中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

Person Re-Identification Using Kernel Metric Learning

作者姓名：Syed Muhammad Adnan

指导教师：焦建彬教授

中国科学院大学电子气与通信工程学院

学位类别：工学博士

学科专业：计算机应用技术

培养单位：中国科学院大学电子电气与通信工程学院

2018 年 6 月

Person Re-Identification Using Kernel Metric Learning

A dissertation submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in Computer Applications and Technology

By

Syed Muhammad Adnan

Supervisor： Professor Jianbin Jiao

School of Electronic, Electrical and Communication Engineering

University of Chinese Academy of Sciences

June 2018

中国科学院大学
研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：

日期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：导师签名：

日期：日期：

摘要

行人再识别是计算机视觉领域研究的热门问题之一，其主要研究在多摄像机网络间的行人匹配。由于观测到的行人可能经历多种非线性变换（包括姿态、视角、光照和背景变换等），因此，给多摄像机网络间的行人匹配问题带来了巨大的挑战。非重叠视域下的多重图像变化不仅是非线性的，而且还会因为不同的人在不同视图中经历了不同的非线性变化集，从而累加成为复杂的非线性变换。为了在图像集中获得可靠的正确的匹配，选择能有效处理这些非线性变换的、鲁棒的特征和匹配函数是十分重要的。

学习多核的目的是为了更好地利用不同的内核对一个人的特征空间进行建模，因为在复杂环境影响之下，人的非线性变换遵循的分布是并不确定的。此外，不同行人的非线性变换方式也是不同的。多核投影能够有效地模拟现实世界中的复杂变化，从而在特征空间提供更好的判别特征。本文的主要工作，如下：

1、 基于多核，将特征进行投影及级联，有效地提升了度量矩阵的针对于样本变化多样性的判别能力；

2、 针对行人再识别样本集合较小的情况，提出了一种基于小样本的局部多核学习方法用于行人再识别，进一步解决了样本变化多样性的问题；

3、 针对行人再识别领域中的难负样本问题，采用一种多模态的难负样本挖掘的度量学习方法，很大程度上剔除了难负样本，并提高了匹配的准确率。

关键词：行人再识别、多核学习、度量学习、局部多核学习、难负样本挖掘

Abstract

Person Re-Identification matches persons observed in non-overlapping camera views. The person re-identification is a challenging problem due to the fact that the observed pedestrians undergo different random and non-linear changes including pose, viewpoint, illumination, and background changes. In such situation, the images of multiple non-overlapping views not only become non-linear, but also multi-modal. These changes are also different in a given intra-view, as well as, in inter-views. To obtain the actual match of the query at rank@1 from the *Gallery* view, it is then necessary to obtain both discriminative features and matching function robust against these complex non-linearities.

Therefore, taking all these challenging issues in re-identification: feature space, the number of samples, as well as, the complex non-linearities in feature space, we have adopted to learn multiple kernel projection for re-identification. The aim of learning multiple kernel is to better utilize different non-linear kernels in modeling the complex non-linearities in both global image space, as well as, image space in each disjoint view. The main contribution of the paper is listed as:

1. Multiple kernel learning finds the weighted combination of different non-linear kernels to optimally address complex changes.

2. Further, in real world scenes the non-linearities among different persons are different; multiple kernel learning helps in modeling the complex changes in the observed persons from real world scenes, and provide much better discrimination among persons in the feature space.

3.The metrics in re-Identification suffer badly due to presence of impostor samples; therefore, the learned metric must be robust against impostors. Therefore, we have learned an impostor resistant metric that can both address multi-modal feature space, as well as, impostor resistance, simultaneously. The learned impostor resistant metric can largely reject the impostors from real world and can attain

maximum matching with the actual Gallery sample.

Key Words: Person Re-Identification, Multiple Kernel Learning, Metric Learning, Localized Multiple Kernel Learning, Impostor Resistance.

目录

图目录

表目录

Chapter 1. **Introduction**

1.1 Background and Motivation

Person re-identification is the task of finding the correct match of a query person from a large *Gallery* of unknown persons [1, 2]. It is a recently emerged problem in computer vision. The objective of re-identification is to retrieve or search, identify or recognize a specific person observed in a non-overlapping camera network. The observed persons in disjoint views undergo complex changes in pose, viewpoint, background, and also experience occlusion and illumination changes, making re-identification a very challenging problem.

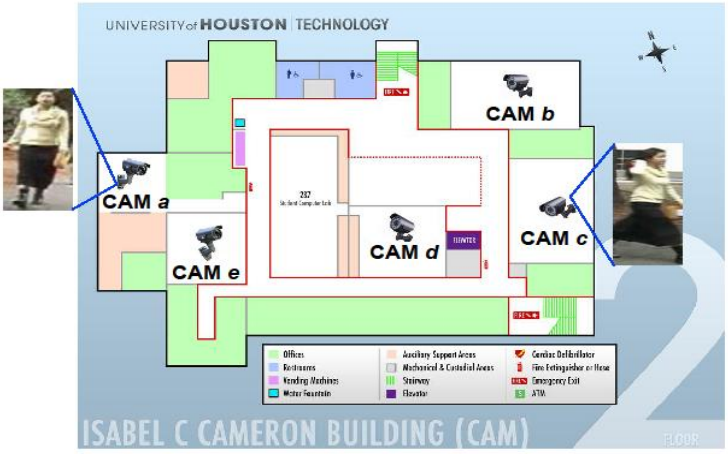

Figure 1.1 Re-Identification Camera Network in Real World

In Fig.1.1, a typical re-identification network is shown with five disjoint camera views (i.e. cam *a* to cam *e*). This camera network is installed in Isabel C. Cameron Building of Houston University, and in one view of the network a lady is observed in cam *a* when walking around the building. And then later, the same lady is observed and identified in cam *c*. Thus, the possible applications of re-identification in real world are person identification, monitoring, and tracking in public and private spaces including university campuses, public bus stations, railway stations, airports and etc. In recent years, it has gained a lot of attraction from scientific community.

Re-Identification can be of two types, which are:

- Close Set Re-identification
- Open Set Re-identification

Close set is a kind of re-identification where each query sample in *Probe* view has one or more gallery samples available in the *Gallery* view, some of the typical examples are shown in Fig.1.2.

Figure 1.2 Probe and Gallery sets in Close Set Re-Identification

While, in the open set there are more gallery images in the *Gallery* view than the number of query samples observed in *Probe* view, it means there are some other additional identities observed in the *Gallery* view that are not observed in the *Probe* view. Open set condition [3] is more like real world scenario. In Fig.1.3, an open set condition can be seen, where the gallery samples of the observed query images are enclosed within green rectangles. While, the extra identities observed in *Gallery* view are shown without green rectangles.

Figure 1.3 Probe and Gallery sets in Open Set Re-Identification

Close set re-identification is mostly performed in person re-identification.

Close set re-identification means that both the *Probe* and *Gallery* views have the same number of identities observed and each identity observed in *Probe* view has at least one sample observed in *Gallery* view.

1.2 Related Work

Based on the close set identification, re-identification is seen as a cross camera matching problem, where the persons undergo complex non-linear changes in each disjoint view. Most of the research in re-identification can be divided into two categories, (i) Feature engineering, and (ii) Distance metric learning. In Fig. 1.4, the general architecture for matching persons in close set re-identification is given.

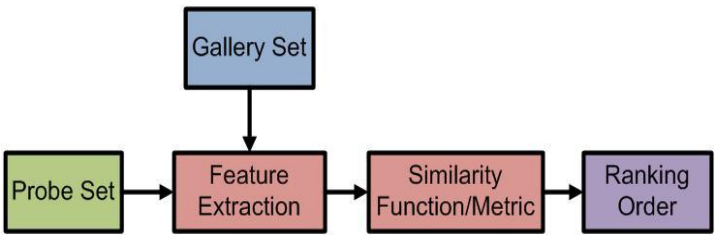

Figure 1.4 General Framework of Matching Persons in Re-Identification

In Fig.1.4 for feature extraction many state of the art features have been proposed including Color naming [4], LOMO [5], SCNCD [6], and GoG [7]. The objectives of all these feature representations is to find stable and invariant representation of a person in each disjoint view. Though, these features have gained success in identifying a person, but still, feature learning and designing is an open problem in re-identification due to the fact that persons in re-identification are multi-modal and experience several random non-linear changes all together at the same time. Therefore, the re-identification image space, and consequently feature space is complex, non-linear and multi-modal.

Further, several different similarity matching functions are also proposed in re-identification to maximize the cross camera person matching [5, 8, 9, 10, 11, 12]. These state of the art matching functions obtain a low dimension subspace where the

objective is to maximize the matching between the positive pairs, while, minimizing the matching against the negative samples. However, none of these metrics have given due consideration to the complex and non-linear feature space that exists in the re-identification.

Recently, kernel based metrics are learned to address non-linear feature space in re-identification. Xiong et.al [13] has applied different kernel functions including Linear, RBF, and Chi^2 to project the features of a person (where features are formed by concatenating all the different extracted features), and then later used the projected features to train a metric. D. Chen et.al in [14] proposed a kernel based metric for person re-identification. Their work learned kernelized metric by utilizing only polynomial kernel, however, different than [13] they proposed an ensemble of different metrics trained with different polynomial kernel functions (i.e. different degree of polynomial kernel). In total, their work uses ensemble of six polynomial kernels, and thus, has attempted to model all the different complex non-linear changes in different disjoint views.

Though, these kernel based metrics are robust, however, these metrics have ignored to realize the fact that each person in disjoint views undergoes complex, non-linear and random changes, i.e. random non-linear changes in pose, viewpoint, and illumination. While, these kernel based metrics during learning have assumed that all the observed images in disjoint views globally lie on the same non-linear manifold, and the non-linearity remains uniform over this entire global image space of re-identification. Thus, these methods projected all the observed persons in re-identification globally into a single chosen kernel projection without taking into care the non-linearity of each single identity.

1.3 Challenges and Objectives

In Fig.1.4, the general re-identification architecture can be challenged by many different issues, which are related to the image acquisition and its quality, feature

extraction, learning distance metrics for addressing complex non-linear and multi-modal feature space. Mainly, all these challenging issues can be categorized into three types which are listed below in Table.1.1. And some of the challenges are shown in Fig1.5.

Table 1.1List of Challenges in Re-Identification

Sources	Problems
Cameras, Sensors, and Environment	• Color and Light Changes; • Background Changes; • Viewpoint Changes;
Human Appearance and Attributes	• Pose Changes; • Common Attributes among different persons; • Mis-alignment of Body Parts;
Learning Matching Functions and Feature Designing	• Pose Invariant Feature Designing; • Color or illumination Invariant Feature Designing; • Multi-modal and Non-Linearity; • Small Sample Size; • Over fitting.

Figure 1.5 General The challenges of the person re-identification. a. Pose Change and Body Part Mis-alignment; b. Illumination and Viewpoint changes, plus Body Part Mis-alignment; c. Viewpoint and Background changes, plus Body Part Mis-alignment; d. Illumination and Viewpoint changes, plus shadowing; e. Pose change, plus Body Part Mis-alignment and shadowing; f. Pose and Viewpoint changes, plus Background changes; g. Pose and Background changes, plus Body Part Mis-alignment. (Source from VIPeR Dataset, and each column represents the same person).

Due to the above challenges, the objective of our work and this thesis is to develop method for feature projection and learning global metric that can resolve complex re-identification issues, and therefore, can maximize the matching among the same persons with changing poses, viewpoints, and illuminations. For this purpose, we will perform a detail study over the re-identification feature space, and also develop different feature projection methods that can maximize the discriminating capabilities of different features. Further, re-identification suffers from complex changes; therefore, one single feature is not discriminative enough to be invariant against all the different non-linear changes. Thus, our objective is to utilize several discriminative features, and then develop an integration method that is designed specifically for re-identification scenario to optimally combine all the different features.

1.4 Thesis Organization

The thesis is organized as follows:

1. In chapter 2, we will cover in detail the recent progress in person re-identification and kernel based methods in re-identification.

2. Chapter **3** discusses in detail all the steps that we have performed to learn global multiple kernel, and then using global multiple kernel in feature projection for global metric learning. We will provide a thorough insight of our multiple kernel learning approach, and learn a global metric for person re-identification.

3. In Chapter **4,** we have explained the extension of global multiple kernel into a local multiple kernel for each single person, and we have described how we can obtain weights of pre-selected kernels for small sample size problem of re-identification, where convex based multiple kernel learning cannot be converged.

4. In Chapter **5**, the insight detail of multi-modal metric learning for person re-identification is provided, and then later using impostors from the *Gallery*, as well as, *Probe* views to further improve the discriminating power of the learned

multi-modal metric is described in detail. In last, we have provided the conclusion of this work, along with the future work that we have intended to do.

Chapter 2. Related Work for Kernel Metric Learning Based Person Re-Identification

This chapter covers the details of recent literature in re-identification. First, we describe the proposed feature extraction methods to address illumination changes, pose changes, as well as, describe the human appearance and human attributes to maximize the matching. Then, we cover the kernel based methods to learn distance metrics for re-identification. Finally, we discuss the details of multiple kernel learning in person re-identification and its challenges.

2.1 Feature Extraction for Re-Identification

2.1.1 Hand-Engineered Features for Re-Identification

In re-identification, person appearance changes in different disjoint views, therefore, many state of the art features are engineered to obtain invariant features of a person in a pair of non-overlapping views. These hand engineered features include LOMO [5], SCNCD [6], GoG [7], BiCov [16], biologically inspired covariance features [17], and mOM [18].

LOMO (Local Maximal Occurrence) features [5] are designed to address viewpoint changes in re-identification. Their methods uses sliding window over a horizontal strip to obtain overlapping patches. From each horizontal patch of size 10 x 10 pixels local features are obtained. Then the extracted feature patterns over all the patches along the horizontal strip are checked to maximize the local occurrence of feature pattern to be used as the feature of the horizontal strip.

In another work in [7] a hierarchical Gaussian feature distribution, denoted as GoG, is proposed to obtain both locally and globally view invariant features. GoG [7] features are designed to address the shortcoming in covariance features which contain no mean information. Similar to LOMO [5] local image region is used to obtain features, while, in each local region local patches are densely extracted to obtain color and texture features [19]. After, extracting features a patch Gaussian

distribution computed, then, later using the patch Gaussians of all the patches a region Gaussian is computed for image region. Finally, powerful Gaussian features are obtained to represent the person containing both mean and covariance information.

Further, there is one shortcoming in GoG [7], which is that its assumption is based on Gaussian distribution. However, if the distribution is not Gaussian, then the features obtained do not provide discriminative feature. Therefore, an empirical moment matrix, denoted as mean of moments mOM [18] is proposed. Moment matrix has the ability to address any arbitrary non-Gaussian distribution, and provides much richer statistics than mean and covariance of pixel patch.

N. Gheissari et.al [20] proposed to extract global features for the whole person body, in contrast, to just pay attention on only passive biometrics features such as face and gait. In their work, novel spatiotemporal segmentation algorithm is employed to generate salient edgels that are robust to changes in appearance of clothing. The invariant signatures are obtained by combining normalized color and salient edgel histograms. T. Lorenzo et.al [21] proposed another feature extraction and feature correspondence method using a graph based method. Their work developed a new approach for establishing correspondences between sparse image features of the two images observed in two disjoint views. After, correspondence among patches and features they formulated the matching task as an energy minimization problem by defining a complex objective function of the appearance and the spatial localization of the features.

Further, as different features carry different information about the person, as well as, it is also possible that some features are more discriminating for a person than the other extracted features. Therefore, recently feature weighting methods are also proposed in re-identification to obtain better feature representation. C. Liu et.al, [22], presented the study and experimental results on feature weighting and its importance. Their study showed that certain features play more important roles than others under

different circumstances. Consequently, they proposed a novel unsupervised approach for learning a bottom-up feature importance, so features extracted from different individuals are weighted adaptively driven by their unique and inherent appearance attributes. D. Figueira et.al [23], also presented feature learning technique for person re-identification. Their motivation lies in the fact that there have been already different hand crafted novel features proposed to highlight discriminant parts of human bodies, as well as, there have been already different feature learning methods proposed to extract pedestrian features. However, there has been no work to combine the two feature extraction methods to further obtain more discriminating features among all the different individuals.

2.1.2 Deep Learned Features for Re-Identification

A convolutional neural network in [24] is trained to compute cross-input neighborhood differences between two images, and capture their local relationships to obtain higher-level features. Following the framework of cross images, many state of the art deep feature extraction methods are proposed recently, including SIR-CIR [25], MCP-CNN [26], Spindle-Net [27], Deep Context aware framework [28], domain guided dropout framework [29], and deeply part aligned network [30].

A novel filter pairing neural network FPNN is proposed in [34] that can jointly handle misalignment, photometric and geometric transformations, and background clutter. FPNN automatically learns features from the training data that are optimal for the re-identification task. Further, in re-identification due to the changing poses and viewpoints the human body parts get displaced, and thus for matching the observed query and gallery of the same person a robust method is needed that attain correspondence between the pair of images semantically. Therefore, Zhao et.al in [27] proposed a deep convolutional network, called as "Spindle Net" to learn human body parts guided multi-stage features. Spindle Net learns the semantic macro and micro body features which remain stable across views, and then, it uses a weighting method to discriminate among the features obtained from different body regions. In another

work, Li et.al [28] proposed a deep context aware feature learning network, where multi-scale deep convolutional network is designed to hierarchically obtain powerful features for human body both globally, as well as, locally.

2.1.3 Feature Transformation and Mapping for Re-Identification

In conventional re-identification methods, discriminative features are designed or learned from the training set. However, there exist complex changes in illumination in a pair of disjoint views. These complex changes do affect the learned features from the training set. Therefore, some researchers have been research about camera transformation function or feature mapping to overcome these effects [31, 32, 33, 34, 35].

L. An et.al [32] proposed a robust canonical correlation analysis (ROCCA) to match people from different disjoint views by projecting them into common subspace. Given a small training set, canonical correlation analysis (CCA) may lead to poor performance due to the computation issue in the covariance matrices. The proposed ROCCA has solved this issue by using the shrinkage estimation and smoothing technique.

G. Lisanti et.al [33] addressed the problem of person re-identification across non-overlapping views by proposing a discriminative, however, robust kernel descriptor to encode the appearance of a person. Then the matching between the pair of images is obtained by projecting the extracted kernel descriptors into a common subspace. This subspace is learned by applying Kernel Canonical Correlation Analysis (KCCA).

Brightness transfer function in [36] is proposed to model the appearance changes by using a novel Weighted Brightness Transfer Function (WBTF), which assigns unequal weights to different observations. WBTF is applied to high-dimensional color and texture features for feature normalization and then later used for matching.

In [37], the implicit camera transfer function is proposed which models the camera transfer by using a binary relation that jointly maps the illumination of the pair

of images of a person to a common illumination map.

2.2 Metric Learning for Person Re-identification

2.2.1 Metric Learning for Re-Identification

Although there have been many discriminative novel features are learned in person matching, there is still one more challenge for re-identification, which is how to learn a similarity function that can give perfect matching between the extracted features. Therefore, there have been a lot of different metrics used and proposed for person re-identification, including XQDA [5] , LMNN [8], ITML [9], KISSME [10], and VAML[11] metrics.

In [10], the authors propose to learn a matching function named KISSME (KISSME stands for keep it simple and straightforward) for re-identification in contrast to the traditional distance metric methods that are based on nearest neighbor approach. Their metric is simple and is obtained by computing the maximum likelihood between two covariance matrices, which are the covariance matrices of pair of dissimilar samples and pair of similar samples.

Further, [11] learned a view adaptive metric learning is proposed to overcome pose and viewpoint problems. View-adaptive metric learning (VAML) method adapts different metrics for different image pairs with changing viewpoints. Initially, VAML estimates the view angles of each of the given pair of images, and then adaptively obtain a view specific metric for the pair of images.

Li et.al. in [38] proposed an adaptive decision function for matching the observed persons in non-overlapping views. Unlikely, LMNN [8] that learns a global metric using fixed threshold for each single identity, [38] uses the local threshold for each single identity. This local threshold is obtained through local function that uses the distance between the given pair of samples, and is used to impose a more stringent learning constraint.

2.2.2 Kernel Based Metric Learning for Re-Identification

Conventional mahalanobis metrics have performed well, but, because the re-identification image space is generated from network of cameras, each with varying illumination and view angle, therefore, the obtained images are multi–modal and non-linear in the feature space. Simple mahalanobis metric learning is not discriminative enough to discriminate among different non-linear image pairs. Therefore, kernelization method is needed to address the non-linearity in the feature space [13, 14, 39, 40].

Xiong et. al. in [13] extensively evaluate the performance of kernel Local Fisher Discriminant Analysis and a ranking ensemble voting scheme on state of the art re-identification datasets. In [14] the learning-to-rank methodology is proposed to learn a similarity function to maximize the difference between the similarity scores of matched and unmatched images for a same person.

However, all the above kernelized metric learning approaches used single global kernel for projecting all the different observed non-linear persons, which cannot discriminate well all the different observed persons in re-identification dataset. In contrast, different persons undergo different non-linear changes, and thus, these approaches in a principle way lack to model the multi-modal feature space.

Recently, proposed multiple kernel learning can well address the complex heterogeneous and multi-modal feature space by learning a linear weighted combination of different non-linear kernels. Further, after projection in multiple kernels the discrimination among samples is further improved. The multi-modal image space in re-identification is address by learning view specific mappings. In [31] across views metric learning method is proposed, and in [41] feature mappings are learned between disjoint cross views. The proposed approach in [41] uses a gating function to automatically partitions the image spaces of two camera views into different modality spaces and then a metric for each different modality space is learned. L. An et.al [42], proposed a reference-based method for cross camera person

re-identification. In the training stage, a subspace is learned in which the correlations of the reference data from different cameras are maximized using Regularized Canonical Correlation Analysis (RCCA). For re-identification, the gallery samples and the probe samples are projected into the RCCA subspace and the reference descriptors (RDs) of the gallery and probe are constructed by measuring the similarity between them and the reference data. Further, the multi-modal feature space in re-identification is augmented by "depth" features. F. Pala et.al [43] experimented RGB-D multi modal re-identification. It is inspired with the idea of fusing clothing appearance cues with other modalities could be exploited as additional information sources, such as gait.

In contrast to these recent multi-modal methods in re-identification, our work does not add any further modality information source in re-identification, instead, the objective is to exploit well the different non-linear modality spaces in re-identification using multiple kernel learning.

2.3 Kernel Spaces

To obtain a discriminative kernel projection, it is a very critical step to choose a kernel that can well model the non-linearity. It is also a critical step of choosing the best kernels for multiple kernel learning. Inspired with the previous multiple kernel based methods [44, 45], several recent methods in classification have shown RBF, χ^2, and polynomial kernels are the best performing kernels in many computer vision problems. These kernels in our work are defined as:

$$K_{RBF}(x_i^a, x_i^b) = \exp^{-(\frac{\|x_i^a - x_i^b\|_2^2}{2\sigma^2})} \tag{2.1}$$

$$K_{RBF}(x_i^a, x_i^b) = \exp^{(-\alpha\sum_i \frac{\|x_i^a - x_i^b\|_2^2}{x_i^a + x_i^b})} \tag{2.2}$$

$$K_{Polynomial}(x_i^a, x_i^b) = (< x_i^a, x_i^b > +1)^m, m \in N \tag{2.3}$$

σ values in RBF kernels are set to values of 0.2, 0.3, 0.4, 0.6 and 0.8. While, the values for α in χ^2 kernels are set to values of 0.3, 0.5, 0.7, and 1. For the polynomial kernels the degree of polynomial *m* is set to values 2, 3, and 5.

2.4 Summary

This chapter covers the detail review of recent literature in person re-identification with respect to feature learning and metric learning. We have given detail literature review for the kernel based metric learning in re-Identification; multi-modal metric learning, and the multiple kernels based re-identification.

Chapter 3. Multiple Kernel Metric Learning Person Re-Identification

In re-identification the observed images are not high quality, therefore, the human faces and other biometric cues are not clear. Hence, appearance based matching methods are widely used to match pedestrians.

In re-identification the observed persons from real world scenes undergo complex non-linear changes in pair of non-overlapping views. These non-linear changes include changes in pose, viewpoint, and illumination [13]. Further, in each disjoint view different persons undergo different non-linear changes. Therefore, the naive assumption in the previous distance metrics [12,13,46,47] that all the observed persons in the disjoint views reside globally on a single non-linear space, as well as, the non-linearity over this entire space remains uniform is not true in real world scenes. Therefore, to handle this situation where the global image space is multi-modal, as well as, has different complex non-linear distributions in each disjoint view, a carefully learned feature projection method is needed.

3.1 Motivation

With detail analysis of re-identification image space, it is clearly evident that re-identification image space is complex, non-linear and multi-modal. This complex non-linear image space can largely affect the person matching across disjoint views. Further, from the previous proposed kernel based metrics [13,14], it is also evident that even after kernel projection the complex non-linear image space still remains challenging for person matching. This is mainly due to the fact that none of these proposed kernel metrics have carefully investigated the non-linearity in re-identification feature space.

To handle such complex non-linear global feature space from a pair of non-overlapping views where intra-view, as well as, inter-views have different non-linear changes, we have opted to use a multiple kernel learning for metric

learning [44,45].

3.2 Methodology and Working Details

In this section we give the methodology and the implementation details for multiple kernel learning and multiple kernel learning based global metric learning. The detail procedure of multiple kernel metric learning is shown in Fig.3.1. In our method the dataset is randomly partitioned into two sets, one for training and one for testing. The training set is used for learning the multiple kernel and the metric.

First, different features are extracted from the training samples as shown in Fig.3.1. Then, the weights of multiple kernels are calculated. In our method, the weights of all the chosen kernels (i.e. RBF, Chi^2, and Polynomial kernels) are obtained globally for the training set. Finally, a global distance metric is learned after projecting the features into the globally learned multiple kernel space. After learning the global metric, we then perform the matching between test query and test Gallery.

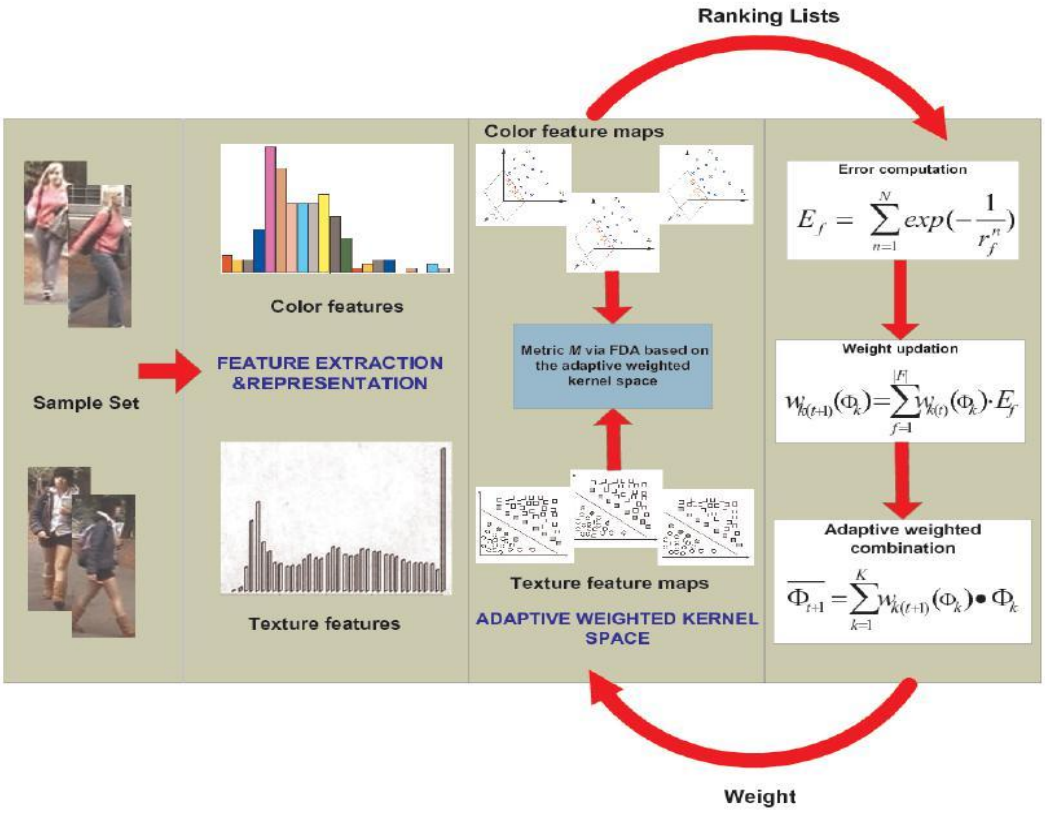

Figure 3.1 Methodology of Multiple Kernel Metric Learning

3.2.1 Feature Extraction

In our method, the image is first divided into six horizontal bands and then patch wise (each patch 6 x 6) features from each horizontal band are extracted, similar to the method in Hu et. al. [48]. In each patch, color feature histograms of HSV, RGB, YUV channels are extracted where each channel uses32-bins histogram. Then, a hand engineered feature, denoted as Color Naming (CN) (Shet et al. [4]) is also extracted from each patch of the image. The CN uses eleven basic color names, i.e. black, blue, brown, grey, green, orange, pink, purple, red, white, and yellow, to describe color naming of all the image pixels. The idea of color naming is that given the mapping from RGB values, it will find the probability of each pixel value to be associated to a certain color. Thus, it obtains a semantic color association histogram descriptor of each pixel in an image.

After extracting color features, we extract texture features to further augment the features about the person for higher matching. The texture feature is obtained by Local Binary Pattern proposed by An et al. [42] that includes uniform 8 neighbors of radius 1 and uniform 16 neighbors of radius 2 from each patch. In Fig.3.2, different persons are shown wearing textured clothing where the texture features can help in distinctively identifying these persons.

In our work, all these different extracted features are stored into a set F as:

$$F = \{F_f\}_{f=1.....|F|}, \tag{3.1}$$

where f is a one single type of extracted feature, while $|F|$ is the total number of extracted features. After extracting all these features the next step is to learn the weights of multiple kernels (already pre-selected) using these extracted features.

Figure 3.2 Persons Wearing Textured Clothing

3.2.2 Multiple Kernel Learning

In our work, instead of using the conventional convex optimization [44,45] to learn weights of multiple kernel, we used a qualitative measure approach to obtain the weights via an analytical method.

With the obtained |*F*| feature maps of a kernel *k*, the weight of the kernel *k* is then computed globally for all the *N* training samples using our proposed qualitative method. In our work, the qualitative method takes one feature map of kernel *k* at a time (i.e. in memory only a single feature map is needed to be loaded) and finds its discriminating capability to match each individual training sample *n* (i.e. query x_f^i) with its gallery sample x_f^j at rank@1 position. This qualitative measure in our work is referred as, Ranking Error *E*, and is computed for the given kernel *k* using the feature *f* as:

$$E_{f,k} = \sum_{n=1}^{N} e^{-\frac{1}{r_{f,k}^n}} \tag{3.2}$$

where $r_{f,k}^n$ is the rank order of query x_f^i obtained for the feature map *f* using the given kernel *k*. This rank order is computed by using the similarity matching function M_{ini} (i.e. a pre-learned initial metric obtained using LFDA [12]). Metric M_{ini} computes the matching between query x_f^i and gallery x_f^j as:

$$Dist_{\phi_k(x_f^i),\phi_k(x_f^j)} = (\phi_k(x_f^i) - \phi_k(x_f^j))^T M_{ini} (\phi_k(x_f^i) - \phi_k(x_f^j)) \tag{3.3}$$

where M_{ini} is a pre-computed metric learned using L-FDA [12], and is trained using the *N* randomly selected training samples, $\Phi_k(x_f^i)$ is the kernel Φ_k projection of feature type *f*. In Eq.3.2, the ranking error *E* of a kernel *k* is computed over the whole training set (i.e. *N* samples) using the rank orders of each individual training sample *n*, and thus, the obtained ranking error *E* for the feature *f* and kernel *k* not only gives the discriminating weight to the given kernel *k*, but, also indirectly gives weight to the extracted feature *f*. Further, using Eq.3.2 we can compute ranking error E_f for all the remaining |*F*|-1 features from the feature set *F* for the given kernel *k*. Then, the

weight w_k^t of the kernel k is then updated after getting all the ranking errors E_f as:

$$w_{k(t+1)}(\Phi_k) = \sum_{f=1}^{|F|} w_{k(t)}(\phi_k) \bullet E_f \tag{3.4}$$

where $w_{k(t+1)}$ is the new weight. After computing the weight of kernel k, we then obtain the new updated weights of all the remaining K-1 pre-selected kernels for the training set by repeating the same procedure. Finally, when the updated weights of all the K kernels are obtained, the global weighted multiple kernel space, referred as **Φ,** for the training set containing N samples is formed as:

$$\bar{\Phi}_{t+1} = \sum_{k=1}^{K} w_{k(t+1)}(\phi_k) \bullet \phi_k \quad . \tag{3.5}$$

In Eq.3.5,we have obtained the global multiple kernel space **Φ**, but, still there could be biasing effect due to pre-learned metric M_{ini}, and therefore, to minimize the biasing effect due to this pre-learned metric M_{ini} we would re-learn the weights of kernels for further t=2 times. In each iteration, a new pre-trained metric M_{ini} is learned after projecting the features into the previously learned multiple kernel space. M_{ini} is re-learned using Eq.3.6 as:

$$\max_M Tr\left(\frac{M^T \bar{\phi}_{t+1}(S_{bc}) M}{M^T \bar{\phi}_{t+1}(S_{wc}) M} \right) \tag{3.6}$$

Finally, after we have learned the reliable and non-biased weights of multi kernel space **Φ** from Eq.3.5, we can then learn a global multi-kernel metric M using the above Eq.3.6 for robustly matching the non-linear multi-modal persons in re-identification.

3.2.3 Impostor Resistance

After learning a multiple kernel metric M, the problem of matching non-linear and multi-modal pedestrian samples is considerably solved. However, there will be many impostor samples in real world scenes that could have higher similarity than the actual gallery sample of a person. In such situation, instead using a fixed threshold, it

would be better to adopt a sample wise threshold, i.e. for each person a suitable threshold needed to be computed that can maximize its matching in different disjoint views and minimize impostors.

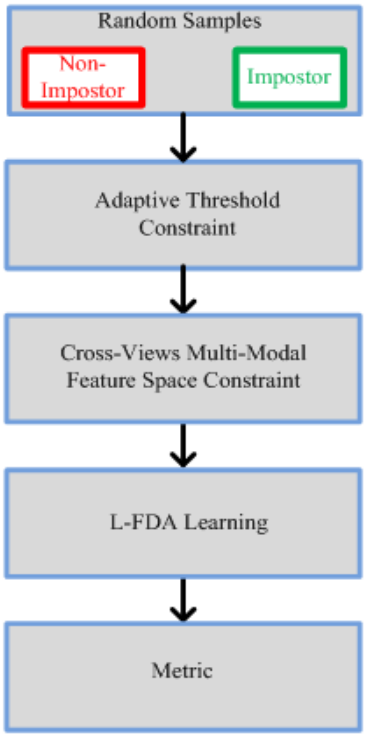

Figure 3.3 Adaptive Threshold for Learning Robust Metric

The methodology of learning metric with adaptive threshold is shown in Fig.3.3. First, the dataset is randomly partitioned into training and test sets: T_{train} and T_{test}, respectively. Then two sets of negative samples for each individual person x_i^a are obtained. These sets of negative samples contain the non-imposter samples of a person x_i^a , referred as S_{NIM}, and impostor samples of a training person x_i^a , referred as S_{IM}. The reason to form these two sets for each sample x_i^a is to obtain an adaptive threshold during learning the metric. Hence, to obtain these two sets S_{NIM} and S_{IM} for x_i^a we first compute the reference similarity value, which is referred as $Sim_{x_i^a, x_j^b}$, between the query x_i^a and its gallery x_j^b using an initially trained metric M_{ini} as:

$$Sim_{x_i^a, x_j^b} = (x_i^a - x_j^b)^T M (x_i^a - x_j^b) \tag{3.7}$$

After we have obtained the reference similarity $Sim_{x_i^a, x_j^b}$, next the similarity between x_i^a and all the gallery samples in cam *b* are computed using Eq.3.7, and are then compared with the reference similarity value $Sim_{x_i^a, x_j^b}$ to decide whether the cam *b* gallery sample is distinctively separated from the actual gallery sample x_j^b or not. Then, after comparison with all the cam *b* samples, the samples which are distinctively separated from the actual gallery x_j^b are then stored into the non-impostor set S_{NIM} as:

$$S_{NIM} = \{S_{NIM,i}^s\}_{s=1....|m|} \tag{3.8}$$

where |m| is the number of non-imposter gallery samples for query sample x_i^a. After forming set S_{NIM}, all the remaining gallery samples in cam *b* (i.e. other than the actual gallery match x_j^b and all the samples already stored in non-impostor set S_{NIM}) are then stored into the set of impostor samples S_{IM} to form the set of hard negative impostor samples of the given query x_i^a:

$$S_{IM} = \{S_{IM,i}^I\}_{I=1.....|l|} \tag{3.9}$$

where |l| is the number of imposter samples of query x_i^a. Further, using Eqs.3.7-3.9 the procedure of finding the sets S_{NIM} and S_{IM} is repeated for all the other *n*-1 remaining instances in the training set to form their respective S_{NIM} and S_{IM} sets. Finally, when we have obtained the two sets of negative samples for all the *n* training instances, we can learn the adaptive threshold based multiple kernel metric which is explained in the next subsection.

3.2.4 Multiple Kernel Metric Learning with Impostor Resistance

Getting the two negative sets for each training instance, we can now learn a more robust global multiple kernel metric *M*, referred as MKL-M. Thus, during metric learning we could now apply different adaptive threshold for the negative samples

from each different negative set. In Eq.3.10, we have incorporated these adaptive thresholds into the objective of learning a global distance metric using fisher discriminant criteria, and is given as:

$$\begin{aligned} &\min_{M,\xi}\left(\frac{1}{2}\right) r1\bullet \| M \|_2 + C\sum_{i,j}\xi_{i,j} \\ &s.t. \\ &D_{s_{new}}(x_i, x_j) = \min\sum_{i,j=1}^{N}(x_i^a - x_j^b) < \mathrm{f}_i^s \\ &D_{d_{new}}(x_i, x_j) = \max\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i^a - x_j^b) < \mathrm{f}_i^d \\ &\xi_{i,j} \geq 0, \forall(\mathrm{i}, \mathrm{j}), \mathrm{M} \geq 0 \end{aligned} \tag{3.10}$$

where f_i^s and f_i^d are incorporated into objective function as the sample wise adaptive threshold function to differentiate between imposter and non-imposter samples, and to impose a harder constraint for the metric during learning. These adaptive threshold functions can be computed as:

$$f_i^s = D_s - (\frac{D_d}{2}) \tag{3.11}$$

and

$$f_i^d = D_d + (\frac{D_s}{2}) \tag{3.12}$$

where D_s and D_d are the original distances between positive sample pairs (x_i^a and x_j^b) and negative samples pairs. Now, the objective function in Eq.3.10 is solved using fisher discriminant analysis as:

$$M = \min_M\left(\frac{M^T S_{\mathrm{inter\_Modal}} M}{M^T S_{\mathrm{intra\_Modal}} M}\right) \tag{3.13}$$

where $S_{\mathrm{inter\_Modal}}$ is the average inter class modality matrix, and is computed as:

$$S_{\mathrm{inter\_Modal}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\| x_i^a - x_j^b \|_2 \tag{3.14}$$

where N is the total number of sample classes (*ID*s) across views. Similarly, $S_{\text{intra\_}Modal}$ is then defined as the average intra class modality matrix, and is computed as:

$$S_{\text{intra\_}Modal} = \frac{1}{N_i}\sum_{i=1}^{N_i} \| x_i^a - x_j^b \|_2 \tag{3.15}$$

where N_i is the number of samples in class i. Finally the metric M is obtained by solving below Eq.3.16 with adaptive threshold to restrain impostors as:

$$M = \min_M \left(\frac{M^T (S_{\text{inter\_}Modal} > f_i^d) M}{M^T (S_{\text{intra\_}Modal} > f_i^s) M} \right) \tag{3.16}$$

3.2.5 Experimental Setup and Comparisons

3.2.5.1 Datasets

We have used three single-shot datasets including VIPeR, GRID, and CAVIAR4REID. While, for multi-shot testing datasets 3DPes,i-LIDS, i-LIDS VID, and CUHK01 are used in our experiments. The results are measured by cumulative matching curve (CMC), and each cumulative matching curve (CMC) is obtained by averaging the experiment results for twenty trials.

VIPeR[5] dataset is a publicly available dataset. It has 632 pair of images captured from two cameras, with one image of each person in each camera view. All the images are 128x48 dimensions. It is a very challenging single shot dataset due to large variations in camera view angles and illumination. All the images have background clutter, captured with arbitrary different poses and with large color variations that even exist between the same person images.

The **GRID** [49] dataset consists of person images captured from 8 disjoint camera views installed in a busy underground station. The dataset set contains 250 persons, each of them has a pair of images in the two cameras. Further, this dataset contains 775 persons, which do not match with any of the 250 persons.

CAVIAR4REID[50] is another challenging dataset that is captured by two cameras in a shopping center. The images are taken from 26 large image sequences. It

contains 72 persons in total, and for each person there are 10 or 20 images that are collected from one or two video sequence. It has total 1220 images of all 72 persons. All the images have occlusions and large viewpoint and illumination variations. Since, there exist large variations in image resolution; therefore, all the images are resized to 128x48.

3DPeS[51] dataset is specifically designed for the human re-identification problem. The images are captured from 8 different surveillance cameras with non-overlapping fields of view in a university campus. The main challenge in this dataset is the severe lighting condition and viewpoint variation. The dataset includes 193 individuals with a total of 1012 images. All the images are normalized to 64 x 128 pixels.

The **i-LIDS** [52, 53] dataset contains images of 119 persons with total 476 images. These images are captured from multiple non overlapping camera views of airport hall CCTV camera network. Most of the images in this dataset have large illumination variations and are mostly occluded.

CUHK01 [54] dataset is specifically captured for person re-identification. In an indoor scenario, two camera views are used for capturing 972 pedestrians. All these pedestrians have two images per view, and hence, in total there are 3, 884 images of all these 972 pedestrians. All the 3, 884 images are manually cropped, and normalized to 160 x 60 pixels.

The **CUHK01**dataset mainly includes images of the frontal view and the back view, with large viewpoint variations in the two camera views. In addition to this, both background and poses are also varying even among the same persons.

iLIDS-VID[55] is a new dataset that captures 300 pedestrians across two disjoint camera views in a public space. It comprises 600 image sequences of these captured 300 distinct individuals, with one pair of image sequences from two camera views for each person. Each image sequence has variable length ranging from 23 to 192 image frames, with an average number of 73.

iLIDS-VID is very challenging dataset due to clothing similarities among people, lighting and viewpoint variations across camera views, cluttered background and random occlusions.

3.2.5.2 Experiment Protocol

In this subsection, we discuss the evaluation strategy of our experiments and sample selection method to obtain random and unbiased samples for fair results to be comparable with previous state of the art methods.

In single shot setting all the datasets are divided into training and test sets, which are formed by randomly choosing *p* persons from the dataset. For VIPeR and GRID, *p* is chosen to be 316 and 125 respectively for training. GRID dataset has uneven number of samples in the *Gallery* view, therefore, we have randomly chosen *p*=125 for training, and the test set is formed by using 900 persons in the *Gallery* view, and the remaining 125 persons as the query images. Among these 900 persons in the gallery set, only 125 persons have one instance in the query set, while, the rest 775 images are extra images having no pair image in the query set.

Gallery set for both the training and test sets in 3DPes and CAVIAR4REID are formed by randomly choosing five samples per person. In i-LIDS, multi-shot gallery is formed by randomly choosing only those persons that have at least two sample images in the *Gallery* view. Similar to 3DPes, gallery set in i-LIDS VID is also formed by randomly choosing five samples per person, while, the query set also contains five samples per person. In CUHK01, each person has already two samples per view, and hence, the multi-shot gallery is already formed.

For all the experiments the training and test tests are randomly split into 20 partitions, each of these partitions are further partitioned into 20 sub-partitions. We have in total 400 partitions. First 200 sub-partitions are used for training and validation, while the remaining 200 sub-partitions are used for testing and ranking.

3.2.5.3 Experiment Results and Analysis

We evaluate global multiple kernel metric *M* on two single shot datasets and two

multi-shot datasets. The datasets used for evaluation areVIPeR,CUHK01, 3DPes, and i-LIDS.

In Fig.3.4, it is clear that even with the complex viewpoint, posture and illumination changes that exist among samples of different identities; the proposed multi-kernel metric *M* can correctly match the probe with its corresponding gallery sample, thus, showing its robustness against all these challenging changes.

Table 3.1 Comparison with State of the Art Methods on VIPeR

Methods	*r=1*	*r=5*	*r=10*	*r=20*
PRDC [57]	15.66	38.42	53.86	70.09
SDALF [56]	19.87	38.89	49.37	65.73
LADF [38]	29.11	61.39	75.63	87.66
K-LFDA [13]	32.3	65.8	79.7	90.9
K-LFDA [13]	32.8	65.5	79.1	90.0
Proposed (Two Kernels)	**35.44**	**67.97**	**80.31**	**90.44**
Proposed (Three Kernels)	**36.97**	**69.87**	**80.68**	**90.76**

Comparisons in details are also provided in Table.3.1, including K-LFDA[13], LADF[38], SalMatch [46], SDALF [56], and PRDC [57]. Proposed MKL-M has significantly outperformed all these previous approaches and successfully utilizes MKL to enhance discriminating capabilities of different features.

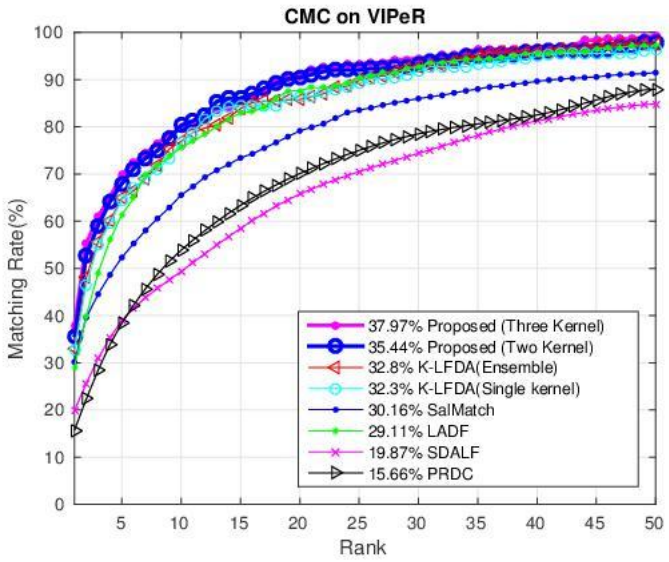

Figure 3.4 CMC Curve for MKL-M on VIPeR Dataset

In Fig.3.5, CMC curve for CUHK01 dataset is shown. In our experiments we

used CUHK01 dataset in single shot mode. Therefore, two image samples are randomly chosen for each person, one sample to be used as probe and the other sample as gallery. It has observed from the results and CMC curves in Fig.3.5 that the proposed MKL-M approach has also outperformed many state of the art approaches including ITML [9], SalMatch [46], SDALF [56], and PatchMatch [58]. Further, empirical comparison is provided in Table.3.2. Empirical analysis clearly reveals the fact that proposed MKL-M has improved re-identification up to 9% compared to SalMatch [46] at rank@1.

Table 3.2 Comparison with State of the Art Methods on CUHK01

Methods	*r=1*	*r=5*	*r=10*	*r=20*
SDALF [56]	9.9	22.57	30.33	41.03
ITML [9]	1598	35.22	45.60	59.81
Patch Match [58]	20.39	34.12	41.09	51.56
SalMatch [46]	28.45	45.85	55.67	67.95
Proposed (Two Kernels)	**29.95**	**56.9**	**71.51**	**81.6**
Proposed (Three Kernels)	**31.19**	**57.93**	**70.66**	**81.39**

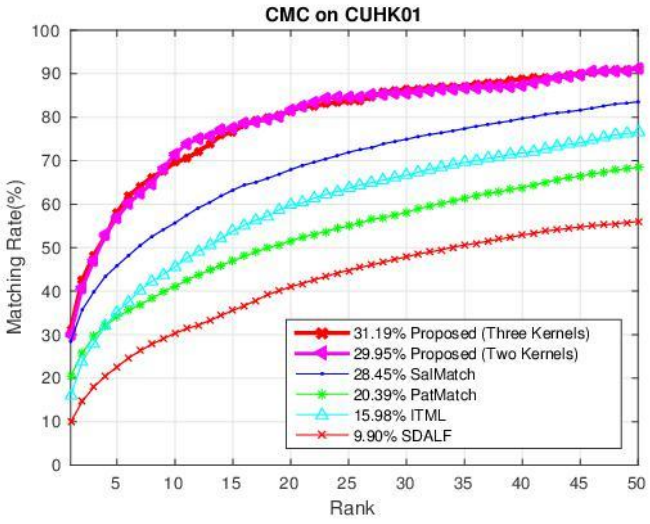

Figure 3.5 CMC Curve for MKL-M on CUHK01 Dataset

We have performed re-identification on two multi-shot datasets to verify the effectiveness of MKL-M metric. Multi Shot experiments are conducted on 3DPes and i-LIDS datasets. For 3DPes dataset, five images in the gallery set are randomly selected for each person. The CMC curve on 3DPes is shown in Fig.3.6. MKL-M has

achieved the highest correct recognition rates at all the ranks (from rank@1 to rank@20) as shown in Fig.3.6, where the rank@1 rate is **57.67%**, and is much higher than many state of the art methods including Martinel et.al [40], Zheng et.al.[59], and Garca et.al [60]. The proposed approach has achieved great increment in the results mainly due to the multiple kernel that enhances the discriminating power of low level features using the weighted combination of discriminating linear and non-linear kernel spaces, thus MKL-M approach finds the robust combination of these linear and non-linear kernels for 3DPes.

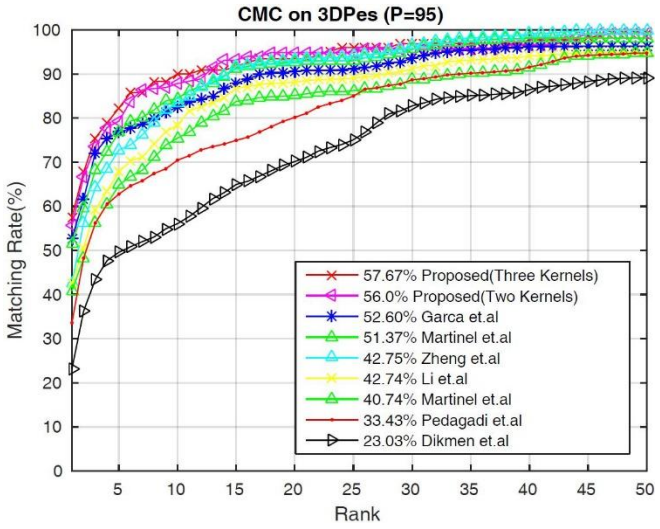

Figure 3.6 CMC Curve for MKL-M on 3DPes Dataset

All the experiments conducted for i-LIDS contain at least two samples per person in the *Gallery* view. In Fig.3.7 the CMC curve is provided for p=50, it is evident from the curve of our approach that our method can improve the rank@1 matching rate when the features are projected into discriminating weighted multiple kernel space. In i-LIDS the learned MKL-M has outperformed all the previous state of the art methods at rank@1 and rank@5. The performance gain is **3**% at rank@1 and **6.5**% at rank@5 using the weighted combination of multi kernels.

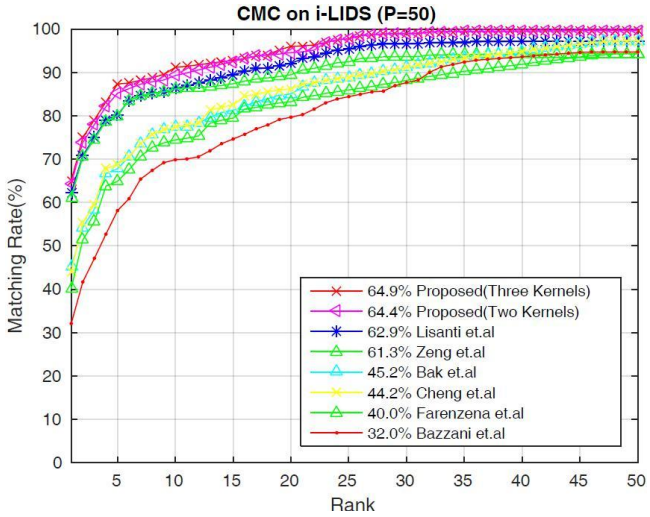

Figure 3.7 CMC Curve for MKL-M on i-LIDS Dataset

3.3 Summary

This Chapter has covered the details of learning multiple kernel for re-Identification, as well as, has given the experiment results with analysis. The learned multiple kernel has shown more improvement at rank@1 than just single kernel which is just chosen randomly without taking into care the nature of distribution in each disjoint view. The learned global multiple kernel is tested on both single-shot and multi-shot settings, and has attained much higher rank@1 than many previous state of the art methods [12,13,14,52].

Chapter 4. Sample Specific Multiple Kernel Metric Learning for Person Re-Identification

The multiple kernel in chapter 3 has addressed the problem of both non-linear and multi-modal feature space, and has improved discrimination among persons, however, it has been learned globally. While, re-identification is a problem to retrieve the exact match of each single person, and hence, it is needed to pay more attention of each single person's individuality.

A real world situation of re-identification is shown in Fig.4.1, where different person images observed in two disjoint (from VIPeR) views are given (images in the same column belong to same identity). All these different persons undergo different complex changes (i.e. pose, illumination, background, and viewpoint changes) in real world.

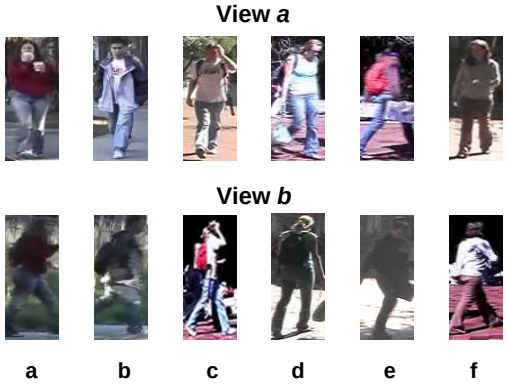

Figure 4.1 Persons Undergoing Complex Non-Linear Changes in VIPeR Dataset.

Many state of the art features [4,5,6,7] are designed to obtain reliable and invariant representation of a person, and to address the differing appearance problem in different views. Therefore, when images undergo complex and random non-linear changes, it becomes difficult even for carefully designed state of the art features [4,5,6,7] to obtain reliable and invariant features of a person in different disjoint views. In this chapter, we propose to learn a localized or sample specific multiple kernel for each individual person. Although local distance metrics have been proposed for person re-identification [61,62], our localized multiple kernel is different with these

previous local metrics in a way that our localized multiple kernel address both the complex non-linear changes in each disjoint view, as well as, address the multi-modal feature space locally for each individual person in a pair of disjoint views.

In Fig.4.2, a pair of samples belong to the same identity, but, are observed in two different disjoint views from the VIPeR dataset, are shown. The observed samples in different views have largely differing appearance due to the fact that the two images in each disjoint view undergo random and non-linear changes in illumination, pose and background. The conventional feature space of the pair of samples is shown in Fig.4.2b, while, the feature space after projection in localized multiple kernel space (i.e. LWMKL) is shown in Fig.4.2c. It is clearly evident from Fig.4.2c that the localized multiple kernel can well address the local complex non-linear changes in different non-overlapping views.

In our work the proposed local or sample wise weighted multiple kernel space is referred as LWMKL, and is learned by computing the weights of each pre-selected kernel using a qualitative weight learning method proposed in chapter 3. The procedure of learning LWMKL for each single person and learning the weights of each per-selected kernel is given in subsection 4.3.2.

Similar to the global multiple kernel projection, when we have learned the local weighted multiple kernel space for each individual person, a global distance metric referred as LWMKL-M is learned to perform the re-identification. LWMKL-M is learned using local fisher discriminant analysis [12] to obtain a discriminative metric to match the non-linear pedestrians in different non-overlapping views. Finally, with the local weighted multiple kernel space and the global metric we can evaluate the metric on several standard datasets, including VIPeR, GRID, CAVIAR4REID, 3DPes, i-LIDS, i-LIDS VID and CUHK01.

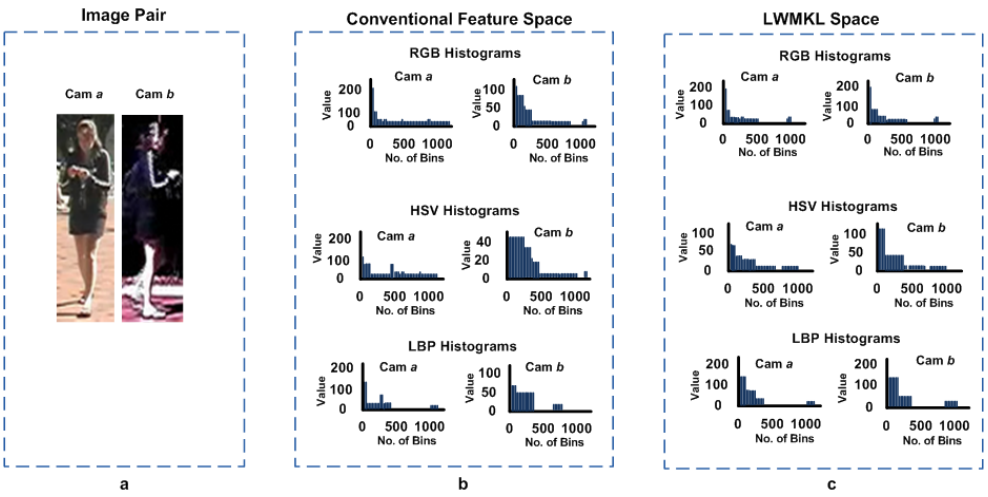

Figure 4.2 (a) Image Pair, (b) Conventional Feature Space, (c) LWMKL Space.

4.1 Existing Work

In re-identification, each single identity is a different person or class, thus, the objective of re-identification is to maximize the matching between the samples of the same identity, and to attain the gallery match of each single person at rank@1. Therefore, recently local methods have been widely proposed in re-identification, which have proposed to learn either local metric or local feature extraction methods.

Further, Ankur Datta et.al in [36] have also proposed feature mapping method. Feature mapping in [36] used weighted brightness transfer function to model only the non-linear illumination changes.

Kai Liu et.al in [61] proposed to learn sample wise feature mapping space to model the non-linear appearance changes between pair of images of the same person in two disjoint views. Ying Zhang et.al in [62] learned sample-specific SVM classifier for each pedestrian. However, the local classifier for each person is learned using very few positive pairs, and considerably hundreds of negative pairs, thus, leading to over fitting in the learned local metric. Kai Liu et.al in [63] proposed datum-adaptive local metric learning method for person re-identification, which approximates local feature projection maps rather than learning it from the actual observed pair of samples of a person.

4.2 Motivation

Recently, local methods have been widely used in re-identification both for matching persons, as well as, in obtaining discriminative features of each individual person. Although the local methods have addressed non-linear changes, however, there are still shortcomings in these recent local methods that make them suffer badly to match persons in non-overlapping person views. The shortcoming is that each person in re-identification have either a single pair or very few pairs, while, the local metrics learned with very few positive pairs and hundreds of negative pairs that would obviously overfit the metric, and during learning it could also discard some of the subtle useful feature about the person.

Therefore, the objective of this chapter is to develop a method for localized multiple kernel learning that can address the non-linear changes, as well as, multi-modal feature space. In addition, the localized multiple kernel is learned by particularly taking into care the small sample size problem of re-identification. The purpose of this localized multiple kernel learning is to model the complex and non-linear changes of each single identity, as well as, address the multi-modal feature space in the two disjoint views.

4.3 Methodology

This section describes the methodology of learning local weighted multiple kernel space (LWMKL), and learning a global distance metric LWMKL for each individual identity. In Fig.4.3, a detail pictorial description of our methodology is provided that shows each single step of learning LWMKL.

In Fig.4.3a, in the training stage the dataset is randomly partitioned into training and test sets, then LWMKL space is learned using the randomly selected training set. First, different heterogeneous features are extracted from the randomly selected "*n*" training samples. Then, using these extracted features a LWMKL space is learned for each training sample using the qualitative method proposed in chapter 3.

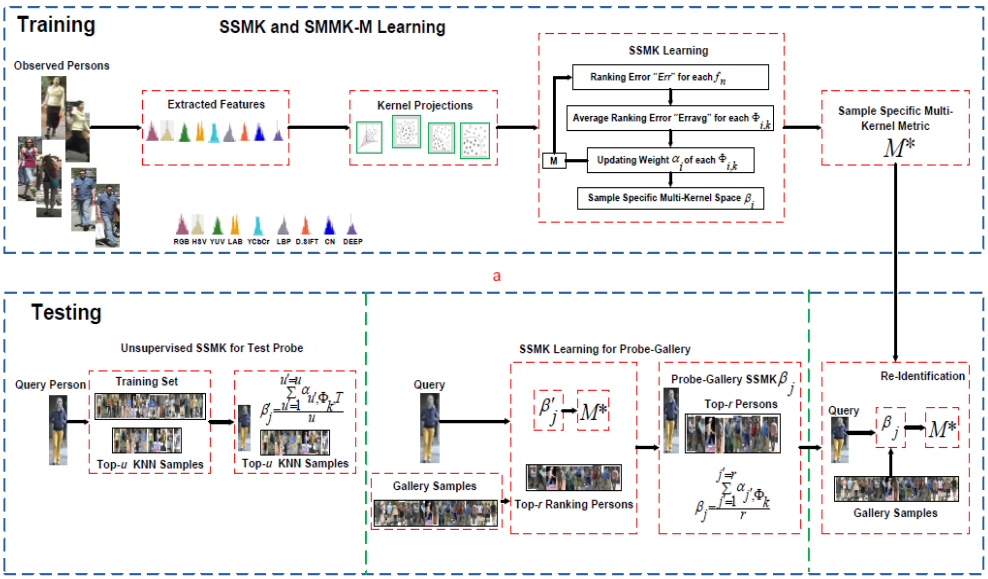

Figure 4.3 (a) Methodology of the Proposed LWMKL space and Learning Metric LWMKL-M (b) Testing Methodology for Person Matching

Finally, when we have obtained the LWMKL space of all the "*n*" training samples, a more robust global matching function is learned to maximize the matching between the observed samples of each person in a pair of non-overlapping views. The learned global matching function in our work is referred as LWMKL-M, and is denoted as *M**.

After we have obtained the global metric in the training stage, as shown in Fig.4.3a, it is then evaluated on test set. But, before evaluating the learned metric it is necessary to obtain the local weighted multiple kernel space for the test samples, since the test samples do also undergo different complex non-linear changes, and is shown in Fig.4.3b (testing stage).

However, the labels are unknown during testing, therefore, the local weights cannot be learned pair-wise. To resolve this issue an unsupervised KNN method is adopted, as shown in Fig.4.3b. It first obtains the top-*u* nearest neighbors of the test query (explained in Sec.4.8), and then obtains the average local weights of all the pre-selected kernels using the top-*u* neighbors. After we have obtained the LWMKL space of the test query, the samples in *Gallery* view are projected into the same

LWMKL space, and then matched with query using the learned metric *M**. Based on *M**, we then performed the matching with the *Gallery*, obtaining a ranking list with the top-*r* gallery samples. Using these top-*r* gallery samples we then re-learn the LWMKL space for the test set to obtain more reliable local weights of all the pre-selected kernels, which is shown in Fig.4.3b.

4.3.1 Feature and Kernel Space

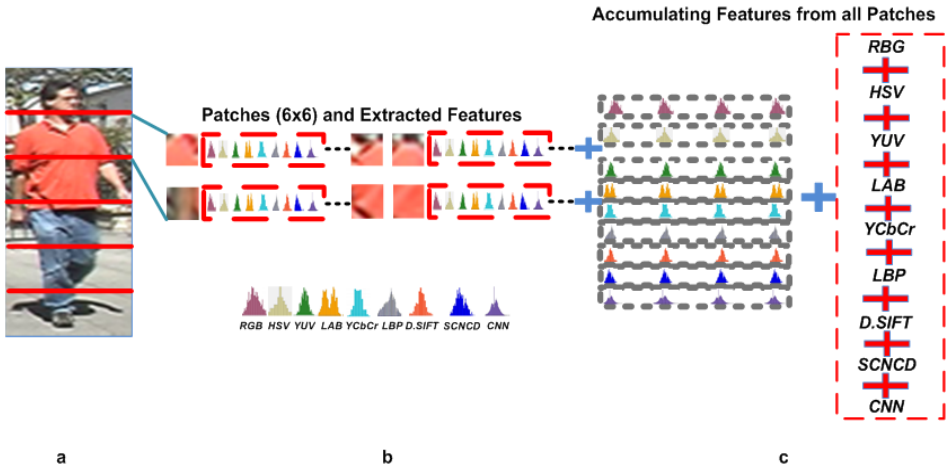

Figure 4.4 Feature Extraction (a) Sample Image from VIPeR Dataset, (b) Patches and Extracted Features, (c) Accumulating Features to form set *F*.

In our work, features are extracted patch wise (i.e. 6x6 pixels size) by first dividing the pedestrian image into six horizontal bands [46,57]. We then extract color and textures features including histograms of RGB, YUV, HSV, LAB and YCbCr for each patch. While for texture features LBP [42] and DenseSIFT [46] histograms are extracted. In Fig.4.4, the methodology of feature extraction is shown in detail. In addition to the above features, color naming [6] and deep features [24] are also extracted to enrich the features about the person. Finally, a set *F* is formed to store all these extracted features, and is given as:

$$F = \{\mathbf{f}_n\}_{n=1.....N} \tag{4.1}$$

Now, as we have stored all the extracted features in the set *F*, we then form a set of pre-selected kernels, which is referred as Φ, and contains RBF, Chi^2 and polynomial kernels with different values of σ and different degree of polynomials.

The pre-selected kernels in our work are given as:

$$K_{RBF}(x_i^a, x_i^b) = e^{-\frac{(\|x_i^a - x_i^b\|_2^2)}{2\sigma^2}} \tag{4.2}$$

$$K_{\chi^2}(x_i^a, x_i^b) = e^{(-\alpha * \sum_i \frac{(x_i^a - x_i^b)^2}{(x_i^a + x_i^b)})} \tag{4.3}$$

$$K_p(x_i^a, x_i^b) = (< x_i^a, x_i^b > +1)^m, m \in N \tag{4.4}$$

Finally, all these kernels are defined in a set Φ as:

$$\Phi = \{\phi_{k,k=1.....K}\} \tag{4.5}$$

4.3.2 Sample Specific Multiple Kernel Learning

When we have obtained both the set of extracted features *F* of all the training persons, as well as, the set of all the pre-selected kernels Φ, we can now learn local weighted multiple kernel space for each individual identity *i*.

The weights of all the pre-selected kernels are then learned one by one by the similar qualitative measure as proposed in chapter 3. We select one training person *i* (whom the local weight is needed to be learned), and then select one pre-selected kernel Φ_k from the set Φ, then we take feature f_n of person *i* one by one from the set *F* to obtain its qualitative measure using each feature f_n.

The qualitative measure similar to global MKL (chapter 3) computes the discriminating power of the pre-selected kernel Φ_k in matching the query x_i^a of person *i* with its gallery x_i^b at rank@1. However, different from the qualitative measure proposed in chapter 3, the qualitative measure for an individual person in LWMKL is obtained by a different analytical method which is given as:

$$Err_{i,f_n,\phi_k} = \frac{d_{rank1}(\phi_k(x_{i,f_n}^a), \phi_k(x_{i,f_n,\text{rank1}}^b))}{d(\phi_k(x_{i,f_n}^a), \phi_k(x_{i,f_n}^b))} \tag{4.6}$$

where *Err* is the ranking error which is computed for the individual person locally. *Err* is basically the ratio between the similarity values $d_{rank1}(.)$ and $d(.)$. $d_{rank1}(.)$ is the similarity value computed between query x_i^a and the rank@1 gallery

sample $x^{b}_{i,f_n,\text{rank1}}$,while $d(.)$ is the similarity value between the actual positive pair (x^{a}_{i}, x^{b}_{i}). And $\phi_k(.)$ is the Φ_k kernel projection of given feature f_n. Based on Eq.4.6, *Err* is then computed for all the extracted features in the set *F* for the given pre-selected kernel Φ_k.

Since *Err* is computed for each extracted feature f_n, an average ranking error *Erravg* is then computed to obtain the local average quality measure of a given pre-selected kernel Φ_k for person *i* using all the *N* extracted features as:

$$Erravg_{i,\phi_k} = \frac{\sum_{n=1}^{N} Err_{i,f_n,\phi_k}}{N} \tag{4.7}$$

With the $Erravg_{i,\phi_k}$, the local weight α_{ϕ_k} of kernel Φ_k for person *i* is then updated as:

$$\alpha_{i,\phi_k} = e^{(-Erravg_{i,\phi_k} \times w_{i,\phi_k})} \tag{4.8}$$

Then we can update the local weights of each pre-selected *k* in set Φ_k for one single person *i*, as well as, for all the "*i*" training persons in the training set by repeating the above procedure. Finally, when the local weights of all the *K* pre-selected kernels are obtained for person *i*, its LWMKL space, denoted as β_i, is obtained as:

$$\beta_i = (\alpha_{i,\phi_k}, \Phi_k) \tag{4.9}$$

With β_i for all the "*i*" persons in the training set, we then repeat the process of local weight learning for further *t*=2 to ensure the weights are unbiased and reliable in a pair of non-overlapping views.

4.3.3 Sample Specific Multiple Kernel Metric

In the previous subsection, we have already obtained LWMKL space β_i for each training sample. In this subsection we will learn the sample specific multi-kernel based metric LWMKL-M, denoted as *M**. Metric *M** in our work is learned via local

fisher discriminant analysis [12], which obtains a low dimension subspace to match the non-linear persons observed in a pair of non-overlapping views:

$$\max_{M^*} \mathrm{Tr}\left(\frac{M^{*T} S_{inter} M^*}{M^{*T} S_{intra} M^*}\right) \tag{4.10}$$

where S_{inter} and S_{intra} are inter class and intra class matrices respectively, which are obtained globally. However, when S_{intra} is computed globally, it may lose significant intra-class information. Since intra class covariance of each person may differ from global S_{intra}, therefore, to incorporate the subtle distinct covariance of each person, S_{intra} is learned by integrating both the global and pair wise averaged class covariance between two identities (i.e. each identity in person re-identification is a single class) [64] as:

$$S^*_{intra} = \gamma S' + (1-\gamma) S_{intra} \tag{4.11}$$

where S' is the pair-wised average co-variance computed between two different persons in the training set, and the S_{intra} is the conventional global co-variance matrix. Here γ ($0 < \gamma < 1$) is a factor that controls the balance between pair wise co-variance S' and global covariance S_{intra}, and is set to 0.5 in this subsection.

Further, the locality information is also incorporated during FDA learning by using locality preserving projection [65], to preserve locality among samples (between samples of the same identity) and restrict large number of outliers (samples for negative identities) in re-identification. Locality information is integrated using local scaling method given in [66] which learns an affinity matrix $A = [a_{ij}]$ by measuring the distance between pair of samples. Each element a_{ij} in the affinity matrix A is then given as:

$$a_{ij} = e^{\frac{-d_p(x_i^a, x_i^b)}{\sigma_i^a \sigma_i^b}} \tag{4.12}$$

where a_{ij} represents the affinity value between samples x_i^a and x_i^b, and is obtained by

computing distance using the d_p (p-norm distance). σ_i^a and σ_i^b represent the local scales of samples x_i^a and x_i^b respectively, and is computed using a similar method in [67]. Affinity matrix A obtained above is then combined with S^*_{intra} and S_{inter} matrices to learn the new modified $S^*_{intra\text{-}new}$ and $S_{inter\text{-}new}$ as:

$$S^*_{intra-new} = \gamma AS' + (1-\gamma) AS_{intra} \tag{4.13}$$

$$S_{inter-new} = AS_{inter} \tag{4.14}$$

Now, the sample specific multi-kernel based metric $M^*$is then learned as:

$$\max_{M^*} \mathrm{Tr}\left(\frac{M^{*T} S_{inter-new} M^*}{M^{*T} S^*_{intra-new} M^*}\right) \tag{4.15}$$

To understand all the steps in learning LWMKL-M metric M^*, algorithm 4.1is given below to describe each learning step in detail.

Algorithm 1 Training: LWMKL and LWMKL-M Learning

Given: $\boldsymbol{F}=[f_1,f_2,.....,f_N]$, $\Phi=[\Phi_1,\Phi_2,.....,\Phi_k]$

Number of persons=n'

Initialize w_{Φ_k} (set to 1/K where K is the number of kernels in set Φ)

for t=1,.....,T **do**

for i=1,.....,n' **do**

for Φ=1,.....,K **do**

for F=1,.....,N **do**

compute $Err_{i,f_n,\phi_k} = \dfrac{d_{rank1}(\phi_k(x^a_{i,f_n}), \phi_k(x^b_{i,f_n,\mathrm{rank1}}))}{d(\phi_k(x^a_{i,f_n}), \phi_k(x^b_{i,f_n}))}$

where $d(.) = (x_i - x_j)^T M (x_i - x_j)$

end for

compute $Erravg_{i,\phi_k} = \dfrac{\sum_{n=1}^{N} Err_{i,f_n,\phi_k}}{N}$

Update $\alpha_{i,\phi_k} = e^{(-Erravg_{i,\phi_k} \times w_{i,\phi_k})}$

where w_{i,Φ_k} is the initial weight of kernel Φ_k

end for

compute $\beta_i = (\alpha_{i,\phi_k}, \Phi_k)$

where β_i is the local weighted multi-kernel space of person *i*

end for

compute $\max_{M^*} \mathrm{Tr}(\frac{M^{*T} S_{inter} M^*}{M^{*T} S_{intra} M^*})$

where *M**is the learned metric, and used in the next iteration

end for

Output Final LWMKL space for person *i* $\beta_i = (\alpha_{i,\phi_k}, \Phi_k)$

and LWMKL-M metric $\max_{M^*} \mathrm{Tr}(\frac{M^{*T} S_{inter} M^*}{M^{*T} S_{intra} M^*})$

4.3.4 Re-Identification for Testing

Similar to the training set, samples in the test set do also undergo complex non-linear changes in disjoint views, and are multi-modal. Therefore, before we perform matching among these test samples with the learned metric *M**, it is necessary to address these complex non-linear changes.

Similar to the training stage, we need to learn LWMKL space β_j for each test sample *j*. However, to learn LWMKL space we require label of the test sample, while, in testing there is no label information available. Therefore, we have adopted an unsupervised method to obtain the local weights for each test sample to form its LWMKL space β_j .

For this purpose, a KNN method is used to obtain top-*u* nearest neighbors of test sample *j* from the corresponding disjoint view in the training set. Now, an initial

multi-kernel space β'_j for sample *j* is then obtained by averaging the *u* local weights of each pre-selected kernel from the *u* neighbors.

Algorithm 2 Re-Identification after Learning Unsupervised LWMKL

Given: $\boldsymbol{F}=[f_1,f_2,.....,f_N]$, $\Phi=[\Phi_1,\Phi_2,.....,\Phi_k]$

$\boldsymbol{X_j}=[x^a_{j,j=1,....,n}]$, $\boldsymbol{G}=[x^b_{j,j=1,....,n}]$

LWMKL for Test Query

compute
$$\alpha'_{j,\phi_k}=\frac{\sum_{f_n=1}^{N}\sum_{i'=1}^{u}\alpha_{i',f_n,\phi_k}}{N\times u}$$

where α'_{j,ϕ_k} is the weight of each Φ_k of query, and is obtained after averaging the corresponding Φ_k values from *u*-neighbors

compute $\beta'_j=(\alpha'_{j,\phi_k},\phi_k)$

Re-Learned LWMKL for Test

Given: Top-r Gallery matches

Repeat Algorithm 1for each $(x^a_j, x^b_{j',j'=1,.....,r})$ pair

compute Average weight α_{j,ϕ_k} of each Φ_k using the above obtained r pairs as:

$$\alpha_{j,\phi_k}=\frac{\sum_{j'=1}^{r}\alpha_{j',\phi_k}}{r}$$

compute β_j as:

$$\beta_j=(\alpha_{j\phi_k},\phi_k$$

Re-Identification:

$$sim=(\beta_j(x^a_j)-\beta_j(x^b_j))^T M^*(\beta_j(x^a_j)-\beta_j(x^b_j))$$

The learned β'_j is used to perform the re-identification of the test probe to obtain top-*r* potential matching candidates from the gallery set, as shown in Fig.4.3b. These

top-r potential gallery candidates are now served as r potential probe-gallery pairs to re-learn the labeled averaged pair wise LWMKL space β_j to further improve the reliability of the learned local weights. Finally, after re-learning LWMKL space β_j for each test probe, the final re-identification can now be performed to acquire the matching of query the samples. The detail procedure of learning LWMKL space for test set and performing re-identification using LWMKL-M metric M^* is provided in the above algorithm 4.2.

4.4 Performance and Results Analysis

In our experiments, six kernels are selected by cross-validation on VIPER, CUHK03, and MARKET1501 datasets. Then, the learned Multi-kernel based metric LWMKL-M, denoted as M^*, is evaluated on different standard single shot and multi-shot datasets including VIPeR, GRID, CAVIAR4REID, 3DPes, i-LIDS, i-LIDS VID, and CUHK01.

The experimental results of VIPeR:

The re-identification results of VIPeR dataset are shown in CMC curve in Fig.4.5. We compared our results with the state of the art methods including [5], [7],[13], [27], [30],[62],[70],[71], [72], [73], and [75].

In Fig.4.5, it is clear that the learned LWMKL space using weighted Chi^2 kernels (referred as LWMKL2) and weighted RBF kernels (referred as LWMKL3) have demonstrated successful re-identification performance from rank@1 to rank@10. The learned metric M^*based on weighted multi-kernels LWMKL3 (i.e. weighted Chi^2 plus RBF) have also significantly outperformed several state of the art approaches including [13], [27], [73] and [75]. Although rank@1 result of LWMKL1and LWMKL2 are lower than [73], the learned LWMKL1and LWMKL2 spaces based metric M^* have obtained much higher re-identification results at rank@5 and afterwards.

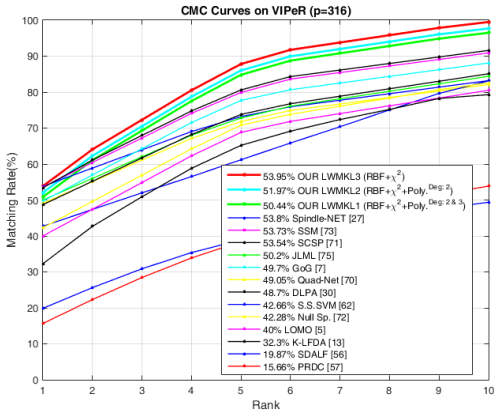

Figure 4.5 Comparison of CMC Curves (VIPeR, p=316)

It is clearly evident from the obtained results and from the CMC curves in Fig.4.5 that even there exists complex non-linear changes in viewpoint, posture and illumination among samples in VIPeR, the learned metric M^* is discriminative and robust, and can correctly match the probe with the corresponding gallery sample.

The experimental results of GRID:

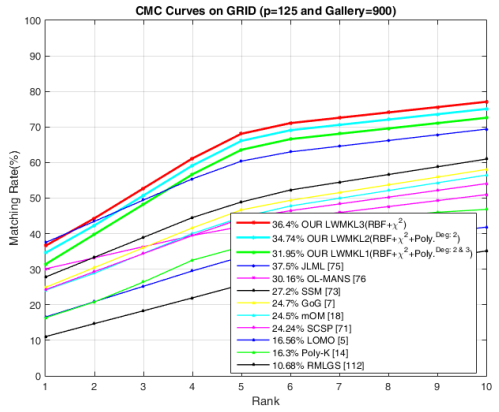

Figure 4.6 Comparison of CMC Curves (GRID, p=125, and Gallery=900)

The CMC curves for GRID are provided in Fig.4.6. Similar to VIPeR, the results obtained on GRID for all the three variants of LWMKL spaces, which are LWMKL1, LWMKL2 and LWMKL3 have attained much higher recognition on complex non-linear pedestrians compared to previous state of the art approaches including [5], [7], [14], [18], [71], [73], [75], and [76]. Although the learned metric M^* has lower

rank@1 result compared to [75], however, LWMKL1, LWMKL2 and LWMKL3 have attained to identify more than **90**% multi-modal persons at rank@5 showing its capability to discriminate each individual person even different persons may have experienced large non-linear changes in viewpoints, illumination, and postures.

The experimental results of CAVIAR4REID:

To further evaluate the discriminating capability of LWMKL space β_j and learned metric M^*, we have evaluated them on CAVIAR4REID in single shot setting. Performance comparison against state of the art methods is provided in Fig.4.7. From the results it is evident that the learned metric M^* has outperformed many state of the art methods, including [12], [13], [38] and [112], on CAVIAR4REID at all ranks (i.e. from rank@1 to rank@10). M^* has obtained rank@1 re-identification of about **47.92**% using LWMKL3, while about **45.72**% using LWMKL2, and about **43.66**% using LWMKL1, respectively. Similar to VIPeR and GRID datasets, M^*on CAVIAR4REID has also shown robustness against complex non-linear changes and against poor quality images (i.e. varying aspect ratio, as well as, blurring).

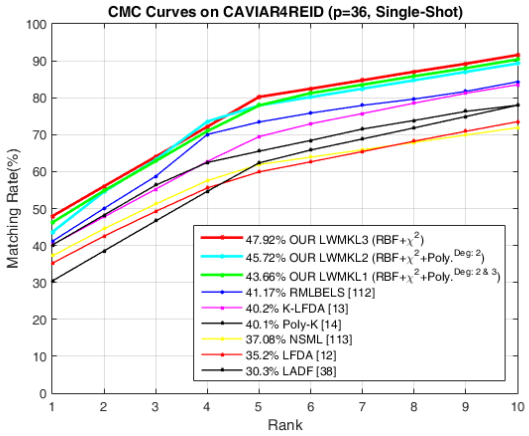

Figure 4.7 Comparison of CMC Curves (CAVIAR4REID, p=36, and S.S.)

The experimental results of 3DPes:

To further evaluate the performance and scalability of the learned space β_j and the learned metric M^*, we have evaluated M^* on 3DPes. The obtained results are shown in CMC curves in Fig.4.8, where the most discriminating LWMKL space is

LWMKL3 (i.e. weighted RBF and Chi^2 only). LWMKL3 has outperformed most of the state of the art methods including [27], [38], [57], [60], [71], [78], [80], and [81], and has attained **73.97**% matching at rank@1. However, LWMKL1 and LWMKL2 have attained lower rank@1 performance compared to [80].The main reason behind this lower matching at rank@1 is due to the fact that the persons observed in 3DPes undergo large viewpoint and illumination changes so that the chosen kernel (i.e. polynomial kernel) in LWMKL1 and LWMKL2 face difficulty in addressing these complex changes, as well as, face difficult in addressing the multi-modal feature space. Therefore, when an optimum local weighted multi-kernel combination LWMKL3 space is learned, i.e. using RBF and Chi^2 kernels, then there is considerable improvement at all ranks.

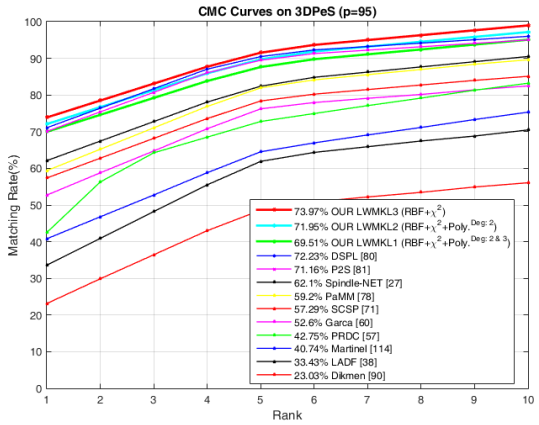

Figure 4.8 Comparison of CMC Curves (3DPes, p=95)

The experimental results of CAVIAR4REID:

Further, CAVIAR4REID dataset is also evaluated in multi-shot setting. The obtained re-identification results on CAVIAR4REID dataset are given in detail in Fig.4.9. Similar, to the single-shot setting all the three variants of LWMKL have outperformed many state of the art methods at rank@1 matching. The obtained results are **66.4**%, **65.75**%, and **63.96**% for LWMKL3, LWMKL2 and LWMKL1, respectively. LWMKL3 variant has particularly outperformed all the previous state of the art approaches including [10], [57], [82], [83], and [105]. These results have

proven that the learned LWMKL is discriminative to model the non-linear changes of different persons in a pair of non-overlapping views, and then the leaned metric $M^*$is also robust against non-linear changes and impostors.

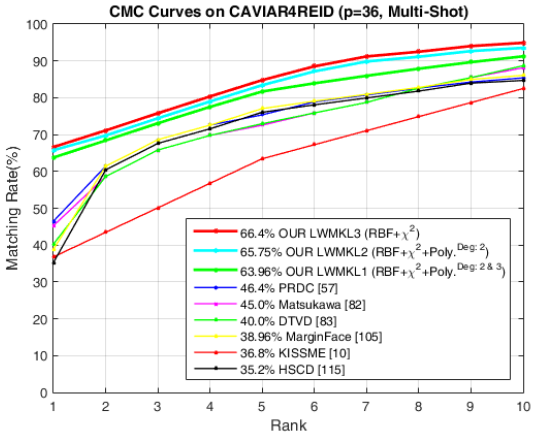

Figure 4.9 Comparison of CMC Curves (CAVIAR4REID, p=36, and M.S.)

The experimental results of i-LIDS:

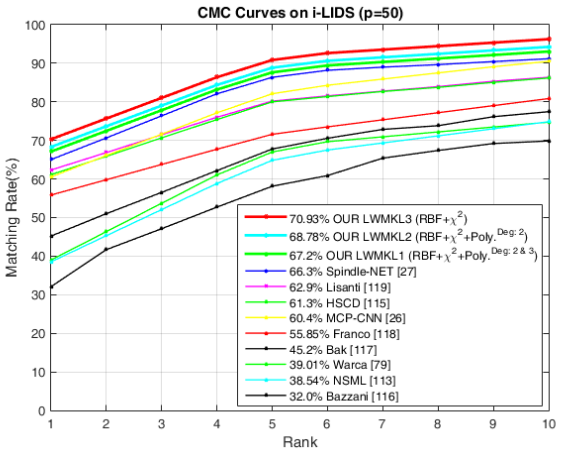

Figure 4.10 Comparison of CMC Curves (i-LIDS, p=50)

CMC curves in Fig.4.10 are obtained for i-LIDS dataset in multi-shot settings. The multi-shot experiments use the samples which have at least more than three samples available, from which one sample is chosen as query.

The experimental results of i-LIDS VID:

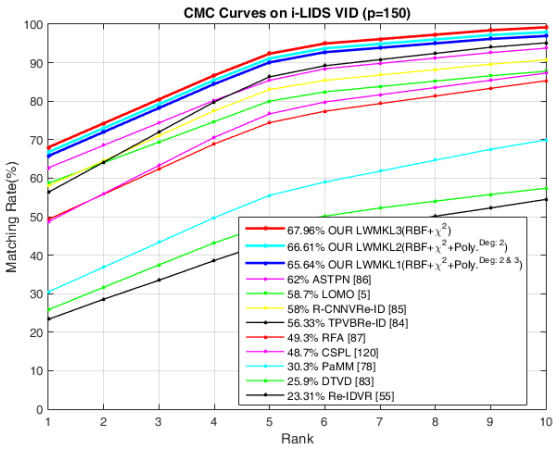

Figure 4.11 Comparison of CMC Curves (i-LIDS VID, p=150)

i-LIDS VID is recently published standard multi-shot dataset, which contains sequence of images of each person in disjoint views. For testing purpose, five samples per person are randomly chosen in each trial to form *Gallery* view. Despite large intra class variation within the images of the same person, caused by changes in poses and viewpoints, the learned metric M^* has still attained remarkably higher matching in identifying large number of query persons correctly, as shown in CMC curves in Fig.4.11. All the three variants LWMKL3, LWMKL2 and LWMKL1 have attained state of the art performance at rank@1 of about **67.96**%, **66.61**% and **65.64**% respectively, which are much higher than state of art methods including [5], [78], [83], [84], [85], [86], and [87]. The main reason to outperform previous state of the art methods is mainly due to the addressing multi-modal, as well as, non-linear feature space using sample specific multi-kernel space β_j .When features are projected into LWMKL space β_j ,the matching between the pair of samples in non-overlapping views is largely maximized, and the learned metric can discriminate well the different complex persons.

The experimental results ofCUHK01:

We evaluated the learned multiple kernel metric M^* on CUHK01 by projecting the features into the locally learned multiple kernel space β_j . Gallery in CUHK01 contains two samples per person. All the three variants (i.e. LWMKL1, LWMKL2,

and LWMKL3) are evaluated for twenty trials, and the averaged results are shown in CMC curves in Fig.4.12. From the obtained results, it is evident that the persons undergoing complex non-linear changes can be well discriminated after projection into LWMKL space β_j .All the three variants have significantly outperformed several previous state of the art methods including [7], [18], [26], [30], [69], [72], and [75]. Compared to [30], LWMKL3, and LWMKL2 have about **6.48**% and **4.4**% rise at rank@1, respectively. This rise in rank@1 is mainly due to the capability of locally handling the complexity, non-linearity and modality in the feature space that cannot be easily modeled using only global single kernel projection in [13].

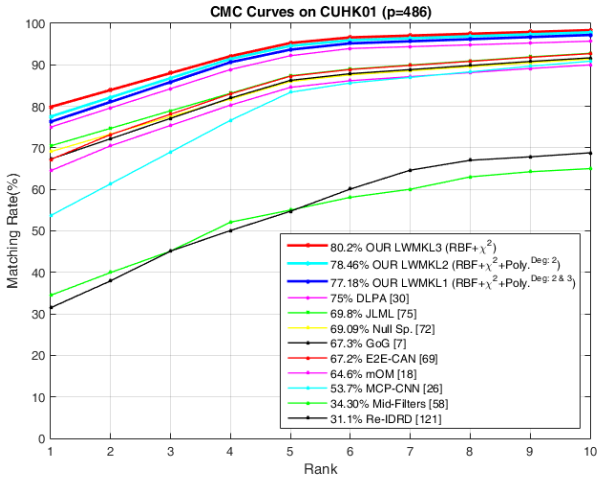

Figure 4.12 Comparison of CMC Curves (CUHK01, p=486)

4.5 Summary

Images in disjoint camera views undergo different non-linear changes. These changes can be different in different view, and even can be different from person to person whether the persons are observed in the same view or in disjoint different views. In this chapter, the proposed local or sample-specific multi-kernel space is learned to address all these complex non-linear changes in each disjoint view. The learned multiple kernel space is then used to learn a more robust and discriminative metric for person matching. The obtained results on single-shot and multi-shot datasets have demonstrated that the learned local multiple kernel space can well

discriminate different persons, and the learned metric can maximize the matching between pair of positive samples in non-overlapping views.

Chapter 5. Multi-Modal Metric Learning with Impostors Resistance for Person Re-identification

In chapter 4 we have proposed a localized weighted multiple kernel space for each individual person to address the non-linear changes it experiences in each disjoint view. In addition, the local weighted multiple kernel space in our work do also take into account that in real outdoor scenes a person may undergo different changes in each different view.

Although, the localized multiple kernel space do address the complex non-linear changes, in real world scenes, when such complex changes exist in a multi-view data(i.e. from different camera viewpoints) , the different observed persons in each disjoint view in re-identification will lie on multi-modal feature space [41]. In actual re-identification, feature space is usually non-linear, multi-modal [41], as well as, multi-view. Though, we have leaned multiple kernel space to address complex non-linear changes, still, the multiple kernel lack to address explicitly the complex multi-modal feature space. In addition, there is one problem arises in the multi-modal feature space, which is that the matching between pair of positive samples further becomes difficult due to the presence of impostor samples which are lying on different and complex modal space other than the positive pair.

In re-identification, both addressing multi-modal feature space and declining impostors during matching are big challenges to obtain maximum matching between a given pair of positive query and gallery images. In re-identification, a modal space is defined as the space which is formed by the joint combination of different changes a given pair of images of the same person undergoes in different disjoint views. These changes are pose, viewpoint, illumination, and background changes. Impostors in re-identification are defined as the persons that belong to the other persons, however, possess higher similarity with the given query than the actual gallery sample.

In past, there have been several methods proposed to solve the problem of impostor persons. These previous methods have either proposed to extract robust

feature or learn discriminative metric [8,90,91,92], while, these methods have totally ignored the intrinsic complex structure of re-identification feature space. Since the intrinsic feature space of re-identification is multi-modal, if the resistance against impostor is learned using the impostor belonging to a different modal other than the given sample itself, then the metric obtained does not perform optimally.

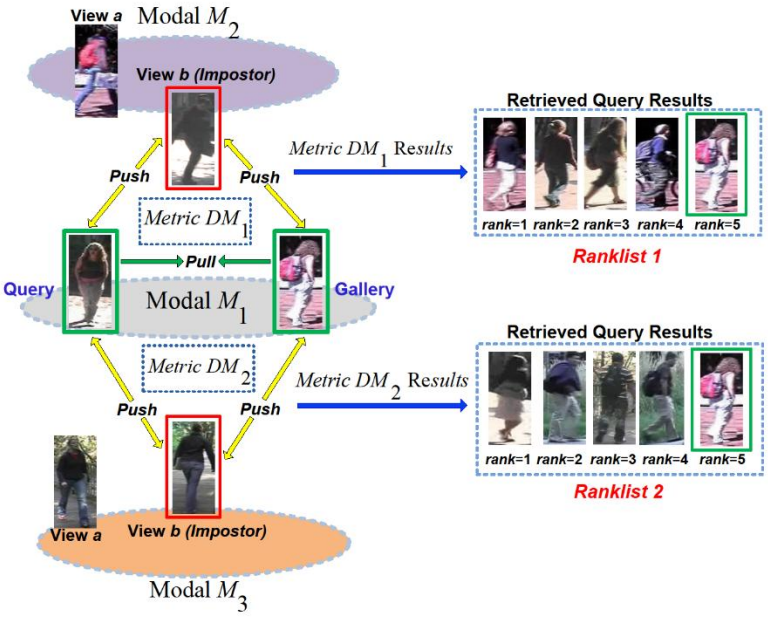

Figure 5.1 Three Modals M1, M2. and M3 in Image Space. Query and Gallery lie in modal M1, while, one impostor for Query lies in Modal M2, and the other in Modal M3.

We have illustrated this situation in Fig.5.1 to provide an insight understanding. Fig.5.1 shows three transform modals M1, M2 and M3 in the global image space. M1 contains a positive pair (query and gallery) enclosed in green rectangles for which a metric is learned, while, there are two more pairs lying in modals M2 and M3, respectively. View ***b*** images (enclosed in red rectangles) in M2 and M3 are similar to query in M1, and thus, are impostors for query sample. In conventional approaches [8,90,91,92], the metric between query and gallery samples in M1 is learned using the impostor sample from M2 (Metric DM1) or M3 (Metric DM2) as a constraint.

Therefore, if the similarity for positive pair is learned under the constraint of an impostor person lying on a different transform modal other than the positive pair, then the learned similarity metric would not be the optimal matching function, which can be proved from poor retrieval results in Ranklist 1 and Ranklist 2 in Fig. 5.1.

Further, in Fig.5.1 previous approaches [8,90,91,92] have used impostor samples for query sample only from the Gallery view, while totally ignored the gallery sample. Therefore, to resolve the above shortcomings in [8,90,91,92], we have proposed an impostor resilient multi-modal metric, referred as IRM3, which eliminates the impostors largely and attains an optimal matching between positive pair. The objective of IRM3 is to maximize the matching of a positive pair against both the negative gallery samples (N.G.S.) (samples which are not impostors and belong to different persons)and the impostors by taking into account the modal a given pair, its negative gallery samples, and its impostors reside. Further, in contrast to [8,90,91,92], it also takes into consideration the impostor samples for both the query and its respective gallery sample. This pair of impostors are referred as Cross views impostors (C.V.I.) which are obtained for query and gallery samples from their opposite views, and help in further maximizing the similarity between given query and gallery samples. The contributions of our impostor resilient multi-modal metric IMR3 are:

- Improving impostors resistance by jointly exploiting the transform modals [41], as well as, impostor samples from both Probe and Gallery views;
- With our IMR3 approach a significant gain in performance is obtained in Multi-Kernel Local Fisher Discriminant Analysis (MK-LFDA) [93].

5.1 Existing Work

In [41] Li et.al, proposed a multi-modal metric. Their work proposed that the re-identification image space is multi-modal due to different and complex non-linear changes in each disjoint view. Therefore, a single matching function that is globally

learned cannot well discriminate all the observed multi-modal persons. Following [41], [61] also proposed a multi-modal metric. But, it assumes that each different person observed in each disjoint view undergoes different complex changes, and thus, lies on a different multi-modal transform in each view. Thus, [61] proposed to learn a view specific metric for each person. Further, there are also metrics proposed to resolve the matching problem against impostors. In [91] an impostors based metric is proposed that uses only impostor samples as negative samples to learn the global metric. The metric is then learned by satisfying the learning constraint to maximize the matching between given positive pairs against the impostor samples only. Then, in [92] another idea is proposed that both easily separable negative samples and impostors have different weights during metric learning, and that could help in learning a more robust metric. The improvement is obtained due to the reason that the metric can now exploit the negative samples, including hard negative samples, to improve the matching of positive pairs.

5.2 Motivation

Though multi-modal metric [41,61]and impostor resistant metrics [91,92] have been proposed in re-identification, none of these methods have taken into care to jointly exploit the benefits in matching the positive by both addressing the multi-modal feature space and the impostor resistance at the same time. Our motivation is based on the fact that if we develop a method to address both the multi-modal feature space and impostor resistance, then we could greatly help in learning a more robust and discriminating global metric. Thus, the matching between positive pairs and their ranking orders can be largely improved. Further, these previous methods have ignored to exploit the cross views impostors, i.e. impostors from both the *Probe* and *Gallery* views, which could deliver more differentiating capability during learning the matching between positive pairs against the impostors. Therefore, using the cross views impostors the positive pair is more strongly pushed

close to each other against both the *Gallery* view impostors, as well as, *Probe* view impostors.

5.3 Methodology

Fig.5.2 shows the framework of our IRM3 approach. In Fig.5.2, first color and texture features are extracted from each training sample, then, different models are discovered in the image space. These models are discovered by using sum of squares clustering which is explained in subsection 5.3.2. Finally, for each modal cross views impostors (C.V.I.) (explained in 5.3.3), negative gallery samples (N.G.S.) (explained in 5.3.4) are generated to train the model metric M_k for each transform model k. In our work, the model metric M_k is learned using MK-LFDA [93], and the learning procedure is explained in 5.3.6. Finally, in 5.3.7 we have explained how we have performed matching between test query and gallery.

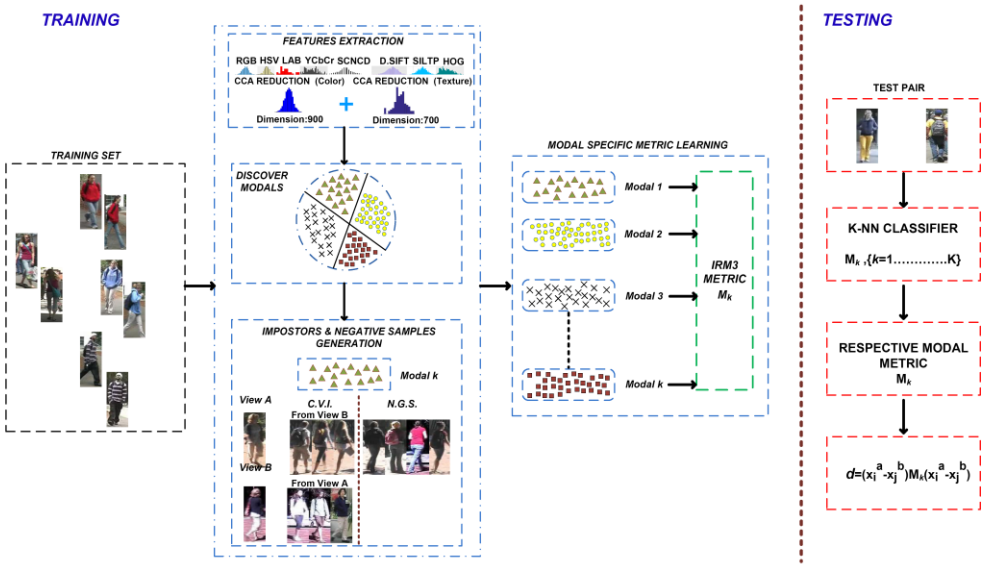

Figure 5.2 Methodology of Impostor Resilient Multi-Modal Metric Learning (IRM3) for Re-Identification.

5.3.1 Feature Extraction

RGB, HSV, LAB, YCbCr, and SCNCD histograms are extracted according to similar settings in [13] using 32 bins per channel, and settings in [6], respectively. Then, all five features are concatenated together. Similarly, SILTP, DenseSIFT, and

HOG features are extracted according to the settings in [5], [94] and [95], respectively, and are concatenated together. Dimension of color and texture features after concatenation become large, and since Re-ID data is multi-view, we used CCA [96] to reduce dimension. However, to keep the local discriminative information of each type of feature, we applied CCA to color and texture features individually. By cross validation on VIPeR and CUHK03 we have obtained optimal dimension for color feature to be 900, and texture feature to be 700. Finally, the reduced color and texture features are concatenated to form a feature vector F of size 1600.

5.3.2 Image Space Partition

Let X be the image space, then X is:

$$X = \{x_i\}_{i=1.....n} \tag{5.1}$$

where x_i is the feature representation F_i of person i, and n is the number of persons in X. Since images in X lie on different transform models, there exists distinct clusters of different models in X. Each of these model clusters has its own unique transformation and visual patterns, thus, all the persons belonging to a model k can be obtained using sum of squares clustering as:

$$S_w = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} z_{k,i}(x_i - m_k)(x_i - m_k)^T \tag{5.2}$$

where K are the number of models in X, S_w is the scatter matrix of within transform models, $z_{k,i}$ is the association of x_i with transform model k, and m_k is the center of the kth transform model. In Eq.5.2, each model center m_k is critical in discovering distinct, stable and non-empty models in X. Thus, choosing any sample (x_i^a ; x_i^b) as center m_k of any given model k, it is necessary to make sure it is a right choice. In order to make sure a chosen modal center is right it has to fulfill two conditions: First, (i) if the chosen sample (x_i^a ; x_i^b)is a center of modal cluster k, then, all the persons in modal k

will be its neighbors, and it has the highest number of nearest neighbors. Second, (ii) center m_k and all its nearest neighbors lie on the same modal, therefore, these neighbors will share similar patterns with the center m_k in both *Probe* and *Gallery* views.

The next step is to compute the number of nearest neighbors for each person in training set by taking into consideration the above two conditions. For this purpose, we have used both probe (x_i^a) and gallery (x_i^b) samples of each person to obtain four lists of neighbors, which are computed from both camera views. To acquire most reliable neighbors we then select only top@40(top@20 for *VIPeR*) neighbors from each list, and then, perform an intersection operation among all the four lists to obtain the cardinality value, as well as, *ID*s of the neighbors which are common in both *Probe* and *Gallery* views of a given person. Here, the reason to choose top@40 neighbors is to maintain maximum reliability with minimum time and memory cost in large datasets. For instance, if we have k=16 modality spaces in *CUHK03*, then, in each modal there will be at least 78 training persons. To obtain a center sample x_i of any modality space it must have at least 51% neighbors in that modal, and thus, we take top@40 neighbors which is in actual 52% proportion of the training persons in a modal to find out whether x_i is a center or not.

Then, the obtained cardinality value and the *ID*s of the obtained neighbors are stored in a matrix. Further, this procedure is repeated for the rest of the remaining n-1 persons in the training set, and then their cardinality values, as well as, *ID*s of the neighbors are also stored in the same matrix.

Using this matrix we can obtain our K initial centers for K modal transforms. These K centers are chosen as the K top persons with highest number of neighbors. However, it could be possible that two or more persons can have the same cardinality value, as well as, share the same nearest neighbors *ID*s. In that condition simply choosing top K persons is not the best solution. Therefore, we chose only those top K persons that do not have any person *ID*s common in their neighbor's lists. In addition,

for situations where more than two persons have the same cardinality and share the same neighbors *ID*s, we randomly chose any one person from them to represent that modal center. Finally, getting the *K* modal centers the optimal partitioning of the image space *X* is obtained by minimizing the variance within the given modal scatter matrix as:

$$\arg\min Tr(S_w) \tag{5.3}$$

Although the image space is partitioned into *K* modals, however, to ensure the obtained modals are distinct and stable (in our work a stable modal is formed when it contains at least 15% training persons) we have updated the modal centers and re-partitioned the space for further *t* =3 times. The modal centers are updated as:

$$m_k = \frac{1}{N'_k} \sum_{i=1}^{n} z_{k,i} * x_i \tag{5.4}$$

where N'_k is the number of persons in modal *k*, and given as:

$$N'_k = \sum_{i=1}^{n} z_{k,i} \tag{5.5}$$

Computing the initial modal centers is computationally tedious in our work, however, it has still moderate computational burden. For the training size of *n* persons the complexity is about $O(t \times K \times n)$,where *t* is the number of iterations, and *K* is the number of modals.

5.3.3 Cross Views Impostors (*C.V.I.*)

After getting the distinct modals in the image space *X*, we can now obtain the set of *C.V.I.* for each positive pair (x_i^a , x_i^b) that are lying in modality space *k* from both of its *Probe* and *Gallery* views. We believe that in real world situation (open set), where a positive pair always has limited or few samples, these *C.V.I.* impostors can be exploited to deliver subtle and differentiating information in metric learning that can differentiate a given pair more efficiently against large number of diverse real world impostors and negative gallery samples. These impostors are obtained by comparing the similarity value of a given person pair against the other persons in *Gallery* and

Probe views. First, the similarity values for a probe sample x_i^a is computed with the whole *Gallery* view using metric M_{ini} and *CCA* reduced feature *F* as:

$$S_{probe}^{i} = (x_i^a - x_j^b)^T M_{ini} (x_i^a - x_j^b) \tag{5.6}$$

where x_i and x_j are CCA reduced feature *F* of person *i* and *j*, while M_{ini} is a globally learned metric with feature *F* using K-LFDA [13]. We have used linear kernel to save memory and computational time. Similarly, the similarity values for gallery person x_i^b are obtained with the whole *Probe* view as:

$$S_{gallery}^{i} = (x_i^b - x_j^a)^T M_{ini} (x_i^b - x_j^a) \tag{5.7}$$

These obtained values S_{probe}^{i} and $S_{gallery}^{i}$ for person (x_i^a , x_i^b) in modal *k* are then stored into two sets as:

$$Sim_{x_i^a} = [S_{probe,i'}^{i}]_{i'=1.....N'_k} \tag{5.8}$$

and

$$Sim_{x_i^b} = [S_{gallery,i'}^{i}]_{i'=1.....N'_k} \tag{5.9}$$

where N'_k refers to the number of persons in a modal *k*. Now, we compare each similarity value in these sets with the reference similarity value $S_{(x_i^a - x_i^b)}^{ref}$ of a given pair (x_i^a , x_i^b) to obtain its *C.V.I.* $Set_{(x_i^a - x_i^b)}^{C.V.I.}$ as:

$$Set_{(x_i^a - x_i^b)}^{C.V.I.} = [x_p], \tag{5.10}$$

where $x_p = S_{probe,p}^{i} < S_{(x_i^a, x_i^b)}^{ref}$ or $x_p = S_{gallery,p}^{i} < S_{(x_i^a, x_i^b)}^{ref}$. And *p* is the index of impostor person, and $S_{(x_i^a - x_i^b)}^{ref}$ is computed as:

$$S_{(x_i^a - x_i^b)}^{ref} = (x_i^a - x_i^b)^T M_{ini} (x_i^a - x_i^b) \tag{5.11}$$

Further, using Eqs.5.6-5.11, C.V.I. set $Set_{(x_i^a - x_i^b)}^{C.V.I.}$ for all the N'_k persons in the

modality space *k* are computed. The computational cost of generating cross views impostors for a modality space *k* is about $O(3\times N'_k)$, where $N'_k \ll n$.

5.3.4 Negative Gallery Samples (*N.G.S.*)

We have also used negative gallery samples (*N.G.S.*) to learn metric M_k. Set of *N.G.S.* is denoted as $Set^{Ng}_{(x_i^a - x_i^b)}$. For person pair (x_i^a, x_i^b) *N.G.S.* are obtained from *Gallery* view only as:

$$Set^{Ng}_{(x_i^a - x_i^b)} = [x_q^b], \tag{5.12}$$

where $q \neq p$ in $Set^{C.V.I.}_{(x_i^a - x_i^b)}$, and $q \neq i$ for probe *i*. Here *q* is the index of *N.G.S.*. Further, the set of N.G.S. $Set^{Ng}_{(x_i^a - x_i^b)}$ for all N'_k persons in modal *k* are then obtained using the above Eq.5.12.

5.3.5 Triplet Formulation

Getting the set of C.V.I. $Set^{C.V.I.}_{(x_i^a - x_i^b)}$ and N.G.S. $Set^{Ng}_{(x_i^a - x_i^b)}$ for all N'_k persons in modality space *k*, we will now generate triplet samples to learn metric M_k. Since the positive samples for each person x_i are too scarce than the number of negative samples, following the protocol of data augmentation in [34] we augment each person pair five times. Similarly, following the protocol in [26] we generate 20 triplets for each positive pair. Now, the triplet samples T_i^{imp} and T_i^{Ng} for person x_i using impostor *p* and negative gallery *q* are given as:

$$T_i^{imp} = [< x_i^a, x_i^b, p >] \tag{5.13}$$

and

$$T_i^{Ng} = [< x_i^a, x_i^b, q >] \tag{5.14}$$

where *p* and *q* are taken from respective sets $Set^{C.V.I.}_{(x_i^a - x_i^b)}$ and $Set^{Ng}_{(x_i^a - x_i^b)}$ of person x_i.

5.3.6 Impostors Resistance Multi-Modal Metric (*IRM3*) Learning

Taking triplets from T_i^{imp} and T_i^{Ng}, metric *IRM3* for modality space *k* is learned using *MK-LFDA* [93].To save both the computational time and memory requirements, we adapted [93] and use three RBF kernels and one χ^2 kernel. The weights for these kernels are learned globally for once for each dataset in our work using the similar method in [93]. The reason to learn weights globally is to both save time and computational burden. Further, there is considerably minor effect on kernel weights, even the weights are learned globally. This is due to the fact that the global space is comprised of all the existing modality spaces, and thus, all the modality spaces contribute in learning the global weights. For learning weights of kernels, all the extracted features are used individually, and the dimension of these features are also individually reduced to 450 by CCA before learning weights.

In all our experiments the obtained weights for VIPeR are 0.3, 0.22, and 0.22 for RBF kernels, while weight for χ^2 kernel is 0.26. For*CUHK01* and *CUHK03*, the obtained weights for RBF kernels are 0.28, 0.24, and 0.24, while weight for χ^2 kernel is 0.24. The σ values in all the datasets for the three RBF kernels are set to the mean value of modal *k*, which is, $(mean\ \text{value} + \frac{mean\ \text{value}}{2})$ and $(mean\ \text{value} - \frac{mean\ \text{value}}{2})$. These values for σ are chosen to model all the different variations in the modal *k*. While the σ value for χ^2 kernel is also set to mean value of modal *k*. The mean value in our work is the similarity value between probe and gallery samples of center m_k. Finally, the metric M_k is learned as:

$$\max_{M_k} Tr\left(\frac{M_k^T S_B M_k}{M_k^T S_W M_k}\right) \tag{5.15}$$

where matrices S_B and S_W are obtained with similar method in [93]. Eq.5.15 is then solved using generalized eigenvalue problem [98] in Eq.5.16 to obtain first *r*'=300 eigenvectors corresponding to eigen values with largest magnitude as:

$$S_B\varphi = \lambda S_W\varphi \tag{5.16}$$

5.3.7 Re-Identification

From Fig.5.2, re-identification between test pair (x_i^a, x_j^b) is performed by first determining the transform modality space the test pair belongs to using *K-NN* classifier. In *K-NN* classifier, the parameter **K** is set to the number of modality spaces in the image space, i.e. in VIPeR the value of **K** is set to the number of modality spaces *k*=7. Then, the features of (x_i^a, x_j^b) are projected into the weighted multi-kernel space of the respective modality space, followed by the respective modality space metric M_k to perform matching as:

$$d_{(x_i^a, x_j^b)} = (x_i^a - x_j^b)^T M_k (x_i^a - x_j^b) \tag{5.17}$$

5.4 Experiments

Our *IRM3* metric is evaluated on three benchmark datasets: VIPeR, CUHK01, and CUHK03. We follow the evaluation protocol of [70] for test/train split for VIPeR, CUHK01, and CUHK03 datasets. In our work we have tested CUHK01 for *p*=486 only, while, CUHK03 is tested for both labeled and detected settings. All the experiments are conducted in single-shot mode, and all the reported Cumulative Matching Curve (CMC) are obtained by averaging the results over 20 trials.

5.4.1 Experiment Protocols

To thoroughly analyze the performance of *IRM3* we have devised three evaluation strategies. These strategies evaluate *IRM3* performance with different number of discovered modality spaces *K* in *X*, with Gallery view impostors (*G.V.I.*), as well as, Cross views impostors (*C.V.I.*). *G.V.I.* are the impostors from *Gallery* view only, and are obtained in similar way as in previous conventional metrics [8,90,91,92].

- **IRM3 Only:** It is basic multi-modal metric, learned with only Negative Gallery Samples (N.G.S.).

- **IRM3+ G.V.I.(*p*'):** IRM3 is learned with impostors from *Gallery* view (*G.V.I.*), as well as, with *N.G.S.*. Here *p*' refers to the number of impostors taken from *Gallery* view to form triplet samples, and have values *p*'=5, 10, and 15. While, the remaining triplets are formed using *N.G.S.*
- **IRM3+ C.V.I.(*p*'):** IRM3 is learned with *C.V.I.*, as well as, with *N.G.S.*. Here *p*' refers to number of *C.V.I.* samples used to form triplets, and have values *p*' =5, 10, and 15. While, the remaining triplets are formed using *N.G.S.*

All the samples from *N.G.S.*, *G.V.I.*, as well as, *C.V.I.* contain most difficult instances for a person and are randomly sampled offline, before, training metric. In all the three strategies above, we have partitioned image space into *k*=3, 5, and 7 for VIPeR, while, for CUHK01 we have used *k*=6, 7, and 10 partitions, and for CUHK03 *k*=13, 14, and 16 partitions are used, respectively.

5.4.2 Results on VIPeR

Comparison with State of the art Features: Results of *IRM3* metric are compared with three state of the art features LOMO [5], GoG [7] and moM_f^{LE} [18] in Table.5.1. All the results in Table.5.1 are obtained for *K*=7 modality spaces, and our IRM3+C.V.I.(*p*'=15) has attained rank@1 **52.81%** and has outperformed all the three features of re-identification, providing evidence that if the metric can address multi-modal transform variations well, as well as, have strong resistance against impostors, then the matching accuracy can be improved. Our learned IRM3+C.V.I.(*p*'=15) considers to optimize all the rank orders simultaneously, and thus, has large improvement at rank@5 and rank@10.

Comparison with Metric Learning: We also compared metric *IRM3* with 7 metrics. From Table.5.1 IRM3+C.V.I.(*p*'=15) has outperformed both multi-modal metric LAFT [41] and impostor resistance metric LISTEN [92]. The prime difference between IRM3 and [41,92] is its capability of addressing both the person modal transform, as well as, capability to further maximize the matching against joint constraint of cross views impostors. All these are the causes of great challenge in

matching pedestrians. In Table.6.1 only SS-SVM [62] is a metric that tries to model the transform modal for each individual person, however, it never paid attention to acquire resistance against impostors, and thus has 19.21% lower rank@1 accuracy than IRM3+C.V.I.(p'=15). Though, IRM3 has successful results, still it has 1.36% lower rank@1 than SCSP [71]. Obviously, VIPeR has large pose, mis-alignment and body parts displacement issues which are especially not addressed in our work, and thus, is necessarily needed to improve the matching and results largely.

Table 5.1 Top Rank Comparison on VIPeR Dataset

	VIPeR (Single-Shot, p=316)			
	Method	***r*=1**	***r*=5**	***r*=10**
F	LOMO[5]	**40.0**	**-**	**80.51**
	mOM[18]	**48.0**	**76.8**	**85.4**
	GoG[7]	**49.7**	**79.7**	**88.7**
DF	SIR-CIR[25]	**35.76**	**68.38**	**82.9**
DMN	GS-CNN[100]	**37.8**	**66.9**	**77.4**
	DGD[29]	**38.6**	**-**	**-**
	LSTM[101]	**42.4**	**68.7**	**79.4**
	Mu Deep[68]	**43.03**	**74.36**	**85.76**
	E2E-CAN[69]	**47.2**	**79.2**	**89.2**
	DLPA[30]	**48.7**	**74.7**	**85.1**
	Quad-NET[70]	**49.05**	**73.10**	**81.96**
	JLML[75]	**50.2**	**74.2**	**84.3**
M	ITL[101]	**15.2**	**34.2**	**45.9**
	LAFT[41]	**29.6**	**-**	**69.3**
	LISTEN[92]	**37.47**	**70.78**	**-**
	WARCA[79]	**39.62**	**69.97**	**82.9**
	L-1 GRAPH[102]	**41.5**	**-**	**-**
	S.SSVM[62]	**42.66**	**70.1**	**84.27**
	SCSP[71]	**53.54**	**82.59**	**91.49**
	***IRM3* Only**	**45.92**	**82.90**	**91.63**
	IRM3+ G.V.I.(p'=15)	**50.39**	**85.79**	**95.73**
	IRM3+ C.V.I.(p'=15)	**52.81**	**87.95**	**97.29**

Comparison with Deep Methods: Though deep features (DF) and deep matching networks (DMN) have no match with conventional metric learning methods, from the results in Table.5.1 it is clearly evident if two major issues of re-identification (i.e. multi-modal transforms, and strong rejection capability against impostors) can be well handled simultaneously, then comparable or even higher performance than deep methods can be attained. Our IRM3+C.V.I.(p'=15) has **7.1%**

and **4.94%** higher rank@1 than Quadruplet-Net [70] and JLML [75], respectively. These obtained results demonstrate the fact that for smaller dataset like *VIPeR* deep matching networks have insufficient training samples to learn a discriminative network.

Fig.5.3 shows the comparison of retrieval results of two queries from *VIPeR* dataset for XQDA[5] and our IRM3+C.V.I.(p'=15) when K=7 modality spaces are used. Retrieval results of *Query1* for XQDA find the correct match at rank=4 enclosed in green rectangle ***b***. While, *IMR3* finds the match at rank=2 enclosed in green rectangle ***e***. Similarly, for *Query2* our *IMR3* finds the match at rank=1 enclosed in green rectangle ***j***, in contrast, XQDA finds the correct match at rank=3 enclosed in green rectangle ***h***. Thus, our *IRM3* approach improves matching, and consequently rank gets higher.

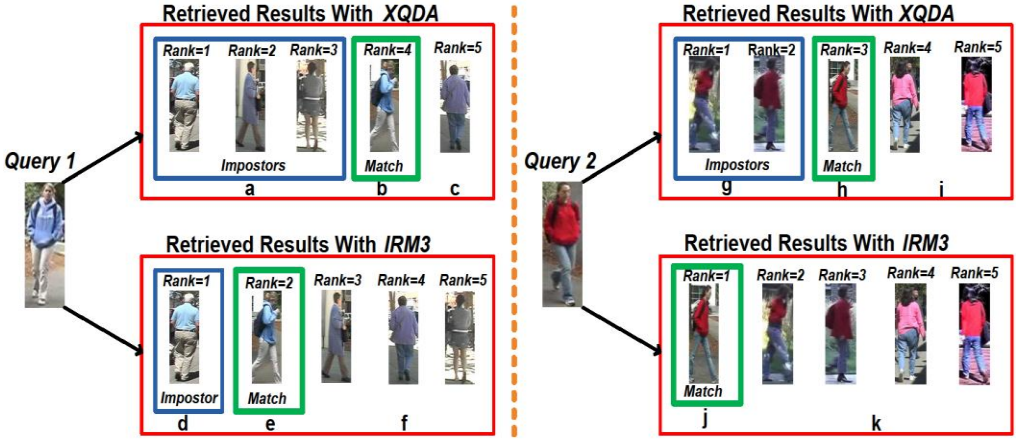

Figure 5.3 Two Queries are shown Query1 and Query2, and their retrievals results using XQDA [34] and using our IRM3. Correct Match is shown in green rectangle, while, blue rectangle shows Impostors.

5.4.3 Results on CUHK01

Comparison with State of the art Features: Table.5.2 summarizes results of *IRM3* for K=10 modality spaces in CUHK01, and compares the obtained results with

LOMO [5], GoG [7], and moM_f^{LE} [18]. Though, the three features are discriminative, however, our *IRM3* approach is better than the three features in solving the two big challenges of Re-ID, i.e. multi-modal pedestrians matching and impostors resistance.

Since, CUHK01 has larger training set than VIPeR, thus, modal transforms can be well learned, and therefore, IRM3+C.V.I.(p'=15) attains larger discrimination than moM_f^{LE} [18]. Our IRM3+C.V.I.(*p*'=15) has **15.15%** higher rank@1 accuracy than moM_f^{LE} due to inherent virtue of handling different modality spaces, person specific variations, as well as, rejecting large number of impostor, all simultaneously.

Comparison with Metric Learning: In Table.5.2 three most recently proposed metrics WARCA [79], L-1 Graph [102], and CVAML [103] are compared with our *IRM3* approach. All the three metrics have the assumption of uni-modal inter-camera transform, rather than, multi-modal image space. Though, WARCA [79] have employed hard negative samples as learning constraint, however, ignoring other negative samples from *Gallery* view, as well as, did not take into consideration a person modality space during learning have made it suffer greatly to attain higher accuracy. On the other hand, IRM3+C.V.I.(*p*'=15) has capability to deal all these challenges, and thus, has attained **76.14%** rank@1 accuracy.

Comparison with Deep Methods: In Table.5.2, we can see several deep matching networks (DMN) have performed much well than conventional metrics on CUHK01. Only *K-LFDA* when trained with moM_f^{LE} [18] feature attains comparable performance than DMN. However, motivated to resolve the challenges for re-identification in real world (i.e. multi-modal image space, and diverse impostors) IRM3+C.V.I.(*p*'=15) has much better results than MCP-CNN [26], E2E-CAN [69], Quadruplet-Net [70], and JLML [75]. While, our IRM3+C.V.I.(p'=15) has **1.49%** higher rank@1 than DLPA [30]. DLPA extracts deep features by semantically aligning body parts, as well as, rectify pose variations. We believe if semantic body parts alignment, and rectification of poses variations are included in our *IRM3* then

the results can be further improved.

Table 5.2 Top Rank Comparison on CUHK01 Dataset

	CUHK01 (Single-Shot, p=486)			
	Method	***r=1***	***r=5***	***r=10***
F	LOMO[5]	**49.2**	**75.7**	**84.2**
	mOM[18]	**64.6**	**84.9**	**90.6**
	GoG[7]	**57.8**	**79.1**	**86.2**
DMN	MPC-CNN[26]	**53.7**	**84.3**	**91.0**
	DGD[29]	**66.6**	**-**	**-**
	E2E-CAN[69]	**67.2**	**87.3**	**92.5**
	DLPA[30]	**75.0**	**93.5**	**95.7**
	Quad-NET[70]	**62.55**	**83.44**	**89.71**
	JLML[75]	**69.8**	**88.4**	**93.3**
M	CVAML[103]	**57.3**	**81.2**	**86.5**
	WARCA[79]	**58.34**	**79.76**	**-**
	L-1 GRAPH[102]	**50.1**	**-**	**-**
	***IRM3* Only**	**68.24**	**88.71**	**94.31**
	IRM3+ G.V.I.(p'=15)	**72.91**	**94.61**	**97.6**
	IRM3+ C.V.I.(p'=15)	**76.14**	**96.90**	**98.35**

5.4.4 Results on CUHK03

Comparison with State of the art Features: Table.5.3 compares LOMO [5] and GoG [7] features with our *IRM3* metric in both Labeled and Detected settings. All the results in Table.5.3 are obtained for K=16 modals. In Table.5.3, obtained results are much higher than the two features. Our *IRM3* and [5,7] mainly differs in motivation. In [5,7] a universal feature representation is proposed for all the different persons, which may not be optimal for all the persons at the same time residing on different modality spaces, in contrast, our motivation is based on discovering distinct modals in the image space and then addressing each modality space specifically with empowerment of large number of impostors rejection. Therefore, our IRM3+C.V.I.(p'=15) (in Labeled setting) has rank@1 accuracy of about **86.17%.**

Comparison with Metric Learning: In Table.5.3, recently proposed SSM [73] and WARCA [79] are compared with our *IRM3* approach. WARCA [79] differs with our *IRM3* approach in a way that it only addresses hard negative samples, while, SSM [73] differs in a way that it has no measure to account for different modality transform spaces, as well as, have no resistance against impostors. Our IRM3+C.V.I.(p'=15) (in Labeled setting) has surpassed [73] and [79], and has attained **11.1%** and **9.04%** rise

at rank@1 accuracy, respectively.

Comparison with Deep Methods: Interestingly, in Table.5.3 all the deep methods in Labeled and Detected settings have very high performance on CUHK03. These high results demonstrate the fact that CUHK03 is the largest dataset among all, and thus, can help in learning a more discriminative DMN. Even though, both DLPA [30] and JLML [75] learn deep body features with global and local body parts alignment, as well as, pose alignment, however, our *IMR3* approach benefitted with transform specific metrics empowered with impostors rejection have still maintained to attain better results. Our *IRM3* considers optimizing all the rank orders simultaneously, and thus, have large gain atrank@5 and rank@10 in Labeled setting.

Table 5.3 Top Rank Comparison on CUHK03 Dataset

	CUHK03 (LABELED, p=100)			
	Method	***r*=1**	***r*=5**	***r*=10**
F	LOMO[5]	**52.2**	**82.23**	**92.14**
	GoG[7]	**67.3**	**91.0**	**96.0**
DF	DCAF[28]	**74.21**	**94.33**	**97.54**
DMN	DGD[29]	**75.3**	-	-
	Quad-NET[70]	**75.53**	**95.15**	**99.16**
	Mu Deep[68]	**76.87**	**96.12**	**98.41**
	E2E-CAN[69]	**77.6**	**95.2**	**99.3**
	JLML[75]	**83.2**	**98.0**	**99.4**
	DLPA[30]	**85.4**	**97.6**	**99.4**
M	S.SSVM[62]	**57.0**	**85.7**	**94.3**
	Null Sp.[72]	**62.55**	**90.05**	**94.80**
	SSM[73]	**76.6**	**94.6**	**98.0**
	WARCA[79]	**78.38**	**94.55**	-
	***IRM3* Only**	**78.83**	**95.97**	**98.37**
	IRM3+ G.V.I.(p'=15)	**83.32**	**98.70**	**99.54**
	IRM3+ C.V.I.(p'=15)	**86.17**	**99.02**	**99.68**
	CUHK03 (DTECTED, p=100)			
	Method	***r*=1**	***r*=5**	***r*=10**
F	LOMO[5]	**46.25**	**78.9**	**88.55**
	GoG[7]	**65.5**	**88.4**	**93.7**
DF	SIR-CIR[24]	**52.17**	**83.7**	**90.4**
	DCAF[28]	**67.99**	**91.04**	**95.36**
DMN	LSTM[100]	**57.3**	**80.10**	**88.3**
	GS-CNN[99]	**68.1**	**88.1**	**94.6**
	E2E-CAN[69]	**69.2**	**88.5**	**94.1**
	Mu Deep[68]	**75.64**	**94.36**	**97.46**
	JLML[75]	**80.6**	**96.9**	**98.7**
	DLPA[30]	**81.6**	**97.3**	**98.4**
M	L-1 GRAPH[102]	**39.0**	-	-
	S.S-SVM[62]	**51.2**	**80.8**	**89.6**
	Null Sp.[72]	**54.70**	**84.75**	**94.80**
	SSM[73]	**72.7**	**92.4**	**96.1**
	***IRM3* Only**	**72.98**	**91.7**	**93.02**
	IRM3+ G.V.I.(p'=15)	**78.68**	**95.60**	**98.09**
	IRM3+ C.V.I.(p'=15)	**80.77**	**96.94**	**98.67**

5.5 Summary

This chapter provides the detail analysis about the multi-modal metric learning with impostor resistance in person re-identification. In addition, we have analyzed several other factors in obtaining stable and reliable modals. The experiments clearly provide evidences that if the person is well identified that which modal it belongs to then we could learn a robust and discriminative metric to match all the different multi-modal persons.

Chapter 6. **Conclusion and Future Work**

In our work we have presented an approach to learn metric using multiple kernels. The person re-identification images are acquired from multiple views which undergo several complex non-linear and random changes, and thus, we have multi-modal feature space. In such situation, we have presented a global, as well as, local multiple kernel learning based metric learning methods to deal with the multi-modal feature space and complex non-linear changes. Therefore, after multiple kernel projection the discrimination among persons is further maximized.

The purpose of using multiple kernel learning for the re-identification problem is to differentiate well all the observed persons by projecting the features into the learned space. Then, after feature projection in a carefully learned feature space, the projected features can help in well distinguishing the persons, and thus help in learning a more robust metric.

Our proposed multiple kernel method is based on the intuition that re-identification needs discriminative features and feature learning is still an open problem, and the state of the art features already proposed in re-identification are still not fully exploited due to the complexity in feature space, as well as, in the past there are none carefully designed feature projection methods proposed for re-identification feature space. Therefore, to exploit these proposed features well we have opted to model the feature space using multiple kernel rather than attempting to learn another new feature.

With the obtained promising results using multiple kernels, it is well evident that if both the non-linearity and multi-modality can be well addressed with a carefully learned projection space, then the learned metric could attain much higher matching performance, and can match the observed non-linear pedestrians observed in non-overlapping views. However, still we believe there is much work needed to be done for higher re-identification accuracy at rank@1, i.e. it is needed to design new kernels specific for re-identification and then learn more robust kernelized metrics for

person matching. Further, it is needed that the learned metric is scalable, fast, and require lower memory.

In addition, there are still many unsolved challenges for real world implementation of re-identification which are also need to be taken into care. Our future aim is to work on end to end person re-identification in real world environment, as well as, its implementation of logic devices, i.e. on microprocessor or FPGA.

REFERENCES

[1] U. Park, Anil K. Jain, I. Kitahara, K. Kogure and N. Hagita, "ViSE: Visual Search Engine Using Multiple Networked Cameras", IN ICPR, 2006, pp. 1204-1207.

[2] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person Re-identification: Past, Present and Future", CoRR, 2016.

[3] S. Liao, Z. Mo, J. Zhu, Yang Hu, and Stan Z. Li, "Open-set Person Re-identification", arXiv:1408.0872, 2014.

[4] S. Vinay, K. Sameh, and K. C. Hao, "Person Re-identification Using Semantic Color Names and RankBoost", in WACV, 2013, pp. 281-287.

[5] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by Local Maximal Occurrence representation and metric learning ", in CVPR,2015,pp. 2197-2206.

[6] Y. Yang, Y. Jimei, Y. J. Jie, L. Shengcai Y. Dong, and L. Stan Z. ,"Salient Color Names for Person Re-identification", in ECCV, 2014.

[7] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification", in CVPR 2016, pp. 1363-1372.

[8] Weinberger K.Q.,and Saul L.K, "Distance metric learning for large margin nearest neighbor classification", Journal of Machine Learning Research, vol.10,2009, pp 207-244.

[9] D. Jason V., K. Brian, J. Prateek, S. Suvrit, and D. Inderjit S., "Information-theoretic Metric Learning", in ICML, 2007.

[10] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large Scale Metric Learning from Equivalence Constraints", in CVPR, 2012.

[11] Y. C. Shan, S. W. Dan, and C. Xilin "View-Adaptive Metric Learning for Multi-view Person Re-identification", in ACCV, 2014.

[12] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian,"Local Fisher Discriminant Analysis for Pedestrian Re-identification", in CVPR, 2013.

[13] X. Fei, G. Mengran, C. Octavia, and S. Mario, "Person Re-Identification Using Kernel-Based Metric Learning Methods", in ECCV, 2014.

[14] 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 CVPR, 2015.

[15] D. Gray, and H. Tao, "Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features," in ECCV, 2008,pp. 262-275.

[16] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification", in BMVC, 2012, pp. 1–11.

[17] 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,pp 379-390,2014.

[18] G. Mengran, C. Octavia, and S. Mario, "moM: Mean of Moments Feature for Person

Re-Identification", in ICCV 2017.

[19] T. Kobayashi, and N. Otsu, " Image feature extraction using gradient local auto-correlations", In ECCV, 2008, pp. 346–358.

[20] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person Re-identification Using Spatiotemporal Appearance ", in CVPR, 2006,pp. 1528-1535.

[21] T. Lorenzo, K. Vladimir,and R. Carsten, "Feature Correspondence Via Graph Matching: Models and Global Optimization ", in ECCV, 2008, pp. 596-609.

[22] C. Liu, S. Gong, C. C. Loy, and X. Lin., "Person Re-Identification: What Features are Important? ", in ECCV,2012.

[23] D. Figueira , L. Bazzani, H. Q. Minh, and M. Cristani , "Semi-supervised multi-feature learning for person Re-identification", in AVSS, 2013, pp.111-116.

[24] A. Ejaz, J. Michael, and M. Tim K.,"An Improved Deep Learning Architecture for Person Re-Identification", in CVPR, 2015.

[25] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single image and cross-image representations for person re-identification", in CVPR 2016, pp. 1288-1296.

[26] 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1335-1344.

[27] Z. Haiyu, T. Maoqing, S. Shuyang, S. Jing, Y. Junjie, Y. Shuai, W. Xiaogang, and T. Xiaoou, "Spindle Net: Person Re-Identification With Human Body Region Guided Feature Decomposition and Fusion", in CVPR 2017, pp. 907- 915.

[28] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning Deep Context-aware Features over Body and Latent Parts for Person Re-identification", in CVPR 2017, pp. 7398-7407.

[29] X. Tong, L. Hongsheng, O. Wanli, and W. Xiaogang, "Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification", in 'CVPR2016, pp. 1249-1258.

[30] L. Zhao, Xi Li, Y. Zhuang, and J. Wang, "Deeply Learned Part Aligned Representations for Person Re-identification", in ICCV 2017, pp. 3239-3248.

[31] Y. C. Chen, W. S. Zheng, J. H. Lai, and P. Yuen, "An Asymmetric Distance Model for Cross-view Feature Mapping in Person Re-identification", IEEE Transactions on Circuits and Systems for Video Technology, 2016.

[32] L. An, S. Yang, and B. Bhanu, " Person Re-Identification by Robust Canonical Correlation Analysis", IEEE Signal Processing Letters, vol. 22, no. 8, 2015, pp. 1103-1107.

[33] G. Lisanti, I. Masi, and A. D. Bimbo, "Matching People across Camera Views using Kernel Canonical Correlation Analysis", in ACM ICDSC, 2014.

[34] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep Filter Pairing Neural Network for Person Re-identification", in CVPR, 2014.

[35] Y. C. Chen, W. S. Zheng, and J. Lai, "Mirror Representation for Modeling View-Specific Transform in Person Re-Identification", in International Joint Conference on Artificial Intelligence, 2015, pp. 3402-3408.

[36] A. Datta, L. M. Brown, R. Feris, and S. Pankanti, "Appearance modeling for person re-identification using Weighted Brightness Transfer Functions", in ICPR, 2012.

[37] A. Tamar, G. Ilya, L. Michael, and M. Shaul, "Learning Implicit Transfer for Person Re-identification", in ECCV, 2012.

[38] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith ,"Learning Locally-Adaptive Decision Functions for Person Verification", in CVPR, 2013.

[39] M. S. Biagio, A. Ulaş, M. Crocco, M. Cristani, U. Castellani, and V. Murino, "A multiple kernel learning approach to multi-modal pedestrian classification", in ICPR 2012, pp.2412 -2415.

[40] N. Martinel, C. Micheloni, and G. L. Foresti, "Kernelized saliency-based person re-identification through multiple metric learning", in IEEE Transactions on Image Processing , Vol. 24, Issue 12, 2015, pp. 645-5658.

[41] W. Li, and X. Wang, "Locally Aligned Feature Transforms across Views", in CVPR, 2013.

[42] L. An, M. Kafai, S. Yang, and B. Bhanu, "Reference-based person re-identification", in AVSS, 2013, pp. 244-249.

[43] F. Pala, R. Satta, G. Fumera, and F. Roli, "Multimodal Person Re-identification Using RGB-D Cameras", IEEE Transactions on Circuits and Systems for Video Technology, Vol.26 , Issue.4, pp. 788-799.

[44] G. Mehmet, and A. Ethem, "Localized Algorithms for Multiple Kernel Learning", Pattern Recognition , vol. 46, no.3, 2013, pp 795-807.

[45] G. Mehmet, and A. Ethem, "Localized Multiple Kernel Learning", in ICML, 2008.

[46] R. Zhao, W. Ouyang, and X. Wang, "Person Re-identification by Salience Matching", in ICCV 2013.

[47] A. Mignon, and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints", in CVPR2012, pp. 2666-2672.

[48] Y. Hu, S. Liao, Z. Lei, D. Yi, and S. Z. Li, "Exploring Structural Information and Fusing Multiple Features for Person Re-identification", in CVPR, 2013, pp. 794-799.

[49] C. Liu, S. Gong, and C. C. Loy, "On-the-fly Feature Importance Mining for Person Re-Identification", Pattern Recognition, Vol. 47, no. 4, 2014, pp. 1602-1615.

[50] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification", In BMVC, 2011, pp. 68.1--68.11.

[51] D. Baltieri, R. Vezzani, and R. Cucchiara, "3DPes: 3D People Dataset for Surveillance and Forensics", in ACM Workshop on Multimedia access to 3D Human Objects, 2011, pp. 59-64.

[52] W. S. Zheng, S. Gong, and T. Xiang, "Associating Groups of People", In BMVC, 2009, pp. 23.1-23.11.

[53] U. H. Office, i-LIDS multiple camera tracking scenario definition, 2008.

[54] L. Wei, Z. Rui, and W. Xiaogang, "Human Reidentification with Transferred Metric Learning", in ACCV, 2012, pp. 31-44.

[55] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person Re-Identification by Video Ranking", In ECCV, 2014, pp 688-703.

[56] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features", in CVPR, 2010, pp. 2360-2367.

[57] W.S. Zheng, S.G. Gong, and T. Xiang, "Reidentification by relative distance comparison", in IEEE Trans. PAMI, vol. 35, no. 3, 2013, pp.653–668.

[58] R. Zhao, W. Ouyang, and X. Wang, "Learning Mid-level Filters for Person Re-identification", in CVPR 2014, pp.144-151.

[59] W. S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison", in CVPR 2011, pp. 649-656.

[60] J. Garca, N. Martinel, G. L. Foresti, A. Gardel, and C. Micheloni, "Person orientation and feature distances boost re-identification", in ICPR 2014, pp.4618-4623.

[61] K. Liu, Z. Zhao, and A. Cai, "Parametric Local Multi-modal Metric Learning for Person Re-identification", in ICPR, 2014, pp. 2578-2583.

[62] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-Specific SVM Learning for Person Re-identification", in CVPR2016, pp.1278-1287.

[63] K. Liu, Z. Zhao, and A. Cai, "Datum-Adaptive Local Metric Learning for Person Re-identification", in IEEE Signal Processing Letters, Vol. 22, Issue 9, 2015, pp.1457-1461.

[64] D. Kong, and C. H. Q. Ding, "Pairwise-covariance linear Discriminant analysis", in AAAI, 2014, pp. 1925-1931.

[65] X. He, "Locality preserving projections", Ph.D. thesis, Chicago, IL, USA, aAI3195015 (2005).

[66] H. Wang, F. Nie, and H. Huang, "Learning robust locality preserving projection via p-order minimization", in AAAI, 2015, pp. 3059-3065.

[67] L. Zelnikmanor, and P. Perona, "Self-tuning spectral clustering", in NIPS, 2004, pp. 1601-1608.

[68] Q. Xuelin, F. Yanwei, J. Y. Gang, X. Tao, and X. Xiangyang, "Multi-Scale Deep Learning Architectures for Person Re-Identification", in ICCV 2017, pp.7398-7407.

[69] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-End Comparative Attention Networks for Person Re-Identification", IEEE Transactions on Image Processing, vol.26, no.7, pp.3492-3506, 2017.

[70] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-identification", in CVPR 2107, pp.1320-1329 .

[71] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification", in CVPR 2016, pp. 1268-1277.

[72] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification", in CVPR 2016, pp. 1239-1248.

[73] S. Bai, X. Bai, and Q. Tian, "Scalable Person Re-identification on Supervised Smoothed Manifold", in CVPR 2017, pp.3356-3365.

[74] 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 CVPR 2017, pp.907-915.

[75] W. Li, X. Zhu, and S. Gong, "Person Re-Identification by Deep Joint Learning of Multi-Loss Classification", CoRR, bs/1705.04724,2017.

[76] J. Zhou, P. Yu, W. Tang, and Y. Wu, "Efficient Online Local Metric Adaptation via Negative Samples for Person Re-Identication", in ICCV2017, pp. 2439-2447.

[77] S. Liao, Y. Hu, and S. Z. Li, "Joint dimension reduction and metric learning for person re-identification", CoRR abs/1406.4216.

[78] Y. J. Cho, and K. J. Yoon, "Improving Person Re-Identification via Pose-Aware Multi-Shot Matching", in CVPR, 2016, pp. 1354-1362.

[79] J. Cijo, and F. Fleuret, "Scalable Metric Learning via Weighted Approximate Rank Component Analysis", in ECCV2016, pp. 875-890.

[80] S. Zhou, J. Wang, D. Meng, X. Xin, Y. Li, Y. Gong, and N. Zheng, "Deep Self-Paced Learning for Person Re-Identification", Pattern Recognition, Vol.76, 2017, pp.739-751.

[81] S. Zhou, J. Wang, D. Meng, X. Xin, Y. Li, and Y. Gong, "Point to Set Similarity Based Deep Feature Learning for Person Re-Identification", in CVPR 2017, pp. 5028-5037.

[82] T. Matsukawa, T. Okabe, and Y. Sato, "Person re-identification via discriminative accumulation of local features", in ICPR 2014, pp. 3975-3980.

[83] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries", in ICCV 2015, pp. 4516-4524.

[84] J. You, A. Wu, X. Li, and W.S. Zheng, "Top-push video-based person re-identification", in CVPR 2016, pp. 1345 - 1353

[85] N. McLaughlin, J. M. d. Rincon, and P. Miller, "Recurrent Convolutional network for video-based person re-identification", in CVPR 2016, pp. 1325 - 1334

[86] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly Attentive Spatial-Temporal Pooling Networks for Video-based Person Re-Identification", in ICCV 2017, pp. 4743- 4752.

[87] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person Re-Identification via

Recurrent Feature Aggregation", in ECCV2016, pp. 701-716.

[88] C. Su , F. Yang , S. Zhang , Q. Tian , L. S. Davis, and W. Gao, "Multi-Task Learning with Low Rank Attribute Embedding for Person Re-Identification", in ICCV, 2015.

[89] Y. Li, Z. Wu, S. Karanam, and R. J. Radke, "Multi-shot human re-identification using adaptive Fisher Discriminant analysis", in BMVC2015, pp.73.1-73.12.

[90] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian Recognition with a Learned Metric", in ACCV 2010, pp.501-512.

[91] M. Hirzer, P. M. Roth, and H. Bischof, "Person Re-identification by Efficient Impostor-Based Metric Learning", in AVSS 2012, pp.203-208.

[92] X. Zhu, X. Y. Jing, F. Wu, W. Zheng, R. Hu, C. Xiao, and C. Liang, "Distance learning by treating negative samples differently and exploiting impostors with symmetric triplet constraint for person re-identification", in ICME 2016, pp.1-6.

[93] M. A. Syed, and J. Jiao, "Multi-kernel metric learning for person re-identification", in ICIP 2016, pp.784-788.

[94] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised Salience Learning for Person Re-identification", in CVPR 2013, pp.3586-3593.

[95] L. Giuseppe, M. Iacopo, and D. B. Alberto, "Matching People Across Camera Views Using Kernel Canonical Correlation Analysis", in ICDSC 2014.

[96] D. R. Hardoon, S. Szedmak, and J. S. Taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods", Neural Computation, vol.16,no. 12, 2014, pp.2639-2664.

[97] X. Cai, F. Nie, H. Huang, and F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering", in CVPR, 2011, pp. 1977-1984.

[98] Y. Ying, P. Li, S. Sonnenburg, F. Bach, and C. S. Ong, "Distance metric learning with eigenvalue optimization", Journal of Machine Learning Research, vol.13, no.1, 2012, pp.1-26.

[99] R. R. Varior, H. Mrinal, and W. Gang, "Gated Siamese Convolutional Neural Network Architecture for Human Re-identification, in ECCV 2016, pp. 791-808.

[100] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A Siamese Long Short-Term Memory Architecture for Human Re-identification", In ECCV 2016, pp.135-153.

[101] L. Wentong, Y. Y. Michael, Z. Ni, and R. Bodo, "Triplet-Based Deep Similarity Learning for Person Re-Identification", in ICCV 2017.

[102] E. Kodirov, T. Xiang, Z. Fu, and S. G. Gong, "Person Re-Identification by Unsupervised l1 Graph Learning", In ECCV2016, pp.178-195.

[103] Y. H. Xing, W. Ancong, and Zheng W. S., "Cross-View Asymmetric Metric Learning for Unsupervised Person Re-Identification", in ICCV2017, pp. 994-1002.

[104] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning Deep Context-Aware Features over

Body and Latent Parts for Person Re-identification", in CVPR 2017, pp.7398-7407.

[105] F. Wang, X. Wang, D. Zhang, C. Zhang, and T. Li, "marginface: A novel face recognition method by average neighborhood margin maximization", Pattern Recognition, Vol. 42, issue 11,2009, pp.2863-2875.

[106] A. B. Gala, and S. K. Shah, "A survey of approaches and trends in person re-identification", Image Vision Comput., vol. 32, no. 4, 2014, pp. 270–286.

[107] S.G. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, "The re-identification challenge, in Person Re-Identification", Advances in Computer Vision and Pattern Recognition, 2014, pp. 1–20.

[108] R. Felipe de C. P., and W. R. Schwartz, "CBRA: Color-based ranking aggregation for person re-identification", in ICIP, 2015.

[109] Ma B., L. Qian, and C. Hong, "Gaussian Descriptor Based on Local Features for Person Re-identification", in ACCV, 2014.

[110] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan, "Clothing Attributes Assisted Person Re-identification", IEEE Transactions on Circuits and Systems for Video Technology, vol.25,no.5, pp 869-878,2015.

[111] R. Layne, T. M. Hospedales, and S. Gong, "Re-id: Hunting Attributes in the Wild", in BMVC, 2014.

[112] J. Chen, Z. Zhang, and Y. Wang, "Relevance Metric Learning for Person Re-Identification by Exploiting Listwise Similarities", IEEE Transactions on Image Processing, vol.24,no.12,pp 4741-4755.

[113] W. Li, Y. Wu, and J. Li, "Re-identification by neighborhood structure metric learning", Pattern Recognition, Vol. 61, 2017, pp.327-338.

[114] N. Martinel, C. Micheloni, and G. L. Foresti, "Saliency Weighted Features for Person Re-identification", in ECCV, 2014, pp.191-208.

[115] M. Zeng, Z.Wu, C. Tian, L. Zhang, and L. Hu, "Efficient person re-identification by hybrid spatiogram and covariance descriptor", in CVPRW2015, pp.48-56.

[116] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses", Pattern Recognition Letters, vol. 33, issue 7, 2012, pp.898-903.

[117] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Multiple-shot human re-identification by mean Riemannian covariance grid", in AVSS 2011, pp.179-184.

[118] A. Franco, and L. Oliveira, "Convolutional covariance features: Conception, integration and performance in person re-identification", Pattern Recognition, Vol.61, 2017, pp. 593–609.

[119] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo, "Person re-identification by iterative re-weighted sparse ranking", in PAMI, Vol. 37, Issue 8, 2015, pp.1629-1642.

[120] J. Dai, Y. Zhang, and H. Lu. "Cross-view semantic projection learning for person

re-identification", Vol.75, 2018, pp. 63-76.

[121] L. An, M. Kafai, S. Yang, and B. Bhanu, "Person re-identification with reference descriptor", in IEEE Transactions on Circuits and Systems for Video Technology, Vol. 26, issue 4, pp. 776-787.

[122] S. Riccardo, F. Giorgio, and R. Fabio," Exploiting Dissimilarity Representations for Person Re-identification ", IN SIMBAD, 2011, pp. 275-289.

[123] L. Yang, and R. Jin. "Distance metric learning: A comprehensive survey", Michigan State University, 2006.

[124] A. J. Ma, P. C. Yuen, and J. Li, "Domain Transfer Support Vector Ranking for Person Re-identification without Target Camera Label Information", in ICCV,2013, pp. 3567-3574.

[125] Bazzani L., Cristani M., Perina A., Farenzena M., and Murino V., "Multiple-shot person re-identification by HPE signature", in ICPR, 2010, pp.1413-1416.

[126] Nakajima C., Pontil M., Heisele B., and Poggio T., "Full-body person recognition system", Pattern Recognition , Vol. 36, issue 9, pp.1997–2006.

[127] Javed O., Shafique K., Rasheed Z., and Shah M., "Modeling inter-camera space time and appearance relationships for tracking across non-overlapping views", Computer Vision and Image Understanding, Vol. 109， issue 2, pp. 146–162.

[128] N. Martinel, C. Micheloni, and G. L. Foresti, "A pool of multiple person re-identification experts", Vol. 71, issue 1, pp. 23-30.

[129] K. Rangachar, and E. Rajmadhan, "Person Reidentification and Recognition in Video", in CIARP, 2014, pp. 280-293.

[130] S. Iodice, and A. Petrosino, "Salient feature based graph matching for person re-identification", Pattern Recognition, Vol. 48, Issue 4, 2015, pp.1074–1085.

[131] Y. Li, and T. Ziru, "Person Re-identification Based on Color and Texture Feature Fusion", in ICIC, 2016, pp. 341-352.

[132] D. Yi, Z. Lei, and S. Z. Li, "Deep Metric Learning for Practical Person Re-Identification", arXiv:1407.4979v1, 2014.

[133] R. R. Varior, G. Wang, and J. Lu, "Learning Invariant Color Features for Person Reidentification", IEEE Transactions on Image Processing, Vol. 25, Issue 7, 2016, pp. 3395-3410.

[134] A. Li, L. Liu, and K. Wang, "Clothing Attributes Assisted Person Reidentification", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 25, Issue 5, 2015, pp.869- 878.

[135] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification", IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 7, 2013, pp. 1622-1634.

[136] V. E. Liong, J. Lu, and Y. Ge, "Regularized local metric learning for person

re-identification", Pattern Recognition Letters, Vol. 68 Issue 2, 2015, pp. 288-296.

[137] H. Wang, S. Gong, and T. Xiang, "Highly Efficient Regression for Scalable Person Re-Identification", in BMVC, 2016, pp.1-14.

[138] N. Zhao, R. Hong, M. Wang, X. Hu, and T. S. Chua, "Searching for Recent Celebrity Images in Microblog Platform", in ACM-MM, 2014, pp. 841-844.

[139] S. Cappallo, T. Mensink, and C. Snoek, "Query-by-Emoji Video Search", in ACM-MM, 2015, pp.735-736.

[140] M. Sugiyama, "Local Fisher Discriminant analysis for supervised dimensionality reduction", in ICML, 2006, pp. 905-912.

[141] R. Vezzani, D. Baltieri, and R. Cucchiara. "People reidentification in surveillance and forensics: A survey, ACM Computing Surveys (CSUR), Vol. 46, Issue 2, 2013, pp.29:1-29:37.

[142] M. Uricár, R. Timofte, R. Rothe, J. Matas, and L. V. Gool, "Structured Output SVM Prediction of Apparent Age, Gender and Smile from Deep Features", in CVPRW, 2016, pp.730-738.

[143] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, "Cross-Modal Retrieval With CNN Visual Features: A New Baseline", IEEE Transactions on Cybernetics, Vol. 47, Issue 2, 2017, pp. 449-460.

[144] M. Sun, J. Yang, B. Sun, and K. Wang, "Shape-guided segmentation for fine-grained visual categorization", in ICME, 2016, pp.1-6.

[145] J. V. de Weijer, C. Schmid, and J. Verbeek, "Learning Color Names from Real-World Images", in CVPR, 2007, pp.1-8.

[146] X. Li, A. Wu, M. Cao, J. You, and W. S. Zheng, "Towards more reliable matching for person re-identification", in ISBA, 2015, pp.1-6.

[147] H. Bouma, S. Borsboom, R. J. M. d. Hollander, S. H. Landsmeer, and M. Worring, "Re-identification of persons in multi-camera surveillance under varying viewpoints and illumination", in SPIE C3I, 2012.

[148] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and Appearance Context Modeling", in CVPR, 2007, pp.1-8.

[149] Y. Takahashi, and H. Miyano, "A compact color descriptor for person re-identification with clothing selection from a wardrobe", in ICME 2015, pp.1-6.

[150] Prosser B., Zheng W.S. Gong, and S. Xiang.T, "Person re-identification by support vector ranking", In BMVC 2010, pp.21.1-21.11.

[151] R. Layne, T. Hospedales, and S. G. Gong, "Person Re-identification by Attributes", in BMVC 2012, pp.24.1-24.11.

[152] M. Nishiyama, S. Nakano, T. Yotsumoto, H. Yoshimura, Y. Iwai, and K. Sugahara, "Person Re-identification using Co-occurrence Attributes of Physical and Adhered Human

Characteristics", In ICPR 2016, pp.1-6.

[153] R. Layne, T. M. Hospedales, and S. G. Gong, "Attributes-Based Re-identification", Person Re-Identification, Part of the series Advances in Computer Vision and Pattern Recognition, 2014, pp 93-117.

[154] E. A. Meirom, and P. Kisilev, "NuC-MKL: A Convex Approach to Non Linear Multiple Kernel Learning", in AISTATS 2016, pp.610-619.

[155] T. Nakashika, A. Suga, T. Takiguchi, and Y. Ariki, "Generic Object Recognition Using Automatic Region Extraction and Dimensional Feature Integration Utilizing Multiple Kernel Learning", in ICASSSP, 2011,pp.1229-1232.

[156] A. J. Ma, J. Li, P. C. Yuen, and P. Li," Cross-Domain Person Re-identification Using Domain Adaptation Ranking SVMs", IEEE Transactions on Image Processing, vol. 24, no.5, 2015, pp.1599-1613.

[157] H. Liua, B. Maa, L. Qinb, J. Pangc, C. Zhanga, and Q. Huang, "Set-label modeling and deep metric learning on person re-identification", Neurocomputing, Vol.151, issue 3, 2015, pp.1283–1292.

[158] M. Ring, and B. M. Eskofier, "An approximation of the Gaussian RBF kernel for efficient classification with SVMs", Pattern Recognition Letters, Vol. 84, 2016,pp.107–113.

[159] B. Mocanu, R. Tapu, and T. Zaharia, "Using Computer Vision to See", in ECCV 2016,pp. 375-390.

[160] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search", in CVPR, 2015,pp. 4184-4193.

[161] Y. Li, Z. Wu, and R. J. Radke, "Multi-shot re-identification with random projection-based random forests", in WACV 2015, pp. 373-380.

[162] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thonnat, "Learning to Match Appearances by Correlations in a Covariance Metric Space", in ECCV 2012, pp. 806-820.

[163] S. Karanam, Y. Li, and R. J. Radke, "Sparse re-id: Block sparsity for person re-identification", in CVPRW 2015, pp. 33-40.

[164] L. Wentong, Y. Y. Michael, Z. Ni, and R. Bodo, "Triplet-Based Deep Similarity Learning for Person Re-Identification", in ICCV 2017.

Acknowledgment

I am thankful to Allah (Almighty) for His blessings on me.

Then I am thankful for my Supervisor Professor Jiao Jianbin and all the respected teachers, including Han Zhenjun, Ye Qixiang, and Qin Fei whom all have provided me guidance and support through all my works. Particularly, I am thankful for Professor Jiao Jianbin, Professor Han Zhenjun in helping me to write this thesis.

In addition, I am also thankful of my Family and my Parents that have encouraged me all through my PhD, specially my mother and my wife that both provided me continuous support.

Syed Muhammad Adnan

2018 年 6 月

作者简历及攻读学位期间发表的学术论文与研究成果

已发表（或正式接受）的学术论文：

1. M. A. Syed and J. Jiao, Multi-kernel metric learning for person re-identification, IEEE International Conference on Image Processing (ICIP), 2016, pp.784-788.

2. M. A. Syed, Zhenjun Han, Zhaoju Li and J. Jiao, Imposter Resilient Multi-Modal Metric learning for Person Re-Identification, Accepted in Press Advances in Multimedia, 2017

3.M. A. Syed, Zhenjun Han, and J. Jiao, Sample Specific Multi-Kernel metric Learning for Person Re-Identification, Elsevier Computer Vision and Image Understanding, (Submitted).