Robust Video Object Tracking in Distributed Camera Networks

Younggun Lee

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Jenq-Neng Hwang, Chair

Linda Shapiro

Zicheng Liu

Program Authorized to Offer Degree:
Department of Electrical Engineering

University of Washington

Abstract

Robust Video Object Tracking
in Distributed Camera Networks

Younggun Lee

Chair of the Supervisory Committee:
Professor Jenq-Neng Hwang
Department of Electrical Engineering

We propose a robust video object tracking system in distributed camera networks. The main problem associated with wide-area surveillance is people to be tracked may exhibit dramatic changes on account of varied illuminations, viewing angles, poses and camera responses, under different cameras. We intend to construct a robust human tracking system across multiple cameras based on fully unsupervised online learning so that the camera link models among them can be learned online, and the tracked targets in every single camera can be accurately re-identified with both appearance cue and context information. We present three main parts of our research: an ensemble of invariant appearance descriptors, inter-camera tracking based on fully unsupervised online learning, and multiple-camera human tracking across non-overlapping cameras.

As for effective appearance descriptors, we present an appearance-based re-id framework, which uses an ensemble of invariant features to achieve robustness against partial occlusion, camera color response variation, and pose and viewpoint changes, etc. The proposed method not only solves the problems resulted from the changing human pose and viewpoint, with some tolerance of illumination changes but also can skip the laborious calibration effort and restriction.

We take an advantage of effective invariant features proposed above in the tracking.

We present an inter-camera tracking method based on online learning, which systematically builds camera link model without any human intervention. The aim of inter-camera tracking is to assign unique IDs when people move across different cameras. Facilitated by the proposed two-phase feature extractor, which consists of two way Gaussian mixture model fitting and couple features in phase I, followed by the holistic color, regional color/texture features in phase II, the proposed method can effectively and robustly identify the same person across cameras.

To build the complete tracking system, we propose a robust multiple-camera tracking system based on a two-step framework, the single-camera tracking algorithm is firstly performed in each camera to create trajectories of multi-targets, and then the inter-camera tracking algorithm is carried out to associate the tracks belonging to the same identity. Since inter-camera tracking algorithms derive the appearance and motion features by using single-camera tracking results, *i.e.*, detected/tracked object and segmentation mask, inter-camera tracking performance highly depends on single-camera tracking performance. For single-camera tracking, we present multi-object tracking within a single camera that can adaptively refine the segmentation results based on multi-kernel feedback from preliminary tracking to handle the problems of object merging and shadowing. Besides, detection in local object region is incorporated to address initial occlusion when people appear in groups.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DEDICATION

to my dear wife, Jiyoung

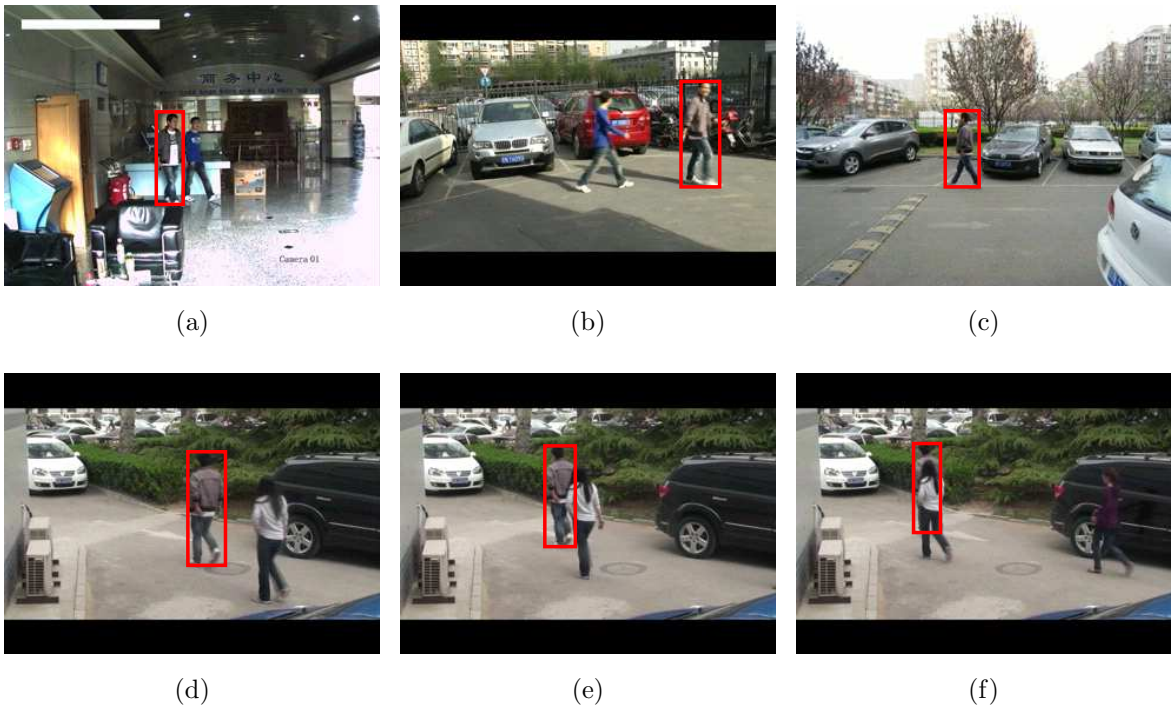
Chapter 1

INTRODUCTION

1.1 Introduction

For security and safety purpose, the demands for surveillance cameras rapidly increase in the world in recent years. Because of limitation of Field Of Views (FOVs) and cost-efficiency, in most cases, multiple cameras are installed in a wide area with no overlap. One of the most important things in intelligent surveillance and monitoring system is automated object tracking for a huge amount of recorded and live streaming video data. The goal of automated object tracking in the camera network is to keep the unique identity of each object within every single camera and across multiple cameras without human intervention. In other words, tasks involved in Multiple Camera Tracking (MCT) include multi-object tracking in every single camera and delivering detected identities to disjoint cameras.

Many MCT approaches exploit a two-step framework, Single-Camera Tracking (SCT) is firstly performed in each camera to create trajectories of multiple targets, and then Inter-Camera Tracking (ICT) is carried out to associate the tracks belonging to the same identity. There are several difficulties in ICT. First, people may exhibit dramatic changes on account of varied illuminations, viewing angles, poses and camera responses, under different cameras. Figures 1.1(a)-(c) show examples of ICT. One approach to solve this problem is person re-identification (re-id), which is to identify the same person in more than two cameras with only some human image-pairs. Many researchers on person re-id focus on extracting discriminative visual features to characterize the appearance of an individual human without taking advantage of context information, *e.g.*, group behavior between a pair of accompanying persons. Second, to achieve a good performance in ICT, a robust SCT, which detects accurate object positions and keeps the same identity on each object, needs to be guaran-



(a) (b) (c)

(d) (e) (f)

Figure 1.1: Examples of MCT in NLPR_MCT Dataset4 (the same identity is marked with red bounding box).

teed. Since ICT algorithms derive the appearance and context features by using SCT results, *i.e.*, detected/tracked objects and segmentation masks, ICT performance highly depends on SCT.

The research in SCT in recent years has shown significant progress and enhances ICT performance as well. Figures 1.1(d)-(f) show examples of SCT, where (e) shows partial occlusion and (f) shows severe occlusion. The task of single-target tracking has been well addressed by following the cues of appearance or silhouette of the selected target. However, for multi-target tracking in a video, the situation is more complex due to the data association problem and the interactions among objects. Furthermore, because of variations in a number of targets, we need to employ effective schemes to initialize the object locations in every frame, which are usually derived from appearance-based object detection [9, 10] or

background subtraction [11]. The former category is called tracking by detection, and the latter tracking by segmentation. In [12], most of the state-of-the-art methods are in the tracking-by-detection school, in which they exploit the continuity in space and time, but the information of object appearance is seldom considered to facilitate tracking.

In this paper, we propose a robust MCT system based on a two-step framework. For SCT, we present multi-object tracking within a single camera that can adaptively refine the segmentation results based on multi-kernel feedback from preliminary tracking to handle the problems of object merging and shadowing. Meanwhile, ICT can benefit from the optimal segmented foreground blobs of each object as well. Besides, detection in local object region is incorporated to address initial occlusion when people appear in groups. Additionally, we follow our previous work [13] to rely on Constrained Multi-Kernel (CMK) tracking to deal with occlusion. The presented SCT method has been partially described in [14]. In addition to giving a more detailed explanation of our proposed SCT method, we introduce how to combine the algorithm with advanced change detection to improve segmentation performance, add the detection module for enhancing robustness against initial occlusion, and conduct new experiments on the benchmark dataset of MCT.

For ICT, we present a fully unsupervised online learning approach which integrates discriminative visual features and context feature efficiently, and systematically builds camera link model without any human intervention. With a two-phase feature extractor, which consists of TWO-Way Gaussian Mixture Model Fitting (2WGMMF) and couple features in phase I, followed by the holistic color, regional color/texture features in phase II, the proposed method effectively and robustly identifies the same person across cameras. To be more specific, illumination variation is dealt with by a fully unsupervised color transfer method [15], and changes of poses and camera viewpoints are overcome with pose-invariant features, 2WGMMF and regional color/texture features. The context information is represented as couple feature, which describes a pair of the person traveling together through the scene. Those features are integrated with fusion feature weights belonging to camera link model.

1.2 Contributions

The contributions of this dissertation are summarized as follows.

- We propose a way of describing distinct visual characteristics and utilizing the dominant color modes from pedestrian images based on human perception. This descriptor, called *Enhanced and Integrated Features* (EIFs), is the core of our approaches. The proposed features consist of the holistic and regional pose-invariant features.

- The invariant holistic features are extracted by using a publicly available pre-trained Deep Convolutional Neural Network (DCNN), which is originally used in generic object classification. The DCNN gives rich and discriminative features including color, texture, shape and other visual cues to describe a human. We investigate the pre-trained model which is suitable to extract features for person re-id.

- The regional features are extracted from partitioned body parts by exploiting Gaussian Mixture Models (GMMs) on the color histograms. Each Gaussian distribution of a GMM is exploited to represent a dominant color mode of a human subject. The proposed GMM based feature matching method is applied to a two-way process, *i.e.*, probe-to-gallery and gallery-to-probe, to compensate different poses of the same identity. A distance between two GMMs is computed by our proposed feature distance metric.

- In the direct-method (without any specific dataset training) category, our method have improved recognition rate over the state-of-the-art performance on the challenging public benchmark datasets, Viewpoint Invariant Pedestrian Recognition (VIPeR) [16] and 3D People Surveillance (3DPeS) [17] datasets.

- We present an ICT method based on online learning, which systematically builds camera link model without any human intervention.

- Three pose-invariant color features are effectively integrated with proposed context-based couple feature.

- We propose a robust two-step MCT method based on SCT by segmentation and local object detection and a two-phase online feature learning ICT framework.

- To preserve precise foreground segmentation, a combination of the advanced change detection algorithm and multi-kernel feedback is utilized.

- We validated superior performance on benchmark NLPR_MCT dataset.

1.3 Dissertation Roadmap

The rest of this dissertation is organized as follows.

Chapter 2: we review the related works on previously proposed tracking systems including person re-identification, single-camera tracking, and inter-camera tracking.

Chapter 3: we present an appearance-based re-id framework, which uses an ensemble of invariant features to achieve robustness against partial occlusion, camera color response variation, and pose and viewpoint changes, etc. The proposed method not only solves the problems resulted from the changing human pose and viewpoint, with some tolerance of illumination changes but also can skip the laborious calibration effort and restriction.

Chapter 4: we present an ICT method based on online learning, which systematically builds camera link model without any human intervention. Facilitated by the proposed two-phase feature extractor, which consists of TWO Way Gaussian Mixture Model Fitting (2WGMMF) and couple features in phase I, followed by the holistic color, regional color/texture features in phase II, the proposed method can effectively and robustly identify the same person across cameras. ICT can benefit from the optimal segmented foreground blobs of each object as well.

Chapter 5: we propose a robust MCT system based on a two-step framework. For SCT,

we present multi-object tracking within a single camera that can adaptively refine the segmentation results based on multi-kernel feedback from preliminary tracking to handle the problems of object merging and shadowing. Besides, detection in local object region is incorporated to address initial occlusion when people appear in groups. In addition to giving a more detailed explanation of our proposed SCT method, we introduce how to combine the algorithm with advanced change detection to improve segmentation performance, add the detection module for enhancing robustness against initial occlusion, and conduct new experiments on the benchmark dataset of MCT.

Chapter 6: conclusions of this dissertation are given by summarizing the main contributions and discussing some extensions to this work that lead to potential research topics in the future.

Chapter 2

BACKGROUNDS AND RELATED WORKS

2.1 An Ensemble of Invariant Features for Person Re-Identification

In this section, we review the state-of-the-art person re-id approaches, which can be roughly divided into two main categories, the learning-based and the direct methods, whose main difference resides on with/without the requirement of the domain data training. The learning-based method uses the specific domain data to train the specific features, classifiers, or metrics. For each camera network environment, the learning-based method needs to collect many images associated with the same identity in different cameras to optimize the re-id performance. Instead of using any target-domain data, the direct method discovers the general and reliable visual descriptors to represent a person. The desired descriptor has the property that the intra-class (same identity) pairs have more similar descriptors than the inter-class (different identity) pairs. The direct method aims to propose the invariant feature extraction and matching metric which can handle the appearance variations caused by the different person pose, camera viewpoint, and environmental illumination change. The occlusion problem is another issue needed to be concerned, however, the existing method does not show the significant solution to this problem. The common solution is to segment one person image into many region parts and measure the similarity by accumulating each part's score to get the final decision.

The first category of approaches [18], [19] select the domain-driven features or obtain specific models trained in advance by a dataset separately to achieve high re-id accuracies if there are enough training data available and the testing environments are similar to the training environments. In [18], the generic descriptive statistical model and the discriminatively learned feature model are combined to attain better results. The generic descriptive statisti-

cal model can generate a rank list initially and the positive and negative training data can be chosen according to the rank list without any manual process for training the discriminating learned feature model. Gray *et al.* [2] propose a viewpoint invariant model by integrating spatial and color information based on the boosting scheme. Prosser *et al.* [20] formulate the re-id problem as a ranking problem, where the Ensemble-Rank SVM is used to learn a better subspace that the potential true-match is given the highest ranking. Wang *et al.* [19] extract the 3D histogram of oriented gradient (HOG) features from the discriminative video segments where the more reliable space-time features are derived to learn a video ranking function for re-id. Zhao *et al.* [21, 22] propose an unsupervised saliency learning framework for human matching. Some of the researchers focus on metric learning to enhance the accuracies using both the similar pairs and dissimilar pairs of training data, *e.g.*, LMNN (Large Margin Nearest Neighbor) [23], KISS metric [24], and PCCA [25]. In contrast to previously mentioned work that exploits hand-crafted features, some novel methods [26, 27] apply the deep convolutional network for learning the features and the similarity metric for person re-id simultaneously. In the existing person re-id dataset, the image number is too less to train a deep convolutional network. To solve this problem, Li *et al.* [26] built the largest person re-id benchmark, CUHK03, which consists of 13,164 images of 1,360 pedestrians. Based on these datasets, the deep convolution network structures with the paired images as input can thus be determined. In the supervised learning framework, each input pair has a binary label representing whether it is the same or different pair. The re-id problem is transformed to a two-class classification problem and the softmax loss minimization is performed in the top layer. When minimizing the loss, the stochastic gradient descent is adopted to update the weights and decrease the loss. Taking advantage of the end-to-end framework of the deep convolutional network, the discriminative and effective features and corresponding decision metric can be learned in the former layers and the top layer at the same time.

Although most learning-based methods achieve better performance than direct methods, it is worthy of noting that learning-based methods are strongly dependent on the training datasets and lacking generalization capabilities. Moreover, it is hard to obtain the sam-

ples with the ground-truth label in the real scenarios where the conditions are dynamically changing, so they are not quite appropriate for practical use in the real-world surveillance applications.

The second category of approaches, the direct methods [3, 4, 7, 28], directly run on each person independently without the training process on a specific dataset. These works mainly focus on proposing the novel and discriminative features. More specifically, an illumination-invariant color feature is proposed by Kviatkovsky *et al.* [4], where the signature is formed by the log-polar quantization in the log-chromatic color space. The SDALF by Farenzena *et al.* [3] divides the human body into head/torso/legs parts and extracts color features based on the horizontal and vertical asymmetries of the human silhouettes. Cheng *et al.* [28] use a pictorial structure to localize the parts, then extract the part descriptor features to match objects. However, the performance of these approaches is quite unreliable, often subject to serious human pose and viewpoint issues. Recently, Baltieri *et al.* [7] propose a framework which exploits a simplified non-articulated 3D model to spatially map 2D appearance descriptors (color and gradient histograms) into the vertices of a regularly sampled 3D body model. This model effectively mitigates the problem of occlusion, partial views and poses changes. However, the image-to-model mapping needs the perspective projection matrix (intrinsic parameters) and extrinsic calibration matrix to estimate the rotation and translation between the world reference coordinate and the model one, resulting in critical inconvenience to the practical use.

2.2 Inter-Camera Tracking Based on Fully Unsupervised Online Learning

Human tracking across multiple cameras has been one of the most active research topics in computer vision [29] and many approaches have been proposed to address this problem. Most of these approaches utilize appearance cues, which include color [30–32], texture [33, 34], and shape [35, 36] of targets, to describe human and match correct correspondence with spatio-temporal reasoning. However, the color appearance is easily influenced by illumination and viewpoint changes across cameras. To solve the problem, Brightness Transfer Functions

(BTFs) [37, 38], which map color information between a pair of cameras, and appearance relationship [39–41] are modeled from training data. For the spatio-temporal feature, transition time distribution, which is the probability of an object entering a camera view with a certain travel time given the location and velocity of its exit from the other camera view, is estimated [32, 42, 43]. However, it requires training data whose correspondences are prelabeled. Since methods relying on human operators are ineffective and lacking in scalability, these supervised learning approaches are less feasible in practice.

For this reason, recently unsupervised [33], graph modeling [44, 45], and online learning [36, 46] methods are exploited. More specifically, Chu *et al.* [33] estimate camera link model as an optimization problem to build the relationship between directly connected camera pairs based on an unsupervised learning scheme. However, they need separate training data for training stage, and transition time distribution included in camera link model is less reliable in case of longer transition time, because the variance of traveling time between connected cameras increases. Chen *et al.* [44] treat multi-camera object tracking as a global tracklet association, which is formulated as a global Maximum A Posteriori (MAP) problem with Piecewise Major Color Spectrum Histogram Representation (PMCSHR) and minimum uncertainty gap measurements. Tracking performance is enhanced with an improved similarity metric, which equalizes the inter-camera similarities in [45]. However, the disappearing points need to be manually selected in each enter/exit area. Chen *et al.* [46] employ an adaptive learning method, which uses the spatio-temporal information and Markov chain Monte Carlo sampling, to learn both spatio-temporal and appearance relationships among cameras. Kuo *et al.* [36] collect the online training samples by observing the spatio-temporal constraints in a time sliding window and use the Multiple Instance Learning (MIL) algorithm to learn a discriminative appearance model online. However, their method is limited to utilize low-level appearance features.

To further improve the tracking performance, group information is also exploited as complementary features in recent approaches [35, 47, 48]. Cai *et al.* [35] propose a relative appearance context model of groups to mitigate ambiguities in individual appearance

matching. However, their relaxed definition of the group-named neighboring set has no social connection, therefore their assumption that the same set of people will reappear in the neighboring camera is not always valid. Wei *et al*. [47] propose a subject-centric group feature to reduce the re-id ambiguity. However, the group feature is limited to improve the accuracy of individual re-id when there is an outlier, who is seen in just one camera without being seen in the other, and sensitive to noise of persons positions and velocities. Chen *et al*. [48] integrate social grouping behavior of an elementary group [49] and an online learned target-specific appearance model by using AdaBoost. The tracking problem is formulated using an online learned Conditional Random Field (CRF) model that minimizes a global energy cost. However, the effectiveness of grouping information is not guaranteed when object detection is not sufficiently robust.

2.3 Online-Learning-Based Multiple-Camera Human Tracking Across Non-Overlapping Cameras

Many MCT approaches exploit a two-step framework, Single-Camera Tracking (SCT) is first performed in each camera to create trajectories of multiple targets, and then Inter-Camera Tracking (ICT) is carried out to associate the tracks belonging to the same identity. Among many SCT techniques for tracking a single target, kernel-based object tracking such as mean shift tracker [50] that searches for similar candidate model around local neighboring regions, has gained lots of popularity, because of its fast convergence and low computation. To improve kernel-based tracking, Chu *et al*. [13] propose to handle occlusion based on adaptive multiple kernels with constraints on their spatial relation, *i.e*., CMK tracking, and the accuracy is comparable to the state-of-the-art trackers. In [51], they embed CMK tracking into a Kalman filtering tracking system to further increase computational efficiency. To extend the application to multi-object tracking, it is necessary to find a way to automatically define the locations of targets. Most of the top-ranked methods in SCT depend on object detection for target initialization [12], but none of them considers to combine the information from segmentation to jointly improve performance.

Robust object segmentation is essential to feature extraction in ICT and supporting intra-camera tracking by segmentation. Many recent works in this field emphasize the concept of adaptation. In [52], a regularized background adaptation for automatically controlling the learning rate of Gaussian Mixture Model (GMM) is presented. Hoffmann *et al*. propose the Pixel-Based Adaptive Segmenter (PBAS) in [53], which utilizes two dynamic controllers to adaptively adjust the decision threshold and learning rate. Self-Balanced SENsitivity SEgmenter (SuBSENSE), as introduced by St-Charles *et al*. [54], further improves PBAS by adding adaptation to local sensitivity and update rate, which allows the technique to rank among the top of the benchmark dataset of change detection, CDnet [55]. Nevertheless, none of the algorithms are designed specifically for supporting tracking, as they can easily fail when the target(s) enter into an area with similar background color (*i.e.*, the problem of object merging), or encounter strong shadowing effect when the subsequent SCT will be negatively influenced as well.

Chapter 3

AN ENSEMBLE OF INVARIANT FEATURES FOR PERSON RE-IDENTIFICATION

3.1 Overview

This chapter describes our proposed Ensemble of Invariant Features (EIFs), which can properly handle the variations of color difference and human poses/viewpoints for matching pedestrian images observed in different cameras with a non-overlapping field of views. Our proposed method is a direct re-identification method, which requires no prior domain learning based on pre-labeled corresponding training data. The novel features consist of the holistic and region-based features. The holistic features are extracted by using a publicly available pre-trained Deep Convolutional Neural Network (DCNN) used in generic object classification. In contrast, the region-based features are extracted based on our proposed Two-Way Gaussian Mixture Model Fitting (2WGMMF), which overcomes the self-occlusion and pose variations. To make a better generalization during recognizing identities without additional learning, the ensemble scheme aggregates all the feature distances using the similarity normalization. The proposed framework achieves robustness against partial occlusion, pose and viewpoint changes. Moreover, the evaluation results show that our method outperforms the state-of-the-art direct re-identification methods on the challenging benchmark VIPeR and 3DPeS datasets.

3.2 Holistic Pose-Invariant Features

The holistic features are desired to extract the meaningful information to describe the whole person. The information should include color, texture, shape and other reliable visual cues. Recently, the DCNNs show its powerful capability to extract the rich and discriminative

features from images [1,56]. Based on the end-to-end structure of a DCNN, the local low-level features (*e.g.*, specific orientation gradient on a person's shoulder) and the global high-level features (*e.g.*, the heads and legs of pedestrians) can be trained and extracted hierarchically from the low-to-high layers [57] during the loss minimization.

Indeed, the DCNN is a supervised learning mechanism which typically requires lots of labeled training data to train model, *i.e.*, learning the network weights parameters to minimize the objective loss. It is difficult and time-consuming to collect thousands of training data for the specific task or domain. Recently, domain transferring method has shown a significant boost performance for object classification [56, 58, 59]. When the labeled training data is scarce and insufficient to train a complicated DCNNs, the domain-specific fine-tuning can transfer the discriminative features learned from a supervised pre-training DCNNs model to the target dataset. Specifically, the transfer procedure can be achieved by the following two steps, pre-training on a large dataset, *e.g.*, ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [60], and fine-tuning on a smaller target dataset. Although this approach shows a significant performance improvement, it still requires hundreds or thousands of target-domain data for performing the fine-tuning. Therefore, it is still very infeasible and unpractical to employ for the person re-id in a real case.

Based on the most common scenario that no image can be obtained for training or fine-tuning the model, we want to discover a potential solution, using the existing pre-trained DCNN models without domain fine-tuning. Since a large-scale DCNN is learned by a great amount of image data, so the extracted features can inherently possess many different types of visual information. The rich features can thus enhance the invariance of different poses, lightings, and viewpoints. Therefore, many researches [56,61,62] for object detection, object segmentation, object tracking, etc., have accommodated the DCNN features to achieve better results. Therefore, we evaluate the different features extracted from three popular large-scale DCNNs, Krizhevsky *et al.* AlexNet [1], Chatfield *et al.* VGG [63] and Szegedy *et al.* GoogLeNet [64]. All these networks are trained by the data of ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [60]. The ILSVRC 2014 classification challenge involves the

task of classifying an image into one of 1,000 leaf-node categories in the ImageNet hierarchy. There are about 1.2 million images for training, 50,000 for validation and 100,000 images for testing. Each image has a label associated with its ground truth category belonged to one of 1,000 categories.

The classifier network is achieved by following the top feature extraction layer with an n-way, *i.e.*, 1,000, softmax layer. Given a training example x that produces n dimensional feature map X_{soft} at the softmax layer, for each possible label $i = 1, ..., n$, this layer computes probability distribution p_i over the n classes by

$$p_i = \frac{\exp(X_{soft}(i))}{\sum_{k=1}^{n} \exp(X_{soft}(k))}. \tag{3.1}$$

When training the model, it minimizes the cross-entropy loss, which we call the classification loss and is denoted as

$$loss_{class}(X_{soft}, y, \theta) = -\sum_{i=1}^{n} y_i \log p_i, \tag{3.2}$$

where $y_i = 1$ if x has label i and 0 otherwise, and θ denotes the weight parameters of softmax layer. Note that this loss does not depend on just θ, because the computation of the feature map X_{soft} involves the weights of the early convolutional and fully-connected layers.

To correctly classify all the classes simultaneously, the early layers will be updated to search the most discriminative and category-related features, *i.e.*, features with large inter-category variations. Generally, the higher layers can provide more global information with better discriminative power as holistic features [56, 57]. Therefore, we extract the activation of the top layer with/without employing the Rectified Linear Unit (ReLU), as visual representations, which is 4,096 dimensions features from FC7 (fully-connected 7^{th} layer) and FC7+ReLU of both AlexNet [1] and Chatfield *et al.* VGG [63], and 1,024 dimensions of Pool5 which represents the Average-Pooling after inception 5 in GoogLeNet [64].

In order to evaluate the capability of representations extracted from different model or setting, we apply a specific distance measurement to compute the distances $d_{\mathrm{DCNN}_k}(I_p(i), I_q(j))$, where k denotes the separate features extracted from the following DCNNs, {Alex FC7, Alex

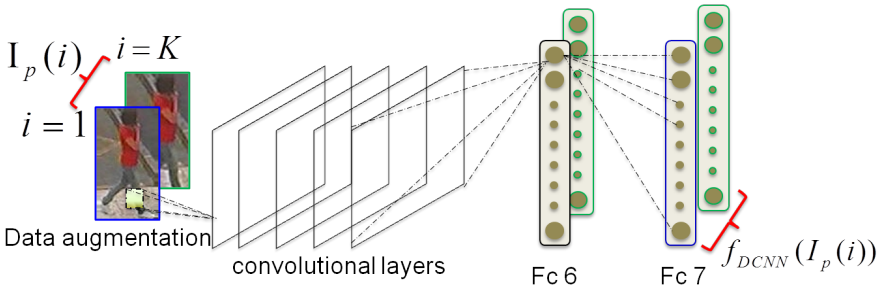


Figure 3.1: An example to show feature from the AlexNet [1]. Data augmentation generates ten sampled images with various geometric distortions before applying the DCNN. The holistic features are extracted the high level representation from the top layers of the pre-trained DCNNs.

FC7+ReLU, VGG FC7, VGG FC7+ReLU, and Pool5 GoogLeNet}. Though the DCNN has pooling layers to eliminate the problem raised by person misalignment in the image, the performance can still easily be affected in the fine-grained recognition.

We thus propose to solve the misalignment problem by additionally generating ten sampled images with various geometric distortions before applying the DCNN. Given two images, the probe image I_p and the corresponding gallery image I_q, the additionally generated sampled images for both probe and gallery images are denoted as $I_p(i)$ and $I_q(j)$, where $1 \leq \{i, j\} \leq 10$, and the extracted DCNN feature vector for an image I is denoted as $f_{\text{DCNN}_k}(I)$. The holistic feature distance $d_{\text{DCNN}_k}(I_p, I_q)$ can be obtained by the minimum of the distance d_{metric} among all the sampled pairs. This data augmentation procedure is able to improve about 40% of accuracy in general.

$$d_{\text{DCNN}_k}(I_p, I_q) = d_{metric}\left(f_{\text{DCNN}_k}\left(I_p(\hat{i})\right), f_{\text{DCNN}_k}\left(I_q(\hat{j})\right)\right), \tag{3.3}$$

where $\hat{i}, \hat{j} = \arg\min_{i,j} d_{metric}\left(f_{\text{DCNN}_k}\left(I_p(i)\right), f_{\text{DCNN}_k}\left(I_q(j)\right)\right)$ and the *metric* can be applied by L_1, L_2, or Chi-squared distance metric. Fig. 3.1 shows an example of data augmentation procedure.

3.3 Regional Invariant Features

We propose regional invariant features based on color appearance for person re-id. Most persons wear two separate clothes for upper and lower bodies, *e.g.*, a shirt and a pair of pants, which have several distinctive dominant colors. To extract discriminative information, we derive features after dividing a human body into three parts, head, torso and legs. Even if the person wears a single piece dress or a long coat, this division can still be applicable.

3.3.1 Body Partition

We assume that bounding boxes and silhouettes of the person in our experiments are acquired and normalized to a fixed template size. Both 3DPeS and VIPeR datasets, which are used in evaluating our proposed algorithm, furnish the pedestrian images and the corresponding paired silhouette masks. Background subtraction is exploited to extract the foreground pixels. In general, a person re-id system, which utilizes color signature, subdivides the human body into salient parts in order to minimize the effects of mixing colors from different clothing articles [3,4]. Since a pedestrian is commonly acquired at very low resolution, it is reasonable to notice that the most distinguishable body parts are three: head, torso and legs [3]. Two boundary lines are taken on both head-torso and torso-legs, respectively. Fig. 3.2 shows an example of the body partition. From the top to the bottom of the rectangular foreground bounding box, we calculate the histogram distance line-by-line between two stripe regions, *i.e.*, the blue stripe and green stripe regions. Each stripe is of the height δ and the width W. Intuitively, we expect that color similarity between two different body parts to be low. Therefore, a boundary line is located at height T_i, which is computed by solving the following problem, for both head-torso and torso-legs regions, respectively:

$$\max_{T_i \in \{A,B\}} d_{Chisq}(\mathbf{h}_{[T_i, T_i+\delta]}, \mathbf{h}_{[T_i-\delta, T_i]}) \tag{3.4}$$

where $d_{Chisq}(\cdot)$ denotes a function that computes Chi-squared distance and $\mathbf{h}_{[a,b]}$ denotes the color histogram derived from the stripes of a to b. Moreover, the boundary line is assumed

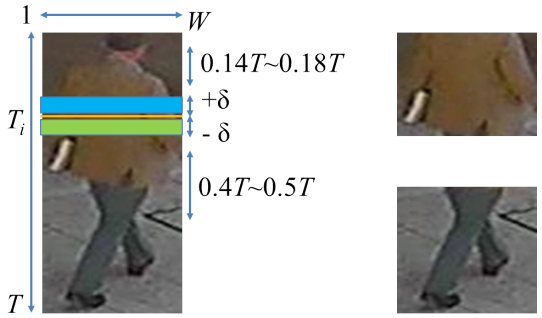


Figure 3.2: A sample body partition using Eq. (3.4), where left side image is an input and right side images are partitioned parts.

to be located within $\{A, B\}$, *e.g.*, $\{0.14T, 0.18T\}$ for head-torso and $\{0.4T, 0.5T\}$ for torso-legs. After partitioning the whole body into three parts, we extract features from the torso and the legs parts and discard the information of the head/face region because standard biometric algorithms usually fail at low-resolution [3]. 8-bin HSV histogram is employed and the height δ value is empirically set to 5 pixels when the entire segmented foreground are normalized with the height of $T = 128$ pixels.

3.3.2 2WGMMF Features

The person re-id problem can be regarded as a task of discriminative feature design to decrease the distance between two snapshots of the same identity. The most difficult challenge is to handle the large variations of poses and viewpoints in feature extraction. Since a human body is a 3D object in reality, some body parts can be occluded by other parts of the same body and the unseen parts in a probe image can be visible in a gallery image when their poses or camera viewpoints are changed. As shown in Fig. 3.3, (a) and (b) are the same person, however, color appearances of the upper body are not similar with respect to the existing color representation due to the viewpoint is changed. The existing color-based signatures such as color histogram, Earth Movers Distance (EMD), and GMM method [65] are not use-

ful to deal pose variations across disjoint camera views. On the other hand, human eyes can recognize them as the same identity based on small distinctive regions, *e.g.*, the same jacket. In this section, we propose a new feature utilizing these human perception characteristics to solve this problem. Human identifies the same person of different poses through matching dominant color compositions, which sometimes occupy only a small portion of body [21]. More specifically, a pose-invariant color feature is proposed to describe the dominant color information of each identity by using GMMs. To calculate a distance between two GMMs, we do not use conventional distance metric, such as Kullback-Leibler (KL) divergence [66], but propose a new fitting model.

2WGMMF with concatenated histogram

In Fig. 3.3, there are three pairs of unmasked and masked images. Masked images on the right side of each pair are the results of Section 3.3.1, *i.e.*, torso and legs part, (a) and (b) are the same person with ID 57 on the VIPeR dataset, captured in different views, and (c) is another person with ID 554. Although (a) and (b) are the same person, and (c) is a different person, torso parts of (a) and (c) look more similar with respect to color composition, since the torso part mainly shows a shirt and a jacket in (a), but only jacket is visible in (b). Our goal is to match (a) and (b) as the same identity with a color appearance model. In Table 3.1, we show the performance of several approaches on Fig. 3.3. The Chi-squared distance of 32-bin RGB histogram is 1.29 between (a) and (b), and 1.07 between (a) and (c). The EMD distance also shows that inter-personal variation, (a)-to-(c), is smaller than intra-personal variation, (a)-to-(b). Intuitively, to solve this problem with color features, we need to extract main color modes from each image and utilize them to identify a correct pair of images. In case of Fig. 3.3(a), the main color mode of the jacket region should be exploited to match with (b) and the shirt region's color mode needs to be ignored. For extracting key information, [21, 22] tried to find salient patches and [28] proposed to adopt Pictorial Structures (PS). However, the saliency matching method needs to build dense correspondence in advance with additional training datasets, which have to be a part of the target dataset.

(a) Probe (ID 57) (b) Gallery (ID 57) (c) Gallery (ID 554)

Figure 3.3: Examples of the VIPeR dataset.

In PS, decomposing a body into 6 parts, *i.e.*, chest, head, thighs, and legs, is not easy because pedestrian image, as captured by a surveillance camera, is normally low resolution and too small to detect each body part exactly. Even though body parts are partitioned perfectly, we should still deal with mixed color components, *i.e.*, a shirt region merged with a jacket in the torso part of Fig. 3.3(a). Naturally, color histogram includes main color modes and the other color components of body parts as well. Instead of cropping the salient part in an image domain, we propose to find main color modes in color histogram domain. We contribute extracting main color modes among several color modes by using GMM and fitting model. In [65], GMM is used as a parametric technique for representing the color distribution in a person's clothing for person re-id. Furthermore, the number of Gaussians is fixed as only one to fit a Gaussian distribution to the upper and the lower parts of the target, respectively, and calculate the distance simply between the overall appearance descriptions using the KL divergence [66]. On the contrary, our proposed method systematically determines the number of Gaussian components and successfully distinguishes main color modes from the others to achieve the re-id.

Table 3.1: Distance on Fig. 3.3 with Hist, EMD, KL and 2WGMMF

Method	$d\left(\mathbf{h}_{part}^{(a)}, \mathbf{h}_{part}^{(b)}\right)$ Total (*torso*, *legs*)	$d\left(\mathbf{h}_{part}^{(a)}, \mathbf{h}_{part}^{(c)}\right)$ Total (*torso*, *legs*)
Hist	1.29 (0.56, 0.73)	1.07 (0.55, 0.52)
EMD	18.74 (9.12, 9.62)	14.76 (5.99, 8.77)
KL	4.12 (2.14, 1.98)	4.00 (0.27, 3.73)
2WGMMF(1D)	1.03 (0.60, 0.43)	1.33 (0.67, 0.66)
2WGMMF(3D)	33432 (16285, 17147)	72103 (26063, 46040)

We construct a color histogram from torso and legs parts of the probe and the gallery images separately as follows:

$$\mathbf{H}^p = \left[\mathbf{h}_{torso}^p, \mathbf{h}_{legs}^p\right], \quad \mathbf{H}^g = \left[\mathbf{h}_{torso}^g, \mathbf{h}_{legs}^g\right] \tag{3.5}$$

where $\mathbf{h}$ denotes a concatenated histogram, which contains m bins and n channels, $\mathbf{h} \in \mathbb{R}^{mn\times 1}$. To build color appearance model of an image, GMM is employed because we believe that dominant color modes can be universally represented as a mixture of Gaussian components in the color histogram. Thus, we have to estimate Gaussian parameters from the color histogram. We assume that each histogram, $\mathbf{h}$, in $\mathbf{H}$ is a mixture of K Gaussian probability density functions with heteroscedastic components, *i.e.*,

$$p(\mathbf{h}|\theta_1, ..., \theta_K) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3.6}$$

where $\theta_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$ denotes the set of parameters for component k, π_k denotes the mixing proportion where $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$, $\boldsymbol{\mu}_k$ denotes the mean vector for component

k, $\mathbf{\Sigma}_k$ denotes the covariance matrix and $\mathcal{N}(\cdot)$ denotes the Gaussian distribution. K denotes the number of Gaussian components, *i.e.*, the number of dominant color modes.

Generally, GMM tries to find θ_k that maximize $p(\mathbf{h})$, this is equivalent to minimizing the negative log likelihood function:

$$-\ln p(\mathbf{h}) = -\ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{h}|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k\right)\right). \tag{3.7}$$

Then, color appearance model can be formulated as follows:

$$\begin{aligned} &\min_{\boldsymbol{\mu},\mathbf{\Sigma},\pi} -\ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\right) \\ &s.t. \quad \pi_k \geq 0, \ \sum_{k=1}^{K} \pi_k = 1. \end{aligned} \tag{3.8}$$

This model fitting problem can be efficiently solved via the Expectation Maximization (EM) algorithm [67]. In the finite mixture models, determining the number of components K is an important but very critical problem, which has not been completely resolved. Akaike information criterion (AIC) and Bayesian information criterion (BIC) have been commonly used for choosing the number of components for a suitable density estimation [67]. Both of them are effective measures of the relative quality of statistical models for a given set of data with the incorporation of a penalty term to discourage overfitting. Since the density estimate that uses BIC to select the number of components in the mixture has performed very well in previous simulation studies [68–70], BIC is employed in this paper. The preferred model is the one with the minimum BIC value, $-2\log\mathcal{L}(\theta) + K\log n$, where $\mathcal{L}(\theta)$ denotes the maximized value of likelihood function, which is equal to the inverse of final prediction error for the estimated model, K is the total number of parameters and n is the sample size [67]. In Fig. 3.4, BIC score is shown with respect to K from 1 to 20 and the preferred model is chosen when K equals 4 in case of the torso in Fig. 3.3(a). Fig. 3.5(a) shows the result of Eq. (3.8), where the blue curve is the color histogram of Fig. 3.3(a), the red curve is the histogram of GMM and the other curves are the components of GMM on the blue channel.

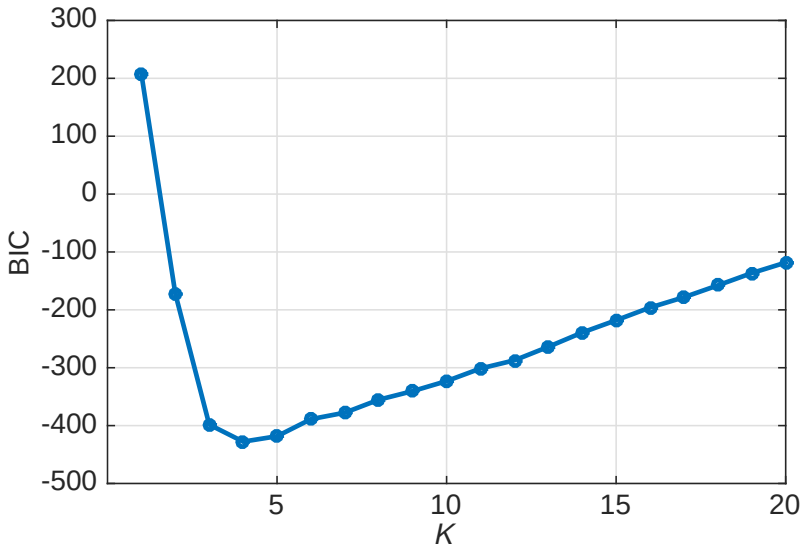
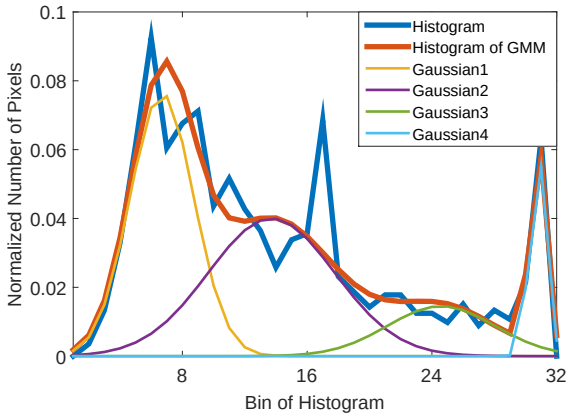


Figure 3.4: BIC curve for GMM component estimation of the torso in Fig. 3.3(a).

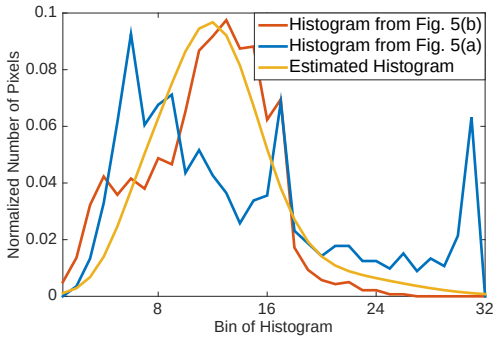
After modeling GMMs from a torso and a legs part of both a probe and a gallery image, respectively, we utilize chosen models to measure the difference between each part. The mean vectors and covariance matrices associated with the Gaussian components of the GMM are exploited to evaluate mixing weights, $\boldsymbol{\pi}$, by solving the least-squares curve fitting problem as follows:

$$\hat{\boldsymbol{\pi}} = \arg\min_{\boldsymbol{\pi}} \|G(\mathbf{h}_i^p) \cdot \boldsymbol{\pi} - \mathbf{h}_i^g\|_2^2 \quad s.t. \quad 0 \leq \pi_k,\ k = 1, ..., K \tag{3.9}$$

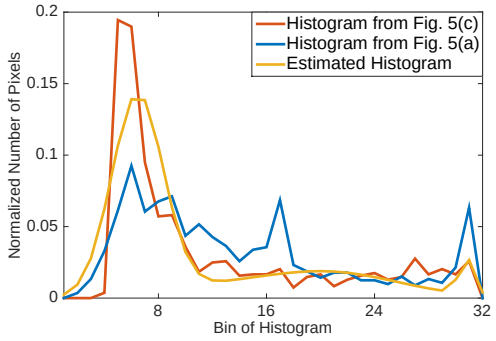
where $\mathbf{h}_i$ is normalized color histogram, $G(\mathbf{h}_i^p)$ denotes a GMM of a probe's i^{th} part and is defined as $G(\mathbf{h}_i^p) = [\mathcal{N}(\mathbf{h}_i^p|\mu_1, \Sigma_1), ..., \mathcal{N}(\mathbf{h}_i^p|\mu_K, \Sigma_K)]$. Note that $\mathcal{N}(\mathbf{h}|\mu_k, \Sigma_k) \in \mathbb{R}^{\text{mn}\times 1}$ for $i = \{torso, legs\}$. The mixing weights are $\boldsymbol{\pi} \in \mathbb{R}^{K\times 1}$ and $||\cdot||_2^2$ denotes the squared L_2 norm. Eq. (3.9) is a formulation of applying the GMM of a gallery on a probe. In this step, a different pose of the same identity can be compensated with changeable mixing proportion, $\boldsymbol{\pi}$. In Fig. 3.5(b) and (c), the red curves denote the 32-bin blue channel histogram of the torso part of Fig. 3.3(b) and (c), respectively. The yellow curves show the fitting result of Eq. (3.9) with the GMM components in Fig. 3.5(a). In other words, blue curves are transformed



(a) GMM of Fig. 3.3(a)



(b) GMM fit of Fig. 3.3(b) with (a)



(c) GMM fit of Fig. 3.3(c) with (a)

Figure 3.5: Example of the 2WGMMF application. (a) Result of Eq. (3.8) with Fig. 3.3(a). (b) Result of Eq. (3.9) with Fig. 3.3(b). (c) Result of Eq. (3.9) with Fig. 3.3(c).

to yellow curves by fitting GMM in Eq. (3.9). More specifically, we find that the second component, purple curve, in Fig. 3.5(a) increases its scale and the fourth component, sky blue curve, in the same figure decreases to 0 in Fig. 3.5(b). The residual calculated based on Eq. (3.10) decreases large enough for us to conclude that these are the same person. In contrast, the first component, yellow curve, in Fig. 3.5(a) increases its scale π_1 and the second component, purple curve, in Fig. 3.5(a) decreases its scale π_2 in Fig. 3.5(c). Thus, we can conclude that the second peak represents the unique jacket region of ID 57 and the

fourth peak represents the shirt region. The residual between a target color histogram and a summed distribution of GMM is computed as a distance:

$$r(\mathbf{h}_i^g) = d_{Chisq}(G(\mathbf{h}_i^p) \cdot \hat{\boldsymbol{\pi}}, \mathbf{h}_i^g). \tag{3.10}$$

We also need to proceed Eqs. (3.8)-(3.10) on the Gaussian mixture components of Fig. 3.3(b) and (c) to (a). That is why this feature is named as 2WGMMF since we apply the same method on both ways, *i.e.*, probe-to-gallery and gallery-to-probe. The sum of residual is used as the final distance for this probe-gallery pair as follows:

$$d_{\text{2WGMMF}}^{\text{concatenate}}(\mathbf{H}^p, \mathbf{H}^g) = r(\mathbf{h}_{torso}^p) + r(\mathbf{h}_{legs}^p) + r(\mathbf{h}_{torso}^g) + r(\mathbf{h}_{legs}^g). \tag{3.11}$$

Algorithm 1 2WGMMF with concatenated histogram

Input: a probe and a gallery image with silhouette masks

Output: a feature distance

1: Construct the concatenated color histogram

$$\mathbf{H}^p = \left[\mathbf{h}_{torso}^p, \mathbf{h}_{legs}^p\right], \quad \mathbf{H}^g = \left[\mathbf{h}_{torso}^g, \mathbf{h}_{legs}^g\right]$$

2: Solve the problem of model fit:

$$\min_{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \pi} -\ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$$

$$s.t. \quad \pi_k \geq 0, \ \sum_{k=1}^{K} \pi_k = 1.$$

3: Solve the least-squares curve fitting problem:

$$\min_{\boldsymbol{\pi}} \|G(\mathbf{h}_i^p) \cdot \boldsymbol{\pi} - \mathbf{h}_i^g\|_2^2$$

$$s.t. \ 0 \leq \pi_k, \ k = 1, ..., K$$

4: Compute the residuals

$$r(\mathbf{h}_i^g) = d_{Chisq}(G(\mathbf{h}_i^p) \cdot \hat{\boldsymbol{\pi}}, \mathbf{h}_i^g)$$

5: Compute the feature distance

$$d_{\text{2WGMMF}}^{\text{concatenated}}(\mathbf{H}^p, \mathbf{H}^g) = r(\mathbf{h}_{torso}^p) + r(\mathbf{h}_{legs}^p) + r(\mathbf{h}_{torso}^g) + r(\mathbf{h}_{legs}^g)$$

By using 2WGMMF, the distance between Fig. 3.3(a) and (b) is 1.03 and the distance between (a) and (c) is 1.33, while KL method gives 4.12 between (a) and (b), and 4.00 between

(a) and (c) in Table 3.1. When we consider d_{2WGMMF} to be a two-way distance, both distances decrease because 2WGMMF utilizes the Gaussian mixture components on histogram fitting, resulting in the better matching of jackets with dominant color components. Note that intra-personal variation becomes smaller than inter-personal's through 2WGMMF. Algorithm 1 summarizes the complete 2WGMMF procedure with concatenated histograms.

2WGMMF with joint histogram

Unlike the concatenated color histogram (we refer to this as the 1D histogram), which is constructed and exploited to build GMM in Section 3.3.2, we also generate a three-dimensional joint color histogram (we refer to this as the 3D histogram), which takes three channels jointly. For example, RGB 3D histogram takes red, green and blue channel jointly, so it can offer unique and distinct information from the 1D histogram. The procedure of feature extraction is similar to Algorithm 1. The difference is the dimension of color histogram $\mathbf{h} \in \mathbb{R}^{m^n}$, instead of $m \times n$, where m denotes the number of bins and n denotes the number of channels. The method for calculating distance is different as well. To compute the difference between joint color histograms of a probe and a GMM of a gallery, we utilize negative log likelihood as follows:

$$\begin{aligned} d_{NL}(\mathbf{h}_i^g, G(\mathbf{h}_i^p)) &= -\ln p(\mathbf{h}_i^g | \theta_1^p, ..., \theta_K^p) \\ &= -\ln \left(\sum_{k=1}^{K} \pi_k^p \mathcal{N}\left(\mathbf{h}_i^g | \boldsymbol{\mu}_k^p, \boldsymbol{\Sigma}_k^p\right) \right) \end{aligned} \tag{3.12}$$

where $G(\cdot)$ denotes GMM from inside color histogram and parameters of GMM, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and π_k, are obtained from color histogram of a part of a gallery where $i = \{torso, legs\}$. Eq. (4.4) computes the likelihood function of how much a GMM of a probe in response to the data of a gallery. The result from Eq. (4.4) is regarded as a one-way distance of i-part and a small value indicates that they are likely to be the same identity. The final distance can

be represented as follows:

$$\begin{aligned} d^{\text{joint}}_{\text{2WGMMF}}(\mathbf{H}^p, \mathbf{H}^g) = d_{NL}\Big(\mathbf{h}^p_{torso}, G(\mathbf{h}^g_{torso})\Big) + d_{NL}\left(\mathbf{h}^p_{legs}, G(\mathbf{h}^g_{legs})\right) \\ + d_{NL}\Big(\mathbf{h}^g_{torso}, G(\mathbf{h}^p_{torso})\Big) + d_{NL}\left(\mathbf{h}^g_{legs}, G(\mathbf{h}^p_{legs})\right) \end{aligned} \tag{3.13}$$

The complete 2WGMMF procedure with joint histogram is summarized in Algorithm 2.

Algorithm 2 2WGMMF with joint histogram

Input: a probe and a gallery image with silhouette masks

Output: a feature distance

1: Construct the joint color histogram

$$\mathbf{H}^p = \left[\mathbf{h}^p_{torso}, \mathbf{h}^p_{legs}\right], \quad \mathbf{H}^g = \left[\mathbf{h}^g_{torso}, \mathbf{h}^g_{legs}\right]$$

2: Solve the problem of model fit:

$$\min_{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \pi} -\ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$$

$$s.t. \quad \pi_k \geq 0, \ \sum_{k=1}^{K} \pi_k = 1.$$

3: Compute the negative log likelihood

$$d_{NL}(\mathbf{h}^g_i, G(\mathbf{h}^p_i)) = -\ln\left(\sum_{k=1}^{K} \pi^p_k \mathcal{N}\left(\mathbf{h}^g_i|\boldsymbol{\mu}^p_k, \boldsymbol{\Sigma}^p_k\right)\right)$$

4: Compute the feature distance

$$\begin{aligned} d^{\text{joint}}_{\text{2WGMMF}}(\mathbf{H}^p, \mathbf{H}^g) = d_{NL}\Big(\mathbf{h}^p_{torso}, G(\mathbf{h}^g_{torso})\Big) + d_{NL}\left(\mathbf{h}^p_{legs}, G(\mathbf{h}^g_{legs})\right) \\ + d_{NL}\Big(\mathbf{h}^g_{torso}, G(\mathbf{h}^p_{torso})\Big) + d_{NL}\left(\mathbf{h}^g_{legs}, G(\mathbf{h}^p_{legs})\right) \end{aligned}$$

3.4 Aggregation of Features

We propose mainly three representations to describe person appearance, a holistic signature based on the DCNN, regional color with concatenate histogram ($\text{2WGMMF}_{\text{1D}}$) and joint histogram ($\text{2WGMMF}_{\text{3D}}$). To perform the person re-id, we should combine these three features effectively. Since the dynamic ranges of three distances are quite different (see in Table 3.1), it is not so meaningful to add them directly. Some form of normalization is necessary to balance the contributions of individual features.

In case of d_{2WGMMF}, it can integrate several kinds of color space, *e.g.*, RGB, HSV, and YCbCr. When more than two color spaces are exploited, we collect them according to their dimensions. For example, RGB, HSV and YCbCr color spaces are used in 2WGMMF, they are combined as follows:

$$\begin{aligned} d^{\text{1D}}_{\text{2WGMMF}} &= d^{\text{1D RGB}}_{\text{2WGMMF}} + d^{\text{1D HSV}}_{\text{2WGMMF}} + d^{\text{1D YCbCr}}_{\text{2WGMMF}} \\ d^{\text{3D}}_{\text{2WGMMF}} &= d^{\text{3D RGB}}_{\text{2WGMMF}} + d^{\text{3D HSV}}_{\text{2WGMMF}} + d^{\text{3D YCbCr}}_{\text{2WGMMF}} \end{aligned} \tag{3.14}$$

Aggregating similarity scores promises more convincing result than minimizing accumulated distances [71, 72]. Thus, the distances are converted into the similarity as in inverse proportion, respectively.

$$sim\left(\mathbf{H}^p, \mathbf{H}^g_j\right) = 1/d\left(\mathbf{H}^p, \mathbf{H}^g_j\right) \quad \text{for } j = 1, ..., N \tag{3.15}$$

where N denotes the size of gallery set. In case of 2WGMMF with joint histogram, their feature distance can be negative because $d^{\text{joint}}_{\text{2WGMMF}}$ denotes the sum of negative log-likelihood. To make them positive, we add the minimum value of them before converting into similarity scores.

In order to balance the contributions of separate features, min-max normalization is performed to transform the feature distances into 0-to-1.

$$\widehat{sim}\left(\mathbf{H}^p, \mathbf{H}^g_j\right) = \frac{sim\left(\mathbf{H}^p, \mathbf{H}^g_j\right) - \min S}{\max S - \min S} \tag{3.16}$$

where $S = \{sim\left(\mathbf{H}^p, \mathbf{H}^g_1\right), ..., sim\left(\mathbf{H}^p, \mathbf{H}^g_N\right)\}$.

To stick to direct methods and to show the intrinsic quality of the proposed descriptors, we have set equal weights as follows:

$$sim_{Final}\left(\mathbf{H}^p, \mathbf{H}^g_j\right) = \frac{1}{3}\sum_{n=1}^{3} \widehat{sim}_n\left(\mathbf{H}^p, \mathbf{H}^g_j\right) \tag{3.17}$$

where $j = 1, ..., N$. Finally we choose the identity of a probe image to be the one with maximum similarity among N gallery images as follows:

$$identity(\mathbf{H}^p) = \arg\max_j S_{Final} \tag{3.18}$$

where $S_{Final} = \left\{ sim_{Final}\left(\mathbf{H}^p, \mathbf{H}_j^g\right)\right\}$ for $j = 1, ..., N$. Algorithm 3 shows all the steps of classifying the same identity of a probe in the gallery based on calculating the distance scores.

Algorithm 3 Person re-identification with aggregation of distance scores

Input: distance scores $(d_{\mathrm{DCNN}}, d^{\mathrm{1D}}_{\mathrm{2WGMMF}}, d^{\mathrm{3D}}_{\mathrm{2WGMMF}})$

Output: identity

1: Add the distances of the same dimensional 2WGMMF feature together

$$d^{n\mathrm{D}}_{\mathrm{2WGMMF}} = d^{n\mathrm{D\ color_1}}_{\mathrm{2WGMMF}} + d^{n\mathrm{D\ color_2}}_{\mathrm{2WGMMF}} + ... + d^{n\mathrm{D\ color_M}}_{\mathrm{2WGMMF}}$$

2: Transform the distance into the similarity:

$$sim\left(\mathbf{H}^p, \mathbf{H}_j^g\right) = 1/d\left(\mathbf{H}^p, \mathbf{H}_j^g\right) \quad \text{for } j = 1, ..., N$$

where N denotes the size of gallery set.

3: Min-max normalization

$$\widehat{sim}\left(\mathbf{H}^p, \mathbf{H}_j^g\right) = \frac{sim\left(\mathbf{H}^p, \mathbf{H}_j^g\right) - \min S}{\max S - \min S}$$

where $S = \{sim\left(\mathbf{H}^q, \mathbf{H}_1^g\right), ..., sim\left(\mathbf{H}^p, \mathbf{H}_N^g\right)\}$

4: Compute the final similarity

$$sim_{Final}\left(\mathbf{H}^p, \mathbf{H}_j^g\right) = \frac{1}{3}\sum_{n=1}^{3} \widehat{sim}_n\left(\mathbf{H}^p, \mathbf{H}_j^g\right)$$

for $j = 1, ..., N$

5: Classify the same identity

$$identity(\mathbf{H}^p) = \arg\max_j S_{Final}$$

where $S_{Final} = \left\{ sim_{Final}\left(\mathbf{H}^p, \mathbf{H}_j^g\right)\right\}$ for $j = 1, ..., N$

3.5 Experimental results

This section presents the evaluation results of our approach on the challenging benchmark datasets, VIPeR [1] [16] and 3DPeS[2] [17] , which are video surveillance datasets designed for

[1]VIPeR dataset is available at https://vision.soe.ucsc.edu/node/178

[2]3DPeS dataset is available at http://www.openvisor.org/3dpes.asp

person re-id between multiple cameras with non-overlapping FOVs. We follow the same cross-validation protocol as reported in [3, 7], where ten random splits of the snapshots for the probe and gallery sets. There is no overlapped image between the probe and gallery set. The results are obtained by averaging ten corresponding outcomes. The performances are represented using the Cumulative Matching Characteristic (CMC) curve and the normalized Area Under Curve (nAUC). CMC curve represents where the rank-k recognition rate is the expectation of correct matches within the top k ranks. The nAUC is the area under the CMC curve and is exploited to summarize the overall performance. The higher nAUC indicates the better performance.

3.5.1 Analysis of Individual Features

For the pre-trained DCNNs, we apply three popular large scale convolutional neural networks: AlexNet [1], VGG [63] and GoogLeNet [64]. Since the higher layers can provide more global information with better discriminative powers [56, 57], we select activations of top layer and compare the performance with/without employing the unit rectifier, ReLU, for the AlexNet [1] and VGG [63]. As summarized in Table 3.2, we evaluate the features 4,096 dimensional features from FC7 and FC7+ReLU (ReLU7) of both AlexNet [1] and VGG [63], and 1,024 dimensional features from Pool5 representing the Average-Pooling after inception 5 in GoogLeNet [64]. To compare the performance of different features, we exploit Eq. (3.3) based on the same squared L_2 distance ($metric$ = sqeuclidean) metric to find the best matched r instances (top-rank r) in the gallery set. The CMC curves for VIPeR and 3DPeS datasets are shown in Fig. 3.6(a) and 3.6(b), the performance using features from ReLU7 in AlexNet [1] outperforms than other forms of VGG [63], and GoogLeNet [64]. Furthermore, based on the promising features of the ReLU7 in AlexNet [1], Fig. 3.6(c) and 3.6(d) show that the data augmentation allows to improve about 4% on rank-1 accuracy and it performs better using Chi-squared distance than the squared L_2 or L_1 distance on both VIPeR and 3DPeS datasets. Therefore, we decide to adopt the feature vector of AlexNet ReLU7 with Chi-squared distance metric to generate the holistic invariant features. This holistic invariant

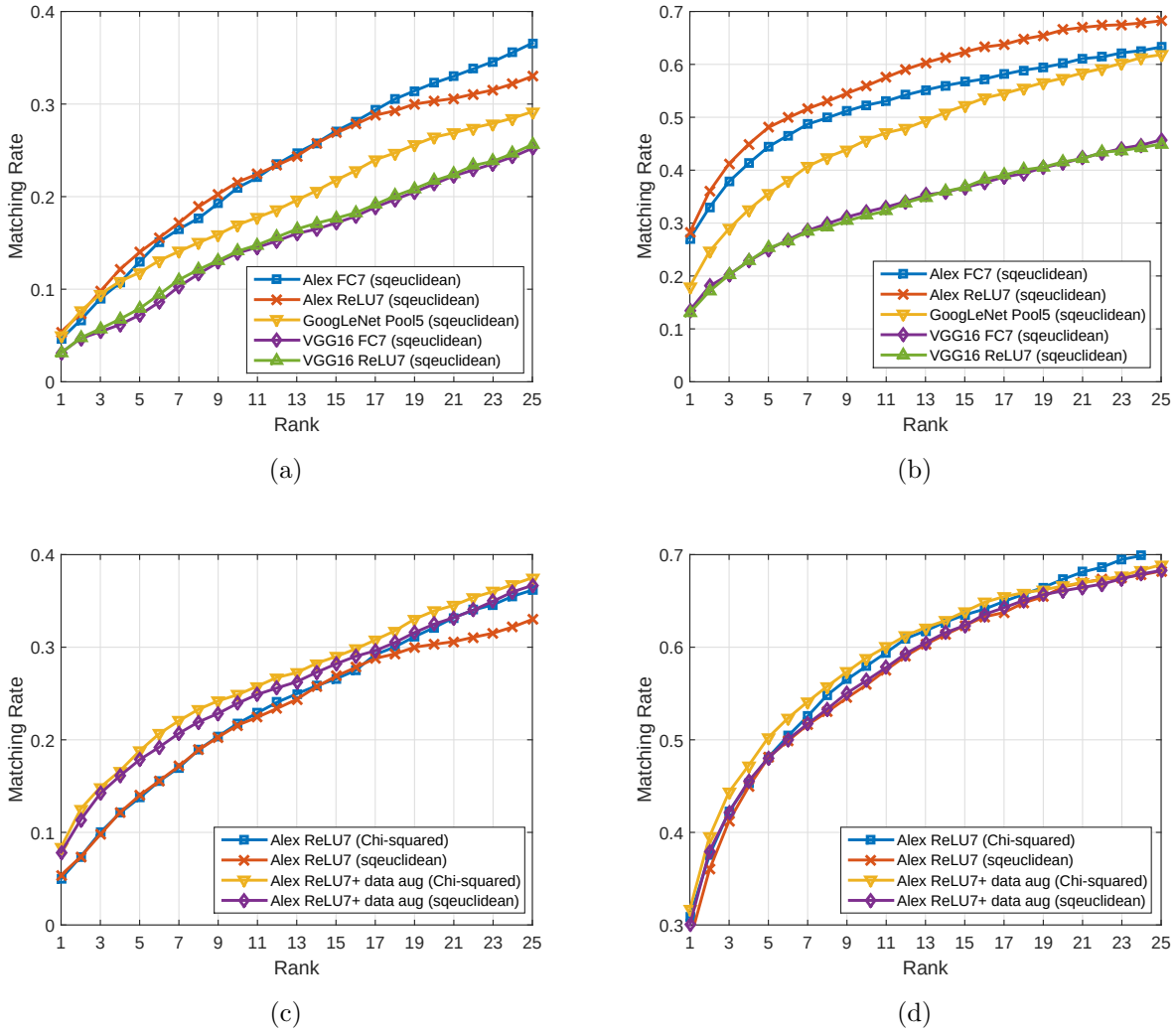
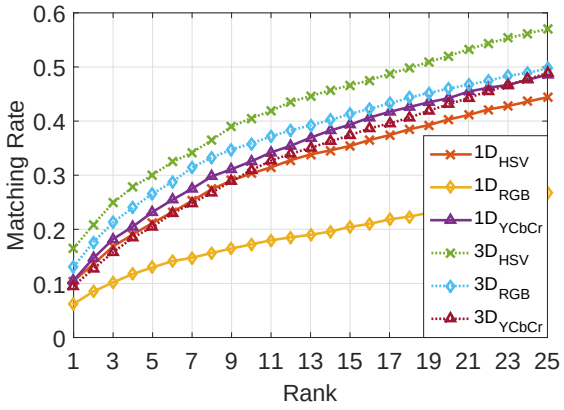
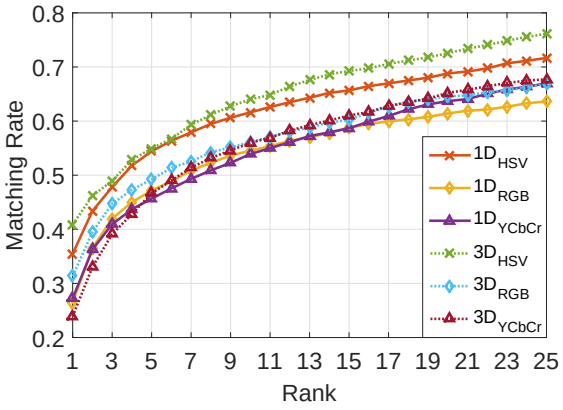


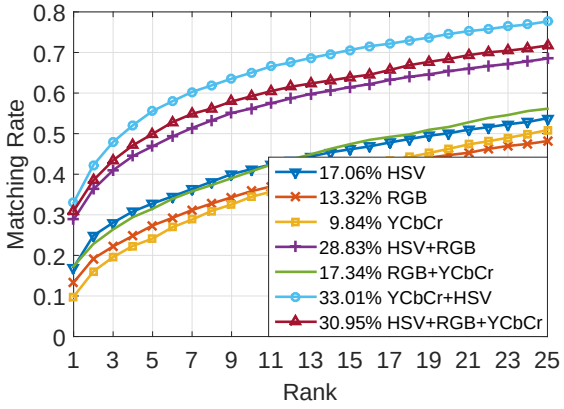
Figure 3.6: DCNN comparisons on (a) VIPeR and (b) 3DPeS dataset. The features from FC7+ReLU (ReLU7) layer in AlexNet [1] have a better discriminative power than the other pre-trained net models. The data augmentation with Chi-squared distance metric can improve about 4% on rank-1 accuracy.



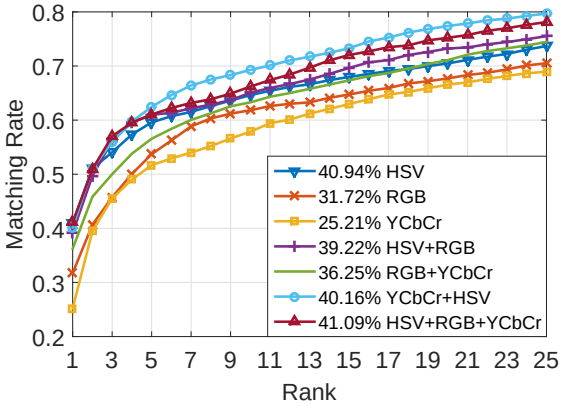
(a) VIPeR dataset



(b) 3DPeS dataset



(c) VIPeR dataset



(d) 3DPeS dataset

Figure 3.7: (a)-(b) Comparison of individual 2WGMMF feature with a various of color spaces and dimensions. (c)-(d) Comparison of color space with 2WGMMF feature (marked values are rank-1 accuracies).

Table 3.2: The selected pre-train models and layer for comparison.

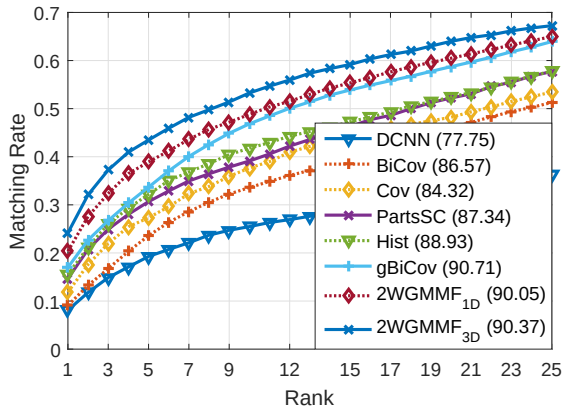
Selected Layer	Feature Dimension
AlexNet FC7 [1]	4,096
AlexNet ReLU7 [1]	4,096
VGG FC7 [63]	4,096
VGG ReLU7 [63]	4,096
GoogLeNet Pool5 [64]	1,024

features will combine with the regional invariant features for further comparison.

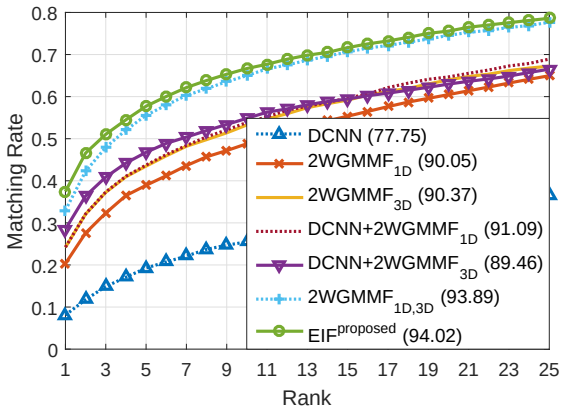
To investigate the characteristics of the 2WGMMF feature, several kinds of experiments are performed and analyzed. In Fig. 3.7, we show the CMC curve of several 2WGMMF features on the two datasets. In particular, Fig. 3.7(a) and (b) show the performances of individual 2WGMMF features with varied color spaces and dimensions. Fig. 3.7(c) and (d) show the results of 2WGMMF features with both concatenated and joint color histograms on the marked color space. On both datasets, 2WGMMF features with HSV joint histogram performs the best consistently as shown in Fig. 3.7(a) and (b). However, the other results are not consistent. In Fig. 3.7(c) and (d), the best color space is not the same with respect to rank-1 accuracy. 2WGMMF features with the combination of YCbCr and HSV color spaces outperforms the others on the VIPeR. In contrast, when three color spaces are combined, rank-1 accuracy is the highest on the 3DPeS in Fig. 3.7(d). To sum up, the performance of features in each color space and dimension is not consistent on two different datasets because 2WGMMF features are designed to deal with varied viewpoints and poses, and three kinds of color spaces are not illumination-invariant. Thus, we expect that the performance can be improved further when 2WGMMF features are integrated with invariant color space.

3.5.2 VIPeR Dataset

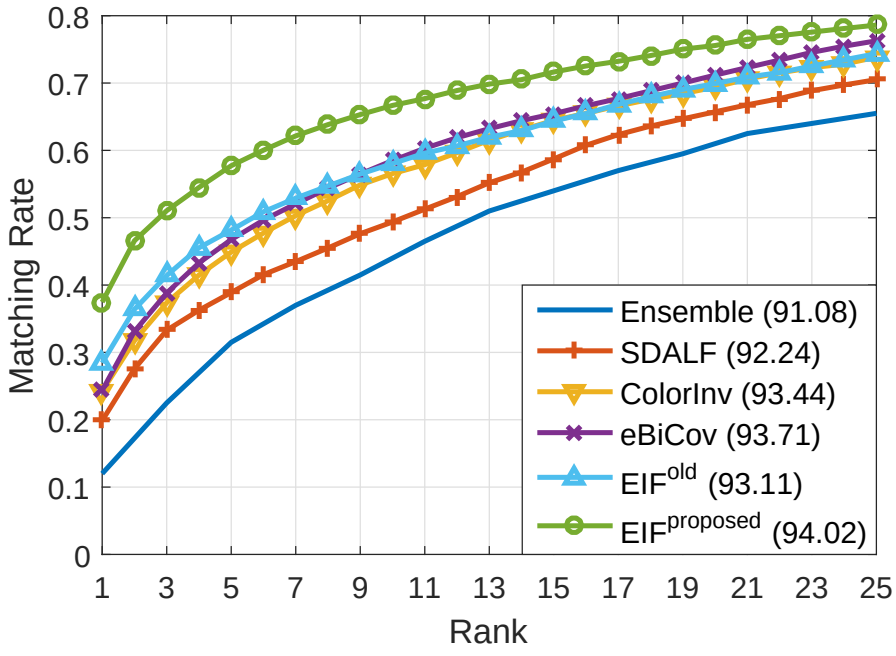
The VIPeR dataset contains 632 pedestrian image pairs taken from two different cameras. Most of the image pairs contain a viewpoint change of 90° and are taken under varying



(a)



(b)



(c)

Figure 3.8: VIPeR dataset. (a) Comparison of the state-of-the-art single set of features. (b) Comparison of DCNN, 2WGMMF and their combinations. (c) Comparison to Ensemble [2], SDALF [3], ColorInv [4], eBiCov [5] and EIF$^{\text{old}}$ [6].

illumination conditions. Each pair is randomly split into two sets. Images from one set are considered as the probe set and images from the other set are regarded as gallery set. All the images of VIPeR are scaled to the size of 128×48 pixels for experiments. For each image on the dataset, a silhouette mask is automatically extracted by using the STEL model [73] and provided in [3].

To evaluate the performance, we compare the proposed features with the state-of-the-art methods [2–6, 74]. Farenzena *et al.* [3] divides a human body into head/torso/legs based on the horizontal and vertical asymmetries of a human silhouette. They extract three complementary descriptors including weighted HSV histogram, the spatial arrangement of colors into stable regions, Maximally Stable Color Regions (MSCR), and Recurrent High-Structured Patches (RHSP). Similar to our proposed scheme, the Ensemble method by Gray *et al.* [2] emphasizes the viewpoint-invariant features. They propose an ensemble of color-based features (RGB, HSV and YCbCr histograms) and gradient-based features (the histogram generated by Gabor filter response). The ensemble features are computed for each of the three fixed size stripes of a persons silhouette. For the fair comparison, a distance between feature vectors is computed through the distance metric between the descriptors rather than applying the additional metric learning as proposed in [2]. Kviatkovsky *et al.* [4] proposes illumination-invariant color features. They combine their features with several standard signatures, the parts shape context descriptor (PartsSC), color histogram (Hist), and the region covariance descriptor (Cov). BiCov [74] is a representation relied on the combination of Biologically Inspired Features (BIF) and covariance descriptors used to compute the similarity of the BIF features at neighboring scale. gBiCov [5] is the extended work of BiCov, combining covariance similarity with BIF. Furthermore, Ma *et al.* propose eBiCov (enriched gBiCov) by combining gBiCov with weighted color histograms and MSCR defined in [3]. In [6], we proposed EIF by combining 2WGMMF with concatenated histogram, DCNN, and CLBP (completed local binary pattern), which represents texture information.

In Fig. 3.8(a), the proposed features are compared with BiCov, Cov, PartsSC, Hist and gBiCov for analyzing the sensitivity of the single set of features. These results are obtained

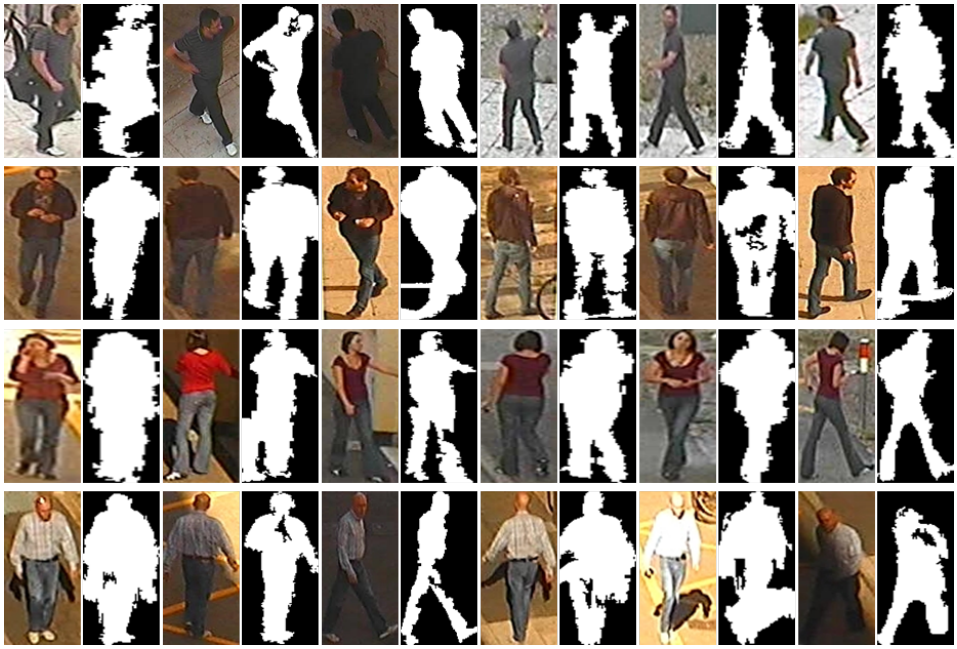
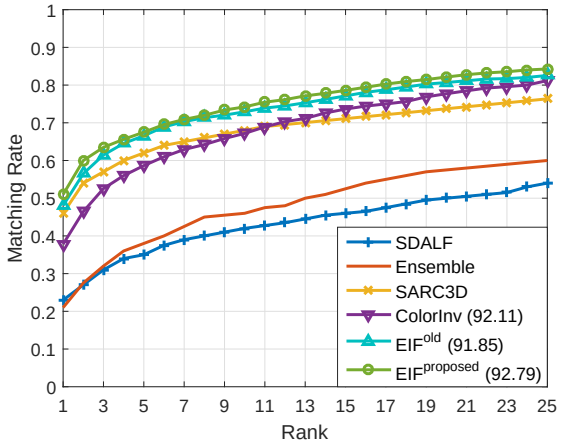
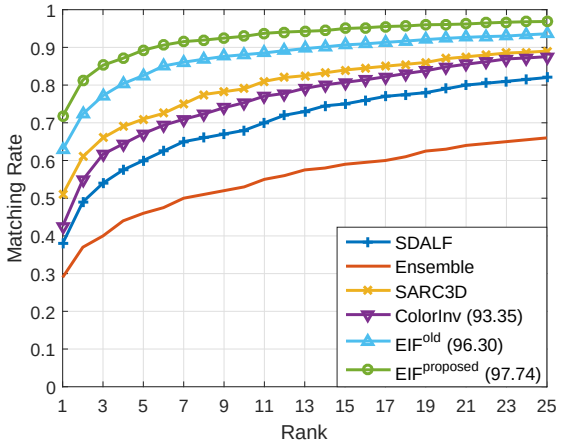


Figure 3.9: Image and mask examples from 3DPeS dataset.

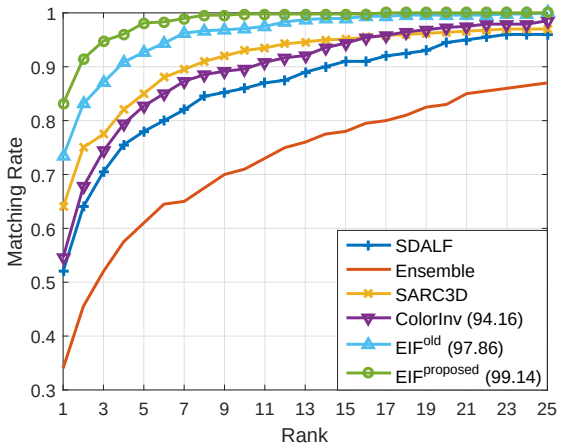
from authors' reported materials [75, 76]. The 2WGMMF features with joint histogram show the best performance among the state-of-the-art single set of features both on CMC curve and nAUC, as marked in the legend. On the VIPeR dataset, we exploit HSV and YCbCr for 2WGMMF features since they show the best performance with respect to rank-1 accuracy (see Fig. 3.7(c)). In Fig. 3.8(b), we compared DCNN, 2WGMMF and their combinations. Among all combination of features, 2WGMMF features with concatenated and joint histogram outperform all others. In Fig. 3.8(c), the proposed method is compared to the state-of-the-art, eBiCov, ColorInv, SDALF, Ensemble and EIF^{old}. In case of ColorInv, we adopt the best signature that combines PartsSC, Hist, and Cov features together as mentioned in [4]. Our proposed EIF achieves 37.47% at rank-1, which is 13% higher than eBiCov [5] and 9% higher than the old EIF [6]. EIF^{old} has texture information, which is not effective in case of viewpoint change. In contrast, $\text{EIF}^{\text{proposed}}$ exploits joint histogram in 2WGMMF, which is the most accurate single set of features (see Fig. 3.8(a)).



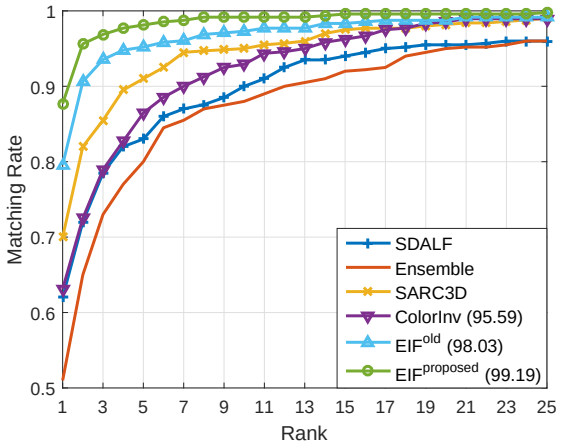
(a) 1vs1



(b) 3vs1



(c) 5vs1



(d) 3vs3

Figure 3.10: Comparison to the SDALF [3], Ensemble [2], SARC3D [7], ColorInv [4] and EIF$^{\text{old}}$ [6] on the 3DPeS dataset (Nvs$M$: N gallery shots vs M probe shots).

3.5.3 3DPeS Dataset

The 3DPeS dataset contains short video sequences instead of still images, which increase the diversity of person poses and camera viewpoints. A collection of snapshots (6 different shots) are collected for 4 different persons (one person per row) as shown in Fig. 3.9. The complete dataset is composed of 1,012 snapshots of 192 different people. 3DPeS images are normalized to the size of 128×64 pixels. The silhouette mask is automatically extracted by [77] and provided in [17]. In contrast to VIPeR dataset, 3DPeS contains multi-shots for each person. We evaluate the capability to integrate multi-shots on the 3DPeS.

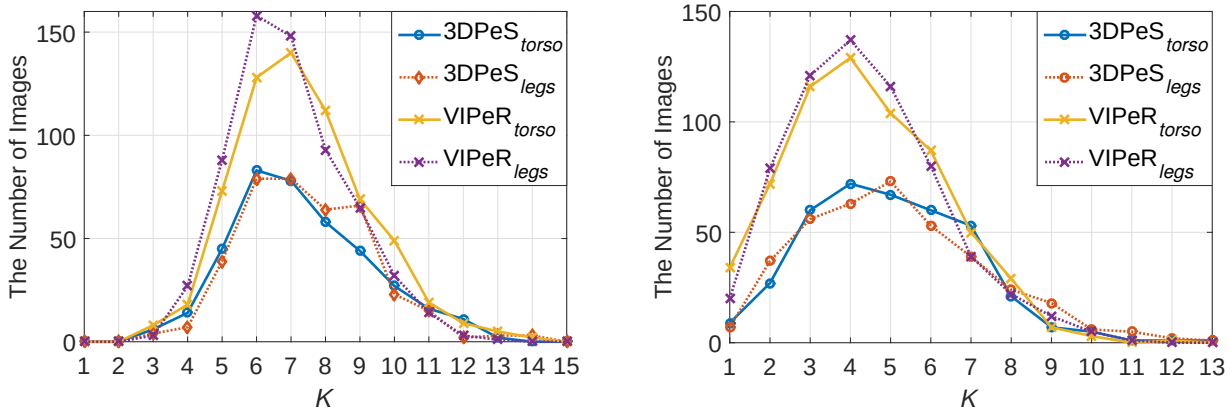
We compare CMC performance with five approaches, SDALF, Ensemble, SARC3D [7] ,ColorInv and EIF^{old} on the 3DPeS. Baltieri *et al.* [7] proposes a simplified non-articulated 3D body model to spatially map appearance descriptors (color and gradient histograms) into a vertex-based 3D body surface. Due to the 3D information, they can directly handle the issues of occlusions, partial views or pose changes, which normally cause performance degradation by using 2D descriptors. However, the Image-to-Model mapping needs the perspective projection matrix (intrinsic parameters) and extrinsic calibration matrix to estimate the rotation and translation information between the world reference and the model coordinates. Fig. 3.10 shows the top-rank results of our proposed algorithm, in comparison with five other state-of-the-art methods, where 2WGMMF features are applied on HSV, RGB and YCbCr color spaces (see Fig. 3.7(d)). For multi-shot integration, we take a pair of images, which has maximum similarity, as matching identity. Overall, the performance is improved by the proposed method about 3%, 9%, 10% and 8% better accuracies at rank-1 with 1vs1, 3vs1, 5vs1 and 3vs3 dataset, respectively. Nvs$M$ means a test dataset has N gallery images and M probe images for an identity. It is noteworthy that the proposed method achieves a better result than SARC3D, even though our method is purely based on 2D features without requiring any additional information, *e.g.*, intrinsic parameters, extrinsic calibration matrix, etc, as needed in SARC3D. Specifically, our method achieves 51% rank-1 recognition rate and SARC3D obtains 46% in 1vs1 scenarios as shown in Fig. 3.10(a). In the experiments

of multi-shot, the performance of our algorithm is still superior. At rank-1, our proposed scheme achieves about 72%, 83% and 88%, while SARC3D achieves 51%, 64% and 70% with 3vs1, 5vs1 and 3vs3, respectively. For the 3DPeS, we only mark nAUC of ColorInv since implementation software codes or feature representations of the other methods are not available. The CMC curves of the rest are acquired in [7].

3.5.4 Discussions

In 2WGMMF, there is no manual tweaking of parameters for experiments. The number of components K is automatically decided by BIC. We present the results of K on two datasets in Fig. 3.11. Each figure shows the distributions of K on each body part with both concatenated RGB-based color histogram and joint color histogram, respectively. On the VIPeR and 3DPeS datasets, 632 images of 316 identities and 384 images of 192 identities are exploited to derive K, respectively. Most of the concatenated color histogram has 6 or 7 components in both body parts and the distribution of two parts are very similar on both datasets. In case of the joint histogram, the number of components is less than that of the concatenated histogram and the distributions of two parts are comparable on both datasets. The only difference between the distributions of VIPeR and 3DPeS is the magnitude due to the difference of dataset sizes (*e.g.*, in case of $K = 7$, the number of images are 140 and 78 on the VIPeR and the 3DPeS datasets, respectively, in Fig. 3.11(a)). In other words, the distribution of K is consistent on the disparate datasets of pedestrian images. Since person appearance, *i.e.*, the composition of clothes, is similar on the street, these results are reasonable. Thus, we can conclude that our proposed method, 2WGMMF, is robust to extract the dominant color modes.

We have also conducted experiments to compare the separate importance of holistic and regional features on the VIPeR. In Fig. 3.12, the use of regional features alone shows the better performance than that of holistic features. More specifically, 2WGMMF with torso part gives better result than legs part. That means torso part gives more discriminative information for re-id. In addition, the green curve in Fig. 3.12 shows the results of no body



(a) Concatenated histogram

(b) Joint histogram

Figure 3.11: Distribution of the number of components K by 2WGMMF.

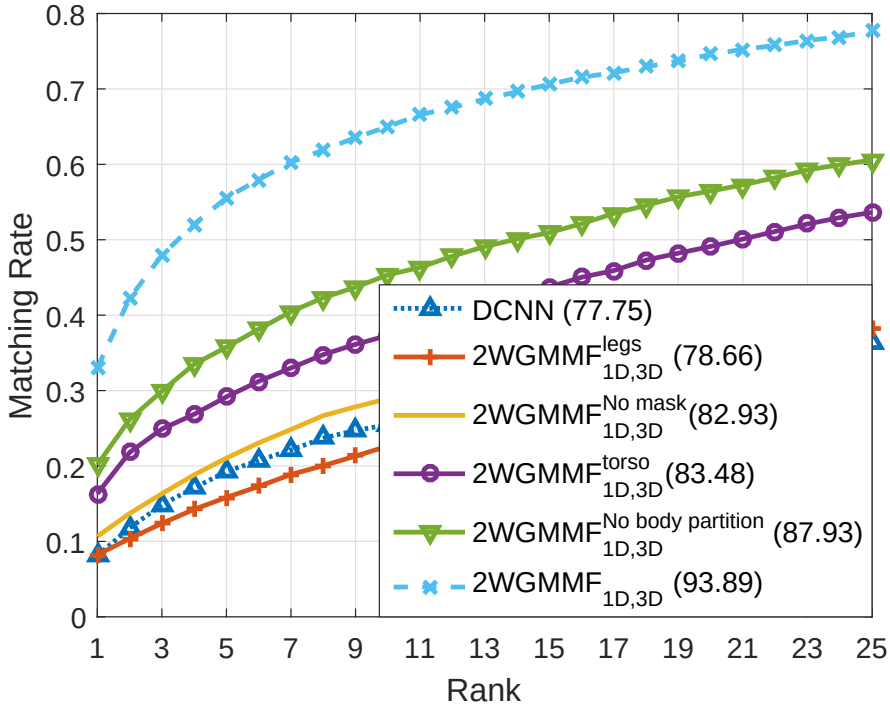
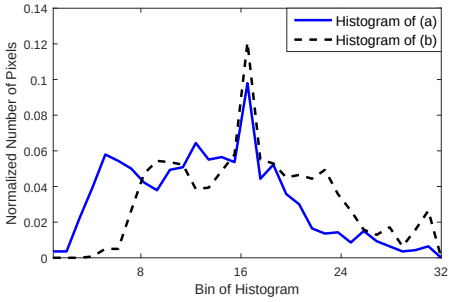
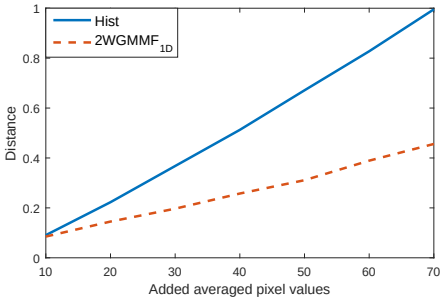


Figure 3.12: Comparative results of holistic and regional features.

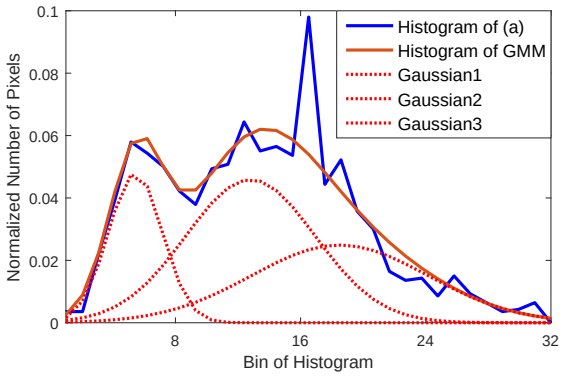
(a) (b)



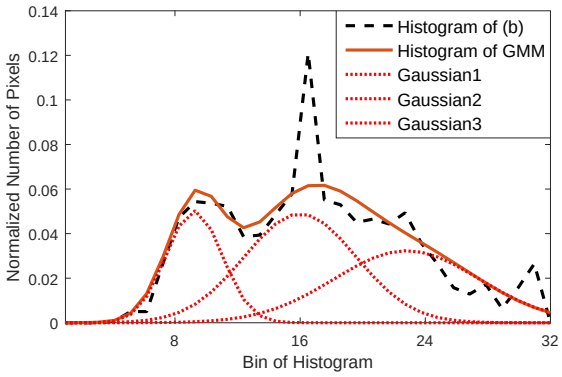
(c)



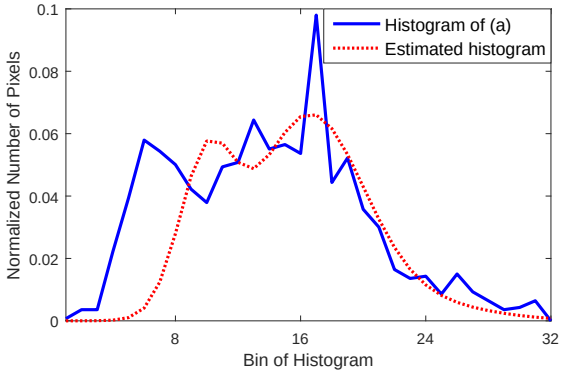
(d)



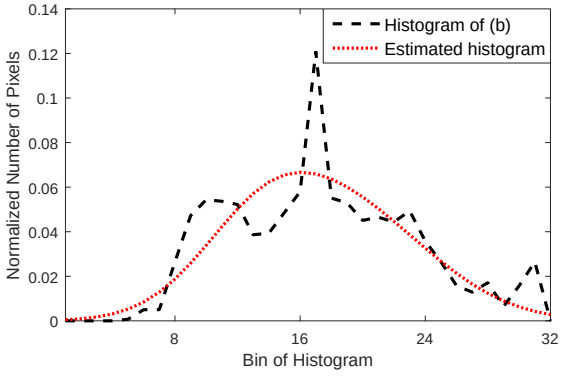
(e)



(f)



(g)



(h)

Figure 3.13: (a) ID 81 foreground image of VIPeR. (b) Averaged 30 pixel value added image of (a). (c) Red channel histograms of torsos of (a) and (b). (d) Comparative results of histogram and 2WGMMF distances with concatenated histogram. (e) GMM of (a). (f) GMM of (b). (g) GMM fit (a) with (b). (h) GMM fit of (b) with (a).

partition. When we compare green curve and sky blue curve, we can find that body partition indeed increases the rank-1 score about 13%. We conjecture that in the process of building histograms without body partition, pixel values from different colored clothes are overlapped and mixed in the histogram domain. That prevents us from extracting genuine color components, resulting in degraded accuracy. To find out the impact of background removal, we have evaluated the performance with no mask. Yellow curve shows the performance of 2WGMMF without the mask. As expected, the performance without masks degrades significantly compared to that with the mask, which is shown in sky blue curve. However, 2WGMMF still outperforms DCNN when no silhouette masks are available. In the proposed method, masks are exploited in body partition (Section IV-A) and feature extraction (Section IV-B). If background appearance inside the detected bounding box is included, it can result in less accurate boundary lines between head-torso and torso-legs in body partition, and disturb dominant color modes extraction in feature extraction. In other words, color components from the background can be regarded as dominant color modes of person appearance and that creates more mismatches.

We also conduct some experiments to show some extent of tolerance on illumination changes of our method. In person re-id, illumination and viewpoint changes of two non-overlapping cameras are the most difficult problems to solve. For example, strong light makes the image in Fig. 4.2(a) brighter, which shown in (b) where red rectangles represent torso parts and color histogram shifts to the right side in (c). Fig. 4.2(c) shows examples of red channel histograms of torso parts when we increase the value by 30 on average to each pixel. Even though color histogram is shifted, Gaussian components are almost retained in Fig. 4.2(f) compared to (e). Fig. 4.2(d) shows distance curve with respect to several added averaged values on the Chi-squared distance of histogram and 2WGMMF with the concatenated histogram. It can be seen that the slope of 2WGMMF is smaller than histogram distance, which implies 2WGMMF can tolerate more significant color changes caused by lighting change rather than histogram distance. 2WGMMF fits Gaussian components from probe to gallery and vice versa. Then, the scale change of each Gaussian is able to mitigate

the effect of the shifted histogram (see Fig. 4.2(g) and (h)), so that 2WGMMF tolerates more significant light changes.

In this paper, since we assume that we do not have any additional training set to train/update our model, it is difficult to increase the model adaptation. But once we are capable of collecting some training data from a specific camera network, we can improve from two aspects. One thing is improving ensemble/combination strategy and another one is enhancing the similarity measurement. The effective feature ensemble or combination have still not been conclusively addressed in the existing researches [78]. Most reported methods concatenate all the feature vectors from different cues, which may cause some issues. Firstly, the total feature dimension is increased when adding new cues, resulting in higher computational load. Secondly, the direct concatenation does not consider the importance among different image cues so high dimensional features will probably dominate the low dimensional ones. In fact, different features may carry complementary information, *e.g.*, our proposed pre-trained DCNN features and 2WGMMF, and they require effective fusion instead of direct concatenation, which may lose some useful information for re-id. To tackle these problems, ensemble learning methods could be applied to get better predictive performance, for example, the linear combination of similarity measurements scheme where the weight parameters w are learned by the SVM, AdaBoost or nonlinear combination by Bagging (*e.g.*, random forest algorithm).

Chapter 4

INTER-CAMERA TRACKING BASED ON FULLY UNSUPERVISED ONLINE LEARNING

4.1 Overview

In this section, we present main components of our proposed ICT methodology. An overview of the proposed approach is shown in Fig. 4.1. First, the SCT and segmentation results are acquired from each disjoint surveillance camera as input to ICT. The features are extracted on image domain, so more precise masks lead to better ICT results (see Table 4.4). Examples are shown in Fig. 4.2, our proposed method gives more accurate segmentation masks compared to the SuBSENSE method. To mitigate variations of illumination and color response among cameras, we transfer color characteristics of a source image to a target image before extracting features (Section 4.2.1). For extracting appearance features, an object is divided into three parts, head, torso, and legs (Section 4.2.2). Our feature extractor consists of two phases, and the phase change occurs after having at least two good matches to build camera link model, which includes region mapping matrix, region matching weights and feature fusion weights. In phase I, ICT relies on 2WGMMF (Section 4.2.4) and couple features (Section 4.2.6). Subsequently, holistic color (Section 4.2.3) and regional color/texture features (Section 4.2.5) are further incorporated with feature fusion weights in phase II after the camera link model is systematically and continuously learned and updated (Section 4.2.7).

4.2 Inter-Camera Object Tracking

4.2.1 Color Transfer

The appearance of the same person may appear differently under two cameras because of illumination changes and different cameras color responses. In [15, 79], an algorithm

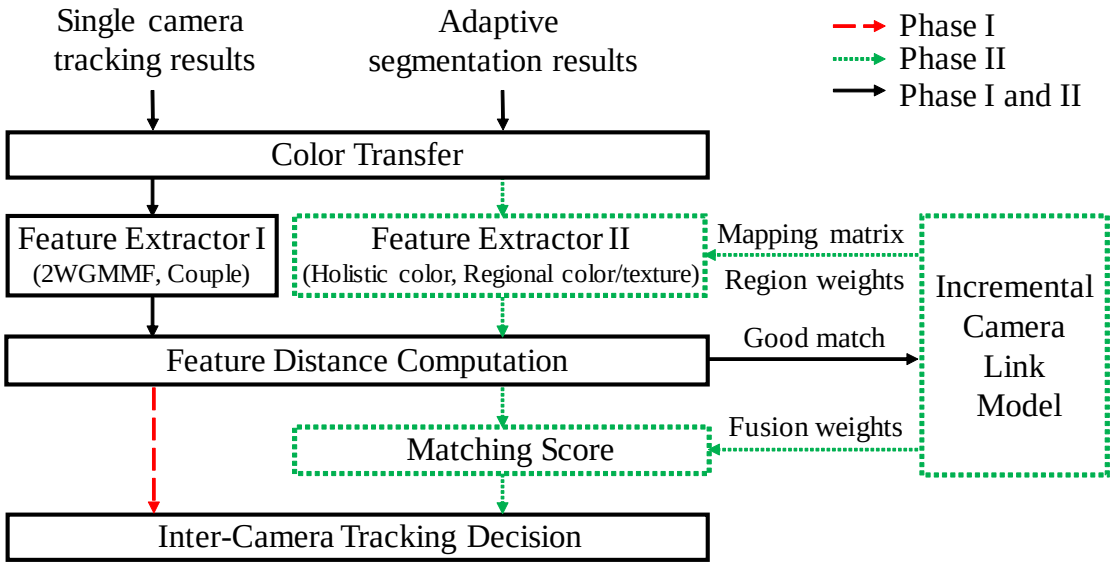


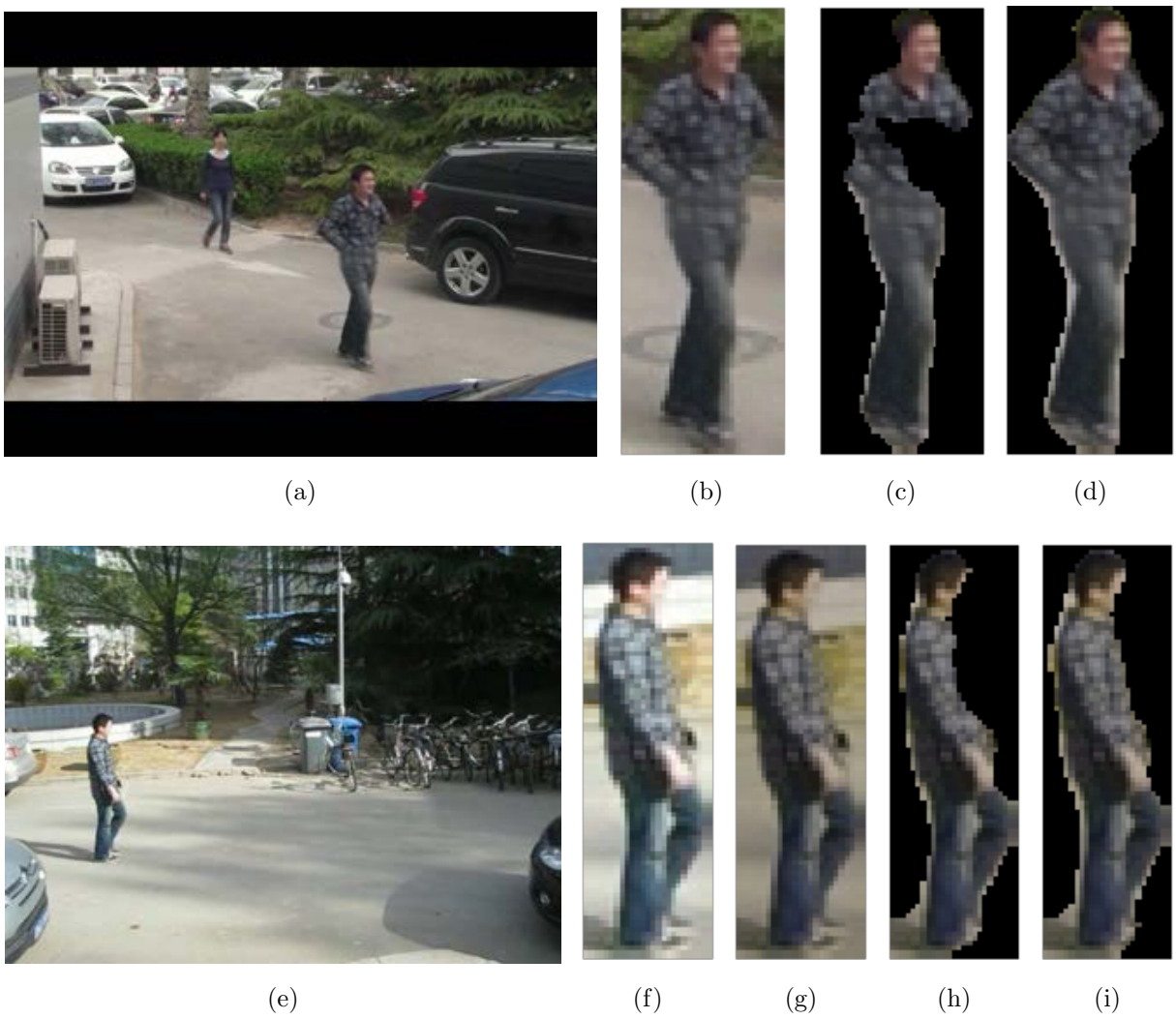
Figure 4.1: An overview of our inter-camera multiple target tracking approach.

corrects/transfers one images color characteristics to the other by de-correlating color space and statistical computation. More specifically, the RGB color space is transformed to the $l\alpha\beta$ color space and the data points composing the color transformed image are scaled by factors determined by the respective standard deviations in each channel as follows:

$$\begin{aligned} l_s^{'} &= \frac{\sigma_t^l}{\sigma_s^l}(l_s - \mu_s^l) + \mu_t^l, \\ \alpha_s^{'} &= \frac{\sigma_t^\alpha}{\sigma_s^\alpha}(\alpha_s - \mu_s^\alpha) + \mu_t^\alpha, \\ \beta_s^{'} &= \frac{\sigma_t^\beta}{\sigma_s^\beta}(\beta_s - \mu_s^\beta) + \mu_t^\beta, \end{aligned} \tag{4.1}$$

where μ and σ denote mean and standard deviation, and subscripts s and t denote source and target images, respectively.

We apply color characteristics transfer method between bounding boxes of two objects. Fig. 4.2(g) shows such an example of color transfer.



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

(e) (f) (g) (h) (i)

Figure 4.2: (a) Source frame. (b) Global ID 6 in CAM4. (c) Masked image of (b) with SuBSENSE segmentation. (d) Masked image of (b) with the proposed segmentation. (e) Target frame. (f) Global ID 6 in CAM5. (g) Color transferred result of (f). (h) Masked image of (g) with SuBSENSE segmentation. (i) Masked image of (g) with the proposed segmentation.

4.2.2 Body Partition

Since a pedestrian is commonly acquired at very low resolution in surveillance cameras, it is reasonable to notice that the most distinguishable body parts are three: head, torso, and legs [3, 80]. Two boundary lines are systematically located and used to separate head-torso and torso-legs parts, respectively, as shown by the red lines in Fig. 4.3(b) and 4.3(c). From the top to the bottom of the rectangular foreground bounding box, we calculate the histogram distance line-by-line between two block regions, *i.e.*, the blue and green blocks in Fig. 4.3(a). Each block is of height δ_h and width W from a line T_i. Intuitively, we expect that color similarity between two different body parts to be low. Therefore, a boundary line is located at height T_i computed by solving the following problem, for both head-torso and torso-legs regions, respectively:

$$\max_{T_i \in \{S,E\}} d(\mathbf{h}_{[T_i, T_i+\delta_h]}, \mathbf{h}_{[T_i-\delta_h, T_i]}), \tag{4.2}$$

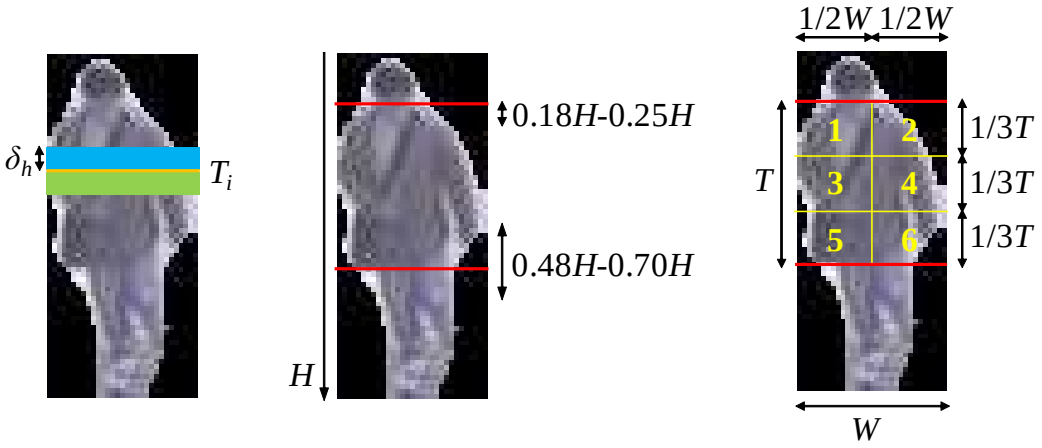
where $d(\cdot)$ denotes Euclidean distance and $\mathbf{h}_{[a,b]}$ denotes the color histogram derived from the region from a to b. Moreover, the boundary line is empirically assumed to be located within $\{S, E\}$, *i.e.*, $\{0.18H, 0.25H\}$ for head-torso and $\{0.48H, 0.70H\}$ for torso-legs. In our experiments, 8-bin RGB histogram is employed and the height δ_h value is empirically set as 5 pixels.

4.2.3 Holistic Color Feature

A Color histogram is widely used for representing color distributions [29]. In case two camera viewpoints are similar, the color histogram is effective to match the same person. So we utilize it as the holistic color feature to describe person clothing from head-torso boundary to the bottom. The total cost function for the holistic color feature is

$$d_{\text{holistic color}}(A, B) = d(\mathbf{h}^A, \mathbf{h}^B), \tag{4.3}$$

in which $\mathbf{h} \in \mathbb{R}^n$ denotes the holistic color histogram of the observation concatenating all color channels. In this paper, we also use 8-bin histogram for each RGB channel.



(a) (b) (c)

Figure 4.3: An example of body partition. (a) Two blocks on masked image to find boundary lines using (4.2). (b) Two boundary lines on masked image. (c) Seven body regions for regional features.

4.2.4 2WGMMF Feature

Since some parts of a human body in an image frame can be occluded by other parts of a human body, and the unseen body parts in an image can be visible in another image when their poses or camera viewpoints are changed. To handle the variation of poses and viewpoints, 2WGMMF feature [80] is thus employed.

The main idea of this feature is that main color modes of the same identity in color histogram domain should be consistent regardless the changes of poses and viewpoints. So 2WGMMF feature represents main color modes of a query person and candidates as GMMs, and computes the two-way distances (*i.e.*, query-to-target and target-to-query) between the color histograms and GMMs. In detail, the feature distance between color histogram of person A and the GMM of person B can be computed by Negative Loglikelihood (NL) as

follows:

$$\begin{aligned} d_{NL}(\mathbf{h}_i^A, G(\mathbf{h}_i^B)) &= -\ln p(\mathbf{h}_i^A|\theta_1^B, ..., \theta_K^B) \\ &= -\ln\left(\sum_{k=1}^{K} \pi_k^B \mathcal{N}\left(\mathbf{h}_i^A|\boldsymbol{\mu}_k^B, \boldsymbol{\Sigma}_k^B\right)\right), \end{aligned} \tag{4.4}$$

where $\mathbf{h} \in \mathbb{R}^{m^c}$ denotes joint color histogram of m-bin and c-channel, which is obtained from either of the body part $i = \{\text{torso}, \text{legs}\}$, $G(\cdot)$ denotes GMM from given color histogram and $\theta_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$ indicates the set of parameters for component k. Moreover, π_k denotes the mixing proportion, $\boldsymbol{\mu}_k \in \mathbb{R}^c$ denotes the mean vector, $\boldsymbol{\Sigma}_k \in \mathbb{R}^{c\times c}$ denotes the covariance matrix and $\mathcal{N}(\cdot)$ denotes the Gaussian distribution. K is the number of Gaussian components, *i.e.*, the number of dominant color modes. Equation (4.4) computes the likelihood function of the GMM of B in response to the histogram of A.

The result from (4.4) is regarded as a one-way distance of i-part, and a small value resulting from (4.4) indicates that they are likely to belong to the same identity. The 2WGMMF feature distance is represented as follows:

$$\begin{aligned} &d_{\text{2WGMMF}}(A, B) = \\ &d_{NL}\Big(\mathbf{h}_{\text{torso}}^A, G(\mathbf{h}_{\text{torso}}^B)\Big) + d_{NL}\Big(\mathbf{h}_{\text{legs}}^A, G(\mathbf{h}_{\text{legs}}^B)\Big) \\ &+ d_{NL}\Big(\mathbf{h}_{\text{torso}}^B, G(\mathbf{h}_{\text{torso}}^A)\Big) + d_{NL}\Big(\mathbf{h}_{\text{legs}}^B, G(\mathbf{h}_{\text{legs}}^A)\Big). \end{aligned} \tag{4.5}$$

Here, we use 32 bins for each channel in RGB color space.

4.2.5 Regional Color and Texture Features

To enhance the ability of ICT through a more detailed comparison, we divide a human torso into multiple regions, since torso part usually carries richest and the most discriminant appearance. After body partition (Section 4.2.2), the torso part is further divided into six equal-size regions (r_1, r_2,..., r_6) as shown in Fig. 4.3(c). Because the region of legs (r_7) usually changes little under different perspectives, we do not further divide the region of legs. Since each specific region normally covers different area of the human torso due to different

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

Figure 4.4: (a) Frame 241 in CAM3 of Dataset3. (b) Seven body regions of Global ID 4 in CAM3. (c) Frame 67 in CAM4 of Dataset3. (d) Seven body regions of Global ID 4 in CAM4.

viewpoints (see Fig. 4.4(b) and 4.4(d)), and as observed in Chu *et al.* [33] that a walking human is usually captured at the similar viewing perspective of body by a fixed camera on either the exit/entrance point, so the histogram extracted from one region of human torso can be modeled as a linear combination of the histograms extracted from multiple regions of human torso in the other camera,

$$\mathbf{h}^A_{map_k} = \left[\mathbf{h}^A_{r_1} \dots \mathbf{h}^A_{r_6}\right] \mathbf{w}_k, \tag{4.6}$$

where $\mathbf{h}^A_{r_k} \in \mathbb{R}^n$ denotes the regional color histogram extracted from the region k of the observation A and $\mathbf{w}_k \in \mathbb{R}^6$ is the mapping matrix of region k for linear combination.

Furthermore, because some regions may be visible only under one cameras view, they should have small weights in the feature distance computation. The distance of regional color feature is the weighted sum of the distances from all seven regions derived from torso and legs as

$$d_{\text{region color}}(A,B) = \sum_{k=1}^{6} q_k \times d\left(\mathbf{h}^A_{map_k}, \mathbf{h}^B_{r_k}\right) + q_7 \times d\left(\mathbf{h}^A_{r_7}, \mathbf{h}^B_{r_7}\right), \tag{4.7}$$

where $\mathbf{q}= [q_1 \dots q_7]^T$ denote the weights for all seven region distances. The computation of region matching weights is discussed in Section 5.3.4. Note that all the seven regions are included in the feature distance computation, however only the torso regions are considered for the region mapping by using the mapping matrix $\mathbf{W}_{map} = [\mathbf{w}_1 \dots \mathbf{w}_6]$.

The texture feature distance can be computed similarly to that of the color feature. The Local Binary Pattern (LBP) [81] is exploited as the texture feature and is represented as l-dimensional LBP histograms of observation A, $\mathbf{h}^A_{r\text{LBP}_k} \in \mathbb{R}^l$, in which k is from 1 to 7. Hence, the distance of regional texture feature is

$$d_{\text{region texture}}(A,B) = \sum_{k=1}^{6} q_k \times d\left(\mathbf{h}^A_{map\text{LBP}_k}, \mathbf{h}^B_{r\text{LBP}_k}\right) + q_7 \times d\left(\mathbf{h}^A_{r\text{LBP}_7}, \mathbf{h}^B_{r\text{LBP}_7}\right), \tag{4.8}$$

where $\mathbf{h}^A_{map\text{LBP}_k} \in \mathbb{R}^l$ is the linear combination of torso region LBP histograms with the same weights $\mathbf{w}_k$ defined in (4.6) as follows:

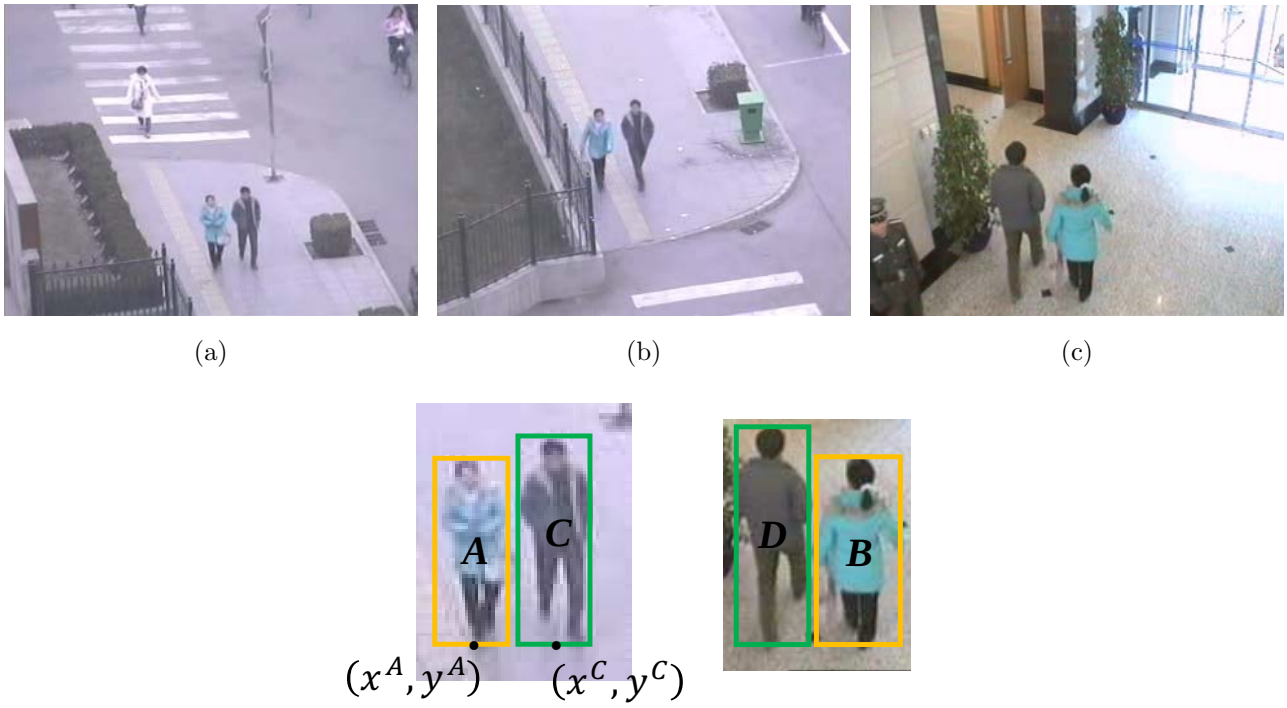
$$\mathbf{h}^A_{map\text{LBP}_k} = \left[\mathbf{h}^A_{r\text{LBP}_1} \dots \mathbf{h}^A_{r\text{LBP}_6}\right] \mathbf{w}_k. \tag{4.9}$$

4.2.6 Couple Feature

We present a simple and effective group feature to improve the accuracy of ICT. Figures 4.5(a)-(c) show examples of a couple on three different cameras. In this paper, a couple is defined as a pair of persons traveling together through the scene and formulated as

$$\left|x^A - x^C\right| < \delta_x, \quad \left|y^A - y^C\right| < \delta_y, \tag{4.10}$$

$$\left|t^A_{\text{exit}} - t^C_{\text{exit}}\right| < \delta_t, \quad \left|t^A_{\text{entry}} - t^C_{\text{entry}}\right| < \delta_t, \tag{4.11}$$



(a) (b) (c)

(d) (e)

Figure 4.5: Examples of couple across multi-cameras. (a) Couple in CAM1. (b) Couple in CAM2. (c) Couple in CAM3. (d)(e) Cropped and enlarged couples in CAM1 and 3 (A-B and C-D denote the same person, respectively).

where x and y denote the 2-D coordinate of center of bottom line of the bounding box (see Fig. 4.5(d)), and t_{exit} and t_{entry} denote time stamps when the person exits and enters FOV, respectively. With these spatio-temporal conditions, couples are detected in each camera.

To identify the same couple across cameras, 2WGMMF feature is again utilized as

$$\begin{aligned} & d_{\text{couple identifier}}(AC, BD) \\ & = \min\left(d_{\text{2WGMMF}}(A, B), d_{\text{2WGMMF}}(A, D)\right) \\ & + \min\left(d_{\text{2WGMMF}}(C, B), d_{\text{2WGMMF}}(C, D)\right). \end{aligned} \tag{4.12}$$

In phase I, it is the negative of 2WGMMF feature distance between one and couple person

of target that is used as follows,

$$\begin{aligned} d^{\mathrm{I}}_{\mathrm{couple}}(A,B) &= -d_{\mathrm{2WGMMF}}(A, B_{\mathrm{couple}}) \\ &= -d_{\mathrm{2WGMMF}}(A, D), \end{aligned} \tag{4.13}$$

and the couple feature distance exploits other feature distances to match person-to-person in a couple in phase II. The combination of four feature distances with feature fusion weights in phase II is shown as follows,

$$\begin{aligned} & d^{\mathrm{II}}_{\mathrm{couple}}(A,B) = \\ & -\alpha_1 d^{Norm}_{\mathrm{2WGMMF}}(A,D) - \alpha_2 d^{Norm}_{\mathrm{holistic\ color}}(A,D) \\ & -\alpha_3 d^{Norm}_{\mathrm{region\ color}}(A,D) - \alpha_4 d^{Norm}_{\mathrm{region\ texture}}(A,D), \end{aligned} \tag{4.14}$$

where α_j denote feature fusion weights (see Section 5.3.5). Note that (19) only shows a scenario of Comb1 (see Table II for different feature combinations). Moreover, to normalize feature distance, min-max normalization is used:

$$d^{Norm}_j(A,D) = \frac{d_j(A,D) - \min d_j}{\max d_j - \min d_j}, \tag{4.15}$$

where $\min d_j$ and $\max d_j$ represent the smallest and largest values of each feature j's distance, respectively. They are obtained from computing feature fusion weights with training data. In our experiments, δ_x, δ_y and δ_t are set to 15 empirically.

4.2.7 Final Score

In order to further improve the discriminative power, we utilize a combination of features for distance measures. Since the value range of each feature distance is different, we use normalization and fusion methods to get the final score. Feature fusion weights are derived by exploiting d-prime metric [82] (see Section 5.3.5). In phase I, the final score is a combination of 2WGMMF and couple features as follows:

$$d^{\mathrm{I}}_{\mathrm{Final}}(A,B) = d_{\mathrm{2WGMMF}}(A,B) + d^{\mathrm{I}}_{\mathrm{couple}}(A,B). \tag{4.16}$$

In phase II, final score is a combination of normalized feature distances of 2WGMMF, holistic color, regional color/texture and couple features with feature fusion weights. The following is formulation of the first category of combinations, Comb1:

$$\begin{aligned} d^{\mathrm{II}}_{\mathrm{Final}}(A,B) &= \alpha_1 d^{Norm}_{\mathrm{2WGMMF}}(A,B) \\ &+ \alpha_2 d^{Norm}_{\mathrm{holistic\ color}}(A,B) + \alpha_3 d^{Norm}_{\mathrm{region\ color}}(A,B) \\ &+ \alpha_4 d^{Norm}_{\mathrm{region\ texture}}(A,B) + d^{\mathrm{II}}_{\mathrm{couple}}(A,B). \end{aligned} \tag{4.17}$$

4.3 Experimental Results

This section presents the evaluation results of our approaches on the benchmark dataset, NLPR_MCT [8], which is collected for multi-camera pedestrian tracking over non-overlapping cameras.

4.3.1 Dataset and Evaluation Criteria

The NLPR_MCT dataset consists of four sub-datasets. Every sub-dataset includes 3-5 cameras with non-overlapping FOVs and details of them are summarized in Table 4.1, where GT^s is the number of ground truths in a single camera and GT^c is the number of ground truths across cameras. FOVs and the topological relationships of all the cameras are shown in Fig. 4.6. We assume that the connectivity between entry/exit zones in multiple cameras has already been specified [32, 35, 79].

The evaluation metric adopted is called Multi-Camera object Tracking Accuracy (MCTA) [45]:

$$\begin{aligned} &\mathrm{MCTA} \\ &= \mathit{Detection} \times \mathit{Tracking}^{\mathrm{SCT}} \times \mathit{Tracking}^{\mathrm{ICT}} \\ &= \left(\frac{2 \times \mathit{Precision} \times \mathit{Recall}}{\mathit{Precision} + \mathit{Recall}}\right)\left(1 - \frac{\sum_t mme^s_t}{\sum_t tp^s_t}\right)\left(1 - \frac{\sum_t mme^c_t}{\sum_t tp^c_t}\right) \\ &= \mathrm{SCTA} \times \left(1 - \frac{\sum_t mme^c_t}{\sum_t tp^c_t}\right), \end{aligned} \tag{4.18}$$

where mme_t and tp_t denote the number of mismatches and ground truths, respectively at time t. MCTA ranges from 0 to 1, and the higher value indicates better performance. The metric

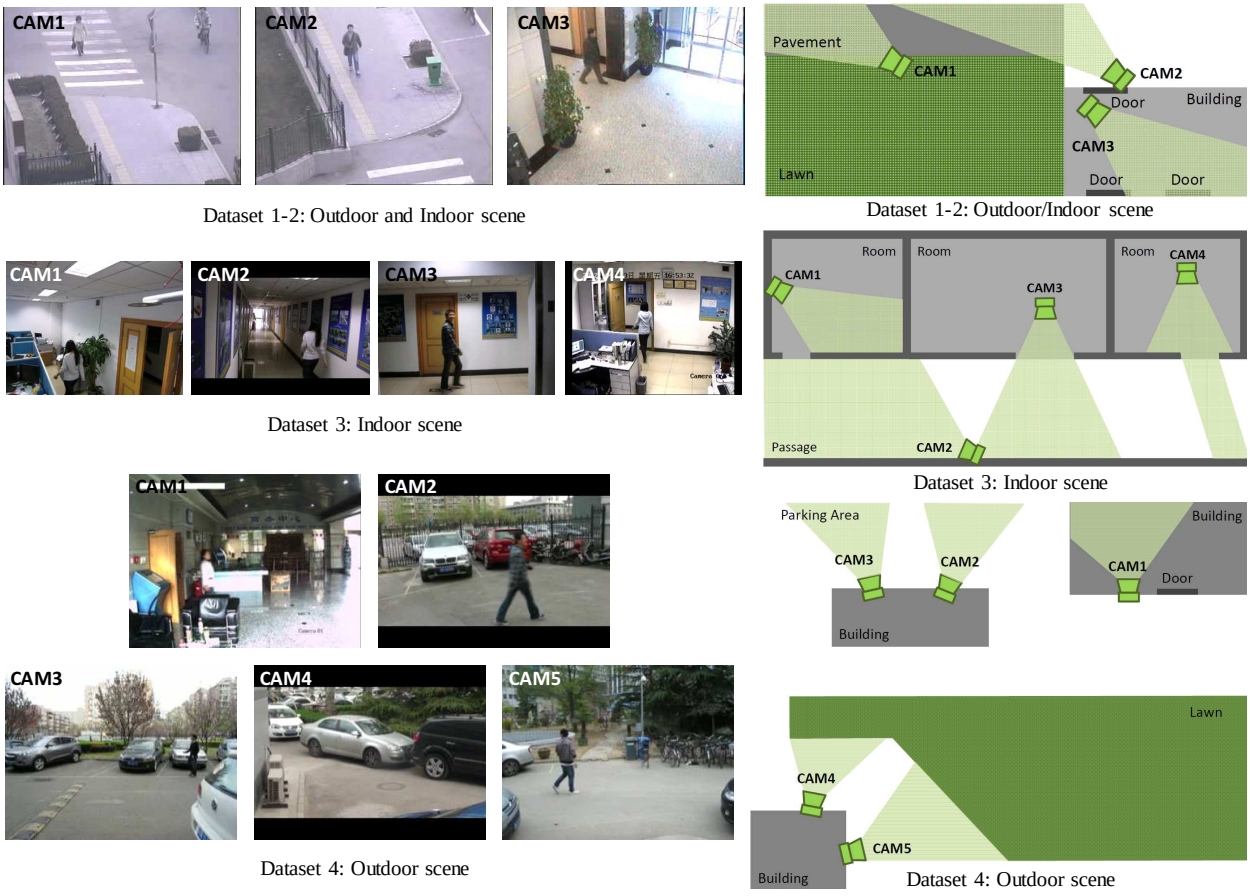


Figure 4.6: Illustration of the topological relationship during tracking.

Table 4.1: Details of NLPR_MCT Dataset [8].

Sub-dataset	Dataset1	Dataset2	Dataset3	Dataset4
# of cameras	3	3	4	5
Duration	20 min	20 min	3.5 min	24 min
Resolution	320×240	320×240	320×240	320×240
Frame rate	20 fps	20 fps	25 fps	25 fps
# of persons	235	255	14	49
GT^s	71853	88419	18187	42615
GT^c	334	408	152	256

Table 4.2: Description of feature combination in evaluation.

Denotation	Feature combination
Comb1	2WGMMF, holistic color, regional color/texture, couple
Comb2	2WGMMF, regional color/texture, couple
Comb3	2WGMMF, holistic color, couple
Comb4	holistic color, regional color/texture, couple

can be divided into three parts, detection, SCT and ICT abilities, which are corresponding to the three brackets in (4.18). *Detection* in (4.18) is also known as F_1-score, which reaches its best value at 1 and worst at 0. The new object is counted as inter-camera ground truth, tp_t^c, by default in this criterion. The evaluation kit is available in [8]. Moreover, to evaluate the performance of SCT specifically, we define the Single-Camera object Tracking Accuracy (SCTA) by discarding the term of $Tracking^{\mathrm{ICT}}$.

4.3.2 Tracking results

We have experiments based on the ground-truth of SCT. More specifically, *Detection* and $Tracking^{\mathrm{SCT}}$ are both 1, thus, MCTA depends only on mme_t^c, which represents the number of mismatches in time t across different cameras in (4.18). In Table 4.3, the experimental results of the proposed method are compared with the state-of-the-art [35,44,45,48]. Chen *et al.* [48] formulate ICT as an inference problem using the CRF framework. They first obtain the initial labels using Hungarian algorithm. Then, a global appearance model and an online learned target-specific appearance model using AdaBoost are combined with grouping information as high-level context feature to formulate the tracking task as an energy minimization problem. The problem is solved by their proposed iterative approximation algorithm. Our proposed method achieves the best result as shown in Table 4.3. With regard to average MCTA, Comb1 to Comb4 all perform better than the state-of-the-art methods. Moreover, Comb1 outperforms the other methods in every sub-dataset.

Table 4.3: Performance comparison of inter-camera tracking with ground-truth single camera tracking. The best results are highlighted in colors (Underlined red font is rank-1 and *italicized green* font is rank-2).

Sub-dataset	Evaluation metric	Comb1	Comb2	Comb3	Comb4	USC-Vision [35]	CRF [48]	NLPR [45]	CRIPAC-MCT [44]	Hfutdsp-mct
Dataset1	mme^c	13	14	19	17	27	54	55	113	86
	MCTA	0.9611	*0.9581*	0.9431	0.9491	0.9152	0.8383	0.8353	0.6617	0.7425
Dataset2	mme^c	30	46	31	36	34	81	121	167	141
	MCTA	0.9265	0.8873	*0.9240*	0.9118	0.9132	0.8015	0.7034	0.5907	0.6544
Dataset3	mme^c	32	35	41	36	70	51	39	44	40
	MCTA	0.7895	*0.7697*	0.7303	0.7632	0.5163	0.6645	0.7417	0.7105	0.7368
Dataset4	mme^c	62	69	72	80	72	70	157	110	155
	MCTA	0.7578	*0.7305*	0.7188	0.6875	0.7052	0.7266	0.3845	0.5703	0.3945
Average MCTA		0.8587	*0.8364*	0.8291	0.8279	0.7625	0.7577	0.6662	0.6333	0.6321

Table 4.4: Performance comparison of inter-camera tracking with single features. The best results are highlighted in colors (Underlined red font is rank-1 and *italicized green* font is rank-2).

		Holistic color		2WGMMF		Regional color		Regional texture	
Sub-dataset	Evaluation metric	SuB-SENSE	Proposed	SuB-SENSE	Proposed	SuB-SENSE	Proposed	SuB-SENSE	Proposed
Dataset1	mme^c	34	25	35	23	36	24	44	37
	MCTA	0.8982	0.9132	0.8952	0.9311	0.8922	*0.9281*	0.8683	0.8892
Dataset2	mme^c	52	49	60	55	82	64	88	78
	MCTA	*0.8725*	0.8800	0.8529	0.8652	0.7990	0.8431	0.7843	0.8088
Dataset3	mme^c	63	59	46	42	77	45	77	45
	MCTA	0.5855	0.6118	0.6974	0.7237	0.4934	*0.7039*	0.4934	*0.7039*
Dataset4	mme^c	95	87	75	72	87	73	94	90
	MCTA	0.6289	0.6602	0.7070	0.7188	0.6602	*0.7148*	0.6328	0.6484
Average MCTA		0.7457	0.7663	0.7881	0.8097	0.7112	*0.7975*	0.6947	0.7587

Table 4.5: Performance comparison of couple feature.

Sub-dataset	Evaluation metric	2WGMMF		Comb1	
		w/o couple	w/ couple	w/o couple	w/ couple
Dataset1	mme^c	23	19	20	13
	MCTA	0.9311	0.9431	0.9401	0.9611
Dataset2	mme^c	55	37	46	30
	MCTA	0.8652	0.9093	0.8873	0.9265

To validate the effectiveness of every single feature towards the final results, we also compare the performance of them in Table 4.4. Compared to the performance of their combinations in Table 4.3, the performance based on each individual feature are worse. Note that 2WGMMF feature shows the best performance and regional color feature is the second-best in terms of average MCTA. However, the performance of holistic color feature is better than 2WGMMF feature in Dataset2 because many people are crossing the cameras, between CAM1 and CAM2, which have similar viewpoints. In addition, the performance of the proposed segmentation method and SuBSENSE are compared. From the performance of every single feature, it can be seen that the proposed segmentation gives better results. Because erroneously included background or cropped body part makes feature representation inaccurate, more precise masks lead to better ICT results. Specifically, the performances of regional color and texture features with SuBSENSE segmentation are much degraded. Because regional color and texture features are extracted from small areas comparably, they are more sensitive to the accuracy of segmentation.

In Table 4.5, the performance of couple feature in ICT is compared with 2WGMMF feature and Comb1. In all cases, the performance is improved when combining with the couple feature. Since couples only appear in Dataset1 and 2 only, there are no such comparisons for Dataset3 and Dataset4. USC-Vision [35] exploits relative appearance context learning, which

Table 4.6: Performance comparison of feature fusion weights and uniform weights on Comb1.

Sub-dataset	Evaluation metric	Uniform weights	Feature fusion weights
Dataset1	mme^c	19	13
	MCTA	0.9431	0.9611
Dataset2	mme^c	33	30
	MCTA	0.9191	0.9265
Dataset3	mme^c	37	32
	MCTA	0.7566	0.7895
Dataset4	mme^c	70	62
	MCTA	0.7266	0.7578
Average MCTA		0.8364	0.8587

is motivated by the fact that the same sets of people tend to re-appear in the neighboring camera. However, it is not applicable when FOV is limited with few people being captured. As a result, their performance is the worst in Dataset3 compared to the other methods in Table IV.

For showing the effect of feature fusion weights, we put uniform weights in (4.17) in case of Comb1 and the experimental results are shown in Table 4.6. With feature fusion weights, the overall accuracy is enhanced. It is because some features become more discriminative and other features are less effective according to the relative difference of camera viewpoint. For example, the link between CAM1 and CAM2 is almost the same in Dataset1 and Dataset2 (see Fig. 4.6). Then, the holistic color histogram is the most effective and its weight becomes larger. However, in case of the link between CAM2 and CAM3, or CAM3 and CAM4 in Dataset3, regional features, *i.e.*, 2WGMMF and regional color/texture features, should have larger weights.

Chapter 5

ONLINE-LEARNING-BASED MULTIPLE-CAMERA HUMAN TRACKING ACROSS NON-OVERLAPPING CAMERAS

5.1 Overview

Due to the expanding scale of camera networks, multiple camera tracking of human has received higher attention in recent years. In this chapter, we present a novel approach to track each human within a single camera and across multiple disjoint cameras. Our framework includes a multi-object tracking and segmentation system, a two-phase feature extractor, and an online-learning-based camera link model estimation. For tracking within a single camera, we apply tracking by segmentation and local object detection with multi-kernel feedback to adaptively improve the robustness of the algorithm. In inter-camera tracking, we introduce an effective integration of appearance and context features. Automatically couples are detected, and the couple feature is also integrated with existing features. The proposed algorithm is scalable by a fully unsupervised online learning framework. In our experiments, the proposed method outperforms all the state-of-the-art in the benchmark NLPR_MCT dataset.

5.2 Single-Camera Tracking and Object Segmentation

Since both accurate segmentation and SCT results are necessary for supporting ICT, we develop a robust tracking and segmentation system to achieve the goal. The proposed system is coined as the MAST, short for "Multi-kernel Adaptive Segmentation and Tracking", because we make use of multi-kernel feedback to adaptively control the thresholding parameters in segmentation for preserving more foreground around object region. Figure 5.1 shows the overview flow diagram of the MAST architecture. Note that this framework is extendable for

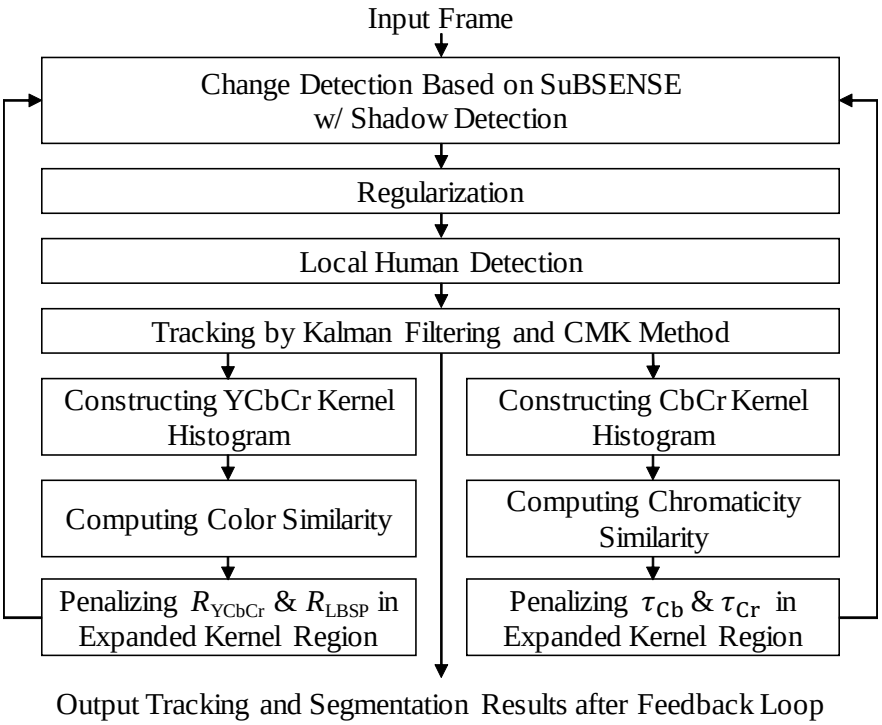


Figure 5.1: Flow diagram of MAST for SCT and segmentation. The role of each block is detailed in Section 5.2.

the use of any object segmentation method, tracking-by-segmentation method, and object detection method.

To begin with, the state-of-the-art change detection scheme, SuBSENSE [54], is adopted for object segmentation. Each pixel in the input frame is represented by color (here we choose to use YCbCr space instead of RGB space since it will facilitate the process of shadow detection) and Local Binary Similarity Patterns (LBSP) feature [83]. The background model is constructed by a set of background samples $B_n(x, y)$ at each pixel location (x, y), which is updated according to an automatically adjusted learning rate. When each new pixel arrives for background/foreground classification, it will be compared with all background samples at the corresponding location. The comparisons are based on two distance thresholds, R_{YCbCr}

and R_{LBSP}, in the color space and feature space, respectively. If the number of matching samples (with sufficiently short distance to the input pixel) is smaller than a specific minimum, the pixel is labeled as foreground. To further enhance robustness of SuBSENSE, we add a shadow detection block based on YCbCr color space that starts to function after a pixel is classified as foreground,

$$Q_t(x,y) = \begin{cases} 1, & \#\{\left(\alpha_{\text{Y}} \leq I_t^{\text{Y}}(x,y)/B_n^{\text{Y}}(x,y) \leq \beta_{\text{Y}}\right) \\ & \wedge \left(\left|I_t^{\text{Cb}}(x,y) - B_n^{\text{Cb}}(x,y)\right| \leq \tau_{\text{Cb}}\right) \\ & \wedge \left(\left|I_t^{\text{Cr}}(x,y) - B_n^{\text{Cr}}(x,y)\right| \leq \tau_{\text{Cr}}\right), \forall n\} \\ & > N_{\max}^{Q} \\ 0, & \text{otherwise} \end{cases} \tag{5.1}$$

where $Q_t(x,y)$ indicates shadow when the value is 1, $I_t(x,y)$ is a pixel from current frame t, the superscripts of $I_t(x,y)$ and $B_n(x,y)$ indicate the YCbCr channels, $N_{\max}^{Q}$ is the maximum number of matches required for the shadow detection, and α_{Y}, β_{Y}, τ_{Cb} and τ_{Cr} are the thresholds for their corresponding color channels. If a pixel is detected as shadow, it is discarded from foreground and will be used for updating the background model. After segmentation, morphological operations, *e.g.*, closing, opening, and flood-filling, are further applied on the derived foreground mask for shape refinement.

In the segmented foreground, each object blob may contain more than one target, *i.e.*, the problem of initial occlusion. Thus, a Histogram of Oriented Gradient (HOG) human detector [9] is run on the cropped frame image within each object bounding box. If multiple targets are detected and their overlapping area with each other is small enough, they will be initialized separately for SCT. Different from traditional tracking by detection that needs to process each entire frame image, the computation complexity is much reduced since only the local region around each foreground blob is considered. Other than HOG human detector, we have also tested Deformable Part Model (DPM) human detector [10] and C^4 pedestrian detector [84]. HOG is chosen for its simplicity and efficiency.

Based on the initialization of object positions from segmentation and local object detection, we can start tracking each target. The preliminary tracking results are generated by

the method proposed by Chu *et al.* that combines Kalman filtering and CMK tracking [51]. Kalman filter prediction is first conducted on all the objects tracked in the previous frame. Then we detect whether there is an abnormality in size change of each foreground blob, which can be caused by occlusion or failure in segmentation. The abnormal targets and those initialized by object detection are tracked by the CMK method, which relies on multiple inter-related kernels to represent different parts of human, so that we can add weights of trust on different kernels depending on their severity of occlusion. Multiple measurements are produced from CMK tracking that is handled by probabilistic data association. On the other hand, the normal foreground blob with a single object is directly selected as the measurement for Kalman filtering.

From preliminary tracking results, we follow the concept of multiple kernels to measure the similarity between current frame and background in object regions. In our experiments, each human target is described by two kernels that cover half of his/her body on the top and bottom respectively, as people usually wear differently in these two body parts. Moreover, since the bounding box of each object may include background area, we can use kernel histogram to emphasize on the central region that the object occupies. Two kernel histograms are constructed within each kernel region for both current frame and background model: one of them is built in the YCbCr color space, and the other only uses the Cb and Cr channels to represent the chromaticity information. Note that the kernel histograms for background use all background samples and is normalized for comparison. To emphasize the object region that usually covers the central area of each kernel, a Gaussian kernel function is added for constructing kernel histograms,

$$w_{\text{ker}} = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\frac{(x-x_m)}{2\sigma_x^2} - \frac{(y-y_m)}{2\sigma_y^2}\right], \tag{5.2}$$

in which σ_x and σ_y are set as half of the width and height of the kernel bounding box respectively, while x_m and y_m locate the mean point of the foreground shape within the kernel.

Afterwards, the color similarity and chromaticity similarity are computed as the recipro-

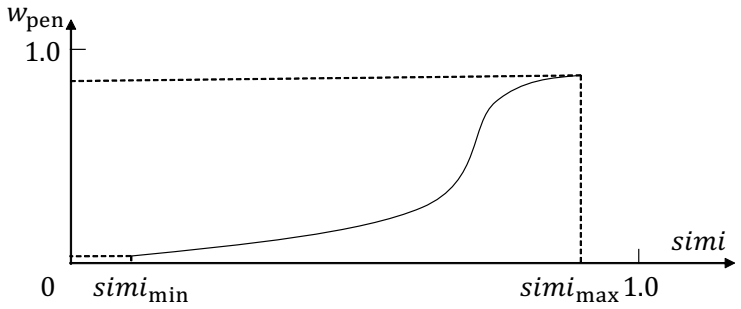


Figure 5.2: The shape of fuzzy Gaussian penalty weighting function for adaptation of thresholding parameters in object segmentation.

cals of Bhattacharyya distances [85] between corresponding kernel histograms, *i.e.*,

$$simi_{\text{color}} = 1 \Big/ \sum\nolimits_c \sqrt{h^I_{\text{YCbCr}}(c) \cdot h^{BG}_{\text{YCbCr}}(c)}, \tag{5.3}$$

$$simi_{\text{chrom}} = 1 \Big/ \sum\nolimits_c \sqrt{h^I_{\text{CbCr}}(c) \cdot h^{BG}_{\text{CbCr}}(c)}, \tag{5.4}$$

where superscripts I and BG denote the kernel histograms in current frame and background respectively, and c is the index of channel bin. We have also tested other measurements such as correlation, Kullback-Leibler (KL) distance [66], dual KL distance [4], schemes in Automatic Reference Color Selection (ARCS) [86], etc. The Bhattacharyya distance is selected for its superior performance in our scenarios. The higher the color similarity of object region with the background, the more likely the object will mistakenly merge into the background during segmentation. Likewise, if the object region shares high similarity in chromaticity with the background, *e.g.*, a human wearing black pants is walking on a grey ground plane, it is easy for his/her body parts to be wrongly recognized as shadow and removed from the foreground. Next, a second segmentation using thresholding parameters penalized by $simi_{\text{color}}$ and $simi_{\text{chrom}}$ is performed in order to preserve more foreground in the local region around tracking targets. Under the consideration of smoothness of segmentation, the penalty weights on segmentation thresholds are computed by a fuzzy Gaussian penalty weighting function

(a) (b)

Figure 5.3: Comparison of segmentation performance. (a) Segmentation from the preliminary result of SuBSENSE with shadow detection. (b) Segmentation after the application of multi-kernel feedback loops (foreground in red, and detected shadow in blue).

as shown in (5.5),

$$w_{\text{pen}} = \begin{cases} \exp\left[-\frac{9\cdot(1.0-simi)^2}{4\cdot(1.0-simi_{\text{min}})^2}\right], & simi_{\text{min}} \leq simi \leq simi_{\text{max}} \\ 0, & \text{otherwise} \end{cases} \tag{5.5}$$

in which $simi$ is the color or chromaticity similarity computed from (5.3) or (5.4), respectively, while $simi_{\text{min}}$ and $simi_{\text{max}}$ indicate the region of $simi$ value to perform re-segmentation. As shown in the fuzzy Gaussian curve in Fig. 5.2, when $simi$ is smaller than the lower bound $simi_{\text{min}}$, the preliminary segmentation is considered successful, and there is no need for further adaptation, which is based on the concept of fuzzy set. On the contrary, if $simi$ is too large, it is highly likely to be caused by tracking error, where CMK tracking wrongly shifts to a background area. Hence, to prevent propagation of errors, an upper bound $simi_{\text{max}}$ is necessary for the similarity between current frame and background. The w_{pen} computed based on $simi_{\text{color}}$ is used to penalize R_{YCbCr} and R_{LBSP} in SuBSENSE, while the one for $simi_{\text{chrom}}$ is applied on τ_{Cb} and τ_{Cr} in shadow detection, where the penalization is defined by multiplying $(1 - w_{\text{pen}})$. Meanwhile, since the preliminary foreground

blob may fail to cover the entire object body, the kernel region to conduct re-segmentation is expanded by a factor of $w_{\text{pen}}/2$. In summary, the adaptive segmentation is operated in a larger kernel region with lower thresholds for background subtraction and less shadow detected, therefore, the segmented foreground area is expanded to maintain continuity of tracking by segmentation. The final foreground mask is created by a union combination of the first segmentation across the entire frame and local adaptive segmentation in selected kernel regions.

Lastly, the tracking module is called again to generate the final tracking results from the updated foreground mask. Note that Kalman filter update is not performed until after re-segmentation. The optimized segmentation results will also be used in ICT for feature extraction. The superiority of adding multi-kernel feedback loops to adaptively control the segmentation thresholds can be seen from Fig. 5.3(b), in which more foreground belonging to the target is retained, compared to Fig. 5.3(a), when the chromaticity of her clothing is similar to background.

5.3 Camera Link Model Estimation

After collecting some video samples online (Section 5.3.1), camera link models including transition time distributions for time window (Section 5.3.2), region mapping matrix (Section 5.3.3), region matching weights (Section 5.3.4), and feature fusion weights (Section 5.3.5) are estimated, and phase change occurs.

5.3.1 Online Sample Collection

In a FOV of a surveillance camera, a pedestrians appearance is usually captured in dozens of frames. So one good matching pair in ICT is equal to dozens of positive samples. If we have two good matching pairs, *e.g.*, *A*-*C* and *B*-*D*, we can collect negative samples as well by cross-matching pairs, *e.g.*, *A*-*D* and *C*-*B*.

In our framework, good matching pairs are determined by the distance value calculated in (4.16). According to the characteristics of the 2WGMMF feature, it has a negative value

when query and target images are very similar. In other words, there is rarely false positive pairs with a negative value of the 2WGMMF feature. Thus, good matching pairs are selected as follows:

$$d^{\mathrm{I}}_{\mathrm{Final}} < 0, \tag{5.6}$$

for all the camera links. After obtaining two good matching pairs, we can start to build camera link models and update them by adding good matching pairs continuously.

5.3.2 Estimation of Time Window

People tend to walk in similar paths in most cases considering available pathways, obstructs, and shortest routes. Hence, the transition time t forms a certain distribution, and it can be exploited to infer and model the camera network topology [32, 42, 43]. We utilize the estimated distribution to determine the time window, which helps to reduce the number of candidates. Transition time distribution is modeled as a Gaussian distribution,

$$f(\forall t \in \mathbf{T} | \mu_T, \sigma_T) = \frac{1}{\sigma_T \sqrt{2\pi}} e^{-\frac{(t-\mu_T)^2}{2\sigma_T^2}}, \tag{5.7}$$

where $\mathbf{T} = [t_1, ..., t_N]$ represents a set of valid transition time values, μ_T and σ_T indicate the mean and standard deviation of transition time distribution, respectively. Before having transition time distributions, we fix the length of time window $\tau = 120$ seconds, which can be set as different values for different known camera topologies. In all experiments, after estimating them, we set $\tau = \mu_T \pm 6\sigma_T$.

5.3.3 Estimation of Region Mapping Matrix

Both the regional color and texture features are exploited to estimate region mapping matrix as follows:

$$\mathbf{R}^A = \begin{bmatrix} \mathbf{h}^A_{r_1} \dots \mathbf{h}^A_{r_6} \\ \mathbf{h}^A_{r\mathrm{LBP}_1} \dots \mathbf{h}^A_{r\mathrm{LBP}_6} \end{bmatrix} = [\mathbf{r}_1 \dots \mathbf{r}_6], \tag{5.8}$$

where $\mathbf{r}_k \in \mathbb{R}^{n+l}$ for $k = 1, 2, ..., 6$. We minimize the following objective function to get each vector $\bar{\mathbf{w}}_k$,

$$\begin{aligned} \mathbf{w}_k &= \arg\min_{\bar{\mathbf{w}}_k} g_{\mathbf{w}}(\bar{\mathbf{w}}_k) \\ \text{s.t.} \quad g_{\mathbf{w}}(\bar{\mathbf{w}}_k) &= \sum_{i=1}^{N_{\text{exit}}} \sum_{j=1}^{N_{\text{entry}}} \left\| \mathbf{R}_j^A \bar{\mathbf{w}}_k - r_{ik}^A \right\|_2^2, \\ \bar{\mathbf{w}}_k &\geq 0, \quad \|\bar{\mathbf{w}}_k\|_1 = 1, \end{aligned} \tag{5.9}$$

where N_{exit} and N_{entry} denote the number of exiting and entering observations respectively.

5.3.4 Estimation of Region Matching Weights

The matching weights method in [82] is employed to determine the region matching weights, which are inversely proportional to the corresponding estimation error as follows:

$$q_k = \frac{1/\varepsilon_k^{error}}{\sum_{i=1}^{7} 1/\varepsilon_i^{error}} \quad k = 1 - 7, \tag{5.10}$$

where the estimation error of the mapping vector is defined as $\varepsilon_k^{error} = g_{\mathbf{w}}(\mathbf{w}_k)$ for $k = 1, 2, ..., 6$ and $\varepsilon_7^{error} = \sum_{i=1}^{N_{\text{exit}}} \sum_{j=1}^{N_{\text{entry}}} \left\| \mathbf{r}_{j7}^A - \mathbf{r}_{i7}^B \right\|_2^2$. A large weight implies that the body region in one camera is well visible in the other camera as well.

5.3.5 Estimation of Feature Fusion Weights

Since there are four features, holistic color, 2WGMMF, region color, and region texture features, used in ICT, we need an efficient method to fuse them together. The feature fusion weights are systematically determined based on the degree of separation between the distributions of the values in the positive and negative sets. The separation is measured by the d-prime metric [82, 87],

$$d_j = \frac{\mu_j^{\text{N}} - \mu_j^{\text{P}}}{\sqrt{\left(\sigma_j^{\text{N}}\right)^2 + \left(\sigma_j^{\text{P}}\right)^2}}, \tag{5.11}$$

where μ_j and σ_j denote the mean and standard deviation of the distribution of the feature distance for each feature j; and superscripts P and N represent positive and negative sets,

respectively. The feature fusion weights, α_j, $j = 1, ..., 4$, are calculated as

$$\alpha_j = \frac{d_j}{\sum_{i=1}^{4} d_i}. \tag{5.12}$$

5.4 Experimental Results

To evaluate the performance, several combinations of the proposed features are compared with the state-of-the-art methods [35, 44, 45, 88] in Table 5.1. Experimental dataset and evaluation metric is the same as in Chapter 4.3.1. The details of feature combinations are described in Table 4.2. USC-Vision team [35, 88] is the winner of the MCT challenge in conjunction with ECCV 2014, and they utilize a two-step approach. They employ a four-body-part-based pedestrian detector [89] and a detection-based three-level hierarchical association approach [88, 90] for SCT. During the three-level association, detections are connected into the final trajectories by selecting detections discretely and complementing some missing detections. For ICT, they use two kinds of context information, spatio-temporal context, and relative appearance context, to improve the ICT performance [35]. Spatio-temporal context is used to collect positive/negative training samples for each tracked object. Relative appearance context is used to model inter-object appearance similarities between the query person and the people in the neighboring set, which is defined as other people who enter/exit this zone in a time window with the query person. NLPR team [45] uses the DPM detector [10] to get the detection results, and applies an equalized global graph model that combines PMC-SHR with an accurate similarity equalizer to compensate the weak invariance of appearance representation for ICT. CRIPAC-MCT team [44] exploits a head-shoulder detector [91] and an adaptive integrated feature (AIF) tracker [92] to get all the tracklets from each camera. Then, tracklets are merged into trajectories by a global tracklet association, which models ICT as a global MAP problem. Hfutdspmct team utilizes the Visual Background extractor (ViBe) algorithm [93] for detecting foreground. For SCT, their method is based on center location bi-directional matching. As for ICT, adjacency constrained patch matching, as well as bi-directional weighted matching, are applied.

Table 5.1: Performance comparison of multiple camera tracking without ground-truth of object detection. The best results are highlighted in colors (Underlined red font is rank-1 and *italicized green* font is rank-2).

Sub-dataset	Evaluation metric	Comb1	Comb2	Comb3	Comb4	USC-Vision [88] + [35]	NLPR [45]	Hfutdsp-mct	CRIPAC-MCT [44]
Dataset1	*Precision*	0.7724				0.6916	0.7967	0.7113	0.1488
	Recall	0.6088				0.6061	0.5929	0.3465	0.2154
	Detection	0.6809				0.6460	0.6799	0.4660	0.1760
	$Tracking^{SCT}$	0.9981				0.9981	0.9744	0.9229	0.9955
	SCTA	0.6796				0.6448	*0.6625*	0.4301	0.1752
	$Tracking^{ICT}$	0.8851	0.8851	0.8665	0.8789	0.9288	0.6220	0.6534	0.7111
	MCTA	0.6015	0.6015	0.5889	0.5973	*0.5989*	0.4120	0.2810	0.1246
Dataset2	*Precision*	0.8334				0.6948	0.7977	0.7461	0.1431
	Recall	0.7091				0.7843	0.6332	0.3669	0.1933
	Detection	0.7662				0.7368	0.7060	0.4919	0.1645
	$Tracking^{SCT}$	0.9991				0.9986	0.9779	0.9347	0.9945
	SCTA	0.7655				*0.7358*	0.6904	0.4598	0.1636
	$Tracking^{ICT}$	0.8842	0.8793	0.8818	0.8768	0.8691	0.6942	0.6122	0.7510
	MCTA	0.6769	0.6732	*0.6751*	0.6713	0.6260	0.4793	0.2815	0.1075
Dataset3	*Precision*	0.6597				0.4750	0.8207	0.3342	0.0853
	Recall	0.7260				0.6615	0.5345	0.0986	0.1206
	Detection	0.6913				0.5529	0.6474	0.1523	0.0999
	$Tracking^{SCT}$	0.9864				0.9904	0.9749	0.9682	0.9715
	SCTA	0.6819				0.5476	*0.6312*	0.1475	0.0971
	$Tracking^{ICT}$	0.5461	0.5329	0.5329	0.5000	0.1014	0.2953	0.2432	0.1143
	MCTA	0.3724	*0.3634*	*0.3634*	0.3410	0.0555	0.1864	0.0359	0.0111
Dataset4	*Precision*	0.8758				0.5216	0.8355	0.7720	0.0606
	Recall	0.8600				0.7938	0.6193	0.1210	0.0944
	Detection	0.8678				0.6295	0.7113	0.2092	0.0738
	$Tracking^{SCT}$	0.9977				0.9948	0.9275	0.9865	0.9762
	SCTA	0.8658				0.6262	*0.6597*	0.2064	0.0720
	$Tracking^{ICT}$	0.6270	0.6151	0.5992	0.6071	0.5437	0.4308	0.2944	0.2950
	MCTA	0.5429	*0.5326*	0.5188	0.5257	0.3404	0.2842	0.0608	0.0213
Average MCTA		0.5484	*0.5427*	0.5366	0.5338	0.4052	0.3405	0.1648	0.0661

In Table 5.1, all the combinations, Comb1 to Comb4, of our proposed algorithm outperform the state-of-the-art methods in terms of average MCTA. Especially, Comb1, which has a combination of all the features, shows the best results. It proves that the proposed algorithm is robust in various environments.

We adopt the default setting of parameters in [14] for SCT. From the results of SCT in terms of F_1-score (*Detection*), mismatches ($Tracking^{\mathrm{SCT}}$) within a single camera, and SCTA, it can be seen that our proposed method based on MAST also achieves the most robust overall performance in all the four scenarios. The main advantage of MAST over general tracking-by-detection is that we effectively combine the information from segmentation with local object detection so that our method is less affected by the false positives generated by human detector in the entire frame. Moreover, the continuity of object tracking by segmentation is superior over those methods based on connecting trajectories, since there often exist many missing detections during tracking. This explains why the performance of MAST is more robust compared with the other state-of-the-art SCT used by each team. It can also be seen that the improved performance of intra-camera tracking and object segmentation also contribute to robust tracking across cameras. In addition, according to our parameters setting, the average runtime of our SCT together with object segmentation is 16.018 fps. The runtime is estimated on an Intel Core i7 PC with 2.67 GHz processor and 6G RAM in a Windows 7 environment.

Chapter 6

CONCLUSIONS AND FUTURE WORKS

6.1 Conclusions

In this dissertation, we propose a robust video object tracking system in distributed camera networks. The system includes invariant features for appearance-cue, single camera human tracking, and two-phase inter-camera tracking across multiple cameras.

We propose a novel enhanced and integrated person re-identification framework, which consists of three important techniques: holistic invariant feature extraction, regional invariant feature extraction, and aggregation. A pre-trained DCNN is used for extracting features describing holistic person appearance including color, texture, shape and other visual cues. Furthermore, we propose a two-way GMM fitting scheme to model dominant color modes of the target image as GMM in color histogram domain with the partitioned body parts. We also propose the integration scheme to combine three feature distances effectively using min-max normalization. In the experiments, we show that the new framework exceeds the state-of-the-art methods on the challenging benchmarks.

We propose a robust approach for tracking the same identity across multiple cameras based on online learning in a fully unsupervised manner. In SCT, we introduce a method that depends on tracking by multi-kernel adaptive segmentation with the assistance of local object detection, which also generates optimal foreground mask for the extraction of features in ICT. We make use of color transfer method to mitigate the change of illumination in ICT. The pose-invariant appearance features are exploited to overcome variation of poses and camera viewpoints between adjacent cameras. Moreover, the combination with context feature improves the performance of ICT. After collecting some video samples online, camera link models including transition time distributions for the time window, region mapping

matrix, region matching weights, and feature fusion weights are estimated, and the phase change occurs. We demonstrate significant advantages compared to the state-of-the-art methods on the benchmark dataset representing real-world camera network scenarios.

6.2 Future Works

In this dissertation, a robust video object tracking in distributed camera networks is proposed and developed, including invariant appearance descriptor, context-based feature, two-phase online-learning-based human tracking across non-overlapping cameras. However, the proposed approaches are limited to static camera networks.

The first future work is extending the proposed method to moving cameras. In terms of ICT with the real-world positions and motion information, tracking across static cameras or moving cameras is not different at all. By utilizing motion [94] and spatio-temporal cues, we are able to prune the outliers of matching candidates. To get the real-world positions and track the person in 3D space, we consider employing human detector and self-calibration from tracking [95]. In this way, we believe the system can further improve the capability of tracking as well.

In addition to the above, we plan to add the facial feature in the tracking. Face detection [96,97], alignment [98], and recognition [99] methods are well developed. However, most of the methods had experimented with vivid face images. The frames captured by surveillance cameras are blurry and face is sometimes not available in case rear head is taken. With utilizing trajectory, we can decrease false positive of face detection. In surveillance camera network, face poses are not consistent between two different cameras. Then, face frontalization [100] can be useful to improve the accuracy of face identification. By adding a facial feature, human tracking system across cameras can be more complete.

BIBLIOGRAPHY

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[2] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Computer Vision–ECCV 2008*, pp. 262–275, Springer, 2008.

[3] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2360–2367, 2010.

[4] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, 2013.

[5] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image and Vision Computing*, vol. 32, no. 6, pp. 379–390, 2014.

[6] S.-C. Chen, Y.-G. Lee, J.-N. Hwang, Y.-P. Hung, and J.-H. Yoo, "An ensemble of invariant features for person re-identification," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2016.

[7] D. Baltieri, R. Vezzani, and R. Cucchiara, "Mapping appearance descriptors on 3d body models for people re-identification," *International Journal of Computer Vision*, vol. 111, no. 3, pp. 345–364, 2015.

[8] "Multi-Camera Object Tracking (MCT) challenge [online]," http://mct.idealtest.org/index.html.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, IEEE, 2005.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[11] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 2, IEEE, 1999.

[12] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[13] C.-T. Chu, J.-N. Hwang, H.-I. Pai, and K.-M. Lan, "Tracking human under occlusion based on adaptive multiple kernels with projected gradients," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1602–1615, 2013.

[14] Z. Tang, J.-N. Hwang, Y.-S. Lin, and J.-H. Chuang, "Multiple-kernel adaptive segmentation and tracking (mast) for robust object tracking," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1115–1119, IEEE, 2016.

[15] P. Shirley, "Color transfer between images," *IEEE Corn*, vol. 21, pp. 34–41, 2001.

[16] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, vol. 3, 2007.

[17] D. Baltieri, R. Vezzani, and R. Cucchiara, "3dpes: 3d people dataset for surveillance and forensics," in *Proceedings of the joint ACM workshop on Human gesture and behavior understanding*, pp. 59–64, 2011.

[18] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*, pp. 91–102, Springer, 2011.

[19] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Computer Vision–ECCV 2014*, pp. 688–703, Springer, 2014.

[20] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking.," in *British Machine Vision Conference (BMVC)*, vol. 2, p. 6, 2010.

[21] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 2528–2535, 2013.

[22] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3586–3593, 2013.

[23] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Computer Vision–ACCV 2010*, pp. 501–512, Springer, 2011.

[24] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2288–2295, 2012.

[25] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2666–2672, 2012.

[26] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 152–159, 2014.

[27] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3908–3916, 2015.

[28] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification.," in *British Machine Vision Conference (BMVC)*, vol. 1, pp. 1–6, 2011.

[29] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern recognition letters*, vol. 34, no. 1, pp. 3–19, 2013.

[30] G. Lian, J.-H. Lai, C. Y. Suen, and P. Chen, "Matching of tracked pedestrians across disjoint camera views using ci-dlbp," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 7, pp. 1087–1099, 2012.

[31] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views," *Computer Vision and Image Understanding*, vol. 109, no. 2, pp. 146–162, 2008.

[32] D. Makris, T. Ellis, and J. Black, "Bridging the gaps between cameras," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, pp. II–205, IEEE, 2004.

[33] C.-T. Chu and J.-N. Hwang, "Fully unsupervised learning of camera link models for tracking humans across nonoverlapping cameras," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 979–994, 2014.

[34] Y.-G. Lee, J.-N. Hwang, and Z. Fang, "Combined estimation of camera link models for human tracking across nonoverlapping cameras," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2254–2258, IEEE, 2015.

[35] Y. Cai and G. Medioni, "Exploring context information for inter-camera multiple target tracking," in *IEEE Winter Conference on Applications of Computer Vision*, pp. 761–768, IEEE, 2014.

[36] C.-H. Kuo, C. Huang, and R. Nevatia, "Inter-camera association of multi-target tracks by on-line learned appearance affinity models," in *European Conference on Computer Vision*, pp. 383–396, Springer, 2010.

[37] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions.," in *BMVC*, vol. 8, pp. 164–173, Citeseer, 2008.

[38] T. D'Orazio, P. L. Mazzeo, and P. Spagnolo, "Color brightness transfer function evaluation for non overlapping multi camera tracking," in *Distributed Smart Cameras, 2009. ICDSC 2009. Third ACM/IEEE International Conference on*, pp. 1–6, IEEE, 2009.

[39] O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 26–33, IEEE, 2005.

[40] H. Lim, O. I. Camps, M. Sznaier, and V. I. Morariu, "Dynamic appearance modeling for human tracking," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, pp. 751–757, IEEE, 2006.

[41] B. C. Matei, H. S. Sawhney, and S. Samarasekera, "Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3465–3472, IEEE, 2011.

[42] A. Gilbert and R. Bowden, "Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity," in *European conference on computer vision*, pp. 125–136, Springer, 2006.

[43] C.-C. Huang, W.-C. Chiu, S.-J. Wang, and J.-H. Chuang, "Probabilistic modeling of dynamic traffic flow across non-overlapping camera views," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 3332–3335, IEEE, 2010.

[44] W. Chen, L. Cao, X. Chen, and K. Huang, "A novel solution for multi-camera object tracking," in *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 2329–2333, IEEE, 2014.

[45] L. Cao, W. Chen, X. Chen, S. Zheng, and K. Huang, "An equalised global graphical model-based approach for multi-camera object tracking," *arXiv preprint arXiv:1502.03532v2*, 2016.

[46] K.-W. Chen, C.-C. Lai, Y.-P. Hung, and C.-S. Chen, "An adaptive learning method for target tracking across multiple cameras," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[47] L. Wei and S. K. Shah, "Subject centric group feature for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 28–35, 2015.

[48] X. Chen and B. Bhanu, "Integrating social grouping for multi-target tracking across cameras in a crf model," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[49] X. Chen, Z. Qin, L. An, and B. Bhanu, "An online learned elementary grouping model for multi-target tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1242–1249, 2014.

[50] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[51] C.-T. Chu, J.-N. Hwang, S.-Z. Wang, and Y.-Y. Chen, "Human tracking by adaptive kalman filtering and multiple kernels tracking with projected gradients," in *Distributed Smart Cameras (ICDSC), 2011 Fifth ACM/IEEE International Conference on*, pp. 1–6, IEEE, 2011.

[52] H.-H. Lin, J.-H. Chuang, and T.-L. Liu, "Regularized background adaptation: a novel learning rate control scheme for gaussian mixture modeling," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 822–836, 2011.

[53] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 38–43, IEEE, 2012.

[54] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Subsense: A universal change detection method with local adaptive sensitivity," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 359–373, 2015.

[55] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection. net: A new change detection benchmark dataset," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–8, IEEE, 2012.

[56] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, 2014.

[57] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014*, pp. 818–833, Springer, 2014.

[58] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3626–3633, 2013.

[59] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1717–1724, 2014.

[60] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, pp. 1–42, 2014.

[61] P.-A. Savalle, S. Tsogkas, G. Papandreou, and I. Kokkinos, "Deformable part models with cnn features," in *European Conference on Computer Vision (ECCV), Parts and Attributes Workshop*, 2014.

[62] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[63] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *British Machine Vision Conference (BMVC)*, 2014.

[64] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[65] K. Jeong and C. Jaynes, "Object matching in disjoint cameras using a color transfer approach," *Machine Vision and Applications*, vol. 19, no. 5-6, pp. 443–455, 2008.

[66] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[67] G. McLachlan and D. Peel, *Finite mixture models*. John Wiley & Sons, 2004.

[68] K. Roeder and L. Wasserman, "Practical bayesian density estimation using mixtures of normals," *Journal of the American Statistical Association*, vol. 92, no. 439, pp. 894–902, 1997.

[69] J. G. Campbell, C. Fraley, F. Murtagh, and A. E. Raftery, "Linear flaw detection in woven textiles using model-based clustering," *Pattern Recognition Letters*, vol. 18, no. 14, pp. 1539–1548, 1997.

[70] A. Dasgupta and A. E. Raftery, "Detecting features in spatial point processes with clutter via model-based clustering," *Journal of the American Statistical Association*, vol. 93, no. 441, pp. 294–302, 1998.

[71] S. Franzini and J. Ben-Arie, "Speech recognition by indexing and sequencing," in *International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, pp. 93–98, 2010.

[72] K. Ma and J. Ben-Arie, "Vector array based multi-view face detection with compound exemplars," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3186–3193, 2012.

[73] N. Jojic, A. Perina, M. Cristani, V. Murino, and B. Frey, "Stel component analysis: Modeling spatial correlations in image class structure," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2044–2051, 2009.

[74] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *British Machive Vision Conference (BMVC)*, pp. 1–11, 2012.

[75] "Evaluation results of bicov, cov and gbicov."

[76] "Evaluation results of partssc and hist."

[77] R. Vezzani, C. Grana, and R. Cucchiara, "Probabilistic people tracking with appearance models and occlusion classification: The ad-hoc system," *Pattern Recognition Letters*, vol. 32, pp. 867–877, Apr. 2011.

[78] X. Liu, H. Wang, Y. Wu, J. Yang, and M.-H. Yang, "An ensemble color model for human re-identification," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pp. 868–875, IEEE, 2015.

[79] X. Chen, K. Huang, and T. Tan, "Object tracking across non-overlapping views by learning inter-camera transfer models," *Pattern Recognition*, vol. 47, no. 3, pp. 1126–1137, 2014.

[80] Y.-G. Lee, S.-C. Chen, J.-N. Hwang, and Y.-P. Hung, "An ensemble of invariant features for person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 470–483, 2017.

[81] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[82] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain, "Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 450–455, 2005.

[83] P.-L. St-Charles and G.-A. Bilodeau, "Improving background subtraction using local binary similarity patterns," in *IEEE Winter Conference on Applications of Computer Vision*, pp. 509–515, IEEE, 2014.

[84] J. Wu, C. Geyer, and J. M. Rehg, "Real-time human detection using contour cues," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pp. 860–867, IEEE, 2011.

[85] A. K. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, no. 35, pp. 99–109, 1943.

[86] H.-C. Shih and E.-R. Liu, "Automatic reference color selection for adaptive mathematical morphology and application in image segmentation," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4665–4676, 2016.

[87] K.-W. Chen and Y.-P. Hung, "Multi-cue integration for multi-camera tracking," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 145–148, IEEE, 2010.

[88] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *European Conference on Computer Vision*, pp. 788–801, Springer, 2008.

[89] C. Huang and R. Nevatia, "High performance object detection by collaborative learning of joint ranking of granules features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 41–48, IEEE, 2010.

[90] C. Huang, Y. Li, and R. Nevatia, "Multiple target tracking by learning-based hierarchical association of detection responses," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 4, pp. 898–910, 2013.

[91] M. Li, Z. Zhang, K. Huang, and T. Tan, "Rapid and robust human detection and tracking based on omega-shape features," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 2545–2548, IEEE, 2009.

[92] W. Chen, L. Cao, J. Zhang, and K. Huang, "An adaptive combination of multiple features for robust tracking in real scene," in *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pp. 129–136, IEEE, 2013.

[93] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image processing*, vol. 20, no. 6, pp. 1709–1724, 2011.

[94] K.-H. Lee and J.-N. Hwang, "On-road pedestrian tracking across multiple driving recorders," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1429–1438, 2015.

[95] Z. Tang, Y.-S. Lin, K.-H. Lee, J.-N. Hwang, J.-H. Chuang, and Z. Fang, "Camera self-calibration from tracking of moving persons," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pp. 260–265, IEEE, 2016.

[96] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *European Conference on Computer Vision*, pp. 720–735, Springer, 2014.

[97] S. Wu, M. Kan, Z. He, S. Shan, and X. Chen, "Funnel-structured cascade for multi-view face detection with alignment-awareness," *Neurocomputing*, vol. 221, pp. 138–145, 2017.

[98] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *European Conference on Computer Vision*, pp. 1–16, Springer, 2014.

[99] C. Ding, J. Choi, D. Tao, and L. S. Davis, "Multi-directional multi-level dual-cross patterns for robust face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 518–531, 2016.

[100] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4295–4304, 2015.

VITA

Younggun Lee received the B.S. degree in chemistry and physics from the Republic of Korea Air Force Academy in 2005, the M.S. degree in electrical engineering and computer science from the Seoul National University, Seoul, South Korea, in 2009. He is currently working toward the Ph.D. degree with University of Washington, Seattle, WA, USA.
His research interests include computer vision, image processing and video surveillance.