DOCTORAL THESIS

Person re-identification with limited labeled training data

Li, Jiawei

Date of Award:
2018

HONG KONG BAPTIST UNIVERSITY

Doctor of Philosophy

THESIS ACCEPTANCE

DATE: May 23, 2018

STUDENT'S NAME: LI Jiawei

THESIS TITLE: Person Re-identification with Limited Labeled Training Data

This is to certify that the above student's thesis has been examined by the following panel members and has received full approval for acceptance in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairman:	Dr. Shiu Wai Chee Associate Professor, Department of Mathematics, HKBU (Designated by Dean of Faculty of Science)
Internal Members:	Dr. Cheung Kwok Wai Head, Department of Computer Science, HKBU
	Dr. Lan Liang Assistant Professor, Department of Computer Science, HKBU
External Members:	Prof. Harada Tatsuya Professor Department of Mechano-Informatics The University of Tokyo JAPAN
	Prof. You Jia Jane Professor Department of Computing The Hong Kong Polytechnic University
Proxy:	Dr. Tam Hon Wah Associate Professor, Department of Computer Science, HKBU
In-attendance:	Prof. Yuen Pong Chi Professor, Department of Computer Science, HKBU

Issued by Graduate School, HKBU

Person Re-identification with Limited Labeled Training Data

LI Jiawei

A thesis submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Principal Supervisor: Prof. YUEN Pong Chi

Hong Kong Baptist University

July 2018

Declaration

I hereby declare that this thesis represents my own work which has been done after registration for the degree of PhD at Hong Kong Baptist University, and has not been previously included in a thesis or dissertation submitted to this or any other institution for a degree, diploma or other qualifications. I have read the Universitys current research ethics guidelines, and accept responsibility for the conduct of the procedures in accordance with the Universitys Committee on the Use of Human & Animal Subjects in Teaching and Research (HASC). I have attempted to identify all the risks related to this research that may arise in conducting this research, obtained the relevant ethical and/or safety approval (where applicable), and acknowledged my obligations and the rights of the participants.

Signature: 

Date: July 2018

Abstract

With the growing installation of surveillance video cameras in both private and public areas, it is an immediate requirement to develop intelligent video analysis system for the large-scale camera network. As a prerequisite step of person tracking and person retrieval in intelligent video analysis, person re-identification, which targets in matching person images across camera views is an important topic in computer vision community and has been received increasing attention in the recent years. In the supervised learning methods, the person re-identification task is formulated as a classification problem to extract matched person images/videos (positives) from unmatched person images/videos (negatives). Although the state-of-the-art supervised classification models could achieve encouraging re-identification performance, the assumption that label information is available for all the cameras, is impractical in large-scale camera network. That is because collecting the label information of every training subject from every camera in the large-scale network can be extremely time-consuming and expensive. While the unsupervised learning methods are flexible, their performance is typically weaker than the supervised ones. Though sufficient labels of the training subjects are not available from all the camera views, it is still reasonable to collect sufficient labels from a pair of camera views in the camera network or a few labeled data from each camera pair. Along this direction, we address two scenarios of person re-identification in large-scale camera network in this thesis, i.e. unsupervised domain adaptation and semi-supervised learning and proposed three methods to learn discriminative model using all available label information and domain knowledge in person re-identification.

In the unsupervised domain adaptation scenario, we consider data with sufficient labels as the source domain, while data from the camera pair missing label information as the target domain. A novel domain adaptive approach is proposed to estimate the target label information and incorporate the labeled data from source domain with the estimated target label information for discriminative learning. Since the discriminative constraint of Support Vector Machines (SVM) can be relaxed into a necessary condition, which only relies on the mean of positive pairs (positive mean), a suboptimal classification model learning without target positive data can be those using target positive mean. A reliable positive mean estimation is given by using both the labeled data from the source domain and potential positive data selected from the unlabeled data in the target domain. An Adaptive Ranking Support Vector Machines (AdaRSVM) method is also proposed to improve the discriminability of the suboptimal mean based SVM model using source labeled data. Experimental results demonstrate the effectiveness of the proposed method.

Different from the AdaRSVM method that using source labeled data, we can also improve the above mean based method by adapting it onto target unlabeled data. In more general situation, we improve a pre-learned classifier by adapting it onto target unlabeled data, where the pre-learned classifier can be domain adaptive or learned from only source labeled data. Since it is difficult to estimate positives from the imbalanced target unlabeled data, we propose to alternatively estimate positive neighbors which refer to data close to any true target positive. An optimization problem for positive neighbor estimation from unlabeled data is derived and solved by aligning the cross-person score distributions together with optimizing for multiple graphs based label propagation. To utilize the positive neighbors to learn discriminative classification model, a reliable multiple region metric learning method is proposed to learn a target adaptive metric using regularized affine hulls of positive neighbors as positive regions. Experimental results demonstrate the effectiveness of the proposed method.

In the semi-supervised learning scenario, we propose a discriminative feature

learning using all available information from the surveillance videos. To enrich the labeled data from target camera pair, image sequences (videos) of the tagged persons are collected from the surveillance videos by human tracking. To extract the discriminative and adaptable video feature representation, we propose to model the intra-view variations by a video variation dictionary and a video level adaptable feature by multiple sources domain adaptation and an adaptability-discriminability fusion. First, a novel video variation dictionary learning is proposed to model the large intra-view variations and solved as a constrained sparse dictionary learning problem. Second, a frame level adaptable feature is generated by multiple sources domain adaptation using the variation modeling. By mining the discriminative information of the frames from the reconstruction error of the variation dictionary, an adaptability-discriminability (AD) fusion is proposed to generate the video level adaptable feature. Experimental results demonstrate the effectiveness of the proposed method.

In short, the major contributions of this thesis are summarized as follows.

- An Adaptive Ranking Support Vector Machines (AdaRSVM) method is proposed to incorporated estimated target label information with the labeled data from source domain.

- A semi-supervised region metric learning algorithm is proposed to adapt a prelearned classifier onto target positive regions without target positive samples for training.

- A novel adaptable feature representation learning using video variation dictionary is proposed for semi-supervised video based person re-identification.

Keywords: person re-identification, unsupervised domain adaptation, dictionary learning

Acknowledgements

I would like to take this good opportunity to express my appreciation to all the people who have helped me during my doctorial study. First of all, I would like to express my sincere gratitude to my principle supervisor Prof. Pong Chi Yuen for providing me the precious opportunity to pursue graduate studies in person re-identification in large-scale camera network, which is a very interesting areas of computer vision with great application prospects. And I sincerely express my heartfelt gratitude to Prof. Yuen for his mentoring throughout my long graduate life with great patience and effort. Prof. Yuen leads me to explore the research field in depth and identify research problems as an independent researcher. He also inspire me to interpret common problems in creative ways and seek higher academic objectives in general and challenging scenarios. More important, Prof. Yuen teaches me to achieve useful research instead of doing researches for publication. To be honest, working with Prof. Yuen is not always an enjoyable experience, but I gain much more than I paid and it will be the most valuable experience in my life.

I would like to thank all the faculty members and staffs in the Department of Computer Science at Hong Kong Baptist University for their kind assistance and support to my study. I also would like to thank all my friends and colleagues who helped me during my study in Hong Kong. Especially, I wish to express my thanks to Dr. Jinhua Ma for their cooperations in my research, as well as Dr. Yicheng Feng, Dr. Weiwen Zou, Dr. Lan Xiangyuan, Dr. Meng-Hui Lim, Dr. Shengping Zhang, Mr. Guangcan Mai, Miss. Baoyao Yang, Mr. Siqi Liu, Mr. Mang Ye, Mr. Zexiong Cai, Mr. Rui Shao and Mr. Sheung Wai Chan for their discussions and

help about my life and research.

Finally, I would like to express my heartfelt gratitude to my parents for their sincere love, encouragement and deep understanding all the time.

Table of Contents

List of Tables

List of Abbreviations

AdaRSVM	Adaptive Ranking Support Vector Machines classifier
AD fusion	adaptability-discriminability fusion
BGRM	background removal feature
BoW + Geo + Gauss	Bag-of-Word feature with weak Geometric constraints and Gaussian mask background suppression
CIS	Color Invariant Signatures
CPS	Custom Pictorial Structures features
CPSDA	Cross Person Score Distribution Alignment
CUHK01	CUHK Person Re-identification 01 datasets
DTW	Dynamic Time Warping classifier
DVR	Classfier of Discriminative Video fragments selection and Ranking
eBiCov	enriched Bio-inspired Covariance features
eLDFV	enriched Local Descriptors encoded by Fisher Vectors
eSDC	enriched Salience Correspondence features
GFS	Geodesic Flow Sampling domain adaptation method
GRDL	Classifier of unsupervised Graph Regularized Dictionary Learning
i-LIDS	Imagery Library for Intelligent Detection Systems dataset
iLIDS-VID	Imagery Library for Intelligent Detection Systems Video re-IDentification dataset

ISR	Classifier of Iterative re-weighted Sparse Ranking
IW	Importance reWeighting classifier
KISSME	Classifier of Keep It Simple and Straightforward MEtric Learning
MAD	Mean Approach Distance classifier
MREM	Multiple REgion-to-point Metric learning
PaMM	Pose-aware Multi-shot Matching classifier
PCA	Principle Component Analysis
PRID2011	Person Re-IDentification dataset 2011
PRID2011m	multiple-shot version of Person Re-IDentification 2011 dataset
PRID2011s	single-shot version of Person Re-IDentification 2011 dataset
POM	POint-to-point Metric learning
PU-learning	Learning from Positive and Unlabeled data
RankSVM	Rank Support Vector Machine classifier
RDC	Relative Distance Comparison classifier
RLR	Robust Logistic Regression classifier
RNNVPN	Recurrent Neural Network classifier for Video-based Person Re-identification
RPML	Relaxed Pairwise Metric Learning classifier
SI^2DL	Simultaneous Intra-video and Inter-video Distance Learning classfier
SDALF	Symmetry-Driven Accumulation of Local Features
SLISC	Classifier of Semi-supervised Learning for Imbalanced Sentiment Classification
SREM	Single REgion-to-point Metric learning
SSC	Classifier of Semi-supervised Spectral Clustering

SSR-IPP	Classifier of Semi-Supervised Ranking method with Increased Positive Prior
SVM	Support Vector Machine classifier
TCA	Transfer Component Analysis domain adaptation method
TDL	Top-push Distance Learning classifier
UMDL	Unsupervised Multi-task Dictionary learning classifier
VIPeR	Viewpoint Invariant Pedestrian Recognition dataset
WHOS	Weighted Histograms of Overlapping Stripes features
XQDA	Classifier of Cross-view Quadratic Discriminant Analysis

List of Figures

Chapter 1

Introduction

This chapter gives an introduction to person re-identification with limited labeled training data. The research background and motivation is introduced in Section 1.1. And a literature view on related person re-identification methods is given in Section 1.2. In Section 1.3, the contributions of this thesis are reported. Finally, the overview of this thesis is listed in Section 1.4.

1.1 Background and Motivation

1.1.1 Background

With the growing installation of surveillance video cameras in both private and public areas, closed-circuit TV has evolved from single-camera to multiple-camera systems, and more recently to large-scale camera networks. Current surveillance applications such as those employed in shopping malls and residential buildings consist of multiple-camera networks of up to hundreds of cameras, and many large cities around the world have installed hundreds of thousands of cameras as part of their surveillance apparatus. Whilst large-scale camera network hardware is generally well-designed and well-installed, the development of intelligent video analysis software lags far behind. So it is an immediate requirement to monitor and analysis the huge number of surveillance videos by intelligent video analysis. As a prerequi-

site step of person tracking and person retrieval in multiple camera network, person re-identification, which target in matching person images across camera views is an important part of the intelligent video analysis. So person re-identification become a important topic in computer vision community and be received increasing attention in the recent years.

The person re-identification task is substantially challenging when variations in illumination condition, background, human pose, occlusion and scale are significant among those views. To address this problem, existing schemes focus on developing robust feature presentation (e.g. [15] [17] [24]) and discriminative learning models (e.g. [32] [44][71]). As a person retrieval task, some of the existing person re-identification methods (e.g. [63]) suggest to employ the same feature description and classification model for all camera views. For this purpose, view-invariant feature representation learning method [63] are proposed and trained on labeled data from part of the camera views. However, such features may not be completely invariant to appearance changes across camera views under uncontrolled environment when the labeled data from those views are not available in training stage.

To address the cross-view variation issue, the rest methods (e.g. [44] [71] [72]) learning view-specific feature presentation and/or classification model for each camera pair. In the scheme of unsupervised learning, representative feature representation [24] [40] [43] [62] [69] and discriminative classification model [41] [72] have been reported. However, without labeled data from target camera views, existing unsupervised models are unable to learn the discriminative components from a person under severe appearance changes across views. So the performance of unsupervised models are typically weaker than that of supervised ones. For more discriminative classification model, supervised distance metric learning methods [48] [71] [20] [46] [58] [45] [4] [23] [31] [32] [68] [36] [66] [10] [30] [6] and supervised deep learning methods [29] [3] [53] [44] [8] [51] [54] [59] [73] [7] [70] [57] [25] [67] have been reported.

1.1.2 Motivation of This Project

While the supervised learning methods could achieve encouraging re-identification performance, the assumption that label information is available for all the cameras, could only be practically feasible in a small-scale camera network. Contrarily, in large-scale camera network, collecting the label information of every training subject from every camera can be extremely time-consuming and expensive. Since hundreds of labeled image pairs are typically needed from each camera pair for supervised learning (e.g. [71] [44]), the labeling cost would be prohibitively high in large-scale camera network with hundreds of cameras. Therefore, sufficient labels of the training subjects may not be able to be collected from certain cameras. This renders the supervised learning approach inapplicable, since the person labels are not available. These setbacks pose the need for new methods to handle the afore-described person re-identification issue in the large-scale camera network setting.

Though sufficient labels of the training subjects are not available from all the camera views, it is still reasonable to collect sufficient labels from a pair of camera views in the camera network. Motivated by domain adaptation approach [18], we consider data with sufficient labels as the source domain, while data from target camera pair missing label information as the target domain. Since the person re-identification task in the two camera pairs are positive correlated, the label information in the source domain can be adapted to the target domain by domain adaptation [18]. The existing domain adaptation methods mainly target at a projection such that the source and the target joint distributions are the same after projection. However, it is almost impossible to verify whether the source and the target conditional distributions are the same. As a result, there is no way to guarantee that the distance model learned from the projected data in the source domain is equivalent to the target one. Thus, a new domain adaptive method is needed to estimate the target label information and incorporate the labeled data from source domain with the estimated target label information for discriminative learning. It has been shown that the discriminative constraint of Support Vector Machines (SVM) can

be relaxed into a necessary condition, which only relies on the mean of positive pairs and the negative data. Along this direction, this thesis proposes a reliable target positive mean estimation method and an adaptive ranking support vector machines on the source labeled data, the target negative data and the estimated target positive mean, where the target negative data can be generated easily.

Since the target positive mean cannot represent the whole positive conditional distribution, the above mean based method is not optimal when the variance of positive conditional distribution is large. In this context, a better solution should be estimating additional target positive information besides target positive mean. Due to the imbalanced data problem, it is difficult to accurately estimate positives from the target unlabeled data. Motivated by the region-to-point metric learning method [74], discriminative classification model can be learned using positive regions and negative data. Along this direction, this thesis also proposes to estimate target positive regions by a pre-learned classifier and develop a semi-supervised region metric learning algorithm to learn discriminative target classification model using the estimated positive regions. Here, the pre-learned classifier can be classification model learned from source labeled data or domain adaptive model learned from source labeled data and target unlabeled data.

Besides collecting sufficient labels from a pair of camera views, it is also feasible to collect a few labeled data from each camera pair. To enrich the labeled data from target camera pair, image sequences (videos) of the tagged persons are collected from the surveillance videos by human tracking as shown in [1]. The existing video based person re-identification methods learning intra-view invariant feature presentations are not optimal for matching the person videos with large inter-view variations. To extract the discriminative and adaptable components from all the frames, this thesis proposes a video variation dictionary learning to model the intra-view variations and a video level adaptable feature learning method using multiple sources domain adaptation and an adaptability-discriminability fusion.

1.2 Review of Related Person Re-identification Methods

This section gives an overview of some related works in person re-identification in four settings: supervised learning, unsupervised learning, semi-supervised learning and domain adaptation.

1.2.1 Supervised Learning Methods

Distance Metric Learning Given the hand-crafted or deep features of captured person images/videos, the general idea of distance metric learning methods is to pull feature vectors w.r.t the same person but different views closer while pushing those w.r.t different persons further apart by a distance function. The learned distance can be weighted l_1 norm for Support Vector Machine (SVM) based methods [48] [71] [20] or weighted l_1 norm for metric learning based ones [46] [58] [45] [4] [23] [31] [32] [68] [36] [66] [10] [30] [6]. In [48], person re-identification was formulated as a ranking problem and the Rank Support Vector Machine (RankSVM) model is learned by assigning higher confidence to the matched image pairs and vice versa. Similar to RankSVM, Zheng *et al.* [71] proposed a Relative Distance Comparison (RDC) method using a second-order distance learning model. This method is able to exploit higher-order correlations among different features, compared with RankSVM. In order to solve the computational complexity issues in RankSVM and RDC, a Relaxed Pairwise Metric Learning (RPML) method [20] was proposed by relaxing the original hard constraints, which leads to a simpler problem that can be solved more efficiently. For higher reliability on small training dataset, Pedagadi *et al.* [46] suggested a dimension reduction using Principle Component Analysis (PCA) before local Fisher discriminative analysis. In [58], kernel local Fisher discriminant [46] was fused with another metric from regularized pairwise constrained component analysis [45] by ranking ensemble voting. Besides the subspace, the person re-identification can also perform on a supervised manifold [4]. Formulating the person image match-

ing problem as a likelihood ratio test, Keep It Simple and Straightforward MEtric (KISSME) [23] model was learned under Gaussian distribution assumption using the principle components. In [31], a discriminative low-dimensional subspace was adopted to replace the principle components for KISSME model learning. In [32], Liao *et al.* suggested the positive semidefinite constraint of weight matrix and give a fast solution using accelerated proximal gradient approach. Since the linear model on pre-defined kernel space may not be discriminative enough, structural SVM [68] [36] and sample-specific SVM [66] are proposed as adaptive non-linear solutions. A pose-aware matching [10] was also proposed to learn multiple matching models for four different human poses. Beside the above global matching methods, local matching model was also adopted. In [30], the global distance metric is coupled with a locally adaptive thresholding rule. Chen *et al.* [6] proposed an explicit polynomial kernel feature map for patch-based matching between two person images.

Deep Learning The deep learning based person re-identification models can be divided into two categories, i.e. matching models and classification models. The former ones learn the similarity between person images/videos by siamese neural network [29] [3] [53] [44] [8] [51] [54] [59] [73] [7], while the latter ones target on predicting person IDs [70] [57] [25] [67]. In the scheme of siamese network, a body-partitioning-based patch matching layer was introduced for robustness against positional differences in corresponding features across images in [29]. While in [3], a neighborhood differences layer was proposed alternatively. Different from the two passive matching layers, Varior *et al.* [53] suggested to actively select features relevant to the paired image by a gating function in the mid-level layer. Mean-pooling and max-pooling were also employed to extract invariant Recurrent Neural Networks (RNN) feature over the image sequence [44]. Formulating person re-identification as a ranking problem, a triplet loss function was learned to guarantee the rank order, i.e. the distances to relevant samples are larger than those to irrelevant ones in the learned feature space for each sample in [8]. While in [51], a semi-supervised attribute learning algorithm was incorporated into a triplet loss based deep model.

In [54], Long Short-Term Memory (LSTM) modules were adopted to memorize the spatial dependencies and extract relevant contextual information. The LSTM based model was then extended to video based person re-id problem in [59]. A spatial recurrent model and a temporal attention model were incorporated into an end-to-end deep architecture in [73]. A quadruplet loss function [7] was proposed for a larger inter-class variation and a smaller intra-class variation. A classification model was proposed in [70]. While a domain guided dropout algorithm [57] was proposed to deal with the multi-task learning problem. A multi-scale context-aware network [25] was proposed to enhance the visual context information over full body and body parts. The human body structure information [67] were employed to align body region features across images.

1.2.2 Unsupervised Learning Methods

One of the key issues in unsupervised person re-identification is to learn representative feature representation [40] [24] [43] [62] [69] from the unlabeled data. [40] proposed a set of local descriptors using Fisher vectors. [24] proposed an illumination-invariant feature representation based on log-chromaticity color space and two-part human structure model. A local descriptor via hierarchical Gaussian distribution [43] is proposed to extract the local appearance structure from person images. Discriminative feature descriptors [69]; [62] are also learned using the unlabeled data. In consideration of valuable salient information, [69] proposed an unsupervised salience learning method, in which greater weights are assigned to image parts with salient patterns (colors and textures). The salience feature is good at distinguishing persons with small cross-camera variations, but it may be sensitive to local color changes. For robust color feature representation, [62] made use of semantic analysis and proposed a local color descriptor based on pre-defined salient color name. A robust l_1-norm graph regularized dictionary learning [22] is proposed to estimate view-invariant and discriminative feature representation. Inspired by the unsupervised cross-dataset transfer learning, a multi-task dictionary learning method is proposed

to learn a dataset-shared and target-data-biased feature representation [47].

The another key issue in unsupervised person re-identification is to design discriminative classification model [41] [72]. A time shift dynamic time warping model [41] was derived for person video matching with automatic alignment. A k-reciprocal encoding method [72] was proposed to represent the probe by the k-reciprocal nearest neighbors and re-rank the re-ID results.

1.2.3 Semi-supervised Learning Methods

In semi-supervised person re-identification, the existing methods target on reliable feature representation or tagging unlabeled samples for discriminative classification model learning. [35] proposed a coupled dictionary learning to model the appearance changes across camera views. The basic idea is to learn a representative dictionary for person images captured in the view of each camera using unlabeled data, and to make the matched images share the same sparse representation. To address the problem with low resolution probe images and high resolution gallery images, [21] employed a super-resolution person re-identification method with low-rank dictionary learning. [16] proposed to fuse multiple features for reliable feature representation. Different from the dictionary learning methods, a discriminative subspace [65] was learned by null Foley-Sammon transfer.

To improve performance of pre-learned matching model without any positive samples, [33] proposed a post-rank optimization method to refine the matching result using one-click negative samples and synthetic positive samples. These methods developed semi-supervised learning methods for person re-identification, but they do not solve the imbalanced unlabeled data problem. Considering the imbalanced data problem, [38] proposed an empirical method to select potential positives iteratively and to learn a regression based ranking model using the estimated potential positives and the negatives. But the estimation error of potential positives may be large and thence the influence of mislabeled data may degrade the re-identification performance.

1.3 Contributions of This Thesis

This thesis addresses two scenarios of person re-identification in large-scale camera network. The major contributions of the thesis are summarized as follows:

1. A discriminative learning method is proposed to incorporated estimated target label information with the labeled data from source domain. Since the discriminative constraint can be relaxed into a necessary condition, which only relies on the mean of positive pairs (positive mean), a suboptimal classification model learning without target positive data can be those using positive mean. A reliable positive mean estimation is given by using both the labeled data from the source domain and potential positive data selected from the unlabeled data in the target domain. An Adaptive Ranking Support Vector Machines (AdaRSVM) method is also proposed to rank the individuals for person re-identification. Inspired by adaptive learning methods [60] [14], RankSVM [48] is employed to learn a distance model by the labeled data from the source domain. After that, the estimated target positive mean and target negative data are used to learn the discriminative model for target domain by adaptively refining the distance model learned in the source domain. These works have been published in [39] [37].

2. A semi-supervised region metric learning algorithm is proposed to estimate target positive region without target positive data and then learn target adaptive classification model using the estimated positive regions. For positive information estimation from imbalanced unlabeled data, a new concept of positive neighbors is introduced, which refer to data close to any true positive. An optimization problem for positive neighbor estimation is derived and solved by aligning the cross-person score distributions together with optimizing for multiple graphs based label propagation. The positive regions are then generated by the regularized affine hull of positive neighbors, which close to those of positive data. A reliable multiple region metric learning method is proposed to

learn a target adaptive metric using estimated positive regions. These works have been published in [27].

3. A novel method to extract adaptable feature representation via video variation dictionary learning is proposed for semi-supervised video based person re-identification. First, a novel video variation dictionary learning is proposed to model the large intra-view variations and solved as a constrained sparse dictionary learning problem. Second, frame level adaptable features are generated by multiple sources domain adaptation using the variation modeling. By mining the discriminative information of the frames from the reconstruction error of the variation dictionary, an adaptability-discriminability (AD) fusion is proposed to generate the video level adaptable feature. These works have been published in [26].

1.4 Overview of This Thesis

The rest of this thesis is organized as follow: Chapter 2 presents the proposed adaptive ranking support vector machines learning for person re-identification without target positive data. We show that the discriminative constraint can be relaxed into a necessary condition, which only relies on the mean of positive pairs (positive mean). A classification model learning without target positive data is proposed using positive mean. Then, a reliable positive mean estimation is given by using both the labeled data from the source domain and potential positive data selected from the unlabeled data in the target domain. An Adaptive Ranking Support Vector Machines (AdaRSVM) method is proposed to learn a discriminative distance model by the labeled data from the source domain. Experiment results show that the AdaRSVM method achieves convincing recognition performance for person re-identification. The proposed AdaRSVM not only outperforms non-learning based methods but also is better than the comparing discriminative learning methods using labeled data from the source domain for training.

Chapter 3 presents the proposed semi-supervised region metric learning method to estimate target positive neighbors besides target positive mean for discriminative classification model learning. A cross person score distribution alignment and a multiple graph based label propagation are jointly optimized for positive neighbor estimation. The positive regions are then generated using the positive neighbors. Finally, a region-to-point metric learning method is presented to learn discriminative metric by maximizing the weighted distance between negatives and positive regions. Experiment results show that the region metric learning method improves the per-learned classifier remarkably and achieves convincing recognition performance for person re-identification. It is shown that the proposed region metric learning method not only outperforms comparing semi-supervised methods and classification algorithms robust against label noises, but also is better than the comparing discriminative person re-identification methods.

Chapter 4 presents the proposed adaptable feature representation learning for video based person re-identification. A video variation dictionary learning algorithm is first proposed to model the large intra-view variations. Second, a frame level adaptable feature is generated by multiple sources domain adaptation using the variation modeling. By mining the discriminative information of the frames from reconstruction error of the variation dictionary, an adaptability-discriminability (AD) fusion is proposed to generate the video level adaptable feature. Experimental results on two public video based person re-identification datasets show that classification on the proposed video level adaptable feature achieves better recognition performance for person re-identification than state-of-the-art methods.

Chapter 5 concludes the thesis and discusses some future directions.

Chapter 2

Domain Transfer Support Vector Ranking for Person Re-Identification without Target Camera Label Information

Chapter 3

Semi-Supervised Region Metric Learning without Positive Data for Person Re-identification

Chapter 4

Semi-supervised Adaptable Feature Learning For Video Based Person Re-identification

4.1 Introduction

Since the unlabeled person images can be easily collected by detecting human from surveillance videos, the setting of semi-supervised learning is more practical than that of the supervised methods in real world large-scale camera network.

In the scheme of semi-supervised learning, the unlabeled data are usually used to propagate labeled information from the neighborhood of labeled data to the whole space. Specifically, in semi-supervised person re-identification, the labeled data are used to exploit view-invariant information, while the unlabeled data are used to deal with the intra-view variations. As a result, couple dictionary learning is adapted to learn projective space robust against cross-view variation, while the intra-view variations are modeled simultaneously by joint learning the view-specific local coordinate coding dictionaries [35]. For a discriminative subspace instead of dictionary based representation, a kernelized null space learning is proposed to minimize both the intra- and inter-view variations [65].

(a) pose (b) resolution (c) occlusion (d) illumination

Figure 4.1: person's images captured in the same view can also undergo large appearance changes due to 4.1(a) view angle changes, 4.1(b) resolution changes, 4.1(c) occlusion and 4.1(d) illumination changes. And the discriminative components can be variant to the intra-view changes.

Now, video based person re-identification has received more attention since the image sequence (video) is more informative than the single image and can be easily collected from surveillance video. A pioneer work on semi-supervised video based person re-identification is proposed by Zhu et al [76], in which the cross-frame variations are minimized for each sequence in the scheme of semi-couple dictionary learning.

However, estimating intra-view invariant dictionary representation/subspace in the above methods is not the optimal way to deal with the intra-view variation. Since the discriminative component can change across frames, the estimated intra-view invariant representation loses discriminative information in the cases of large intra-view variations as shown in Fig. 4.1. Instead of estimating the intra-view invariant representation, in the chapter we propose to model the intra-view variation and extract the discriminative and adaptable feature using the intra-view variation modeling. Here, the adaptable feature means that the person's appearance in the testing images are mainly affected by the factors similar to the training images, e.g. training and testing images captured under similar illumination or view angle. So the matching model learned on the training images can be employed on the testing one naturally using adaptable features even though both the training and the testing features are not invariant.

To model the intra-view variations, we propose a novel Video Variation Dictio-

nary Learning (V^2DL) which incorporates the sparse variation dictionary learning [61] and the temporal consistency and motion information in surveillance videos. A video level adaptable feature are estimated by multiple sources domain adaptation and an adaptability-discriminability fusion.

In summary, this chapter proposes a novel method to extract adaptable feature representation for semi-supervised video based person re-identification via video variation dictionary learning. The contributions of this work are two-folds.

• A novel video variation dictionary learning is proposed to model the large intra-view variations and solved as a constrained sparse dictionary learning problem.

• A frame level adaptable feature is generated by multiple sources domain adaptation using the variation modeling. By mining the discriminative information of the frames from the reconstruction error of the variation dictionary, an adaptability-discriminability (AD) fusion is proposed to generate the video level adaptable feature.

4.2 Related Works

4.2.1 Video based Person Re-identification

For additional spatial-temporal information from the video, gait features are extracted as a biometric to measure people's walking style for re-identification [49]. A Swiss-system based cascade ranking model [56] is proposed to improve the robustness of gait based matching against changes of clothes and view angles. But it is still difficult to obtain discriminative gait features under the cluttered background in re-identification due to the inaccurate silhouette extraction.

The other kind of approaches turn to invariant features by taking advantage of the temporal consistence information. The reliable features across frames are extracted to represent a video, i.e. the salient texture feature [17], the stable color region [15] [9], the recurrent structured region [15], and the motion-invariant local body-action features [34]. In addition, two reliable feature descriptors are also com-

puted using frame level features. The means of frame based features are extracted from reliable subsequences [55] to handle the pose variation problem. Mean-pooling and max-pooling are also employed to extract invariant Recurrent Neural Networks (RNN) feature over the image sequence [44]. But the invariant feature cannot exist and the mean/max based descriptor cannot represent the whole feature set when the large intra-view variations occur.

When the intra-view variation is large, one of the approach is to deal with the problem under different variations separately. A pose-aware matching [10] is proposed to learn four matching models for different human poses. In addition, A frame selection algorithm [10] is also used to select reliable frames according to consistency of human motion and occlusion rate. Such method can perform well in problem without cluttered background and crowded environment. But in challenging cases with multiple large intra-view variations, the multiple imperfect variation modelings can result in large accumulative error.

Ignore the temporal structure of frames in videos, the video based re-identification can be considered as traditional matching problem. When a video level feature is obtained, a so-called "top-push" learning algorithm [64] is proposed to ensure a large margin between the videos of different persons in the metric space. To extend the video level large margin constraint to frame level, a Simultaneous Intra-video and Inter-video Distance Learning (SI^2DL) is proposed to further make small intra-video variations in the metric space. However, the reliability of the frame is miss without the temporal structure of the videos, so the learnt metric is sensitive to the unreliable frames.

4.2.2 Variation Dictionary Learning

Though the intra-view variation modeling for person re-id is relative new, a similar problem, i.e. intra-class variation modeling for face recognition, has been studied for years. Since dictionary learning has been extensively studied in computer vision, it is employed in modeling multiple types of intra-class variations [11] [61] [12]. By

introducing an auxiliary intra-class variation dictionary, the samples are represented by the sum of a target-appearance component and an intra-class variation component [11]. Since the intra-class variations of training data may not be the same to those of gallery data, the sparse variation dictionary is adaptive to the gallery set by learning a variation-model projection [61]. To deal with the pose variation problem, a patch-based transformation dictionary is learned to connect corresponding patches across poses under the multitask learning scheme [12].

The existing variation dictionary based methods mainly use the variation modeling as an auxiliary tool for target-appearance dictionary learning. So the performance of those methods are restricted by the representation ability of dictionary learning.

4.3 Adaptable Feature Learning Under Large Intra-view Variations

This section gives a detailed description of the proposed methods. Suppose we have a collection of person image sequences $\left\{I^a_{i,j}\right\}$ and $\left\{I^b_{i,j}\right\}$, where $I^a_{i,j}$ ($I^b_{i,j}$) refer to jth frame of person i captured by camera a (b). $x^{c_{ID}}_{i,j}$ is N_d-dimension feature vector extracted from image $I^{c_{ID}}_{i,j}$, $i = 1, 2, ..., N_{c_{ID}}$, $c_{ID} = \{a, b\}$. For simplification, camera ID a and b may be skipped when all data are captured by the same camera. Then the feature of image $\{I_{i,j}\}$ is denoted as $\{x_{i,j}\}$, where $i = 1, 2, ..., N$, $j = 1, 2, ..., N_i$.

The testing procedure of proposed method is shown in Fig. 4.2. Based on the learned variation dictionary, the frame level feature is decomposed into the person specific component, variation component and reconstruction error. For each testing sample, the training samples are not equally important in matching model learning. Training samples captured similar variations to the testing one are more important. It motivates us to define variation specific domains and to learn domain specific information. So we propose a frame level adaptable feature using similarities between domains. Subsequently, a final video level adaptable feature is estimated

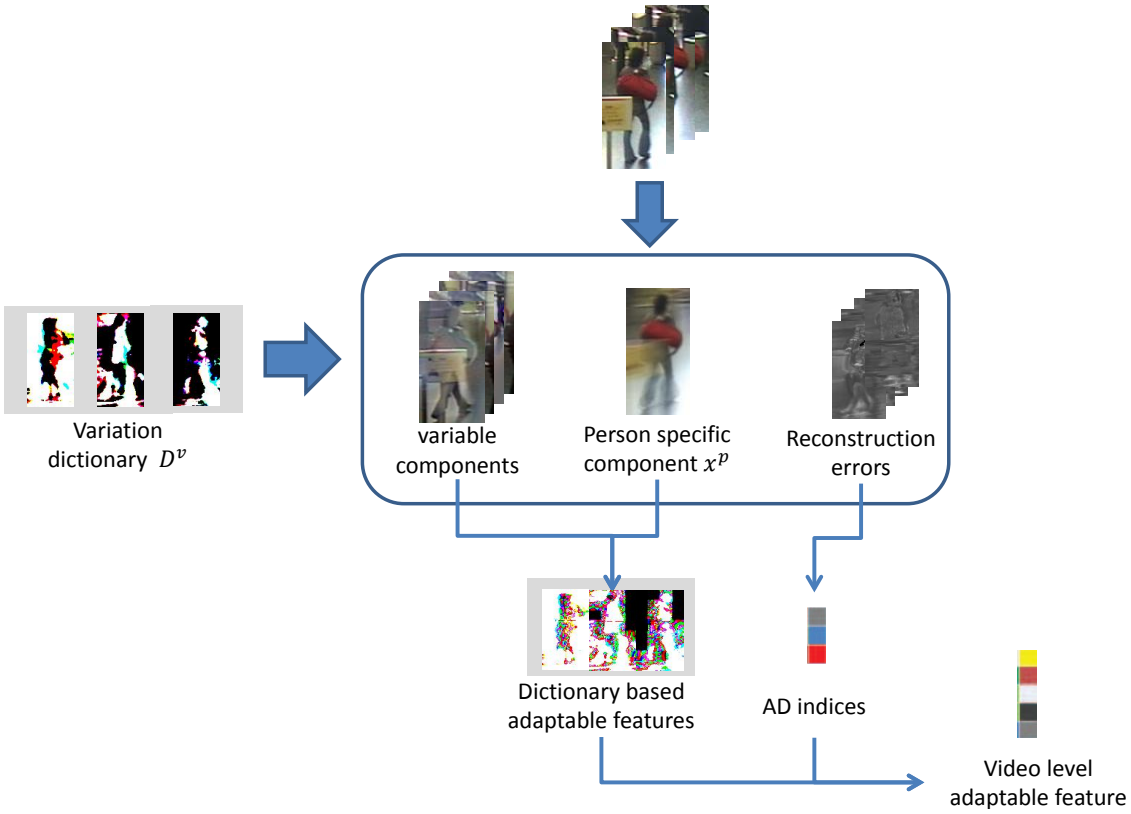

Figure 4.2: Proposed adaptable feature learning for video person re-identification

by an AD fusion of the frame level adaptable features.

4.3.1 Video Variation Dictionary Learning

Following [11], a frame $x_{i,j}$ can be represented by the sum of the target-specific component x_i^p, the variation component $x_{i,j}^v$ and small dense noise z, i.e.

$$x_{i,j} = x_i^p + x_{i,j}^v + z \tag{4.3.1}$$

where intra-view variant term $x_{i,j}^v$ represents lighting changes, pose changes, scaler changes or occlusions that cannot be modeled by the small dense noise z. When the variation component is represented using a dictionary D^v, (4.3.1) can be estimated as follows.

$$x_{i,j} = x_i^p + D^v \omega_{i,j}^v + z \tag{4.3.2}$$

where $\hat{D}^{v*} = \left[\hat{D}_1^{v*}, \hat{D}_2^{v*}, ..., \hat{D}_{N_D}^{v*}\right]$, each column of D^{v*} denotes the appearance under one setting of variation factors, non-negative vector $\hat{\omega}_{i,j}^v$ denotes the responses of $x_{i,j}$ to the variations w.r.t the dictionary.

To learn the sparse variation dictionary $\hat{D}^{v*}$, the noise term z is minimized over all frames, i.e.

$$\begin{aligned} \hat{D}^{v*} &= \arg\min_{\hat{\omega}_{i,j}^v, \hat{D}^v} \sum_{i,j} \frac{1}{2}||x_{i,j} - x_i^p - \hat{D}^v\hat{\omega}_{i,j}^v||_2^2 + \lambda||\hat{\omega}_{i,j}^v||_1 \\ s.t.\ &||D_k^v||_2 = 1, \forall k \\ &\left(\hat{\omega}_{i,j}^v\right)_k \geq 0, \forall k \end{aligned} \tag{4.3.3}$$

where $\left(\hat{\omega}_{i,j}^v\right)_k$ is the kth element of $\hat{\omega}_{i,j}^v$.

Since the person specific component is usually unknown, we also estimate it in (4.3.3), i.e.

$$\begin{aligned} \hat{D}^{v*} &= \arg\min_{\hat{\omega}_{i,j}^v, \hat{D}^v, \hat{x}_i^p} \sum_{i,j} \frac{1}{2}||x_{i,j} - \hat{x}_i^p - \hat{D}^v\hat{\omega}_{i,j}^v||_2^2 + \lambda||\hat{\omega}_{i,j}^v||_1 \\ s.t.\ &||D_k^v||_2 = 1, \forall k \\ &\left(\hat{\omega}_{i,j}^v\right)_k \geq 0, \forall k \end{aligned} \tag{4.3.4}$$

For robust variation dictionary against large variations across frames, the inconsistent components in neighboring frames are considered as unadaptable features. Following [55], we introduce the temporal consistency of variation modelings across frames to remove the inconsistent components. The sparse representation $\hat{\omega}_{i,j}^{v*}$ are assumed to be invariant in some subsequences $S_{i,m}$, i.e.

$$\begin{aligned} &||\hat{\omega}_{i,j_1}^{v*} - \hat{\omega}_{i,j_2}^{v*}||_2 < \mu_2, \\ &\forall j_1 \in S_{i,m}, j_2 \in S_{i,m}, m = 1, 2, ..., M_i, \forall i \end{aligned} \tag{4.3.5}$$

where $S_{i,m}$ denotes the set of frames in the mth view-consistent subsequence. To obtain such subsequences $S_{i,m}$, we divide each sequence by employing an off-line change point detection algorithm [5] which can be solved via expectation-maximization (EM).

Combining constraint (4.3.5), the dictionary learning problem (4.3.4) is converted

to be a constrained sparse dictionary learning problem, i.e.

$$
\begin{aligned}
D^{v*} = \min_{\omega_{i,j}^v, D^v, x_i^p} & \lambda \sum_{i,j} ||\omega_{i,j}^v||_1 + \sum_{i,j} \frac{1}{2} ||x_{i,j} - \left(x_i^p + D^v \omega_{i,j}^v\right)||_2^2 \\
s.t. \ & ||D_k^v||_2 = 1, \forall k \\
& ||\omega_{i,j_1}^v - \omega_{i,j_2}^v||_2 < \mu_2, \forall j_1, j_2 \in S_{i,m}, \forall i \\
& \left(\omega_{i,j}^v\right)_k \geq 0, \forall k
\end{aligned}
\tag{4.3.6}
$$

Optimization method to solve (4.3.6) will be shown in Section 4.3.4.

Given the variation modeling in (4.3.6), a straightforward method is to extract an invariant feature representation for the video. The person specific component $\hat{x}_i^p$ seems a so-called "variation-free" feature [55] [44]. However, the component $\hat{x}_i^{p*}$ can also be affected by intra-view variations that are unchanged over the whole video, e.g. it is difficult to distinguish a person in red and a person in white but under red lighting over the whole video. Furthermore, the imperfect variation modeling can result in additional error, e.g. overdone color correction can result in unfavorable color feature. So the person specific component $\hat{x}_i^{p*}$ is not a reliable feature of the sequence when the intra-view variation is large. Alternatively, we use an aligned feature $x_{i,j}^l$ to represent a frame, which is obtained by sum of the person-specific component x_i^p and the view-dependent component represented by the view-dependent variation dictionary D^{v*}, i.e.

$$
\begin{aligned}
& x_{i,j}^l = D^{v*} \omega_{i,j}^{v*} + x_i^{p*} \\
& \left\{\omega_{i,j}^{v*}, x_i^{p*}\right\} = \arg \min_{\omega_{i,j}^v, D^v, x_i^p} \lambda \sum_{i,j} ||\omega_{i,j}^v||_1 + \sum_{i,j} \frac{1}{2} ||x_{i,j} - \left(x_i^p + D^v \omega_{i,j}^v\right)||_2^2 \\
& s.t. \ ||\omega_{i,j_1}^v - \omega_{i,j_2}^v||_2 < \mu_2, \forall j_1, j_2 \in S_{i,m}, \forall i \\
& \quad \left(\omega_{i,j}^v\right)_k \geq 0, \forall k
\end{aligned}
\tag{4.3.7}
$$

In (4.3.7), the components w.r.t. unseen variations are removed from the aligned feature as the reconstruction error, while the error due to unperfected variation modelling is reduced by summing the person-specific component and the view-dependent component.

4.3.2 Estimating Frame-level Adaptable Feature Using Multiple Sources Domain Adaptation

Although the variation modeling $\hat{\omega}_{i,j}^{v*}$ is not used in the frame level feature directly, it can be used as a side information to indicate the variations for the frame. In this section, we employ the variation modeling in a multiple sources domain adaptation scheme and then generate a frame-level adaptable feature representation.

To deal with the large intra-view variation, we define a domain using aligned features under the same variation modeling, i.e. features $x_{i_1,j-1}^l$ and x_{i_2,j_2}^l belong to the same domain when $\omega_{i_1,j_1}^{v*} = \omega_{i_2,j_2}^{v*}$. Clearly, it is impossible to learn feature representation or recognition model for all the domain Δ_ω one by one. To estimate such a large number of recognition models, we assume that similar domains share similar recognition model following Domain Adaptation Machine (DAM) [13],, i.e.

$$\lambda_{\omega_1,\omega_2}||f_{\omega_1} - f_{\omega_2}|| < C_f \tag{4.3.8}$$

where $f_{\omega_i}, i = 1, 2$ denotes the recognition model for domain Δ_{ω_i}, $\lambda_{\omega_1,\omega_2} = \left(\omega_1^T \omega_2\right)$ denotes the similarity between domains Δ_{ω_1} and Δ_{ω_2}, and C_f is a positive number measuring the cross-domain recognition model dependency.

Let $\{e_k\}$ denotes the N_D-dimensional standard basis vector set. According to (4.3.8), $f_{e_{n_1}}$ and $f_{e_{n_2}}$ are uncorrelated while all the recognition models are correlated to the basis models $\{f_{e_n}\}$. So we use the basis domains $\{\Delta_{e_k}\}$ as source domains and represent recognition models for all the other domains using the basis models. A recognition model f_ω^* for domain Δ_ω can be estimated by minimizing the average matching model differences in (4.3.8) to the source domains, i.e.

$$f_\omega^* = \arg\min_{f_\omega} \sum_n \omega^T e_n ||f_\omega - f_{e_n}|| \tag{4.3.9}$$

When $\{f_\omega\}$ are linear model and $||f - f_{e_n}|| = ||f - f_{e_n}||_2^2$, we obtain $f_\omega = \sum_n \omega^T e_n f_{e_n} / ||\omega||_1$ by solving (4.3.9). So the domain specific recognition score $\hat{s}_{i,j}^l$ for an aligned feature $x_{i,j}^l$ is given by the linear combination of the scores from basis

models, i.e.

$$\begin{aligned}\hat{s}_{i,j}^{l} &= \sum_{n} \left(\omega_{i,j}^{v*}\right)_{n} f_{e_n}^{T} x_{i,j}^{l} / ||\omega||_1 \\ &= \mathrm{Tr}\left(F \cdot \omega_{i,j}^{v*} \left(x_{i,j}^{l}\right)^{T}\right) / ||\omega_{i,j}^{v}||_1\end{aligned} \tag{4.3.10}$$

where $F = \left[f_{e_1}, f_{e_2}, ..., f_{e_{N_D}}\right]$ is the summarized recognition model matrix and $\mathrm{Tr}(\cdot)$ denotes the trace of input matrix.

The multiple source domain adaptation classification problem is transferred to be a single domain classification problem on high dimensional feature space $x_{i,j}^{l}\left(\omega_{i,j}^{v*}\right)^{T}$ in (4.3.10). However, such a high dimensional classifier F is usually noisy and unreliable when the labeled training data are limited. So we reduce the dimension of feature $\omega_{i,j}^{v*}\left(x_{i,j}^{l}\right)^{T}$ to lower down the complexity of recognition model F.

So far we do not consider the similarity between models $\left\{f_{e_n}^{T}\right\}$, while it exists not only because the invariant structure in person re-id but also because the correlation between atoms in over-complete dictionary. So feature $\omega_{i,j}^{v*}\left(x_{i,j}^{l}\right)^{T}$ cannot be directly vectorized for dimension reduction until the common part between $\left\{f_{e_n}^{T}\right\}$ is removed. To extract the common part, we decompose the basis model f_{e_n} into a universal component f_0 and a domain specific component $f_{e_n}^{\delta}$, i.e. $f_n = f_0 + f_n^{\delta}$. The recognition score $\hat{s}_{i,j}^{l}$ in (4.3.10) can be represented by the sum of score from the universal component and the domain specific components, i.e.

$$\hat{s}_{i,j}^{l} = f_0^{T} x_{i,j}^{l} + \mathrm{Tr}\left(F^{\delta} \cdot \omega_{i,j}^{v*}\left(x_{i,j}^{l}\right)^{T}\right) / ||\omega_{i,j}^{v}||_1 \tag{4.3.11}$$

where $F^{\delta} = \left[f_{e_1}^{\delta}, f_{e_2}^{\delta}, ..., f_{e_{D_n}}^{\delta}\right]$ is the summarized domain specific model matrix.

According to (4.3.11), F is decomposed into a low-dimensional common part f_0 and a domain specific part F^{δ}. So the dimension of feature $\omega_{i,j}^{v*}\left(x_{i,j}^{l}\right)^{T}$ can be reduced after vectorization. Here we employ 1D Principal Component Analysis (PCA) on the vectorization of $x_{i,j}^{l}\left(\omega_{i,j}^{v*}\right)^{T}$ for its simplicity.

Let $x_{i,j}^{dr}$ denote the reduced feature of $\omega_{i,j}^{v}\left(x_{i,j}^{l}\right)^{T}$ and F_{dr} denote the counterpart of F_{δ}, the summarized score $s_{i,j}^{l}$ can be estimated as follows.

$$s_{i,j}^{l} = \begin{pmatrix} F_{dr}^{T} & f_0^{T} \end{pmatrix} \begin{pmatrix} x_{i,j}^{dr} / ||\omega_{i,j}^{v}||_1 \\ x_{i,j}^{l} \end{pmatrix} + \delta_{dr} \tag{4.3.12}$$

where δ_{dr} is a small number denoting the error derived from dimension reduction. From (4.3.12), the adaptable feature $x^a_{i,j}$ is given be the concatenation of the weighted reduced feature $x^{dr}_{i,j}/||\omega^v_{i,j}||_1$ and the aligned feature $x^l_{i,j}$.

4.3.3 Video Level Adaptable Feature via AD Fusion

Following weighted sum-rule fusion scheme, we construct an adaptable and discriminative feature vector to represent each sequence by linearly combining the frame level adaptable feature vectors. The video level feature vector x^{ad}_i is given by the weighted sum of the frame level adaptable feature $x^a_{i,j}$, i.e.

$$x^{ad}_i = \sum_j \rho_{i,j} x^a_{i,j} \tag{4.3.13}$$

where $\rho_{i,j}$ is the adaptability-discriminability (AD) index to be determined in the following procedures.

Since the reconstruction error of the dictionary measures the magnitude of unadaptable features, it indicates the reliability of the adaptable feature. Thus, we determine the adaptability component $\epsilon_{i,j}$ as the reconstruction error, i.e.

$$\epsilon_{i,j} = (x_{i,j} - x^{p*}_i) - D^{v*}\omega^{v*}_{i,j} \tag{4.3.14}$$

Since the elements of the adaptability component may not be equally discriminative, we learn discriminative weights of the them to approximate a discriminability measure for the AD index. Inspired by Fisher's linear discriminant, we define the discriminability measure by the ratio $\sigma^{d,c_{ID}}_{i,j}$ of standard deviation between false matching images (inter-class) and correct matching images (intra-class) to represent the discriminability, i.e.

$$\sigma^{d,c_{ID}}_{i,j} = \frac{N^{c'_{ID}}_i \sum_{i'\neq i,j'} d\left(x^{c_{ID}}_{i,j}, x^{c'_{ID}}_{i',j'}\right)}{\left\{\sum_{i'\neq i} N^{c'_{ID}}_{i'}\right\}\left\{\sum_{j'} d\left(x^{c_{ID}}_{i,j}, x^{c'_{ID}}_{i,j'}\right)\right\}} \tag{4.3.15}$$

where $c'_{ID} \neq c_{ID}$ and $d(\cdot,\cdot)$ denotes a distance function between the two input vectors. To approximate the discriminability measure $\sigma^d_{i,m}$, we formulate the learning problem by linear regression on the adaptability component, i.e.

$$\alpha^* = \arg\min_{\alpha} ||\alpha^T \epsilon_{i,j} - \sigma^d_{i,j}||^2_2 \tag{4.3.16}$$

With the learned weight vector α^*, the AD index is determined as the linear combination of elements in $x^p_{i,j}$ and $\epsilon_{i,j}$ using weight α^* learnt in (4.3.17), i.e.

$$\rho_{i,j} = \epsilon^T_{i,j}\alpha^* \tag{4.3.17}$$

4.3.4 Optimization

We first convert (4.3.6) to a optimization problem searching for minimal reconstruction error with fixed sparsity according to [2], i.e.

$$\begin{aligned}
&\min_{D^v,\omega^v_{i,j},x^p_i} \sum_{i,j} \frac{1}{2}||x_{i,j} - \left(x^p_i + D^v\omega^v_{i,j}\right)||^2_2 \\
&s.t.\ ||D^v_k||_2 = 1, \forall k \\
&\quad ||\omega^v_{i,j}||_1 < \lambda', \forall i,j \qquad (**) \\
&\quad ||\omega^v_{i,j_1} - \omega^v_{i,j_2}||_2 < \mu_2, \forall j_1, j_2 \in S_{i,m}, \forall i \quad (*) \\
&\quad \left(\omega^v_{i,j}\right)_k \geq 0, \forall k
\end{aligned} \tag{4.3.18}$$

Then, we simplify the inequality constraint $(**)$ and $(*)$ in (4.3.18) by computing the mean of them in each subsequence, i.e.

$$\begin{aligned}
&\min_{D^v,\omega^v_{i,j},x^p_i} \sum_{i,j} \frac{1}{2}||x_{i,j} - \left(x^p_i + D^v\omega^v_{i,j}\right)||^2_2 \\
&s.t.\ ||D^v_k||_2 = 1, \forall k \\
&\quad \frac{1}{N^S_{i,m}} \sum_{j \in S_{i,m}} ||\omega^v_{i,j}||_1 < \lambda' \qquad (**') \\
&\quad \frac{1}{{N^S_{i,m}}^2} \sum_{j_1,j_2 \in S_{i,m}} ||\omega^v_{i,j_1} - \omega^v_{i,j_2}||_2 < \mu_2 \qquad (*') \\
&\quad \left(\omega^v_{i,j}\right)_k \geq 0, \forall k
\end{aligned} \tag{4.3.19}$$

By relaxing the inequality constraint $(**')$ and $(*')$ in (4.3.19), we obtain the following optimization problem,

$$\begin{aligned}
&\min_{D^v,\omega^v_{i,j},x^p_i} \sum_{i,j} \frac{1}{2}||x^p_i - \left(x_{i,j} - D^v\omega^v_{i,j}\right)||^2_2 \\
&s.t.\ ||D^v_k||_2 = 1, \forall k \\
&\quad \Omega\left(W^v_{i,m}\right) < \min\left(2\lambda', \mu_2\right), \forall i,m \\
&\quad \left(\omega^v_{i,j}\right)_k \geq 0, \forall k
\end{aligned} \tag{4.3.20}$$

where $W_{i,m}^v = \left[\omega_{i,j_1}^v \quad ... \quad \omega_{i,j_{N_{i,m}^S}}^v\right], j_k \in S_{i,m}$.

Since the variable x_i^p in (4.3.20) is unconstrained, the reconstruction error term in (4.3.20) can be considered as the variance of $x_{i,j} - D^v\omega_{i,j}^v$. So the optimal x_i^p is equal to the mean of $x_{i,j} - D^v\omega_{i,j}^v$. According to [2], we convert problem (4.3.20) to a lasso-like formulation and substitute x_i^p by $x_{i,j} - D^v\omega_{i,j}^v$, i.e.

$$\begin{aligned} \min_{D^v, W_{i,m}^v} \; & \lambda_\Omega \sum_{i,m} \Omega\left(W_{i,m}^v\right) + \frac{1}{2}\sum_{i,j} ||x_{i,j} - D^v\omega_{i,j}^v||_2^2 \\ & - \frac{1}{2}\sum_i ||\frac{1}{N_i}\sum_j \left(x_{i,j} - D^v\omega_{i,j}^v\right)||_2^2 \\ s.t. \quad & ||D_k^v||_2 = 1, \forall k \\ & \left(\omega_{i,j}^v\right)_k \geq 0, \forall k \end{aligned} \tag{4.3.21}$$

Problem (4.3.21) is a special case of joint sparse dictionary learning in which the reconstruction error term is replaced by the variation of reconstruction error and the dictionary representations are nonnegative. It can be solved by an iterative algorithm following [42].

4.4 Experiment

4.4.1 Datasets and Settings

Datasets: Extensive experiments to evaluate our method are conducted on two publicly available sequence based re-identification datasets, i.e. the iLIDS-VID dataset [55] and the multiple shot version of PRID 2011 dataset [19]. iLIDS-VID was captured by a real-world multi-camera surveillance camera network at an airport arrival hall. It contains 300 person image sequences under each of the two camera views. The lengths of image sequences are between 23 to 192, and the average number is 73. The dataset is challenging due to occlusions and significant changes of illumination, view angle and background within and across camera views. The multiple shot version of PRID 2011 dataset (short for PRID2011m) consists of person image sequences recorded from two static surveillance cameras outdoor. Each image

sequence contains 5 to 675 frames and averagely 84 frames. The images in PRID 2011 dataset are of low quality and undergo large view angle and pose change across views. Two different subsets of the PRID 2011 dataset, i.e. PRID2011m full and PRID2011m long, are employed in video based person re-identification. In the first setting, total 200 person video pairs captured under the two camera views are used for experiment [44]. In the second one, only 178 person video pairs with more than 27 frames are employed in the experiment [55] since the reliable video feature may not be extracted from the short videos.

Features and classifier: We extract two kinds of features for each person image in a sequence, i.e. Local Maximal Occurrence [31] and robust optical flow [52]. The Local Maximal Occurrence (XQDA) feature analyzes the horizontal occurrence of local feature and thence reliable against viewpoint changes. Optical flow is widely used in video-base visual recognition, and the robust optical flow integrates the flow over large spatial neighborhoods using median filtering and thence more robust in highly noisy videos. The LOMO-OpticalFlow feature vector of each frame is given by concatenating the LOMO and the optical flow features. For the discriminability measure and the cross-view matching model, we deploy the Quadratic Discriminant Analysis (XQDA) [31], which learns a QDA metric on a low dimensional subspace from cross-view quadratic discriminant analysis.

Evaluation settings: From the Two datasets, the sequence pairs are randomly separated into half for training and the other half for testing. The results are shown in Cumulated Matching Characteristics (CMC) curves. For stable statistical results, the experiments was repeated 10 times and the mean accuracy is reported.

4.4.2 Self-evaluation on dictionary size

The size of dictionary is always an important parameter for dictionary based methods and sometimes crucial to the performance. In this section, we evaluate the impact of dictionary size changes for the proposed method in the iLIDS-VID and the PRID 2011 full datasets as shown in Fig. 4.3. We can see that the rank-one

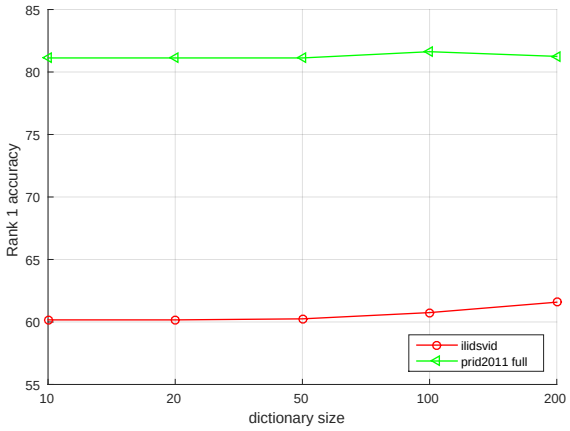

Figure 4.3: Performance of proposed method with different dictionary sizes

accuracy of proposed method in Prid2011 full dataset increases along with the dictionary size when the dictionary size is not larger than 100. That is because the representation ability of the variation dictionary is increasing along with the size. And the performance decreases when the dictionary is larger than 100. It implies an overfitting problem. Since more training data is available in the iLIDS-VID dataset, a larger dictionary size can be employed. In the following section, the dictionary size is set to be 100 in the PRID2011m dataset and 200 in the iLIDS-VID dataset, which is determined by cross-validation.

4.4.3 Evaluation of feature decomposition and AD fusion

We first evaluate whether both the frame level adaptable feature representation and the AD fusion can improve the re-identification performance. For evaluation of the frame level adaptable feature, we compare the original LOMO-OpticalFlow feature vectors $x_{i,j}$ with the frame level adaptable feature (FAF) $x_{i,j}^{dr}$. On the other hand, the proposed AD fusion is compared with the set based distance by calculating the minimum (Min) and Mean of the frame-to-frame distances. The top r rank matching accuracies (%) on the iLIDS-VID and the PRID 2011 full datasets are shown in Table 4.1. The rank one accuracies of the two baseline fusion methods on the original features are 41.9% and 46.9% on iLIDS-VID dataset, respectively. They achieve rank one accuracies of 55.7% and 60.3% on the proposed frame level

Dataset	iLIDS-VID			PRID2011m full		
Rank R	R=1	R=5	R=10	R=1	R=5	R=10
FAF+AD	**61.6**	**84.7**	**92.2**	**81.5**	**95.3**	**99.1**
FAF+Min	55.7	84	92.0	72.9	88.9	92.8
FAF+Mean	60.3	83.7	91.3	70.5	87.8	93.0
Min	41.9	71.8	82.2	65.8	86.8	91.5
Mean	46.9	75.5	85.2	65.9	85.9	90.6

Table 4.1: Top r ranked matching rate (%) comparison in person re-identification.

adaptable feature. This convinces that the proposed frame level adaptable feature is more representative in the complicated person re-identification problem. When the proposed AD fusion method is employed, the rank one accuracy can be further improved to 61.58%. Similar patterns can be observed in results on the PRID2011m dataset. These results indicate that the proposed AD fusion method can extract the adaptable and discriminative features, and thence outperforms both the Min and Mean fusions.

4.4.4 Comparison with multi-shot/video based person re-identifications

Seven state-of-the-art multi-shot/video based person re-identification methods namely Mean Approach Distance (MAD) [28], Dynamic Time Warping (DTW) [50], Discriminative Video fragments selection and Ranking (DVR) [55], Recurrent Neural Network for Video-based Person Re-identification (RNNVPN) [44], Simultaneous Intra-video and Inter-video Distance Learning (SI^{2}DL) [75], Top-push Distance Learning (TDL) [64], and Pose-aware Multi-shot Matching (PaMM) [10], are used for the comparison. Our LOMO-OpticalFlow feature is employed in MAD and DTW for fair comparison, so their performances are better than what were reported by the authors. The top r rank matching accuracies (%) on the two datasets are shown

Dataset	iLIDS-VID			
Rank R	R=1	R=5	R=10	R=20
Proposed	**61.58**	84.67	92.17	98.08
MAD	48.17	75.58	85.42	93.00
DTW	46.42	74.92	85.17	93.33
Salience+DVR	30.9	54.4	65.1	77.1
RNNVPR	58	84	91	96
SI^2DL	48.7	81.1	89.2	97.3
TDL	56.33	**87.60**	**95.60**	**98.27**
PaMM	30.3	56.3	70.3	82.7

Table 4.2: Top r ranked matching rate (%) comparing with multi-shot/video based person re-identification on iLIDS-VID dataset.

in Table 4.2 4.3.

From the results on iLIDS-VID dataset, we can see that the proposed method achieves the best rank-1 accuracy 61.58 and the second best accuracies in rank 5, 10 and 20. On both settings of prid2011 dataset, the proposed method outperforms the other methods remarkably. In the two settings of prid2011 dataset, the proposed method achieves rank one accuracy of 81.52% and 84.83% respectively, which is about 8% higher than other methods. This significant improvement is contributed to our adaptable and discriminative feature representation against noise and large variations within and across camera views. From Table 4.2 4.3, our method outperforms the others on the PRID 2011 dataset under both the two settings.

4.5 Conclusion

In this chapter, we have proposed a new adaptable feature representation method for video based person re-identification via video variation dictionary learning. A video variation dictionary learning algorithm is first proposed to model the large intra-

Dataset	PRID2011m full				PRID2011m long video subset			
Rank R	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
Proposed	**81.52**	**95.25**	**99.13**	**99.75**	**84.83**	**97.20**	**99.46**	**99.86**
MAD	70.50	88.00	92.12	95.38	-	-	-	-
DTW	47.50	68.25	75.88	80.88	-	-	-	-
Salience+DVR	-	-	-	-	41.7	64.5	77.5	88.8
RNNVPR	70	90	95	97	-	-	-	-
SI^2DL	-	-	-	-	76.7	95.6	96.7	98.9
TDL	-	-	-	-	56.74	80.00	87.64	93.59
PaMM	45.0	72.0	85.0	92.5	-	-	-	-

Table 4.3: Top r ranked matching rate (%) comparing with multi-shot/video based person re-identification on PRID2011m dataset.

view variations. Second, multiple sources domain adaptation is adopted to learn a frame level adaptable feature by the variation modeling. Finally, an adaptability-discriminability (AD) fusion is proposed to generate the video level adaptable feature by mining the discriminative information of the frames from the reconstruction error of the variation dictionary,

Experimental results on two public video based person re-identification datasets show that the proposed method achieve better recognition performance for person re-identification than state-of-the-art methods. Comparing the experiments on the two datasets, we observe that the proposed method can achieve more significant improvement in the more challenging dataset, while existing methods deteriorate dramatically in such challenging case. On the other hand, we also demonstrate that both the view-adaptive features and RAD fusion are helpful to improve the re-identification performance.

Chapter 5

Conclusions

5.1 Summary

Person re-identification is an important research topic in computer vision community for its wide range of applications in intelligent video analysis. While the existing person re-identification methods may not fully meet the requirements of person re-identification in large-scale camera network due to the lack of label data and challenging cross-camera variations. As a result, this thesis addresses in two practical scenarios, i.e. unsupervised domain adaptation and semi-supervised learning in large-scale camera network and develops discriminative models for the two scenarios, respectively.

In the unsupervised domain adaptation scenario, labeled data is available from the source views, while only the unlabeled data and unmatched person image pairs (negatives) are available from the target views. This thesis proposes a novel Adaptive Ranking Support Vector Machines (AdaRSVM) method to deal with the unsupervised domain adaptation problem for person re-identification. Without positive image pairs generated by the label information of persons, we relax the discriminative constraint to a necessary condition, which only relies on the mean of positive pairs. In order to estimate the positive mean in the target domain, we make use of the labeled data from the source domain, the negative and unlabeled data from the target domain. With two estimations of the target positive mean, the optimal

combination is determined by the training data. And, the target distance model is trained by adapting the source domain distance model to target domain. Experiment results show that the AdaRSVM method achieves convincing recognition performance for person re-identification. The proposed AdaRSVM not only outperforms non-learning based methods but also is better than the comparing discriminative learning methods using labeled data from the source domain for training.

To further improve the AdaRSVM method, this thesis also proposes a Semi-Supervised Region Metric (SSRM) learning method to estimate target positive neighbors besides target positive mean for discriminative classification model learning. Since the number of positives is very limited in the imbalanced target unlabeled data, we propose to estimate positive neighbors instead of positives. Cross person score distribution alignment and multiple graph based label propagation are jointly optimized for positive neighbor estimation. The positive regions are then generated using the positive neighbors. Finally, a region-to-point metric learning method is presented to learn discriminative metric by maximizing the weighted distance between negatives and positive regions. Experiment results show that the SSRM method improves the per-learned classifier remarkably and achieves convincing recognition performance for person re-identification. The proposed SSRM method not only outperforms comparing semi-supervised methods and classification algorithms robust against label noises, but also is better than the comparing discriminative person re-identification methods.

In the unsupervised domain adaptation scenario, this thesis proposes a novel discriminative feature representation method for video based person re-identification. A video variation dictionary learning algorithm is first proposed to model the large intra-view variations. Second, a frame level adaptable feature is generated by multiple sources domain adaptation using the variation modeling. By mining the discriminative information of the frames from reconstruction error of the variation dictionary, an adaptability-discriminability (AD) fusion is proposed to generate the video level adaptable feature. Experimental results on two public video based person

re-identification datasets show that classification on the proposed video level adaptable feature achieves better recognition performance for person re-identification than state-of-the-art methods.

5.2 Future work

Although the algorithms proposed in this thesis have achieved convincing results in both the two scenarios in large-scale camera network, there is a long way to go for developing a practical person re-identification systems and some issues may be further investigated.

- **Domain adaptive region metric learning** The proposed unsupervised domain adaptive method, AdaRSVM and the metric learning method, SSRM are independent and hence may not be optimal. A better solution should be combining the two methods together and develop a domain adaptive region metric learning. Different from the AdaRSVM that only the positive mean and the negative mean are aligned across domains, the positive regions and the negative regions are aligned across domains in the domain adaptive region metric learning. So the learned classification model is more adaptive and more discriminative to the target domain data.

- **Semi-supervised deep variation modeling for video based person re-identification** The proposed video level adaptable feature is learned using variation dictionary learning. The learned dictionary may not be representative enough in uncontrolled environments with non-linear person's appearance distortion. Deep learning model is considered to be of high representation ability and hence suitable in modeling challenging variations. Moreover, an end-to-end deep model is usually superior to independent feature extraction and classification model learning.

- **Real-Time Person Re-identification** Person tracking in camera network is one of the important applications of person re-identification in intelligent video

analysis. While the proposed methods in this thesis are complicated matching models and hence cannot achieve real-time speed. In order to meet the requirement of person tracking in large-scale camera network, several strategies could be further exploited to increase the training and testing speed. First, a unified framework of the person re-identification methods and the tracking algorithms can be developed such that adaptable features and intermediate results can be shared to avoid repetitive computations. Second, the existing person re-identification methods perform as a person images/videos verification algorithm, which requires a human detector as preprocessing and hence is slow. A better solution should be combining the two ones together and designing a fast target person detector.

Bibliography

[1] *MARS: a video benchmark for large-scale person re-identification*, 2016.

[2] M. Aharon, M. Elad, and A. Bruckstein. k -svd: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Processing*, 54(11):4311–4322, 2006.

[3] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015.

[4] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[5] M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*, volume 104. Prentice Hall Englewood Cliffs, 1993.

[6] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1565–1573, 2015.

[7] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[8] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[9] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *Proc. British Machine Vision Conference*, pages 68.1–68.11, 2011.

[10] Y.-J. Cho and K.-J. Yoon. Improving person re-identification via pose-aware multi-shot matching. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1354–1362, 2016.

[11] W. Deng, J. Hu, and J. Guo. Extended src: undersampled face recognition via intraclass variant dictionary. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34(9):1864–1870, 2012.

[12] C. Ding, C. Xu, and D. Tao. Multi-task pose-invariant face recognition. *IEEE Trans. Image Processing*, 24(3):980–993, 2015.

[13] L. Duan, I. W. Tsang, D. Xu, and T.-S. Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *Proc. Intl Conf. Machine learning*, pages 289–296, 2009.

[14] L. Duan, D. Xu, I.-H. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34(9):1667–1680, 2012.

[15] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2360–2367, 2010.

[16] D. Figueira, L. Bazzani, H. Q. Minh, M. Cristani, A. Bernardino, and V. Murino. Semi-supervised multi-feature learning for person re-identification. In *Proc. IEEE Intl Conf on Advanced Video and Signal-Based Surveillance*, pages 111–116, 2013.

[17] N. Gheissari, T. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, volume 2, pages 1528–1535, 2006.

[18] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Proc. IEEE Intl Conf. Computer Vision*, pages 999–1006, 2011.

[19] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. Scandinavian Conference on Image Analysis*, volume 6688 of *Lecture Notes in Computer Science*, pages 91–102. 2011.

[20] M. Hirzer, P. M. Roth, M. Kostinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *Proc. European Conf. Computer Vision*, pages 780–793. 2012.

[21] X.-Y. Jing, X. Zhu, F. Wu, X. You, Q. Liu, D. Yue, R. Hu, and B. Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 695–704, 2015.

[22] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Person re-identification by unsupervised l1 graph learning. In *Proc. European Conf. Computer Vision*, 2016.

[23] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2288–2295, 2012.

[24] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(7):1622–1634, 2013.

[25] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[26] J. Li, A. J. Ma, and P. C. Yuen. Semi-supervised adaptable feature learning for video based person re-identification. 2017.

[27] J. Li, A. J. Ma, and P. C. Yuen. Semi-supervised region metric learning without using positive data for person re-identification. 2017.

[28] W. Li, Y. Wu, M. Mukunoki, and M. Minoh. Locality based discriminative measure for multiple-shot person re-identification. In *Proc. IEEE Intl Conf on Advanced Video and Signal-Based Surveillance*, pages 312–317, 2013.

[29] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: deep filter pairing neural network for person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[30] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 3610–3617, 2013.

[31] S. Liao, Y. Hu, X. Zhu, and S. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2197–2206, 2015.

[32] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *Proc. IEEE Intl Conf. Computer Vision*, pages 3685–3693, 2015.

[33] C. Liu, C. Loy, S. Gong, and G. Wang. Pop: person re-identification post-rank optimisation. In *Proc. IEEE Intl Conf. Computer Vision*, pages 441–448, 2013.

[34] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *Proc. IEEE Intl Conf. Computer Vision*, pages 3810–3818, 2015.

[35] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 3550–3557, 2014.

[36] X. Liu, H. Wang, Y. Wu, J. Yang, and M. H. Yang. An ensemble color model for human re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 868–875, 2015.

[37] A. J. Ma, J. Li, P. C. Yuen, and P. Li. Cross-domain person re-identification using domain adaptation ranking svms. *IEEE Trans. Image Processing*, 24(5):1599–1613, 2015.

[38] A. J. Ma and P. Li. Semi-supervised ranking for re-identification with few labeled image pairs. In *Proc. Asian Conference on Computer Vision*, pages 598–613, 2014.

[39] A. J. Ma, P. C. Yuen, and J. Li. Domain transfer support vector ranking for person re-identification without target camera label information. In *Proc. IEEE Intl Conf. Computer Vision*, 2013.

[40] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *Proc. Intl Workshop on Re-Identification in conjunction with European Conference on Computer Vision*, pages 413–422. 2012.

[41] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017.

[42] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proc. Intl Conf. Machine learning*, pages 689–696, 2009.

[43] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1363–1372, 2016.

[44] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1325–1334, 2016.

[45] A. Mignon and F. Jurie. Pcca: a new approach for distance learning from sparse pairwise constraints. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2666–2672, 2012.

[46] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 3318–3325, 2013.

[47] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1306–1315, 2016.

[48] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *Proc. British Machine Vision Conference*, pages 1–11, 2010.

[49] A. Roy, S. Sural, and J. Mukherjee. A hierarchical method combining gait and phase of motion with spatiotemporal model for person re-identification. *Pattern Recognition Letters*, 33(14):1891–1901, 2012.

[50] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler. Re-identification of pedestrians in crowds using dynamic time warping. In *Proc. Intl Workshop on Re-Identification in conjunction with European Conference on Computer Vision*, pages 423–432, 2012.

[51] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *Proc. European Conf. Computer Vision*, pages 475–491, 2016.

[52] D. Sun, S. Roth, and M. Black. Secrets of optical flow estimation and their principles. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2432–2439, 2010.

[53] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *Proc. European Conf. Computer Vision*, pages 791–808, 2016.

[54] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *Proc. European Conf. Computer Vision*, pages 135–153, 2016.

[55] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *Proc. European Conf. Computer Vision*, pages 688–703, 2014.

[56] L. Wei, Y. Tian, Y. Wang, and T. Huang. Swiss-system based cascade ranking for gait-based person re-identification. In *Proc. AAAI Conf. Artificial Intelligence*, pages 1882–1888, 2015.

[57] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1249–1258, 2016.

[58] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *Proc. European Conf. Computer Vision*, pages 1–16, 2014.

[59] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *Proc. European Conf. Computer Vision*, pages 701–716, 2016.

[60] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive SVMs. In *Proc. ACM Intl Conf. Multimedia*, pages 188–197, 2007.

[61] M. Yang, L. Van, and L. Zhang. Sparse variation dictionary learning for face recognition with a single training sample per person. In *Proc. IEEE Intl Conf. Computer Vision*, pages 689–696, 2013.

[62] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Li. Salient color names for person e-identification. In *Proc. European Conf. Computer Vision*, pages 536–551, 2014.

[63] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *Proc. Intl Conf. Pattern Recognition*, pages 34–39, 2014.

[64] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1345–1353, 2016.

[65] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1239–1248, 2016.

[66] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan. Sample-specific svm learning for person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1278–1287, 2016.

[67] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: person re-identification with human body region guided feature decomposition and fusion. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[68] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *Proc. IEEE Intl Conf. Computer Vision*, pages 2528–2535, 2013.

[69] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience kearning for person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 3586–3593, 2013.

[70] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian. Person re-identification in the wild. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[71] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2013.

[72] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[73] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[74] P. Zhu, L. Zhang, W. Zuo, and D. Zhang. From point to set: extend the learning of distance metrics. In *Proc. IEEE Intl Conf. Computer Vision*, pages 2664–2671, 2013.

[75] X. Zhu, X.-Y. Jing, F. Wu, and H. Feng. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In *Proc. Intl Joint Conf. Artificial intelligence*, pages 3552–3559, 2016.

[76] X. Zhu, X. Y. Jing, L. Yang, X. You, D. Chen, G. Gao, and Y. Wang. Semi-supervised cross-view projection-based dictionary learning for video-based person re-identification. *IEEE Trans. Circuits and Systems for Video Technology*, PP(99):1–1, 2017.

Curriculum Vitae

Academic qualifications of the thesis author, Mr. LI Jiawei:

- Received the degree of Bachelor of Mathematics and Applied Mathematics from Sun Yat-Sen University, July 2007
- Received the degree of Master of Mathematics from Sun Yat-Sen University, July 2011.

July 2018