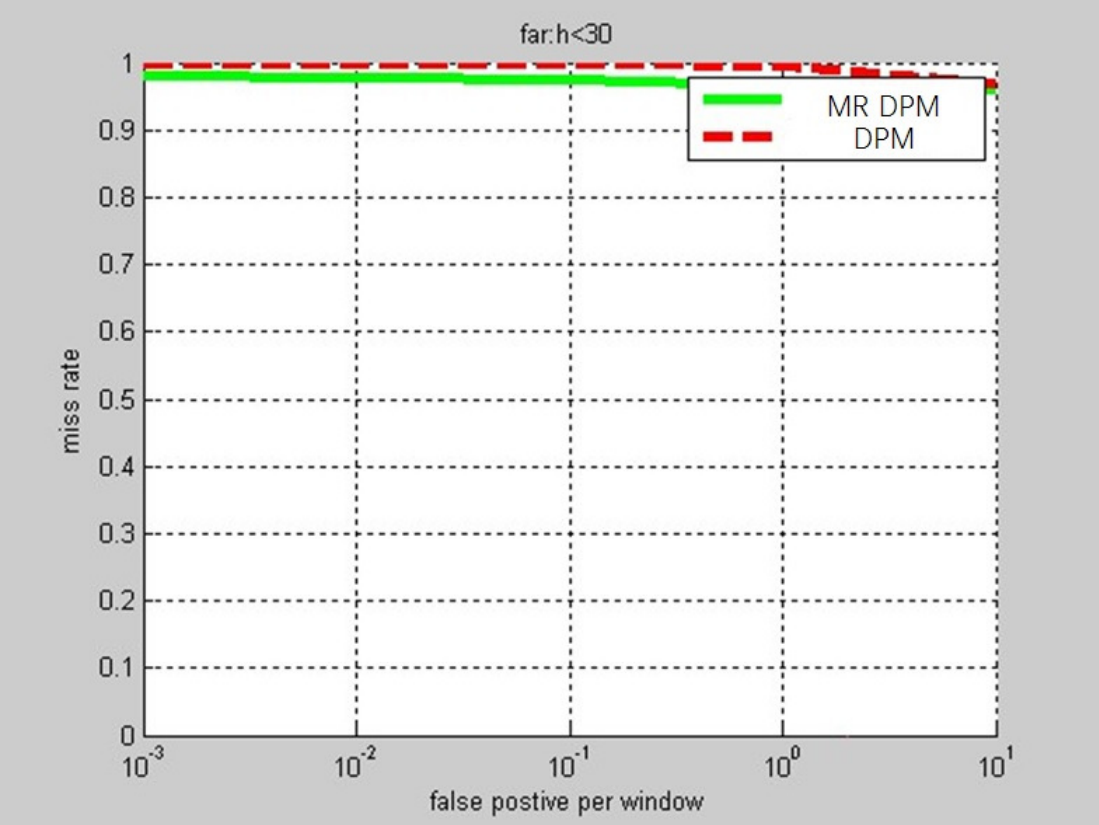

Figure 6.5: The Result on Middle-Distance in Caltech database.

6.2.1 Experiment in Different Color Space

As shown in Table 6.1, the histogram in RGB has a better performance than in HSV in location 1. We listed the top 5 results of RGB histogram for location 1 in Table 5. The quantization method of the RGB color space was RGB_{bin} = 8,8,8. We found the RGB color space can reflect all sorts of colors from the image of result but the color channels contain the background color. This is the important factor affecting the results. Because the scene in the location 2 is too darker, The traditional color histogram method has not spatial information to distinguish between them.

HSV color space The RGB results are based on machine vision, while the HSV results are closer to human visual perception. location 2 are shown in Table 6.2. the quantification of HSV is HSV_{bin}=16,8,8. The wrong result of person wears a similar clothes with right result that a dark red coat and dark green trousers. When we compare the histograms between the results, the wrong result does not perform the histogram of trousers color, but the background color. In location 2, the illumination is dark, and the background is

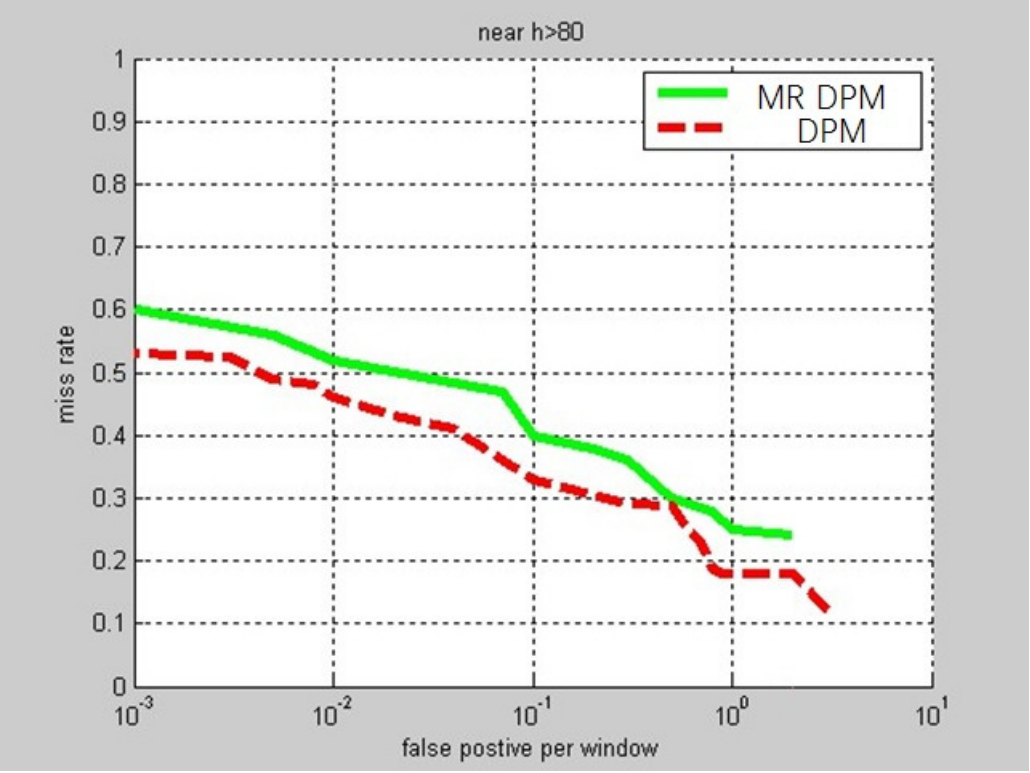

Figure 6.6: The Result on Far-Distance Samples in Caltech database.

	Color spaces		
Persons	RGB(%)	HSV(%)	UVW(%)
P1	80	70	90
P2	80	80	90
P3	70	80	70
P4	80	70	80
P5	60	60	70
P6	80	80	80

Table 6.1: The Precision Rate for Persons Retrieval in Location 1

grey, the results of histogram just showed the clothes, and neglected the background and trousers color.

UVW color space As shown in Table 6.1, the performance of UVW is better than HSV and RGB. The reason is that the results were affected mostly by the color transfer, In different illumination, the color histogram of one??s clothes would be transferred to other color. For example, the red color in a dark environment seems like a black or gray color. The UVW color space is aimed at this problem.

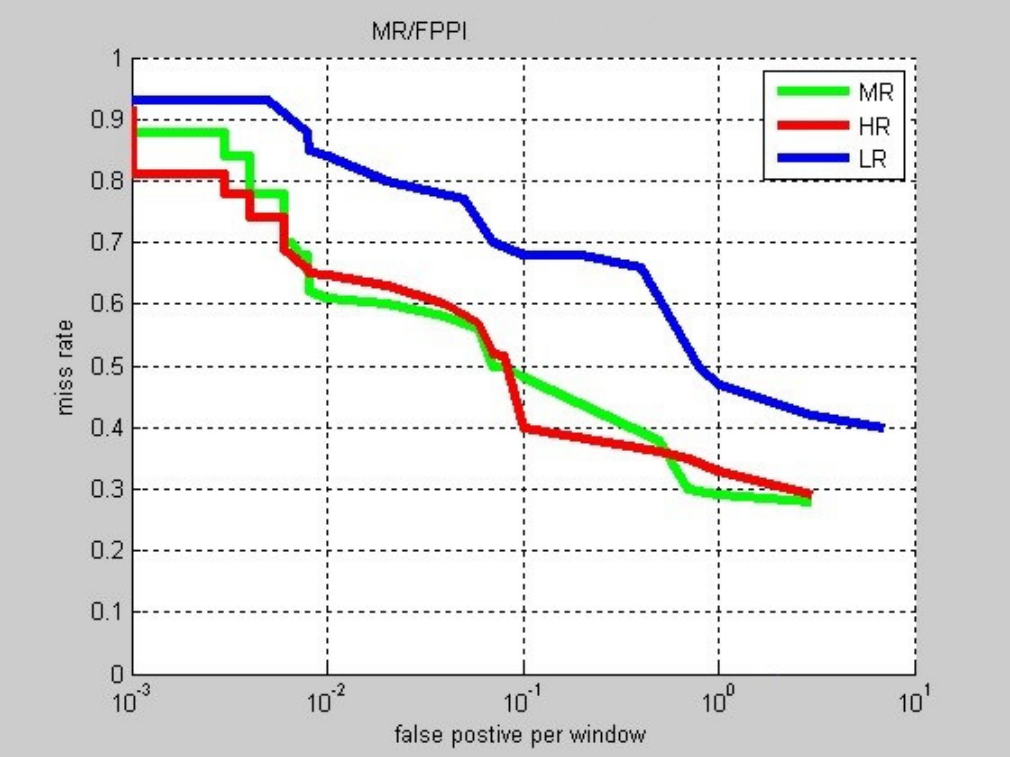

Figure 6.7: Large-scale Target on Subset of the Caltech.

	Color spaces		
Persons	RGB(%)	HSV(%)	UVW(%)
P1	70	80	100
P2	80	70	70
P3	80	80	90
P4	70	70	80
P5	60	70	70
P6	80	80	80

Table 6.2: The Precision Rate for Persons Retrieval in Location 2

6.2.2 The SPM Histogram Experiment

The color space histogram method is presented based on SPM. The SPM histogram provides additional space information in the matching step. The sample person image is evenly segmented as three parts: upper, middle and lower. The three parts are separately computed the color histogram distance, and use average distance to evaluate the results. As shown in Table 6.3 and Table 6.4, the performance of SPM histogram is better than the traditional histogram in RGB and HSV color space. Because the separated parts of retrieval object often contain the different background colors and the proportion of the background is greater than the object??s body in the lower part.

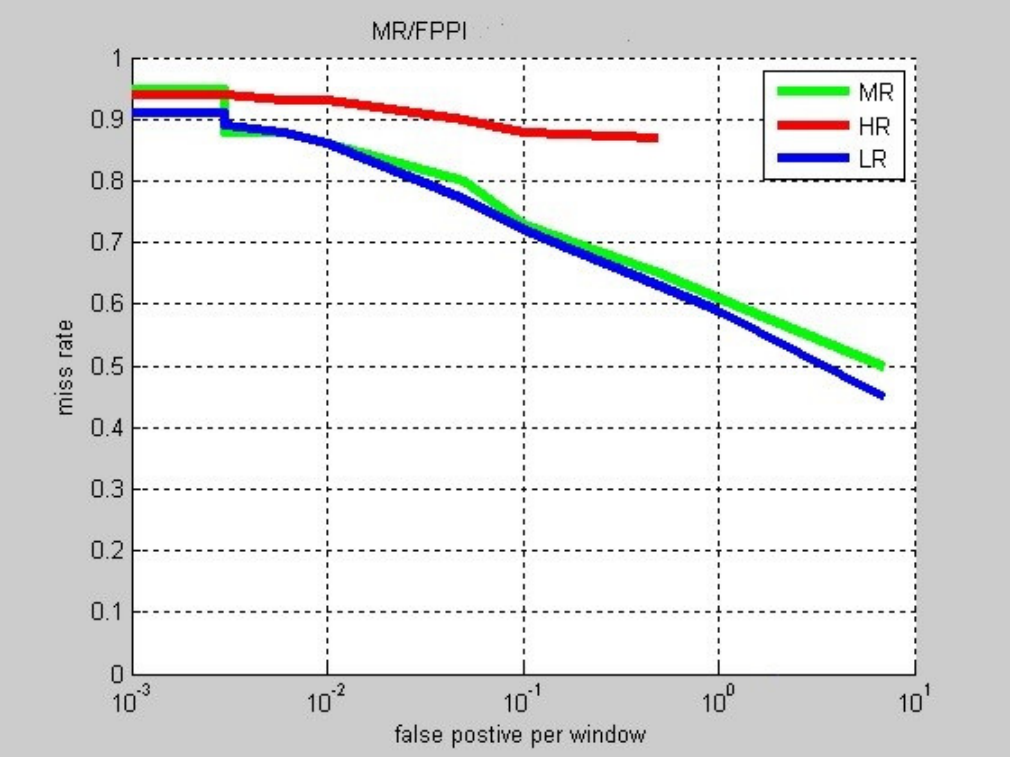

Figure 6.8: Small-scale Target on Subset of the Caltech.

Method	RGB	HSV	UVWS
Histogram	75	73.33	80
SPM histogram	71.66	75	70
GMM	86.66	85	88.33

Table 6.3: The Average Precision for Persons Retrieval in Location 1.

6.2.3 The GMM Main Color Experiment

In case of GMM color modeling, the similarity of two images are the measurement of color Gaussian distribution. In this method, we match the similarity of the main color distribution between the sample and retrieval objects. Because of the weak color discrimination, we use the red blue and green primary colors to constitute an image.The right results have a similar color distribution with the sample person image. In the last image, the irrelevant results of GMM weight distribution are different from the relevant result. the

Method	RGB	HSV	UVWS
Histogram	73.33	75	81.66
SPM histogram	83.33	85	80
GMM	81.66	83.33	85

Table 6.4: The Average Precision for Persons Retrieval in Location 2.

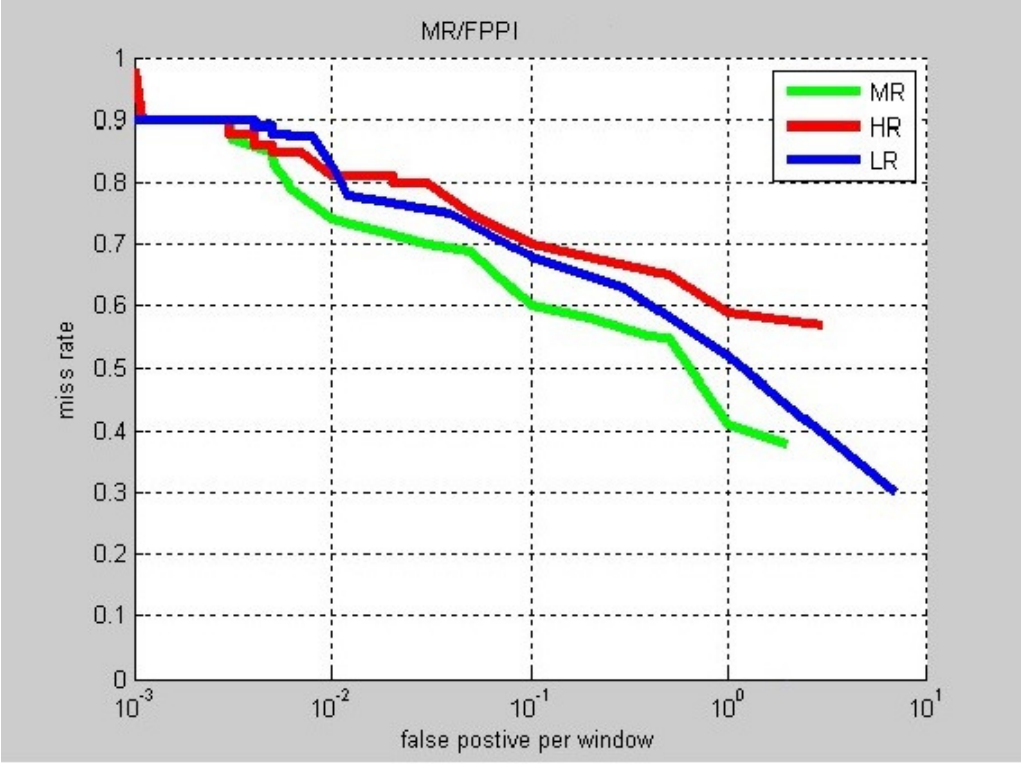

Figure 6.9: The Result on Subset of the Caltech.

GMM method effective solve the problem of color transfer in location 1, but in a dark environment, location 2, GMM has a poor performance.

6.2.4 Experiment on VIPeR Dataset

We examine the appearance model for person re-identification based on the VIPeR dataset, which consists of 632 pedestrian image pairs taken from arbitrary viewpoints under varying illumination conditions. Each image is scaled to 128×48 pixels.

As shown in Fig 6.10, our proposed method outperforms the histogram-based methods in the RGB color space, and the traditional histogram and the SPM histogram methods generate very similar results. We also observe that the proposed method in the HSV space performs better than in the RGB space, as shown in Fig 6.11. This is because that the image illumination in the VIPeR dataset varies significantly. The SDALF method renders a slightly better result than our proposed method, while our method has a great advantage on the calculation cost. Specifically, the SDALF takes about 3850 seconds to extract its features from 1264 images in the VIPeR dataset, while our proposed method takes only 40 seconds to extract and calculate the color histogram features. In addition,

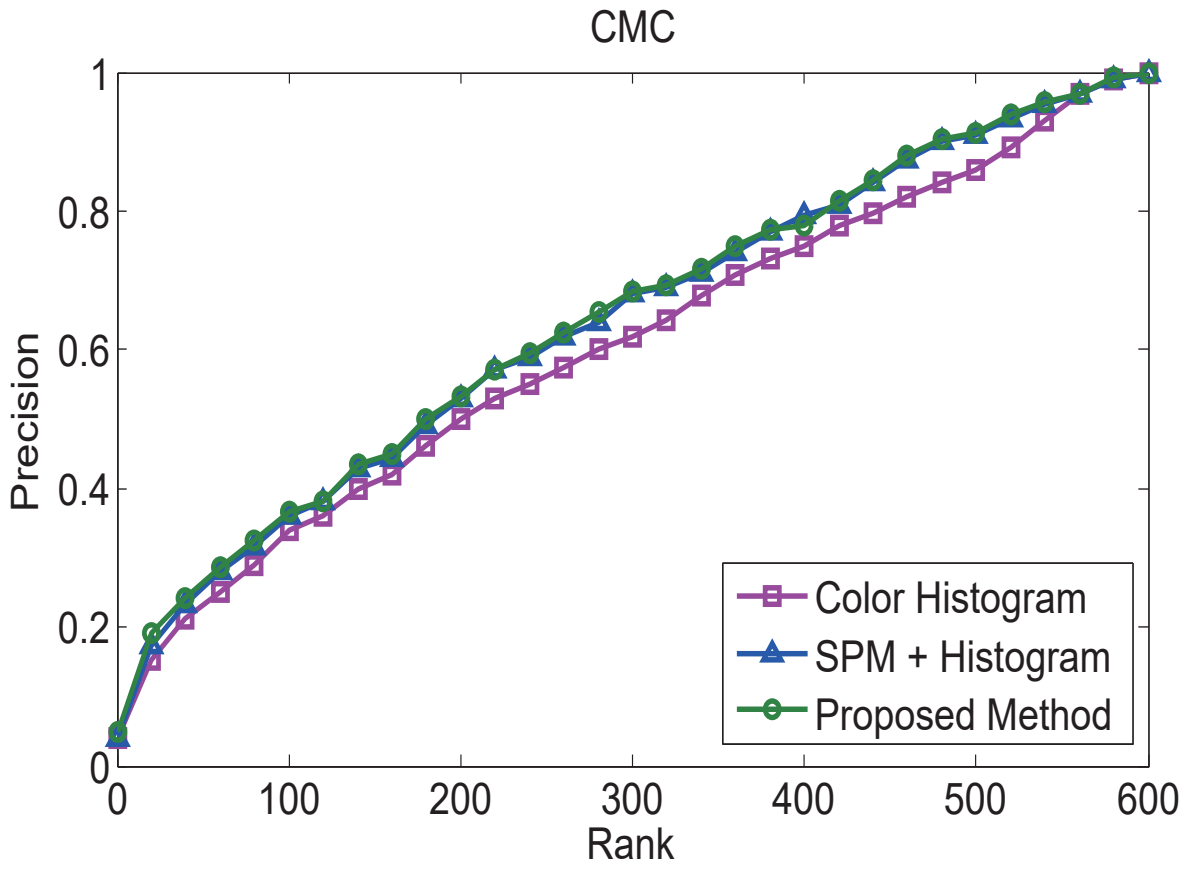

Figure 6.10: CMC curves on the VIPeR dataset for the proposed method and histogram methods in RGB space

the SDALF method needs about 4260 seconds to compare all 399424 pairs of images, while our method needs only 610 seconds to calculate the GMM similarity for comparison in 1264 images. This result suggests that in terms of computational cost our approach significantly outperforms the SDALF method.

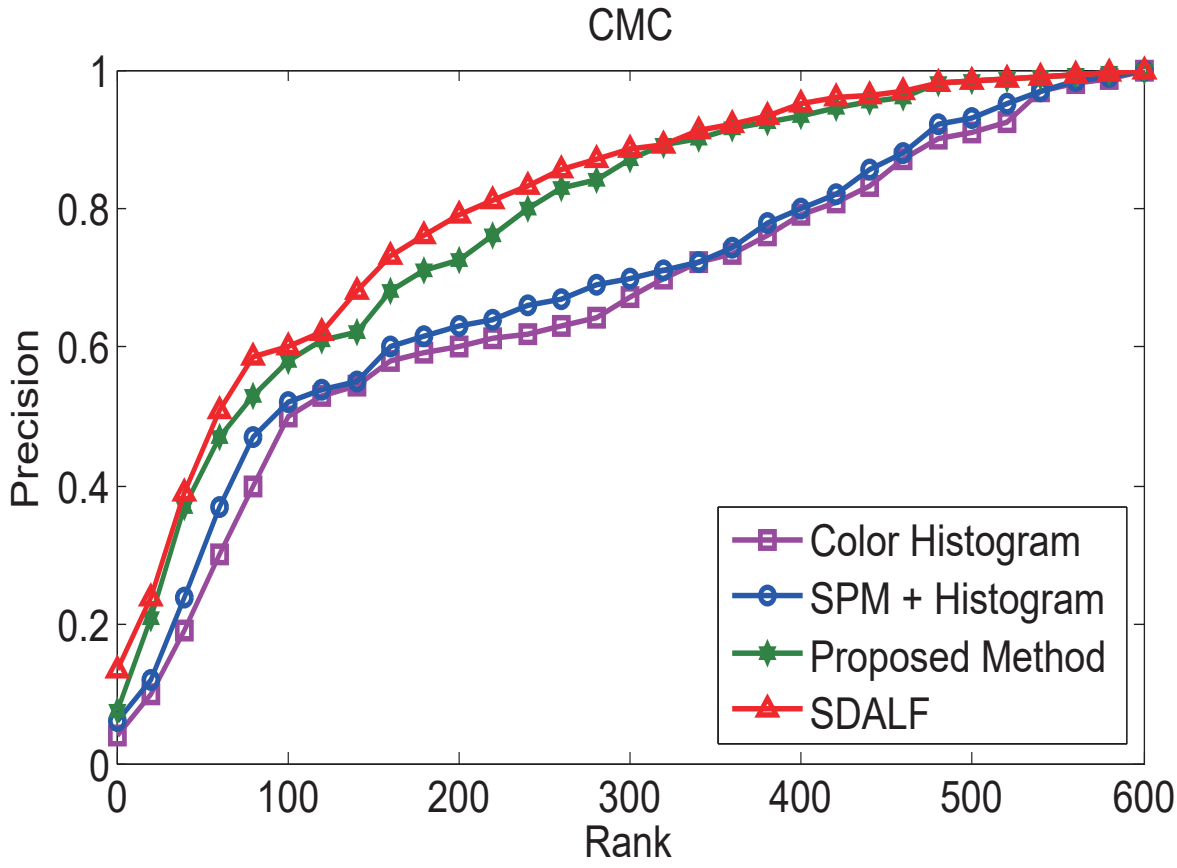

Figure 6.11: CMC curves on the VIPeR dataset for the proposed method and the other methods in HSV space

6.2.5 Experiment on SARC3D Dataset

The SARC3D dataset consists short video clips of 50 people which has been captured with a calibrated camera. We employ the SARC3D dataset to effectively evaluate different person re-identification methods. To simplify the image alignment process, we manually select four frames for each clip which correspond to the predefined positions and postures, i.e. back, front, left, and right, of these people. The selected dataset consists of 200 snapshots with four views for each person. For person re-identification, we randomly choose one of the four views for each person, calculate the similarity scores with all other images, and find the most similar images by sorting their similarities with the chosen image. The images of the same person with different positions and postures should be ranked higher than the other images.

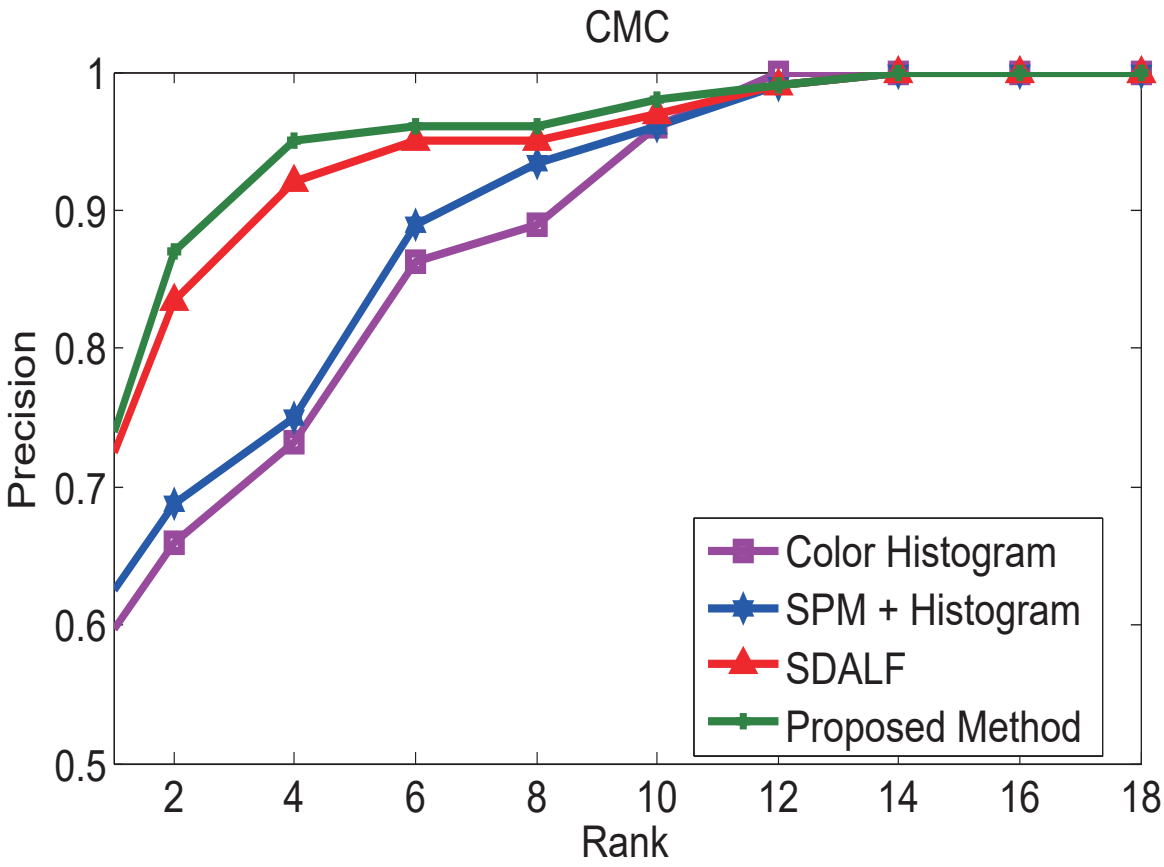

Figure 6.12: CMC curves on the VIPeR dataset for the proposed method and the other methods in RGB space

In the dataset, 6 people are not fully visible in their images and 2 people are observed with the same dressing, i.e. colors and combinations, except for the waling postures. We remove images of these people to avoid the different size of their masks form in the original images. All methods in the experiment are based on the RGB color space. Fig 6.12 shows the average CMC curves for the person re-identification under different methods. Our method significantly outperforms the SDALF method in recognition rate because the backward information in GMM matching has been filtered out given the people annotation template in the dataset. In the meantime, our method significantly outperforms the SDALF method in calculation cost, with only 30 seconds for color histogram feature extraction and image matching in 126 images, while the latter takes about 440 seconds for feature extraction and 70 more seconds for image matching.

Chapter 7

conclusion

7.1 Summary of Pedestrian Detection and Person Re-identification

In the pedestrian detection part we proposed a Pedestrian Detection Method at Multiple Resolution. Especially during pedestrian detection under the high resolution, such an algorithm can generate very significant effects. However, targets in images acquired in the real world are under diverse resolutions in most cases. Considering this, the standard DPM is subjected to great limitations. Here, a multi-resolution DPM algorithm based on the standard DPM algorithm is presented. In this way, pedestrian detection is fixed to different resolutions. For example, pedestrians under the high resolution can be detected through a deformable part model, while those under the low resolution are detected based on the rigid template.

In the person re-identification we present a system to solve the problem of pedestrian re-identification in surveillance video which was limited by low-resolution, high video noise and monitoring scope. Our proposed framework deals with several problems such as variations of illumination conditions, poses and occlusions. Since the appearance of color distribution of pedestrian do not change in the process of monitoring, our paper uses the color histogram as statistic descriptor and chose RGB, HSV, and UVW as color space. Traditional histogram method extracting the global color distribution as the feature and the object color structure information will be neglected. In order to make up for the weakness of the lack of spatial information, the SPM model was used to supplement the structure information for the color feature. We also use the GMM model to verify the

influence of the color feature in pedestrian re-identification. Through the experimental results that the UVW color space could provide an effective matching. The GMM main color model have a good performance in RGB and HSV color.

Bibliography

[1] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 983–990. IEEE, 2009.

[2] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Multiple-shot human re-identification by mean riemannian covariance grid. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, pages 179–184. IEEE, 2011.

[3] A. Bar-Hillel, D. Levi, E. Krupka, and C. Goldberg. Part-based feature synthesis for human detection. In *European Conference on Computer Vision*, pages 127–142. Springer, 2010.

[4] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by hpe signature. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 1413–1416. IEEE, 2010.

[5] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286, 2014.

[6] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool. Seeking the strongest rigid detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3666–3673, 2013.

[7] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998.

[8] G. Chen, Y. Ding, J. Xiao, and T. X. Han. Detection evolution with multi-order contextual co-occurrence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1798–1805, 2013.

[9] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *Bmvc*, volume 2, page 6, 2011.

[10] H. Cho, P. E. Rybski, A. Bar-Hillel, and W. Zhang. Real-time pedestrian detection with deformable part models. In *Intelligent Vehicles Symposium (IV), 2012 IEEE*, pages 1035–1042. IEEE, 2012.

[11] C. Conde, D. Moctezuma, I. M. De Diego, and E. Cabello. Hogg: Gabor and hog-based human detection for surveillance in non-controlled environments. *Neurocomputing*, 100:19–30, 2013.

[12] E. Corvee, F. Bremond, M. Thonnat, et al. Person re-identification using haar-based and dcd-based signature. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 1–8. IEEE, 2010.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[14] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *International journal of computer vision*, 95(1):1–12, 2011.

[15] Y. Ding and J. Xiao. Contextual boost for pedestrian detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2895–2902. IEEE, 2012.

[16] P. Dollar, 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*, 34(4):743–761, 2012.

[17] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing*, 2(2):127–151, 2011.

[18] G. J. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 300–305. IEEE, 1998.

[19] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[20] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person reidentification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE, 2010.

[21] P. Fieguth and D. Terzopoulos. Color-based tracking of heads and other mobile objects at video frame rates. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 21–27. IEEE, 1997.

[22] D. M. Gavrila and V. Philomin. Real-time object detection for" smart" vehicles. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 87–93. IEEE, 1999.

[23] 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*, pages 1–6. IEEE, 2008.

[24] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision*, pages 780–793. Springer, 2012.

[25] M. I. Khedher, M. A. El-Yacoubi, and B. Dorizzi. Probabilistic matching pair selection for surf-based person re-identification. In *Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG-Proceedings of the International Conference of the*, pages 1–6. IEEE, 2012.

[26] J. Kwon and K. M. Lee. Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1208–1215. IEEE, 2009.

[27] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2169–2178. IEEE, 2006.

[28] S. Lee, J. Lee, E. Chon, M. H. Hayes, and J. Paik. Moving object segmentation using motion orientation histogram in adaptively partitioned blocks for consumer surveillance system. In *2012 IEEE International Conference on Consumer Electronics (ICCE)*, pages 197–198. IEEE, 2012.

[29] Z. Lin and L. S. Davis. Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In *International symposium on visual computing*, pages 23–34. Springer, 2008.

[30] Z. Lin and L. S. Davis. Shape-based human detection and segmentation via hierarchical part-template matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):604–618, 2010.

[31] B. Ma, Q. Li, and H. Chang. Gaussian descriptor based on local features for person re-identification. In *Asian Conference on Computer Vision*, pages 505–518. Springer, 2014.

[32] M. P. Manns, J. G. McHutchison, S. C. Gordon, V. K. Rustgi, M. Shiffman, R. Reindollar, Z. D. Goodman, K. Koury, M.-H. Ling, J. K. Albrecht, et al. Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis c: a randomised trial. *The Lancet*, 358(9286):958–965, 2001.

[33] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004.

[34] T. Matsukawa, T. Okabe, and Y. Sato. Person re-identification via discriminative accumulation of local features. In *ICPR*, pages 3975–3980, 2014.

[35] M. J. Metternich, M. Worring, and A. W. Smeulders. Color based tracing in real-life surveillance data. In *Transactions on data hiding and multimedia security V*, pages 18–33. Springer, 2010.

[36] F. Michel, L. Ehm, G. Liu, W. Han, S. Antao, P. Chupas, P. Lee, K. Knorr, H. Eulert, J. Kim, et al. Similarities in 2-and 6-line ferrihydrite based on pair distribution function analysis of x-ray total scattering. *Chemistry of Materials*, 19(6):1489–1496, 2007.

[37] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 26–36. IEEE, 2006.

[38] F. Mokhtarian and R. Suomela. Robust image corner detection through curvature scale space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1376–1381, 1998.

[39] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. *Pattern recognition*, 36(9):1997–2006, 2003.

[40] T. Ojala and M. Pietikäinen. Unsupervised texture segmentation using feature distributions. *Pattern Recognition*, 32(3):477–486, 1999.

[41] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *European Conference on Computer Vision*, pages 28–39. Springer, 2004.

[42] D. Park, C. L. Zitnick, D. Ramanan, and P. Dollár. Exploring weak stabilization for motion feature extraction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2882–2889, 2013.

[43] G. Pass, R. Zabih, and J. Miller. Comparing images using color coherence vectors. In *Proceedings of the fourth ACM international conference on Multimedia*, pages 65–73. ACM, 1997.

[44] C. E. Rasmussen. The infinite gaussian mixture model. In *NIPS*, volume 12, pages 554–560, 1999.

[45] A. Satpathy, X. Jiang, and H.-L. Eng. Human detection using discriminative and robust local binary pattern. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 2376–2380. IEEE, 2013.

[46] A. Satpathy, X. Jiang, and H.-L. Eng. Human detection by quadratic classification on subspace of extended histogram of gradients. *IEEE Transactions on Image Processing*, 23(1):287–297, 2014.

[47] 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*, pages 322–329. IEEE, 2009.

[48] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3633, 2013.

[49] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *BMVC*, volume 2, page 3, 2006.

[50] M. A. Stricker and M. Orengo. Similarity of color images. In *IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology*, pages 381–392. International Society for Optics and Photonics, 1995.

[51] M. J. Swain and D. H. Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991.

[52] D. Thanh Nguyen, M.-K. Tran, and S.-K. Yeung. An mrf-poselets model for detecting highly articulated humans. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1967–1975, 2015.

[53] R. Urtasun, D. J. Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 238–245. IEEE, 2006.

[54] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, 2005.

[55] 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*, pages 1030–1037. IEEE, 2010.

[56] F. Wang, X. Wang, D. Zhang, C. Zhang, and T. Li. marginface: A novel face recognition method by average neighborhood margin maximization. *Pattern Recognition*, 42(11):2863–2875, 2009.

[57] 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*, pages 32–39. IEEE, 2009.

[58] 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*, volume 1, pages 90–97. IEEE, 2005.

[59] H. Wu, X. Liu, and G. Doretto. Face alignment via boosted ranking model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[60] P. Xu, F. Davoine, and T. Denœux. Evidential combination of pedestrian detectors. In *British Machine Vision Conference*, 2014.

[61] Q. Ye, Z. Han, J. Jiao, and J. Liu. Human detection in images via piecewise linear support vector machines. *IEEE transactions on image processing*, 22(2):778–789, 2013.

[62] X. Zeng, W. Ouyang, M. Wang, and X. Wang. Deep learning of scene-specific classifier for pedestrian detection. In *European Conference on Computer Vision*, pages 472–487. Springer, 2014.

[63] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013.

[64] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1491–1498. IEEE, 2006.

[65] P. A. Zuk, M. Zhu, P. Ashjian, D. A. De Ugarte, J. I. Huang, H. Mizuno, Z. C. Alfonso, J. K. Fraser, P. Benhaim, and M. H. Hedrick. Human adipose tissue is

a source of multipotent stem cells. *Molecular biology of the cell*, 13(12):4279–4295, 2002.