Common-Near-Neighbor Information in Discriminative Spaces for Human Re-identification

LI WEI

April 2014

Abstract

Matching people across camera views at different sites, known as human re-identification, is challenging and valuable for both academia and industry. To date, attempts to address this issue have involved feature representation and/or dissimilarity measurement. Although several improvements have been achieved, the problem is still far from being solved because of the real-world complexities. These complexities involve large within-class variations due to the changeable body appearance and environment and small between-class differences arising from the similar body shape and clothing style.

In the literature, there are many works focusing on feature representation. However, to build an accurate correspondence between highly variable elements (sample points or sample sets) of human image data in the feature space, a reliable dissimilarity measurement is indispensable. Conventional dissimilarity directly describes how far apart the concerned pair of elements they are. Although this kind of dissimilarity has been generalized in a variety of ways in the subsequent development, still it solely relies on the pair of measured elements without considering the neighborhood information. And for re-identification, this kind of dissimilarity measurement seems to confront a dilemma to discriminate the isolated elements that have large within-class variations and small between-class differences.

In the thesis, we reformulate re-identification as the problem of finding the correct matches between the elements from the query side and from the corpus side based on a reliable dissimilarity measure. Motivated by the idea to deliver the effectiveness of those well-distributed elements to those badly-distributed elements in a metric space, we creatively propose to quantify the local neighborhood structure of the pair of elements in each other's neighborhood structures into the dissimilarity, as a new trial to enrich the conventional distance notion by considering the neighborhood information. Here, neighborhood structure is defined as the layout relationship between the concerned element and its neighbors in the given met-

ric space. The property of this dissimilarity can be comprehended by the insight of measuring the quantity of common nearness for the pair of elements in each other's neighborhood structure, which we refer to as "common-near-neighbor information". To analyze the common-near-neighbor information, a discriminative metric space is indispensable. Such metric space can be constructed and improved in consideration of intra-class compactness as well as inter-class separation.

Revolving around the philosophy of exploiting the common-near-neighbor information in a discriminative metric space, we have presented various approaches for the re-identification problem in both single-shot and multiple-shot directions. Single-shot and multiple-shot cases have different prerequisites and evaluation principles. For the single-shot case, we need to match each query point to each corpus point one after another, while for the multiple-shot case, we are able to simultaneously decide the matching between the whole set of query points and the whole set of corpus points. Single-shot and multiple-shot re-identification cases own distinct resources and challenges. It is these resources and challenges that necessitate the divide-and-solve strategy for exploiting the common-near-neighbor information. For instance, the point based distance seems more manageable, due to the relatively simple information for each identity in the single-shot case, than the set based distance that is affected by the local within-set variations in the multiple-shot case, while adequacy of the within-set distribution information enhances the reliability of the set based distance in the multiple-shot case instead.

In greater details, against the single-shot re-identification case, not only has the capability of point-level common-near-neighbor information for classifying the complex human image data been confirmed, but also the importance of the metric space where the point-level common-near-neighbor information is exploited has been studied, as elaborated in Chapter 2. Indeed, by making intra-class distances smaller than inter-class distances between sample points, a suitable learned Mahalanobis metric space can benefit modeling the point-level common-near-neighbor information into the dissimilarity. During the specific single-shot vs. single-shot re-identification procedure, to compensate for the insufficiency of the within-class distribution information, we further explore the strengthened metric space by two approaches for the subsequent point-level common-near-neighbor information exploitation, as expounded in Chapter 3. The first approach couples two complementary metric learning schemes together, and the second approach incorporates the constraints of point-level common-near-neighbor modeling dissimilarity comparison into the metric learning framework.

To resolve the multiple-shot re-identification problem, an intuitive idea is to exploit the set-level common-near-neighbor information, by treating each set of

points as a whole, without ignorance of the set based integrity and within-set distribution, as suggested in chapter 4. For effectively measuring this dissimilarity, the metric spaces are selectively constructed by two kinds of representative features. One feature addresses condensing the sets into the representative imaginary points based on covariance descriptors in the Riemannian space. The other feature emphasizes compacting human image sets to depress the outliers and intruders by collaborative representation in the Hilbert space. Though discriminative, these feature spaces are expensive in fact. Representative covariance descriptor requires adequate and typical within-set images. Discriminatory collaborative representation requires diverse and labeled dictionary data. We expect to avoid these expensive requirements for the single-set vs. single-set case, and designs a method based on mining the locality information, as detailed in Chapter 5. In this method, a capable set-to-set distance is crafted by encoding the local minority distribution information between paired sets, and upon this distance, an effective metric field space is constructed to accommodate the local variation of each set, before the set-level common-near-neighbor modeling dissimilarity is measured among all sets.

In addition to theoretic analysis, experimentation across several widely-used benchmark datasets for real-scenario human re-identification has demonstrated not only the philosophical value but also the methodological superiority in the thesis. Future work will cover, but not limited to, developing new models based on this philosophy and applying re-identification to improving cross-camera tracking.

Acknowledgements

This work at Graduate School of Informatics, Kyoto University would not have been possible without grateful helps of many people.

I would like to appreciate my supervisor Professor Michihiko Minoh for supervising this dissertation. Not only has he provided a sound scientific research environment, but also taught me the necessary capabilities and qualities to be a good researcher. I wish to express my gratitude to him for reading the manuscript and making a number of constructive suggestions. I also wish to sincerely thank my dissertation committee, Professor Tatsuya Kawahara and Professor Yuichi Nakamura for their precious comments and suggestions.

I would like to express my gratitude to Associate Professor Masayuki Mukunoki, Dr. Yang Wu, and Dr. Kawanishi Yasutomo. Associate Professor Masayuki Mukunoki advised and guided me during my research journey. Dr. Yang Wu and Dr. Kawanishi Yasutomo tutored and suggested me when I encountered difficulties. My graduation thesis could not complete without their insightful advices and thoughtful help. I am also grateful to all members in Professor Michihiko Minoh's laboratory.

Last but not least, heartfelt thanks go to my parents, Peifeng Li and Meiqin Zhu, for their support, understanding, and encouragement.

Contents

List of Figures

Chapter 1

Introduction

1.1 What Is Human Re-identification?

1.1.1 Issue Description

Matching people across camera views at different sites, known as human re-identification, is challenging and valuable for both academia and industry. This visual surveillance issue can help determine the reappearance of the person of interest who has been observed in the camera network in public places, such as the shopping mall, hospital, airport, and so forth [14]. However, the issue itself is difficult due to the real-world complexities, including pose variations, illumination changes, viewpoint alterations, occlusions, and possibly similar body shapes and clothing styles, as shown in Figure 1.1.

In this thesis, we divide the issue into two directions: single-shot re-identification and multiple-shot re-identification, according to different prerequisites and disparate evaluation principles. For the single-shot case, we need to match each query point to each corpus point one after another, while for the multiple-shot case, we are able to simultaneously match between the whole set of query points and the whole set of corpus points.

These two directions have different resources and challenges to address, which originates from their distinct preconditions and properties, and it is these differences that necessitates the divide-and-conquer strategy. Usually, it is unwise to apply single-shot orientated solutions to the multiple-shot case. Although a set of images can, seemingly, be split into a multitude of single images, this brute-force treatment will break the system integrity. Similarly, it is also impossible to apply multiple-shot orientated solutions to the single-shot case, because the single-shot

Figure 1.1: Some real scenarios for human re-identification.

case cannot satisfy the premise of these solutions.

In this thesis, human re-identification is defined as a matching problem, which aims at building correspondences between human images captured from different cameras, as shown in Figure 1.2. The people in one camera usually have their corresponding matches in the other camera, and this ground-truth assumes the nonoccurrence of turn-back cases. In the single-shot re-identification case, in the corpus side, there is one image per person, while in the query side, there are at least one images for each person in ground truth, though their labels are unknown. In evaluation, we need to match the query image one by one onto the corpus side. The multiple-shot re-identification case is similar to the single-shot situation, except that for each person, there are a set of images sharing the same unknown label. and thereby we can match the whole query set one by one onto the corpus side. Comparatively, set images provide more resources as well as more challenges for the multiple-shot case than the single-shot case.

More concretely, in the single-shot direction, we study both the general single-

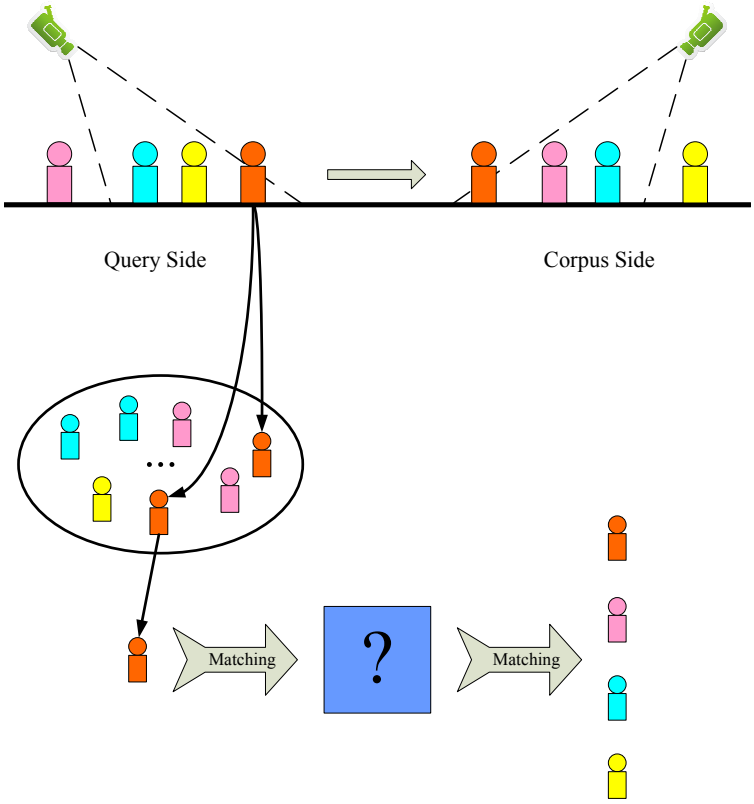


Figure 1.2: Illustration of matching human images between query and corpus sides.

shot and the specific single-shot vs. single-shot situations. In the multiple-shot direction, we study both the general multiple-shot and the specific single-set vs. single-set situations. In the direction of single-shot re-identification, for each person, there is only one image on the corpus side, but possibly more than one images on the query side. Single-shot vs. single-shot re-identification is a specific case in the single-shot re-identification, because it is known beforehand there is only one image for each person on the query side. In the direction of multiple-shot re-identification, for each person, there are a set of images on the corpus side, but possibly more than one image sets on the query side. Single-set vs. single-set re-identification is a specific case in the multiple-shot re-identification, because it

is known in advance there is only one image set for each person on the query side.

In both single-shot and multiple-shot directions, during matching, we let those other unlabeled but available images or image sets in the query side to participate in the matching process, because these images can provide additional useful information representative of the neighborhood for corpus elements (sample points or sample sets). Specifically, in the single-shot re-identification direction, we also can prepare some images that are collected beforehand to adapt the within-camera variation and between-camera transfer, in terms of training a metric to improve the relationship of intra- and inter-class distances. These training data are from the same camera pair, but belonging to different identities than those in the testing stage. Such aspect clearly distinguishes the metric learning for re-identification from the conventional classification training-testing scheme, in which training and testing images come from the same identity in ground truth.

1.2 Background and Related Work

1.2.1 Background

Solutions for the issue of human re-identification chiefly involves two categories: feature representation and dissimilarity measurement. Feature representation projects the human image data into a metric space for distance measurement, and dissimilarity measurement discriminates the human image data based on feature representation. Both of them are important for the issue

Many publications mushroom every year, and we sketch some of the representative ones hereby. For feature representation, ELF (Ensemble of Localized Features) [16] uses the the ensemble of localized features for viewpoint invariant pedestrian recognition. HPE (Histogram Plus Epitome) [29] focuses on general chromatic content and recurrent local patches to extract the complementary global and local features of human appearance. PCS (Part-based Clothing Segmentation) [45] separately segments a person into upper and lower clothing regions, taking into account the person body pose, for re-identification. SDALF (Symmetry-Driven Accumulation of Local Features) [11] accumulates three types of local features, namely weighted color histogram, maximally stable color regions, and recurrent high-structured patches, regarding the symmetric and asymmetric property of body structure. GCC (Global Color Context) [4] combines the spatial distributions of self-similarities with respect to color words to characterize the appearance of pedestrians. MCTCS (Matching Compositional Template with Cluster Sampling)

[52] constructs a simple yet expressive template from a few reference images of a certain individual, which represents the body as an articulated assembly of compositional and alternative parts, and then presents an effective matching algorithm within a candidacy graph. CSPR (Color-Spatial Person Re-identification) [38] involves segmentation of silhouettes into meaningful regions, which are close to human visual categorization of colorful clothes, before extracting the spatial features. MRCG (Mean Riemannian Covariance Grid) [40] extracts covariance descriptors and discriminants to capture the textual information within grid based local patches of a human body. TPCR (Third-Party Collaborative Representation) [49] resorts to the third-party data as the dictionary, and concatenates the reconstructed coefficients from collaborative representation for each sample into a kind of capable relative feature based on this dictionary. RBPR (Reference-Based Person Re-Identification) [1] generates the reference descriptors based on projected features in a common subspace where their correlation is maximized using Regularized Canonical Correlation Analysis. LDFV (Local Descriptors encoded by Fisher Vectors) [31] encodes local descriptors into Fisher Vectors before being pooled to produce a global representation of the image. BiCov (Biologically Inspired Features and Covariance Descriptors) [32] relies on the combination of biologically inspired features and covariance descriptors. CIPR (Color Invariants for Person Reidentification) [25] uses shape context descriptors to represent the intradistribution structure for re-identification. SCN (Semantic Color Names) [24] applies semantic color names to describe a person image, and computes the probability distribution upon those basic color terms as image descriptors, which are then combined with other widely-used features. USL (Unsupervised Saliency Learning) exploits the human salience information in an unsupervised manner, and incorporate this information into patch matching [56].

For dissimilarity measurement, LMNN (Large Margin Nearest Neighbor) [8, 46] learns a metric by maximizing the margin to distinguish the human image samples from different classes. SML (Smooth Metric Learning) [44] replaces the hinge loss function with the logistic loss function to improve LMNN, and thereafter designs a stochastic sampling scheme to accelerate the optimization. RDC (Relative Distance Comparison) [59, 58] maximizes the likelihood that a pair of true identity matches will have a smaller distance than a wrong match pair for more tolerance of appearance changes and less susceptibility to model over-fitting. RankSVM [36] translates the re-identification problem from one of absolute scoring to one of relative ranking, and learns a metric space in which the highest ranking pair, rather than the closest by direct distance measure, is treated as a potential match. JSKML (Jensen-Shannon Kernel Metric Learning) [20]

learns nonlinear distance metrics between color histograms to avoid the naive color histogram matching during re-identification. RPLM (Relaxed Pairwise Learned Metric) [18] introduces a more efficient but still effective metric learning approach by relaxing the constraints to reduce the computational complexity for practical re-identification in large scale scenarios. IML (Impostor-based Metric Learning) [17] targets the computational cost problem by searching for a linear projection that keeps similar pairs together and pushes impostors away. OMRR (Optimizing Mean Reciprocal Ranking) [51] optimizes the list-wise ranking through design of the loss function of MLR (Metric Learning to Rank) [33] to fit the practical re-identification performance expectation. RSML (Robust Structural Metric Learning) [53] aims to learn an optimal distance metric by applying the loss function at the level of rankings, and realizes robustness to noisy information of the extracted features by promoting both input and output sparsity. SBDR (Set Based Discriminative Ranking [50] iteratively constructs convex hulls for set-to-set distance measurements and optimizes the metric for ranking the intra-class human image set pair before the inter-class ones relying on these measurements. KISSME (Keep It Simple and Straightforward MEtric) [23] adopts an effective and efficient strategy to learn the distance metric based on equivalence constraints from a statistical inference perspective. RSKISS (Regularized Smoothing KISS metric learning) [43] integrates smoothing and regularization techniques to enlarge the underestimated small eigenvalues and reduce the overestimated large eigenvalues of the estimated covariance matrix in an effective way. PCCA (Pairwise Constrained Component Analysis) [34] learns a projection into a low-dimensional space where the distance between pairs of data points respects the desired constraints, which exhibits good generalization properties in presence of high dimensional data for re-identification. TML (Transferred Metric Learning) [27] learns an adaptive metric for a specific human image candidate set under the framework of transfer learning, by selecting and re-weighting the human image samples in the training set for the given query sample and its candidate set. SMFL (Semi-supervised Multi-Feature Learning) [12] presents a multi-class learning approach to learn how to fuse arbitrary state-of-the-art set of features, no matter their number or dimensionality. LAFTaV (Locally Aligned Feature Transforms across Views) [26] assigns images to different local experts according to the similarity of their cross-view transforms before projecting them into a common feature space and matching them with a locally learned discriminative metric. MR (Manifold Ranking) [30] propagates the query information along the unlabeled person data manifold in an unsupervised way to obtain the distance score for ranking. CHISD (Convex Image Image Set Distances) [5] measures the set-to-set distance between the convex hulls of image

sets to reduce the negative influence of outliers. SANP (Sparse Approximated Nearest Points) [19] extended the model of CHISD by enforcing the sparsity of samples used for point generation via affine combination. CRNP (Collaboratively Regularized Nearest Points) [48] integrates the simplicity of regularized nearest points method on finding the set-to-set distance and the capability from collaborative representation on human image set based recognition. MC (Matrix completion) [28] performs between-camera information transfer by constructing the corresponding feature to be matched in the corpus camera space from the feature observed in the query camera space based on matrix completion. LDC (Local Distance Comparison) [54] formulates the re-identification problem as a local distance comparison problem, and introduces an energy based loss function that measures the similarity between appearance instances by calculating the distance between corresponding subsets in the feature space.

1.2.2 Related Work

There are several related work in both feature representation and dissimilarity measurement paradigms. Here, related work only indicates the methods that are directly adopted in the thesis, but not the competitors and analogues for each proposed method. Dissimilarity measurement is closely related to the contribution of this thesis. Although feature representation is loosely related to the contribution of this thesis, yet they serve as an indispensable and important platform for the proposed dissimilarity measurement.

Single-shot Image Feature

For each single-shot human image, in the thesis, we adopt some representative features for mining human appearance information from different perspectives. These features are Weighted HSV color histogram (WHSV) [11, 51], Dense Sampled Color Histogram (DSCH) [8, 51], Schmid-Filter-Bank (SFB), Gabor-Transform (GT) [16, 59, 50], and Third-Party Collaborative Representation (TPCR) [49].

WHSV pools the pixel based color information regarding the symmetric and asymmetric property of human body structure. It can handle viewpoint change and pose variation. DSCH can handle illumination variation and occlusion in terms of its cell-based statistical local color description and a global dense sampling of these cells. SFBcan capture the texture information. It is rotationally invariant and has 13 isotropic, which are robust to viewpoint and pose variations. GT uses 8 filters with different frequencies and orientations. It has been found

quite capable for texture representation and discrimination. TPCR is a kind of relative feature. It resorts to the third-party data as the dictionary, and then concatenates the reconstructed coefficients from collaborative representation for each sample based on this dictionary, with summing the coefficients that correspond to axioms of the same class in the dictionary. The effectiveness of TPCR comes from the compression of within-class information by intra-class sum and the usage of diverse third-party data in collaborative representation.

Multiple-shot Image Feature

For the multiple-shot human images, in the thesis, we adopt Mean Riemannian Covariance Grid (MRCG) [40], which is one of the most poweful feature for representing human image sets.

MRCG obtains highly discriminative human signature by condensing information from multiple images. It uses the mean of Riemannian Covariance Grid (RCG) to blend the human appearance information within the densely-tiled grid of the image set. The matric signature keeps the feature distribution information with regard to spatio-temporal change of an appearance. Because this signature condenses the human image sets into representative imaginary points, it provides convenience for bridging the solution gap between single-shot and multiple-shot re-identification.

Set-to-set Distance

For set-to-set distance, in the thesis, we adopt Minimum Point-wise Distance (MPD) [11] and Convex Hull Image Set Distance (CHISD) [5]. They are widely used for measuring the distance between image set for discrimination.

MPD is a kind of baseline set-to-set distance in this thesis. It measures the minimum point-wise distance between paired sets. CHISD is a kind of set-to-set distance. It measures the distance between the convex hulls constructed for paired sets, in order to depress the negative impact from outliers.

Metric Learning Model

For metric learning approaches, in the thesis, we adopt Metric Learning to Rank (MLR) [33] and Maximally Collapsing Metric Learning [13]. They have been proven quite effective for image ranking and classification, respectively.

MLR optimizes the Mahalanobis metric for ranking in a structural SVM framework under the constraints that intra-class distances should be smaller than inter-class distances. MLR has been proved effective for the re-identification issue [51]. Human image data suffers from real-world complexities and multi-class imbalance, and the recognition rate on Rank-1 is usually unsatisfactory. Because ranking can concern several top candidates,it avoids the risk of classification that only depends on Rank-1. Moreover, OMRR emphasizes designing its loss function of Mean Reciprocal Rank (MRR). Here, Reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct match, and MRR is the average of such reciprocal ranks of results over the whole query set. To a degree, this loss function coincides with practical performance expectations. Since only the rank of the first correct match is taken into account, the ranks of both other correct matches and any incorrect matches are arbitrary. Thus, there will be multiple ranking instantiations for a given ground truth. In practice, MRR is quite reasonable, and there seems to be no significantly better options. In the thesis, MLR plays an important role in learning either global or local Mahalanobis metric space for discriminating human image data.

MCML is a metric learning method with the objective function based on Kullback-Leibler divergence, which aims at closing sample within the same class as near as possible while distancing samples from different classes as far as possible, Although MCML is seemingly similar to MLR, while they have different essential meanings. MCML relies on the dipolar balance of intra-class distances to be zeros and inter-class distances to be infinities. but MLR focuses on the relative comparison between intra-class distances and inter-class distances. Honestly, classification oriented MCML is not effective for re-identification, whereas, it can help de-noise the feature space for the better performance of MLR. This property will be addressed in the thesis.

1.3 Philosophy and Organization

1.3.1 Philosophy

Due to the heuristic of handcrafting process, even in a space mapped by representative features, samples still have intra-class variations and large inter-class differences. To cope with this, this thesis mainly focuses on designing the dissimilarity measurement in the feature space, which belongs to second solution paradigm.

Since human image data in a feature space can be treated as elements, their

correspondences can be determined by the distance scores; therefore, the suitable dissimilarity measurement becomes fairly important. So far, literature has witnessed plentiful distances that directly and solely concentrate on the pair of elements in a pre-defined metric space. Although, this kind of classical distance is discriminative sometimes, yet it tends to be incapacitated by the large within-class variations and small between-class differences for the complex real-scenario human image data in general. Accordingly, to address it, we creatively propose to quantify the local neighborhood structure of the pair of elements in each other's neighborhood structures into the dissimilarity, as a new trial to enrich the conventional distance notion by considering the neighborhood information. Here, neighborhood structure is defined as the layout relationship between the concerned element and its neighbors in the given metric space. The property of this dissimilarity can be comprehended by the insight of measuring the quantity of common nearness for the pair of elements in each other's neighborhood structure, which we refer to as "common-near-neighbor information". To exploit the common-near-neighbor information, a reliable and discriminative metric space is indispensable. This space can be constructed by feature selection and then improved by metric learning in consideration of intra-class compactness as well as well inter-class separation.

The discriminability of this dissimilarity comes from delivering the effectiveness of the well-distributed elements to the badly-distributed elements in their neighborhood structures. To carry out this measurement, we use those unlabeled available elements from both query and corpus sides to fill the space, as the possible neighbors of the to-be-measured element pair. When one element resides far from the other element in the same class, but near to another element from a different class, these filled elements can provides additional useful information, to pull elements within the same class stay close together whilst push those from different classes far apart.

For comprehension, we illustrate this idea using a single-shot vs. single-shot re-identification example. Usually, due to large intra-class variations and small inter-class differences, traditional distance, like the l_2-norm, will lose effect to some extent when building the correspondences between the pair of query and corpus elements, as illustrated in Figure 1.3. In this figure, elements' classes can be distinguished by colors and shapes. When c is selected as a query, b and d are assigned in the corpus side. cd are badly-distributed, while ab and ef are well-distributed. Clearly, by Euclidean distance, the element b will rank before the element d with regard to the element c.

To address this problem, we borrow the effectiveness from those well-distributed

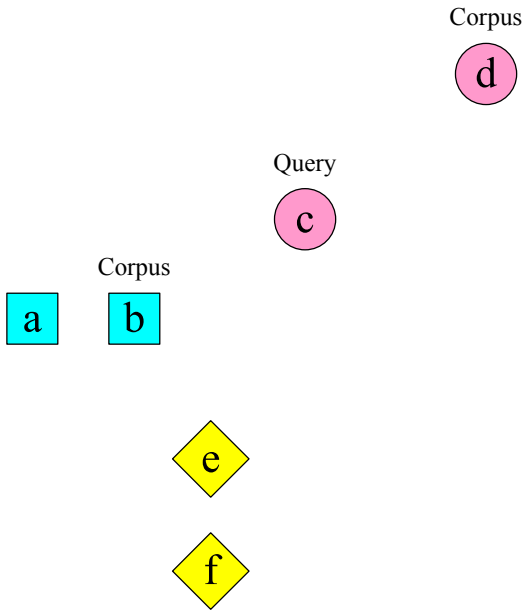


Figure 1.3: An example to show the problem of traditional distance.

elements to improve those badly-distributed elements, and this information is modeled as a novel dissimilarity, Common-Near-Neighbor Modeling (CNNM), which will be instantiated as Point-level and Set-level CNNM in Chapter 2 and Chapter 4 for single-shot and multiple-shot re-identification, respectively.

For each element, the neighborhood structure can be compressed into the rank order list to describe the farness and nearness of the neighbors. In this example, b-list is $baecfd$, c-list is $cbdeaf$, and d-list is $dcbeaf$. If rank orders are defined as 012345, when CNNM counts these rank orders for the paired elements in each other's list, the CNNM dissimilarities will be D(b,c)=1+3=4; D(c,d)=1+2=3. Thus, d will rank before b with regard to c by this dissimilarity measure.

Analytically, let us check the rank order lists for b, c, and d. When comparing b-list and c-list, we can find that, because a and b are close, a rank before c in b-list. This leads up to the increase of the inter-class distance $D(b, c)$. Moreover, when comparing c-list and d-list, we can find that, because e and f are far away from c, they stably reside in the tails of c-list and d-list. This contributes to the decrease of intra-class distance $D(d, c)$.

Actually, for c and d, CNNM considers how c stays in the neighborhood structure of d, and how d resides in the neighborhood structure of c, simultaneously. If

c is the near neighbor to d, and d is the near neighbor to c, we note this situation c and d are "common-near" neighbors to each other. This common-near-neighbor information can be measured by a quantity, which gives birth to the CNNM dissimilarity.

There are some noteworthy aspects on this dissimilarity as well. As the precondition, there needs to exist some well-distributed elements in the feature space. This guarantees the precondition for effective CNNM dissimilarity measure. Actually, this requirement is easy to satisfy in a carefully designed or selected feature space. As the limitation, though the well-distributed elements can help those badly-distributed elements utilizing CNNM, the amount of and the power of the well-distributed elements in the feature space are limited, thus only those badly-distributed elements whose situation is not rather serious can be ameliorated. If the pair of elements to be measured in the same class are quite farther apart than from different classes, CNNM dissimilarity tends to lose efficacy. CNNM encodes the neighborhood structure information, which is based on the direct distance measure on the low level, so the performance of CNNM is more or less influenced by the underlying distances. To guarantee the effectiveness of CNNM, it is indispensable to project and improve a good metric space therefore.

Undoubtedly, concrete analysis should be made according to concrete circumstances. On the whole, exploiting common-near-neighbor information in discriminative spaces makes up the backbone of the solutions in this thesis.

Point based and set based distances are different in the single-shot and multiple-shot cases. The point based distance seems more stable and manageable, due to the relatively simple information for each identity in the single-shot case, than the set based distance that is vulnerable to the local within-set variations in the multiple-shot case, while availability of the within-set distribution information enhances the confidence of the set based distance in the multiple-shot case instead of the point based distance in the single-shot case.

1.3.2 Organization

Based on the philosophy of exploiting common-near-neighbor information in discriminative spaces, stories are expanded and evolved in the single-shot and multiple-shot directions, respectively.

In the main body, this thesis progressively develops into four parts of six methods. Four parts include Point-level Common-Near-Neighbor Information (PCNNI) and Strengthened Metric Space (SMS) in the single-shot re-identification direction; Set-level Common-Near-Neighbor Information (SCNNI), and Locality

Based Discriminative Measure (LBDM) in the multiple-shot re-identification direction. Six methods include Point-level Common-Near-Neighbor Analysis (PCNNA), Coupled Metric Learning (CML), Point-level Common-Near-Neighbor Metric Learning (PCNNML), Riemannian Set-level Common-Near-Neighbor Modeling (RSCNNM), Bi-level Relative Information Analysis (BRIA), and Locality Based Discriminative Measure (LBDM).

The organization of the thesis and the description of each chapter are as below.

Chapter 1 introduces the philosophy and outlines the main content of the whole thesis.

Chapter 2 unwarps the single-shot re-identification story. This chapter proposes the PCNNA method that exploits PCNNI in a discriminative metric space. More concretely, in PCNNA, Point-level Common-Near-Neighbor Modeling (PCNNM) processes the local neighborhood structure information of the paired sample points in each other's neighborhood structure into the dissimilarity in the learned Mahalanobis metric space. By delivering the effectiveness of the well-distributed element pairs to the badly-distributed element pairs, PCNNM can be capable for the complex real-world human image data of large intra-class variations and small inter-class differences, even in the challenging single-shot vs. single-shot case.

For the specific single-shot vs. single-shot case, Chapter 3 suggests using the strengthened metric space to compensate for the deficiency of within-class distribution information for exploiting PCNNI. Two approaches have been tentatively proposed: CML and PCNNML. CML couples two optimization frameworks on account of their complementarity in the relationship between intra-class and inter-class distances: one addresses the dipolar balance, while the other emphasizes the relative comparison. Notwithstanding improving the discriminability of the metric space, CML does not seem to largely boost the performance of PCNNM. To address this, we consider the importance of consistency between metric space projecting and PCNNM dissimilarity measuring, and propose the PCNNML method. This method incorporates the constraints of intra-class and inter-class PCNNM dissimilarity comparison into the metric learning framework, and thus ultimately brings out a substantial performance leap compared with related competitive approaches, and whereupon ends the story of single-shot re-identification.

Chapter 4 unfolds the multiple-shot re-identification story. This chapter suggests the importance of SCNNI, and extends PCNNM to SCNNM in some discriminatve spaces. These spaces are selectively constructed by two types of capable features: the matric descriptor of MRCG and the vectorial signature of T-PCR. These two features are culled due to their compressive abilities. MRCG

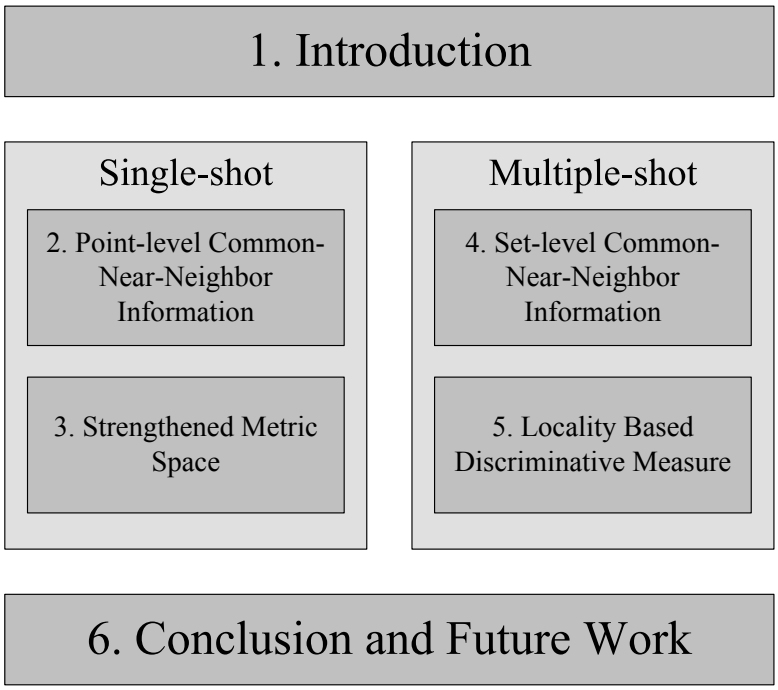


Figure 1.4: Architecture of the thesis.

can condense human image sets into imaginary points in the Riemannian space, thus convenient for extending the power of the single-shot case based PCNNM to the multiple-shot case based SCNNM. Measuring SCNNM in this metric space is called “RSCNNM” [9]. TPCR can compact human image sets to depress the negative impact of outliers and intruders in the Hilbert space, and this feature will complement and be measured by SCNNM from the relative information perspective. Measuring SCNNM in this metric space is called “BRIA”.

Despite effectiveness of RSCNNM and BRIA, they are expensive in fact. RSCNNM requires adequate and typical within-set images to guarantee the representative condensation for MRCG, and BRIA requires diverse and labeled dictionary data to extract the representative signature for TPCR. We try to avoid these expensive requirements in the single-set vs. single-set re-identification case. Chapter 5 designs the LBDM method by means of mining the locality information. In this method, a new set-to-set distance is crafted by encoding the local minority distribution information between paired sets, and upon this distance, an effective metric field space is constructed to accommodate the local variation of each set,

before the set-level common-near-neighbor modeling dissimilarity is measured among all sets. LBDM has leaded to the anticipated performance improvement which inspiringly outperforms the state-of-the-arts, and thereby completes the multiple-shot re-identification story.

Chapter 6 concludes the whole thesis and looks forward to the future work.

The architecture of the thesis is visualized in Figure 1.4.

More generally, all the methods in this thesis can be boiled down to three steps: feature space projecting, metric space improving, and common-near-neighbor information exploiting. And here a table to summarize all the methods in this thesis, as shown Table 1.1.

Table 1.1: Method summary.

	Chapter 2	Chapter 3
FSP	(WHSV+DSCH, ED)	(WHSV+DSCH, ED)
MSI	(WHSV+DSCH, MD improved by MLR)	(WHSV+DSCH, MD improved by CML, PCNNML)
CNNIE	PCNNM	PCNNM
	Chapter 4	Chapter 5
FSP	(MRCG, RD) ; (DSCH+SFB+GF, CHISD)	(DSCH+SFB+GF, MAD)
MSI	– ; (TPCR, CHISD)	(DSCH+SFB+GF, MAD improved by LMF)
CNNIE	SCNNM	SCNNM

The abbreviations are as below: Feature Space Projecting–FSP, Metric Space Improving–MSI, Common-Near-Neighbor Information Exploiting–CNNIE, Euclidean Distance–ED, Mahalanobis Distance–MD, Riemannian Distance–RD, and Local Metric Field–LMF. In Table 1.1, the metric space is written by (M, d), where M denote the feature, and d denote the distance.

1.4 Dataset Description

Experimental demonstration on each proposed method will have been implemented on part of or all the widely-used benchmark datasets including VIPeR [15], ETHZ1, ETHZ2, ETHZ3 [39], i-LIDS [57], i-LIDS-MA, i-LIDS-AA [41], and CAVIAR4REID [6], as exemplarized in Figure 1.5.

The VIPeR dataset consists of 1264 images for 632 unique pedestrians. Each pedestrian image pair has been taken from arbitrary viewpoints under varying illumination conditions. Complicated variations of viewpoint, illumination, and pose make VIPeR one of the most challenging datasets for single-shot vs. single-shot human re-identification [15].

Figure 1.5: Exemplars from datasets VIPeR, ETHZ1,2,3, i-LIDS, i-LIDS-MA, i-LIDS-AA, and CAVIAR4REID, showing the real-world complexities for human re-identification.

The ETHZ dataset is composed of three video sequences of crowded street scenes captured by two moving cameras mounted on a carriage. We utilize three subsets extracted by Schwartz and Davis for human re-identification [39]. ETHZ has smaller variations of pose and viewpoint, yet more occlusions than VIPeR. SEQ1, denoted by ETHZ1, has 83 pedestrians within 4857 images; SEQ2, denoted by ETHZ2, has 35 pedestrians within 1936 images; SEQ3, denoted by ETHZ3, has 28 pedestrians within 1762 images.

The i-LIDS MCTS database is a publicly available video dataset captured at a busy airport arrival hall using a multi-camera CCTV network. From these videos, the i-LIDS dataset, which was extracted by Zheng et al., consists of 479 images for 119 individuals [57]. For each identity, there are 2 to 8 images taken from

non-overlapping cameras. This i-LIDS dataset suffers from more severe pose variations and occlusion than VIPeR and ETHZ1,2,3. We also use the datasets i-LIDS-MA and i-LIDS-AA extracted by Bak et al. [40]. i-LIDS-MA [41] contains 40 individuals extracted from two cameras. Each of them has 46 frames from both cameras annotated manually. i-LIDS-AA [41] is made up of 100 individuals seen from both cameras, and for each individual, there are 21 to 243 images automatically obtained by the detector and tracker, which augments the difficulty for this dataset.

The CAVIAR4REID dataset [6] is manually selected from the less-controlled recorded video of shopping center scenarios within two different viewpoints. For this dataset, there are 72 pedestrians of 10 or 20 image samples. 50 of these pedestrians have two camera views while the remaining 22 have only one camera view. These images include people walking alone, meeting with others, window shopping, entering and exiting shops.

For improved comprehension and comparison, properties of these datasets are displayed in Table 1.2. The abbreviations denote the following: NS–Number of Samples, NP–Number of Persons, NSPP–Number of Samples Per Person, Seq?–Sequential or not, and WCV–Within-Class Variations, where the abbreviations for WCV are: C–Camera parameters, V–Viewpoint, I–Illumination, P–Pose, O–Occlusion, B–Background, and L–Localization.

Table 1.2: Display of dataset properties.

Dataset	NS	NP	NSPP	Seq?	WCV
VIPeR	1264	632	1×2	No	CVIPOBL
ETHZ1	4857	83	7 to 226	Yes	POB
ETHZ2	1961	35	6 to 206	Yes	POB
ETHZ3	1762	28	5 to 356	Yes	POB
i-LIDS	476	119	2 to 8	Partly	CVIPOB
i-LIDS-MA	3680	40	46×2	Yes	CVIPOB
i-LIDS-AA	10329	100	21 to 243	Yes	CVIPOBL
CAVIAR4REID	1220	72	10 or 20	Partly	CVIPOB

Chapter 2

Point-level Common-Near-Neighbor Information

2.1 Introduction

We start to study the single-shot human re-identification in this chapter. Single-shot re-identification is important, fundamental, and sophisticated in visual surveillance. To build an accurate correspondence between human image data requires a suitable dissimilarity measurement. Unfortunately, these real-world image data have large intra-class variations and small inter-class differences, which prevents the traditional dissimilarity measurement from achieving the satisfactory re-identification performance.

Since in a discriminative space, there exist some well-distributed sample points in the same class, which tend to be more tightly clustered and separated from the sample points of different classes, this chapter will propose a novel method, "Point-level Common-Near-Neighbor Analysis (PCNNA)", to deliver the effectiveness of these well-distributed sample points to the badly-distributed ones. In PCNNA, "Point-level Common-Near-Neighbor Modeling (PCNNM)" is designed to exploit the point-level common-near-neighbor information in a metric space learned by OMRR.

2.2 Point-level Common-Near-Neighbor Analysis

2.2.1 Point-level Common-Near-Neighbor Modeling

Human re-identification is recast to the problem of discriminating human image data based on measuring the distances among sample points in feature space. When the feature space has large intra-class variations and small inter-class differences, distributed to-be-measured sample points may make traditional point-wise distance lose effect. To handle this, we reconsider the distance measure from the novel neighborhood information perspective, and propose a neighborhood-wise distance to measure how the local neighborhood structure of one sample point stay in the neighborhood structure of the other sample point. Here, neighborhood structure is defined as the spatial layout relationship between the concerned sample point and its neighbors. Considering neighborhood structure provides a new trial to overcome the vulnerability of traditional point-wise distance, because neighborhood structure provides more additional useful information during measurement.

Neighborhood structure can be compressed by rank order lists in terms of farness and nearness. By using the rank order lists, we can quantify the local neighborhood structure of the paired sample points in each other's neighborhood structures in some metric space, and deliver the effectiveness of those well-distributed sample points to the badly-distributed ones during quantification. To enable the discriminative local neighborhood structure quantification for a feature space having the class structure, we recommend considering the local neighborhood of a moderate fixed size for each paired sample points symmetrically instead of the sole sample point pair, if possible. The moderate and fixed neighborhood can incorporate the within-class sample distribution information for each sample point during measurement, and this can help counteract the stochastic deviation of the sample points from the image feature space of variations. These novel physical meanings can be comprehended by the insight of measuring the quantity of common nearness for the pair of sample points in each other's neighborhood structure.

Technically, to ensure a symmetric neighborhood of the moderate fixed size in the rank order lists for the pair of samples, we suggest a "Fixed-number" n instead of a flexible number, as shown in Figure 2.1, because the top fixed-number samples in the rank order lists tend to contain more within-class distribution information than those ranked behind. We call this "Symmetric dissimilarity", as

given by:

$$D^{\text{PCNNM}}_{\text{Symmetric}}(a,b) = D^{\text{Fixed-number}}_{a}(b) + D^{\text{Fixed-number}}_{b}(a), \tag{2.1}$$

where

$$D^{\text{Fixed-number}}_{a}(b) = \sum_{i=0}^{n-1} O_b(f_a(i)). \tag{2.2}$$

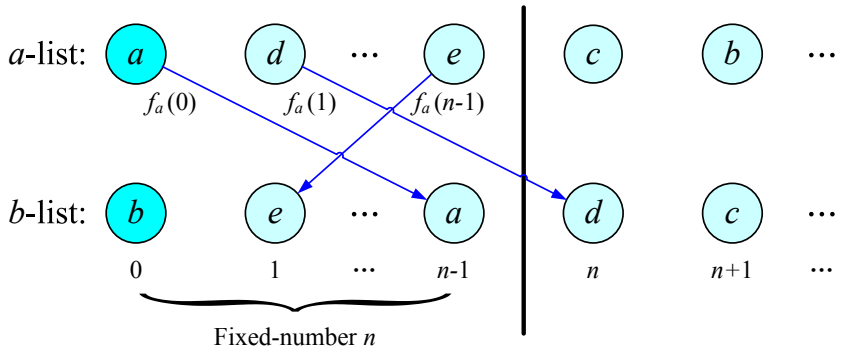


Figure 2.1: Illustration for "Symmetric dissimilarity" of PCNNM. $D^{\text{Fixed-number}}_{a}(b)$ is calculated from the 0^{th} to the $(n-1)^{th}$ nearest neighbor in a-list, and $D^{\text{Fixed-number}}_{b}(a)$ is calculated in a similar way.

Since the "Fixed-number" value can be crucial to the performance of "Symmetric dissimilarity", it deserves further discussion. If the neighborhood is too large, intruder samples from different classes may disturb the encoding of representative locality for the sample. In this case, the "Symmetric dissimilarity" of samples from different classes will have tendency to be lower than that of samples from the same class. If, on the other hand, the neighborhood is too small, local information of the outlier samples may not be sufficient and capable enough to counteract the stochastic deviations. In this case, the "Symmetric dissimilarity" of samples from the same class will have tendency to be higher than that of samples from different classes. Obviously, both cases will negatively affect ranking/classifying performance, and should be avoided. For a properly discriminative "Symmetric dissimilarity", we suggest using a "Fixed-number" n set to approximately $N/2$ of tradeoffs (where N is the approximate average sample number per person known beforehand, or estimated from dividing the total image number by

the identity number) to avoid the cases that the neighborhood is either too large or too small. Note that this suggestion is based on observation and could not be demonstrated through mathematical proof.

In particular, for the single-shot vs. single-shot re-identification problem, there are only two samples for each person and the recommended "Fixed-number" will be "$n = 1$". Then, each sample has to consider the situation of itself within the neighborhood structure of the other sample. In this case, the asymmetric ranking problem is rather obvious. Typically, a given pair of samples will not yield the same rank order for each other in their own rank order lists. Since it would be heuristic and unfair to judge rank order by randomly selecting one of these ranks, or simply averaging the two, we turn to the added discriminative ability of "Asymmetric dissimilarity", which is disregarded in Zhu et al's Rank-Order distance. "Asymmetric dissimilarity", $D^{\text{PCNNM}}_{\text{Asymmetric}}(a, b)$, is given by:

$$D^{\text{PCNNM}}_{\text{Asymmetric}}(a, b) \quad = \quad min(O_a(b), O_b(a)). \tag{2.3}$$

In consideration of the complementarity between "Symmetric dissimilarity" and "Asymmetric dissimilarity", we present a new dissimilarity modeling, PCNNM, as given by:

$$D^{\text{PCNNM}}(a, b) \quad = \quad D^{\text{PCNNM}}_{\text{Symmetric}}(a, b) + 2\lambda n D^{\text{PCNNM}}_{\text{Asymmetric}}(a, b), \tag{2.4}$$

where λ is a trade-off parameter to balance "Symmetric dissimilarity" and "Asymmetric dissimilarity". This makes the model more compatible. Whereas "Symmetric dissimilarity" uses the "Fixed-number" of the rank orders in consideration of symmetry, "Asymmetric dissimilarity" uses only one rank order. Thus, it is reasonable to reformulate these two dissimilarities by doubling n ($2n$) and thereby unifying dimensions.

2.2.2 Comparison between PCNNM and Its Analogue

There exists one similar method, Zhu et al.'s Rank-Order distance, originally designed for clustering during face tagging [60]. Although this method is capable of solving the samples' non-uniform distribution problem by rank order quantization, which also hurts the re-identification performance possibly, yet it may fail to discriminate the human image data with large intra-class variations and small inter-class differences, as the general case in the real world. Note that, in Zhu et al.'s work, clustering addresses the absolute threshold for filtering the distances, while in this paper re-identification emphasizes the relative comparison between

the inter- and inter-class distances. Clustering and re-identification are different in nature, so we should not impose Rank-Order distance on re-identification.

Zhu et al.'s Rank-Order distance is given by:

$$D^{\text{Rank-Order}}(a,b) = \frac{D_a(b) + D_b(a)}{min(O_a(b), O_b(a)),} \tag{2.5}$$

where

$$D_a(b) = \sum_{i=0}^{O_a(b)} O_b(f_a(i)). \tag{2.6}$$

In Equations 2.5 and 2.6, a and b are a pair of samples to be measured. $f_a(i)$ returns the i^{th} sample in a-list, and $O_a(b)$ is the rank order of b in a-list. $O_b(f_a(i))$ is the rank order of $f_a(i)$ in b-list. $D_a(b)$ is calculated by summing the rank orders of the top several samples of a-list in b-list, as shown in Figure 2.2. $D_b(a)$ is calculated in a similar way. Here, rank order lists are generated by sorting Euclidean distances between the measured samples and all the other samples.

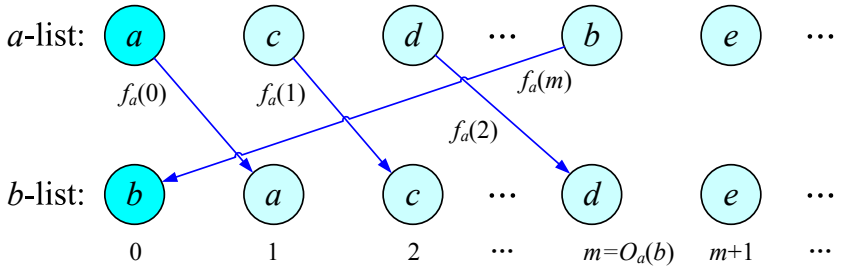


Figure 2.2: $D_a(b)$ is calculated from the 0^{th} sample to b in a-list, while $D_b(a)$ is calculated from the 0^{th} sample to a in b-list. Typically, $D_a(b) \neq D_b(a)$. Note that, here, for conciseness and comprehensibility, only $O_b(f_a(i))$ are visualized by arrows in this figure. Each arrow head points the rank order $O_b(f_a(i))$ in b-list for the sample $f_a(i)$ in a-list that connects the arrow tail.

Despite robustness to clustering, this dissimilarity doesn't consider the potential class information, so it will be non-discriminative for the sample space of the class structure, like in the re-identification situation. From the formulation level, the weakness comes from the flexible neighborhood size and the inappropriate denominator. Evidently, Equation 2.5 is sensitive to $O_a(b)$ and $O_b(a)$, which reflect the position relationship between paired samples a and b. If $O_a(b)$ and $O_b(a)$

are too big or small, the local neighborhood structure of each sample cannot sufficiently represent the class information, thus the comparison will have no discriminatory power. If $O_a(b)$ and $O_b(a)$ are quite different, the local neighborhood structures for each sample pair become incomparable, and worse still, the biased denominator will deteriorate the normalization.

In addition to analysis, we further use some examples to show the advantage of PCNNM.

We first illustrate the ability of "Symmetric dissimilarity". In general, real-world human image data of the limited sample size have large intra-class variations and small inter-class differences, which can be simplified in Figure 2.3. Let d be the query sample. a and e are assigned as the corpus samples. d and a are in the same class, but e is from the different class. By Euclidean distance, e will rank before a with regard to d.

Next, we consider Rank-Order distance and "Symmetric dissimilarity". Each sample ranks itself first, so rank order lists for a, d, and e are $abcdegfh$, $dcbegfah$, and $egfdcbha$, respectively, with integer rank orders from 0 to 7. Zhu et al.'s Rank-Order distance fails because of the large intra-class variations and small inter-class differences: $D(a,d) = [(6{+}2{+}1{+}0){+}(3{+}2{+}1{+}4{+}5{+}6{+}0)]/3 = 10$ and $D(d,e) = [(3{+}4{+}5{+}0){+}(3{+}4{+}5{+}0)]/3 = 8$. However, our "Symmetric dissimilarity" succeeds by exploiting the common-near-neighbor information in the moderate fixed neighborhood: $D(a,d) = (6{+}2){+}(3{+}2) = 13$ and $D(d,e) = (3+4)+(3+4) = 14$.

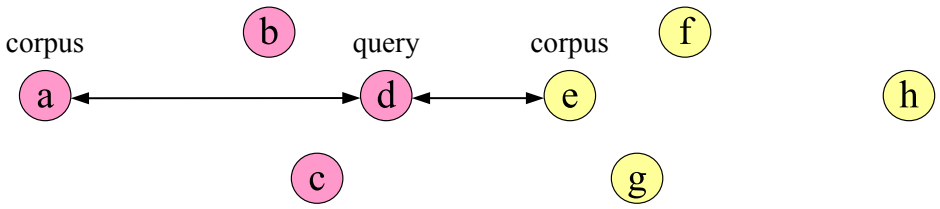


Figure 2.3: An example of comparison between "Symmetric dissimilarity" and Rank-Order distance. Samples' classes can be distinguished by colors.

This example can also be understood by Figure 2.4. In the feature space, some badly-distributed sample points (c_1, c_2, and q) are easy to be re-identified incorrectly. Suppose c_1 and c_2 are the corpus images from different classes and q is a query image. Using Euclidean distance, c_2 will rank ahead of c_1 regarding to q. By delivering the effectiveness of well-distributed points to those badly-distributed ones, "Symmetric dissimilarity" will pull c_1 and q closer together, and

push c_2 and q farther apart, to make intra-class distances smaller than inter-class distances.

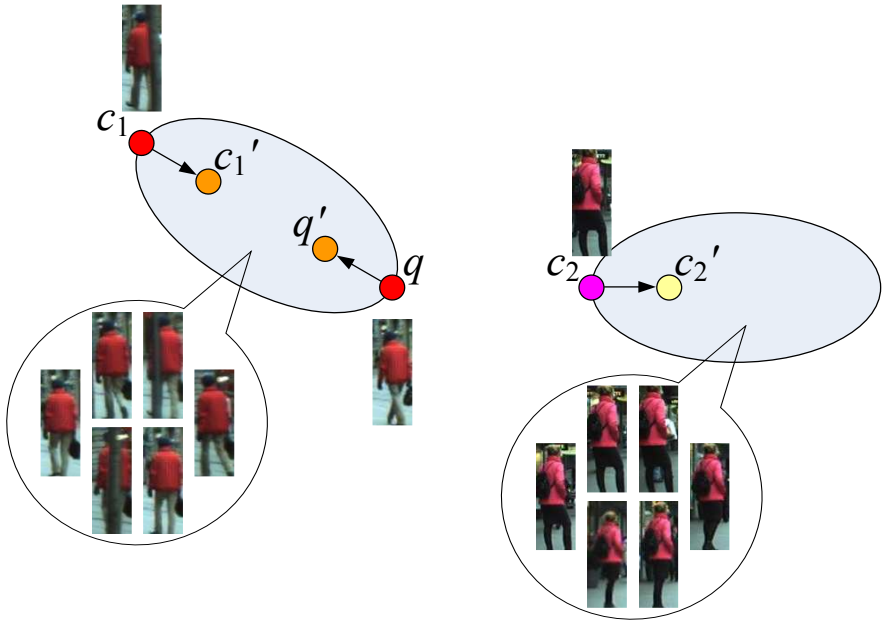
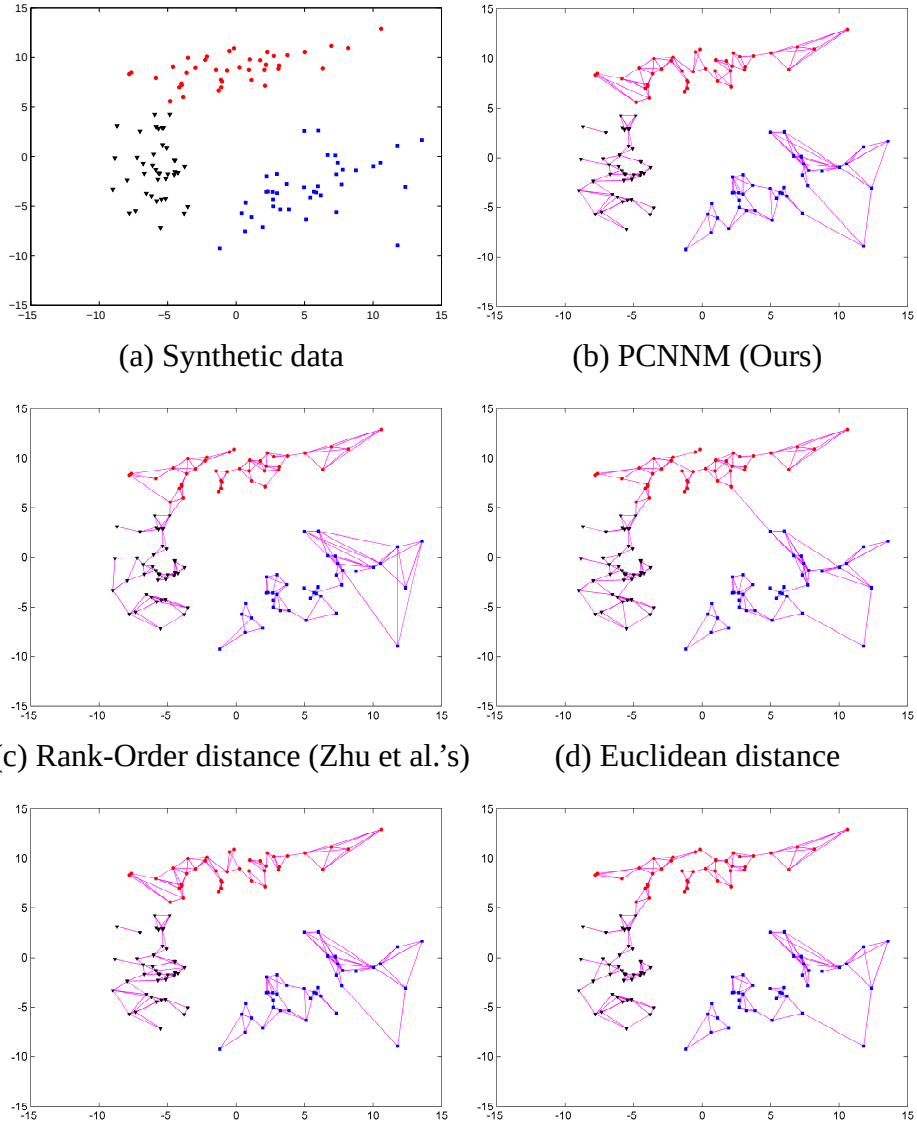


Figure 2.4: Illustration for the effect of "Symmetric dissimilarity". Nodes c_1, c_2, and q denote three samples. c_1 and q belongs to the same class, whereas c_2 is from a different class. Virtual nodes c_1', c_2' and q' are provided to show the effect of applying "Symmetric dissimilarity" to measurements between c_1, c_2 and q.

Besides the concrete example, we now provide a statistical illustration for PCNNM. Synthetic data can help elaborate. We randomly generate three separate, Gaussian-distributed datasets for use as class samples (for a class size of 40). For fairness, corpus and query points are assigned for ten times by randomly halving the data. We connect each query point to its top-ranked corpus point measured by PCNNM, with n set to half of the average sample number in each class and λ tentatively set to 1 and 0, respectively. We then compare PCNNM (for which $n = 20$ and $\lambda = 1$) to the baseline dissimilarity measured by Euclidean distance and to Zhu et al.'s Rank-Order distance $D^{\text{Rank-Order}}$. We also compare our "Fixed-number" based "Symmetric dissimilarity" (for which $n = 20$ and $\lambda = 0$) to Zhu et al.'s flexible-number based $D_a(b) + D_b(a)$ in $D^{\text{Rank-Order}}$. All the experiments use the same generated data and the same corpus-query splits.



(a) Synthetic data (b) PCNNM (Ours)

(c) Rank-Order distance (Zhu et al.'s) (d) Euclidean distance

(e) Symmetric dissimilarity (Ours) (f) $D_a(b) + D_b(a)$ (Zhu et al.'s)

Figure 2.5: Synthetic data is generated randomly to test the performance of PCNNM and "Symmetric dissimilarity". Classes are labeled by distinct colors. A magenta line is used to connect each query point with its top-ranked corpus point.

Note that, our synthetic data based experiments carried out here are essentially different from Zhu et al.'s, though they seem to be similar. Zhu et al. used exper-

iments to demonstrate the robustness of Rank-Order distance during clustering. For clusters with varying density/shape/size, they plotted all the edges between points whose Rank-Order distances are smaller than a given threshold [60]. We use experiments to evaluate the advantage of PCNNM by recognition/rankings. These experiments are relevant to the re-identification issue, but differ from those in Zhu et al.'s work to demonstrate the clustering effect.

The lines shown in Figure 2.5 display the accumulative results for ten-fold cross-validation. From these, we can confirm that PCNNM yields the least false lines among all the compared methods, and that even our "Symmetric dissimilarity" wins both $D_a(b) + D_b(a)$ and $D^{\text{Rank-Order}}$ of Zhu et al.'s as well.

The Gaussian-distributed synthetic data dismissed in Figure 2.5 is used to illustrate the advantage of PCNNM. Real-world human image data are high-dimensional, and it is impossible to visualize these data directly. Synthetic low-dimensional data simulations can be useful to some extent, and there seems to be no better alternative than this approach.

For these synthetic data, most samples in the same class reside closer to each other than those from different classes. Though this is an expected property required for re-identification, the real data distribution rarely satisfies this property. Hence, for those real data, we need to use metric learning to improve the intra-class compactness with regard to the inter-class separation before performing PCNNM. Though being not perfect perhaps, such improvement to a certain degree makes the real data distribution satisfy that property as far as possible, and thus similar to the situation of the synthetic data distribution. From this perspective, the synthetic data visualization can indirectly represent the real distribution of the data.

Furthermore, in the next section, experiments across several widely-used benchmark datasets will have demonstrated the effectiveness of PCNNM in a learned metric space using real data. These synthetic data based experiments opportunely complement them. Overall, synthetic data and real data contribute to a more convincing demonstration on the advantage of our proposed method.

2.2.3 Metric Space Improving

Though PCNNM can encode the common-near-neighbor information into a capable dissimilarity measure, in a noisy feature space, the intruders and outliers of each class may degrade the performance. To avoid this, we need to de-noise the rank order lists by improving intra-class compactness relative to inter-class separation for PCNNM measurement. Coincidentally, OMRR can benefit PCNNM by

optimizing list-wise rankings.

OMRR is an application of MLR. MLR considers a metric good if, when given a query point, sorting the corpus by increasing distance from this point results in good neighbors at the front of the list, and bad neighbors at the end. Hence, MLR casts nearest neighbor prediction as a ranking problem, and the predicted label error rate as a loss function over rankings. retrieval in the query-by-example paradigm.

Given query sample collection $\mathcal{Q} = \{q \mid q \in \mathbb{R}^d\}$ and corpus sample collection $\mathcal{X} = \{x_{qi} \mid x_{qi} \in \mathbb{R}^d\}$, suppose w is the Mahalanobis metric matrix intended to optimize, and $\phi_{qi}(x_{qi}, q)$ is used to denote a kind of matric representation of a corpus sample x_{qi} with regard to q:

$$\phi_{qi} = -(q - x_{qi})(q - x_{qi})^\top. \tag{2.7}$$

A desired ranking model can be defined by

$$g_w(x_{qi}) = w^\top \phi_{qi}(x_{qi}, q) \tag{2.8}$$

for scoring x_{qi}, and the ranking can be done by sorting scores in a descending order.

To learn w, usually, a joint feature map is adopted to represent the whole set of ranked data $\mathcal{X}$. Let $y_q^{\text{ranking}} \in \mathcal{Y}$ be a ranking of $\mathcal{X}$ with respect to q, and $\psi(q, y_q^{\text{ranking}}, \mathcal{X}) \in \mathbb{R}^d$ be a vector-valued joint feature map, which is defined as the partial order feature:

$$\psi(q, y_q^{\text{ranking}}, \mathcal{X}) = \sum_{x_{qi} \in \chi_q^+} \sum_{x_{qj} \in \chi_q^-} \left(\frac{\phi_{qi}(q, x_{qi}) - \phi_{qj}(q, x_{qj})}{\|\chi_q^+\| \|\chi_q^-\|}\right). \tag{2.9}$$

One important property of $\psi(q, y_q^{\text{ranking}}, \mathcal{X})$ is that, for a fixed w, the ranking y_q^{ranking} which maximizes $w^\top \psi(q, y_q^{\text{ranking}}, \mathcal{X})$ can be obtained by sorting $g_w(x_{qi})$ in order of a descending scores.

The best w is expected to be the one that simultaneously makes

$$y_q^* = \arg\max_{y_q^{\text{ranking}}} w^\top \psi(q, y_q^{\text{ranking}}, \mathcal{X}), \tag{2.10}$$

where y_q^* is the ground truth ranking of $\mathcal{X}$ for q. Thus, w can be learned by solving the following optimization problem:

$$\arg\min_w \frac{1}{2}\|w\|^2 + \frac{C}{|\mathcal{Q}|} \sum_q \xi_q \tag{2.11}$$

s.t.

$$w^\top \psi(q, y_q^*, \mathcal{X}) \geq w^\top \psi(q, y_q^{\text{ranking}}, \mathcal{X}) + \Delta(y_q^*, y) - \xi_q, \quad \forall q, y_q^{\text{ranking}} \neq y_q^*;$$
$$\xi_q \geq 0, \forall q,$$

where y_q^* is the ground truth ranking of $\mathcal{X}$ for a given $q \in \mathcal{Q}$, ξ_q is the slack variable, C is the trade-off parameter, and $\Delta(y_q^*, y_q^{\text{ranking}})$ is the loss function to penalize predicting y_q^{ranking} instead of y_q^*, defined by $\Delta(y_q^*, y_q^{\text{ranking}}) = 1 - S_{\text{MRR}}(q, y_q^{\text{ranking}})$, in which

$$S_{\text{MRR}}(q, y_q^{\text{ranking}}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \begin{cases} 1/r_q, & r_q < k; \\ 0, & r_q \geq k, \end{cases} \tag{2.12}$$

where k is a threshold value that can be assigned a priori, and r_q is the rank order of ground-truth corpus sample with regard to the query q.

Conceptually, the reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct match, and MRR (Mean Reciprocal Rank) is the average of such reciprocal ranks of results over the whole query collection. To a degree, this coincides with practical performance expectations. Since only the rank of the first correct match is taken into account, the ranks of both other correct matches and any incorrect matches are arbitrary. Thus, there will be multiple ranking instantiations for a given ground truth. In practice, MRR is quite reasonable, and there seems to be no significantly better options [51].

Given a large number of constraints, it is reasonable to approximate a solution. For this purpose, the cutting-plane algorithm fits nicely. The idea of a cutting-plane algorithm is to approximate either the constraint set or the epigraph of the cost function by intersecting a limited number of half spaces. It usually refines the approximation incrementally, by generating additional half spaces through the use of sub-gradients [21, 42].

With the learned metric, the space will be discriminative, in which, point based intra-class distances should be smaller than inter-class distances, thus helping to improve the performance of PCNNM.

PCNNM is a high-level dissimilarity measure that is based on the rank order lists formed by the low-level dissimilarity measure. Hence, the more reliable low-level dissimilarity measure is, the better performance PCNNM will have. To enhance the quality of rank order lists, we employ OMRR to improve the intra-class compactness with regard to the inter-class separation before measuring the

Algorithm 1 POINT-LEVEL COMMON-NEAR-NEIGHBOR ANALYSIS (PCNNA)

Require: Training data x_ts; corpus samples x_cs and query samples x_qs as testing data.

Ensure: Ranking y^{ranking} of all x_cs for each x_q.

1: Perform OMRR on x_ts to learn a new metric space.
2: Project x_cs and x_qs into the learned new metric space, denoted by $\mathbf{x}_c$s and $\mathbf{x}_q$s, respectively.
3: List all $\mathbf{x}_c$s and $\mathbf{x}_q$s together into $\mathbf{x}$s.
4: Sort $\mathbf{x}$s based on Euclidean distance to acquire the rank order lists for each $\mathbf{x}_c$ and $\mathbf{x}_q$.
5: Measure PCNNM dissimilarity D^{PCNNM} between each pair of $\mathbf{x}_c$ and $\mathbf{x}_q$ based on their rank order lists.
6: Re-rank all $\mathbf{x}_c$s for each $\mathbf{x}_q$ according to D^{PCNNM}s calculated in step 5 to return y^{ranking}.

low-level dissimilarity among samples. From a pre-processing standpoint, OMRR plays a de-noising role for these human image data. If the original data are sufficiently separable, even without OMRR, PCNNM can work effectively. Accordingly, there is no direct relationship between the performance of the distance metric learning and the choice of the "Fixed-number" n, to be frank.

The steps of PCNNA are presented in Algorithm 1.

2.3 Experiments and Results

2.3.1 Experimental Setup

We test our PCNNA method on public available datasets VIPeR [15], ETHZ [10], and i-LIDS [57]. The representative samples in these datasets are shown in Figure 1.5.

We average the results for ten executions against random training-testing data splits. For fairness, we use the same samples from the assigned dataset for comparison. Experiments are conducted in two groups: parameter discussion and method demonstration. In parameter discussion, we seek to confirm the suitable settings of tunable parameters for PCNNM, including the "Fixed-number" n and trade-off parameter λ. In method demonstration, we seek to confirm the impor-

tance of metric space selection and the superiority of our modeling to Zhu et al.'s Rank-Order distance, since both are components of PCNNA. We also evaluate the effectiveness of our PCNNA method in comparison to related state-of-the-art methods.

For all of the above datasets, we normalize images to 48×128 pixels, the same size as images in VIPeR. Because feature representation is not the focus of this chapter, we do not tune it for a better performance. Considering the complexity of the issue and following the state-of-art, PCNNA uses a signature concatenation of two widely-used features: WHSV [11, 51] and DSCH [8, 51], denoted by "WHSV+DSCH". WHSV can handle changes in viewpoint and pose by considering both global color information and structural properties of the human body. DSCH can tackle changes in illumination and occlusion in terms of its cell-based statistical local color description and a global dense sampling of these cells. Since WHSV and DSCH have their own strengths and characterize human appearance information from different perspectives, they can be combined into a powerful signature. It is worth mentioning that, WHSV and DSCH can not only provides a good feature space, but also ensure a fair comparison with the state-of-the-art methods.

Here, to form the WHSV feature representation, we obtain the silhouette masks based on the same STEL (STructure ELement) model as in [11, 2]. A silhouette mask containing only foreground pixel values was acquired for each person by inferring over the STEL generative model [22]. This model captures the general structure of an image class as a blend of several component segmentations, isolating meaningful parts that exhibit tight distributions over the image measurements. This model is customized for foreground/background separation by setting two components and two parts corresponding to the foreground and background, and learned beforehand using a subset of VIPeR not for testing. The segmentation over new samples consists in a fast inference. Additionally, there is one convenient way to acquire the WHSV feature representation for several widely-used benchmark datasets. This feature representation is available on line provided by Loris Bazzani [11, 2].

We randomly halve the persons of each dataset for training-testing use. First, the training data are used to improve the original metric space for the testing data. Then, for each person, one image is randomly selected to form the corpus, leaving the remaining images for the query. We select each image from the query side and match it to each image from the corpus side according to dissimilarity. While calculating the dissimilarity between the query image and corpus image by PCNNM, we use those remaining unlabeled query and corpus images to help form

the rank order lists. We then obtain the correct match upon these dissimilarities. The entire procedure is repeated for 10 times, and the average results are plotted as CMC (Cumulative Match Characteristic) curves [15]. CMC illustrates how the performance (recognition/re-acquisition rate) improves as the number of requested images increases, and it is one of the most widely-used evaluation criteria for human re-identification.

2.3.2 Parameter Discussion

Fixed-number

To discuss how important the "Fixed-number" n is for "Symmetric dissimilarity" of PCNNM in the OMRR metric space, we sample ETHZ datasets randomly to produce five hand-crafted datasets with the same number of images per person: 2, 4, 8, 16, and 32 (denoted by ETHZ_N2, ETHZ_N4, ETHZ_N8, ETHZ_N16, ETHZ_N32, respectively). We also include another hand-crafted dataset with 2 images per person from i-LIDS (denoted by i-LIDS_N2). Combining these with the VIPeR dataset (for which the number of images per person is always 2), we evaluate the possible relationship between the most suitable "Fixed-number" and the number of images per person, with respect to the performance of "Symmetric dissimilarity".

Our results are described by the condensed measure MRR shown in Figure 2.6. MRR can offer an overall evaluation of the rankings [51]. From the compared results, we can see that, coincident with our recommendation, when the "Fixed-number" is set to approximately half the number of images per person, "Symmetric dissimilarity" performs best.

Particularly, when each class has two samples, the best "Fixed-number" is "$n = 1$". If a larger "Fixed-number" is selected, samples from different classes will be easily included. This situation is adverse for characterizing the discriminatory within-class sample distribution for each sample, thereby having a negative impact on "Symmetric dissimilarity". Hence, inviting no additional neighbors seems better than taking the risk of introducing intruders.

As expected, the "Fixed-number" n is related to the sample number of each person rather than the number of the people in the corpus. A suitable neighborhood size is required to exploit the point-level common-near-neighbor information for differentiating samples' classes, considering the representative local information for samples. To model a discriminative "Symmetric dissimilarity", we suggest the use of the "Fixed-number" n to cover the neighboring samples from the

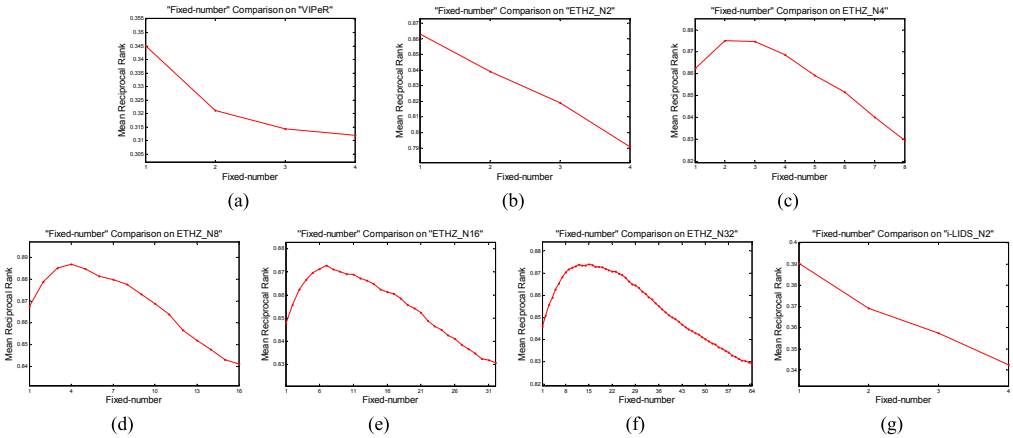


Figure 2.6: Performance of "Symmetric dissimilarity" for different "Fixed-number" recommendation in terms of MRR scores.

same class as far as possible to maximize the capture of the within-class sample distribution information for each sample. Thus, it is necessary to concern the relationship between the "Fixed-number" n and the sample number of each person. Further, though the rank order lists contain the people in the corpus, there is no indicator of a relationship between the "Fixed-number" n and the number of the people in the corpus for producing a most discriminative "Symmetric dissimilarity". This claim has been demonstrated by the experimental results on the datasets of different class sizes and class numbers in Figure 2.6, from which, we can see the relationship between the "Fixed-number" n and the sample number of each person rather than the number of the people in the corpus.

Note that, the discussion and recommendation of the "Fixed-number" parameter setting are within the scope of our examined data.

Trade-off Parameter

Trade-off parameter λ plays a role in balancing "Symmetric dissimilarity" and "Asymmetric dissimilarity" for PCNNM in the OMRR metric space. We conducted experiments on VIPeR, ETHZ_N2, ETHZ_N8, ETHZ_N16, and i-LIDS_N2 using the recommended "Fixed-number" setting to discuss how the number of images per person influences "Symmetric dissimilarity" and "Asymmetric dissimilarity", and possibly to clarify the characteristics of "Symmetric dissimilarity" and "Asymmetric dissimilarity" for different class sizes.

Our results are described by MRR in Table 2.1. Note that when the number of images per person is small, there is not enough common-near-neighbor information to differentiate the contributions of "Symmetric dissimilarity" and "Asymmetric dissimilarity" to the overall performance of PCNNM. As the number of images per person increases, the superior competence of "Symmetric dissimilarity" to "Asymmetric dissimilarity" becomes clearer. From this result, we conclude that the number of images per person should directly impact the selection of λ in PCNNM, favoring "Symmetric dissimilarity" for higher numbers of images.

Table 2.1: Performance of PCNNM for different trade-off parameter settings and different numbers of images per person in terms of MRR scores (%).

λ	VIPeR	ETHZ_N2	ETHZ_N8	ETHZ_N16	i-LIDS_N2
0.001	34.53	86.40	88.70	87.11	47.01
0.01	34.53	86.40	88.72	**87.12**	47.01
0.1	34.58	86.42	**88.75**	87.05	47.03
1	**34.86**	**86.86**	88.47	86.75	**47.73**
10	33.90	86.73	86.95	85.32	47.06
100	33.48	86.38	86.04	84.36	46.98
1000	33.46	86.38	86.01	84.24	46.98

We justified this assertion using the original VIPeR, ETHZ1, ETHZ2, ETHZ3, and i-LIDS datasets, which have approximately averages of 2, 60, 40, 60, and 4 images per person, respectively. Experimental results are described by MRR in Table 2.2. Note that PCNNM in the metric space learned by OMRR performs best on VIPeR when $\lambda = 1$, on ETHZ when $\lambda = 0$, and on i-LIDS when $\lambda = 0.1$.

Indeed, results suggest that the number of images per person is the only factor that significantly influences the performance of "Symmetric dissimilarity". As is well known, VIPeR has an extremely small sample size but a very complicated data distribution, such that when "$n = 1$", there are no near neighbors available in "Symmetric dissimilarity", weakening it relative to "Asymmetric dissimilarity".

PCNNM works best when "Symmetric dissimilarity" does not need to collaborate with "Asymmetric dissimilarity" in the cases of datasets ETHZ1, ETHZ2, and ETHZ3 ($\lambda = 0$). This validates the performance of common-near-neighbor information measurement using "Symmetric dissimilarity" for cases in which the number of images per person is large enough. PCNNM shows its advantages clearly as well when processing the i-LIDS dataset, with its small number of diverse images per person. By using "Asymmetric dissimilarity" to tackle the asymmetric rank-

Table 2.2: Performance of PCNNM against different benchmark datasets for different trade-off parameter settings in terms of MRR scores (%).

λ	VIPeR	ETHZ1	ETHZ2	ETHZ3	i-LIDS
0	34.52	**86.06**	**89.04**	**97.73**	51.55
0.001	34.53	86.06	89.04	97.73	51.57
0.01	34.53	86.05	89.03	97.72	51.60
0.1	34.58	85.97	88.97	97.64	**51.71**
1	**34.86**	84.98	88.17	96.57	51.68
10	33.90	82.33	85.89	93.30	49.78
100	33.48	80.74	84.06	92.05	48.80
1000	33.46	80.52	83.94	91.94	49.68
∞	32.76	80.45	83.85	91.83	48.22

ing problem, PCNNM can still perform robustly against this challenging dataset.

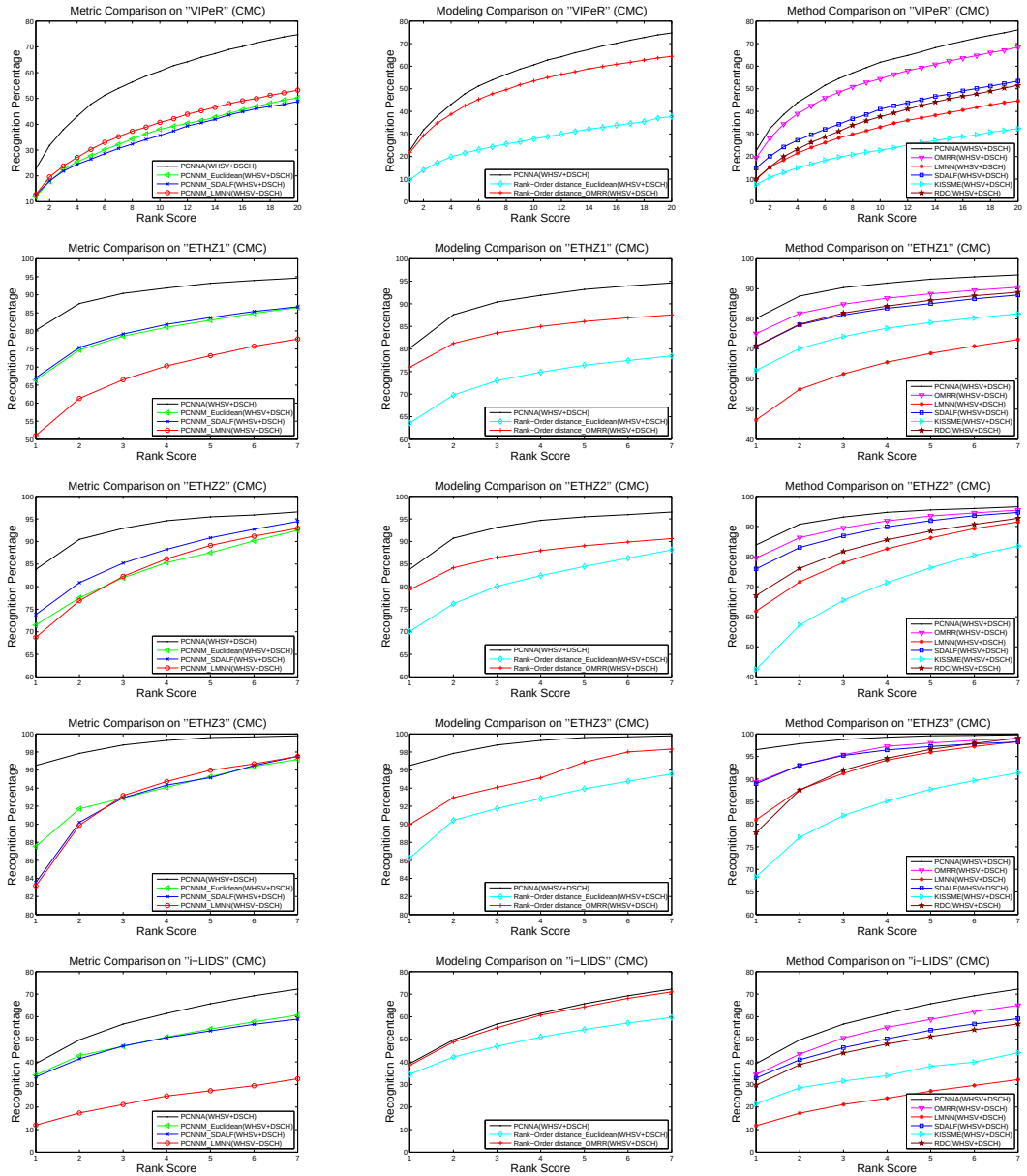
Note that, the recommendation of the trade-off parameter setting is within the scope of our examined data.

2.3.3 Method Demonstration

Metric Space Selection

To verify the suitability of metric space selection, we test PCNNM in different comparable metric spaces, including Euclidean metric space, SDALF metric space, LMNN metric space, and OMRR metric space. Here, combining WHSV and DSCH has been proved more effective than original SDALF [51], so the SDALF metric space is here generalized to indicate measuring distance between suitable feature representations with the Bhattacharyya metric.

From the results shown in Figure 2.7(a), we can see that, for PCNNM, the metric space projected by OMRR significantly outperforms other metric spaces on different datasets. Further, note that, in a metric space learned by LMNN, which aims at optimizing a metric space for classification rather than ranking, PCNNM does not work well. This clearly verifies the importance of metric space selection.



(a) Metric Space Comparison (b) Modeling Comparison (c) Method Comparison

Figure 2.7: CMC performance comparison on the VIPeR, ETHZ, and i-LIDS datasets.

Modeling Validation

To assess and validate the superiority of PCNNM, we compare it with its analogue Zhu et al.'s Rank-Order distance modeling in their Euclidean space. For fairness, we use the same feature representation WHSV+DSCH for PCNNM and Zhu et al.'s modeling. We also tentatively evaluate their modeling into the metric space learned by OMRR.

From the results shown in Figure 2.7(b), we see that, for the issue of human re-identification, our modeling is more effective than Zhu et al.'s even in their Euclidean metric space. Moreover, note that, as expected, the metric space learned by OMRR provides a platform for better performance of Zhu et al.'s Rank-Order distance than the Euclidean metric space. This is because OMRR can optimize list-wise ranking, and Zhu et al.'s Rank-Order distance is based on rank order lists as well.

Method Evaluation

We also compare our PCNNA method to related state-of-the-art methods, including SDALF [11, 2], LMNN [46], OMRR [51], KISSME [23], and RDC [59]. Here, PCNNA uses the recommended parameter-settings (for VIPeR, $\lambda = 1$, $n = 1$; for ETHZ1, $\lambda = 0$, $n = 30$; for ETHZ2, $\lambda = 0$, $n = 20$; for ETHZ3, $\lambda = 0$, $n = 30$; for i-LIDS, $\lambda = 0.1$, $n = 2$).

From the results in Figure 2.7(c), it can be seen that PCNNA always remarkably distances the runner-ups, while the performances of its competitors fluctuate across different datasets. Actually, human images in these datasets suffer from pose variations, illumination changes, viewpoint alterations, occlusions, and so forth, as shown in Figure 1.5. From the experimental results, we can see PCNNA is more robust than other methods to the real-world challenges of human re-identification.

2.4 Summary

This chapter has proposed a novel method, PCNNA, to deliver the effectiveness of the well-distributed samples to the badly distributed-samples for single-shot human re-identification. In PCNNA, a new dissimilarity PCNNM is designed to exploit the point-level common-near-neighbor information in a discriminative metric space learned by OMRR. Advantages of this method are borne out by ex-

perimentation on diverse datasets. Finally, for better understanding PCNNA, we append the framework of it in Figure 2.8.

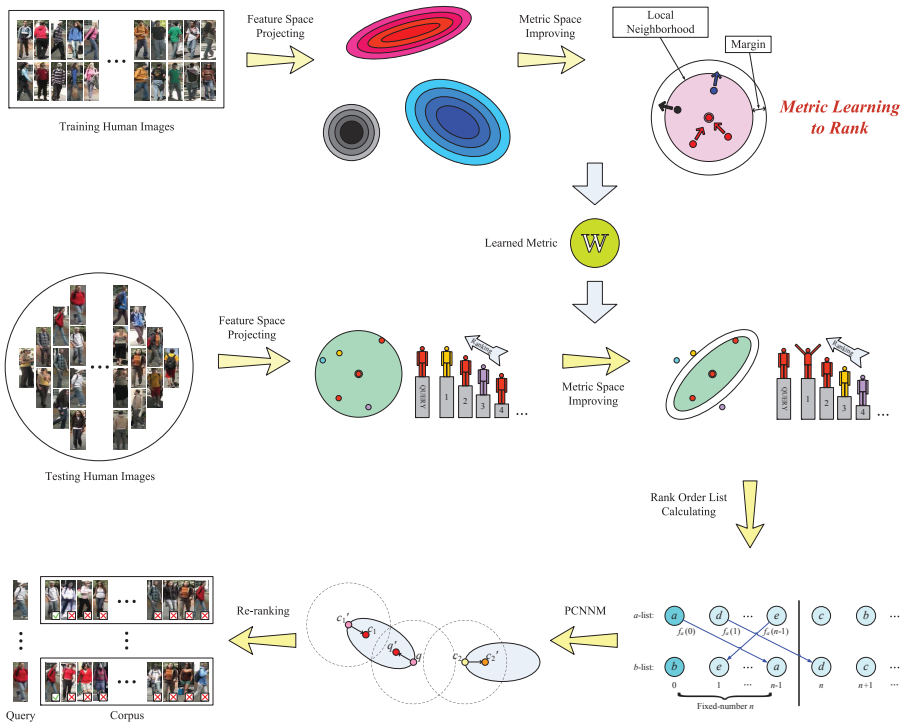


Figure 2.8: The figure illustrates the whole framework of PCNNA. The first row show the training stage. OMRR is implemented on the training human image samples to obtain a discriminative metric space. The second and third rows display the testing stage. In the second row, the rank order list is calculated for each sample in the projected space. Then, in the third row, PCNNM dissimilarity is measured between the testing human image samples.

Chapter 3

Strengthened Metric Space

3.1 Introduction

Chapter 2 has proposed the PCNNA method, which studies the point-level common-near-neighbor information by PCNNM in a learned metric space. Usually, the sufficient number of samples for each person is necessary for the satisfactory performance of PCNNM. However, in the real world, we may not be able to acquire enough samples. The specific and challenging single-shot vs. single-shot case more or less limits the performance of this method.

Recognizing that the discriminative metric space can avail PCNNM, we suggest using the strengthened metric space to compensate for the insufficiency of class samples during the single-shot vs. single-shot re-identification. This chapter will propose two novel methods, “Coupled Metric Learning (CML)” and “Point-level Common-Near-Neighbor Metric Learning (PCNNML)”, to strengthen the metric space for PCNNM. CML takes advantage of the complementarity between MCML and MLR. This method de-noises the original feature space using a learned linear projection from MCML to provide a good platform for the further metric optimization by MLR. PCNNML directly optimizes the neighborhood-wise comparison relationship between sample points. This method incorporates the constraints of the PCNNM dissimilarity comparison into MLR, so as to accord the training stage with the testing stage.

3.2 Coupled Metric Learning

The notion of distance is fundamental for many data mining/machine learning algorithms. Traditionally, the metric matrix for measuring distance has been specified by an a priori assignment. However, metric learning emphasizes that the distance measurement should be learnt from training data.

Dimensional reduction techniques, which exploit the embedding of data, can be categorized as unsupervised metric learning methods such as PCA, Regularized LDA, and so on. Supervised metric learning methods, such as Information-Theoretic Metric Learning [7] and Cosine Similarity Metric Learning [35], utilize objective functions and constraints to optimize the distance measurement. Although these methods have not been directly applied to the issue of human re-identification, they appear to have some potential in this field. When we treat the features of person images as high-dimensional points, the learned Mahalanobis metric matrix is able to map them into a new space to improve their intra-class compactness and inter-class separation.

LMNN [46] has already been introduced to address the issue of human re-identification, and it was shown that optimizing a metric space for ranking is more effective than optimizing a metric space for classification when the sample class size is small [51]. A satisfactory ranking space expects the intra-class distances of all samples to be smaller than the inter-class distances. MLR is based on such an intuitive concept, and OMRR is an application of MLR to the problem of re-identification. It uses a structural SVM framework to optimize the Mahalanobis metric matrix for ranking, paying heed to the design of the loss function.

The performance of MLR is more or less determined by the feature representation and sample class size. Therefore, the main idea of our proposed CML method is that, before metric learning, some projective space is optimally searched for the original feature representation by another metric learning method. We will show that MCML could provide a good platform for ranking optimization by MLR. It will be also shown that these two metric learning methods are in fact different, with a degree of complementary abilities, and their combination could ensure a better performance.

3.2.1 Metric Learning to Rank

The framework of MLR has been given by Equation 2.11. In practise, for a fixed w, the ranking y_q^{ranking} which maximizes $w^\top \psi(q, y_q^{\text{ranking}}, \mathcal{X})$ is obained by simply sorting the score $g_w(x_{qi}) = w^\top \phi_{qi}(x_{qi}, q)$ in a descending order [33, 51].

$\phi_{qi} \triangleq -(q-x_{qi})(q-x_{qi})^\top$ is a kind of matric representation that characterizes the relationship between the query sample q and the corpus sample x_{qi}. It is obvious that ϕ_{qi} describes the differential information for each sample pair and contains the information about one sample relative to the other. Naturally, if such a matrix holds pair-wise information about samples in the same class, we name it the "intra-class representation", and if it describes pair-wise information on samples from different classes, we name it the "inter-class representation". It is reasonable to expect that the representations are of sufficient quality during metric learning by MLR. Considering the property of $\phi_{qi}(x_{qi}, q)$, it is natural to require that samples of the same class should stay as close to each other as possible while samples from different classes should remain far away from one another. Certainly, and notably, such a requirement is sufficient but not necessary for providing a good $\phi_{qi}(x_{qi}, q)$. There may be other ways to improve $\phi_{qi}(x_{qi}, q)$, like designing a capable original feature space. But it is quite difficult for the challenging real-scenario human images, and the heuristic and subjectivity during designing process will unfortunately be delivered to $\phi_{qi}(x_{qi}, q)$. Hence, we recommend learning a suitable projection with respect to the original feature space.

MLR learns a discrimiative space for original features based on the structural SVM framework. Although MLR has advantage in dealing with the high-dimension small-sample problem, yet the real-world challenge of human image data and the limited sample class size may more or less limit its power. Accordingly, we resort to some other auxiliary projection approaches to improve the space for MLR. This improved space is expected to have the property that samples in the same class stay close to each other and samples from different classes are kept far apart. Undoubtedly, MCML [13] coincides with such requirements. However, it is difficult to directly apply MCML for space projection. So we simplify the projection into a linear form and take advantage of the decomposed property of learned the metric matrix. Suppose x is the feature point, and M is the metric matrix learned by MCML. Since M can be decomposed as $M = M_{\text{dec}}^\top M_{\text{dec}}$, the linear projection can be expressed by $h(x) = M_{\text{dec}}x$. Thus, the projected $\phi_{qi}^{\text{projected}}$ is presented as:

$$\begin{aligned} \phi_{qi}^{\text{projected}} &= (h(q) - h(x_{qi}))(h(q) - h(x_{qi}))^\top \\ &= -(M_{\text{dec}}q - M_{\text{dec}}x_{qi})(M_{\text{dec}}q - M_{\text{dec}}x_{qi})^\top \\ &= -(M_{\text{dec}}(q - x_{qi}))(M_{\text{dec}}(q - x_{qi}))^\top. \end{aligned} \tag{3.1}$$

In Equation 3.1, M_{dec} will be competent to make the projected $\phi_{qi}^{\text{projected}}$ more rep-

resentative for learning the metric by MLR.

3.2.2 Maximally Collapsing Metric Learning

MCML relies on the geometric intuition that all points in the same class should be mapped to a single location in the feature space and all points in other classes should be mapped to other locations. Basically, MCML obtains a compact low-dimensional feature representation of the original input space [13].

MCML uses Kullback-Leibler divergence in the objective function to make intra-class distances as small as zero and inter-class distances as large as infinite.

Given some labeled samples $(x_i, y_i^{\text{label}})$, where $x_i \in \mathcal{R}^r$ and $y_i^{\text{label}} \in \{1, ..., H\}$, the distance between any different points indexed by i and j can be defined as:

$$d(x_i, x_j|M) = d_{ij}^M = (x_i - x_j)^T M (x_i - x_j) \tag{3.2}$$

where M is a PSD matrix.

To learn a metric that approximates the ideal geometric intuition, for each training point, a conditional distribution over other points has been introduced. Specifically, for each x_i, a conditional distribution over any other x_j, where $j \neq i$, is defined as:

$$P^M(j|i) = \frac{e^{-d_{ij}^M}}{Z} = \frac{e^{-d_{ij}^M}}{\sum_{k \neq i} e^{-d_{ik}^M}}, \qquad j \neq i. \tag{3.3}$$

where j means any sample other than i. The framework of MCML is as below:

$$\arg\min_M \sum_i KL(P_0(j|i))|P^M(j|i)) \tag{3.4}$$

s.t.

$$M \in PSD, \tag{3.5}$$

where Z is the normalizing factor, and $P^M(j|i)$ takes a pseudo-probabilistic form to describe the conditional distribution over points. $P_0(j|i)$ is the ideal bi-level distribution as in Equation 3.6. If all points in the same class were mapped to a single point and infinitely far from points in different classes, we would have the ideal bi-level distribution:

$$P_0(j|i) = \begin{cases} 1, & y_i^{\text{label}} = y_j^{\text{label}}; \\ 0, & y_i^{\text{label}} \neq y_j^{\text{label}}. \end{cases} \tag{3.6}$$

In the optimizing process, at each iteration, MCML takes a small step in the direction of the negative gradient of the objective function, then the MCML metric is projected back onto the PSD cone by taking the eigen decomposition of M and substituting zero for the components with negative eigenvalues [13].

MCML can perform a dimensional reduction by spectral decomposition. The eigen decomposition of a metric matrix M can be written as:

$$M = \sum_{h=1}^{r_0} \lambda_h v_h v_h^\top, \tag{3.7}$$

where r_0 is the number of eigenvalues, λ_h is the eigenvalues of M, and v_h is the corresponding eigenvectors. The matrix M that has less than full rank can be used to measure the Mahalanobis distance based on the low-dimensional projection. The reduced dimension t_0 can be determined a priori. Hence, we can then select the largest t_0 eigenvalues and their eigenvectors to obtain the low-rank projection matrix:

$$M_{\text{proj}} = diag(\sqrt{\lambda}_1, ...\sqrt{\lambda}_{t_0})[v_1^\top; ...; v_{t_0}^\top]. \tag{3.8}$$

Generally speaking, the low-dimensional space projection is not guaranteed to be the same as the projection corresponding to minimizing the objective function of MCML, subject to a rank constraint on the optimal metric matrix, unless the rank of the optimal metric matrix is less than or equal to t_0. However, as demonstrated in [13], for practical problems, it is often the case that the optimal metric matrix has an eigen-spectrum which is rapidly decaying, so that many of its eigenvalues are indeed very small. This suggests the low rank solution will be close to optimal. So the projection matrix M_{proj} can be used to map the original space to a new low-dimensional space for intra-class compactness as well as inter-class separation.

Usually, the reduce dimension t_0 can be determined a-priori. Although t_0 is determined by heuristic, we need to avoid the case that t_0 is too large or too small. If t_0 is too large, the new feature space of the reduced dimension will be too noisy. If t_0 is too small, the new feature space of the reduced dimension will be too sensitive. Both cases will damage the effect of intra-class compactness and inter-class separation. Here, we suggest t_0 to be approximate 80% to 100% of the original feature dimension.

3.2.3 Modeling Justification

We have discussed that MCML may construct a linear projective space that is beneficial to MLR.

Specially, this implies that the modeling performance of CML could outstrip that of MLR. We denote this by "MCML+MLR $>_{\text{perf}}$ MLR". (Here, $>_{\text{perf}}$ means the former performs better than the latter. Similarly, in the following, $<_{\text{perf}}$ means the latter performs better than the former, $=_{\text{perf}}$ means the former performs almost the same as the latter, $\geq_{\text{perf}}$ means the former performs no worse than the latter, and $\leq_{\text{perf}}$ means the latter performs no worse than the former.)

To further highlight the advantages of MCML+MLR, in this section, we compare it with other modeling choices by discussing the relationship between MCML and MLR.

Analysis of CML

It is impossible to give a direct, mathematical proof for the superiority of MCML+MLR due to the difficulty of unifying MCML and MLR into a single optimized framework. Alternatively, we will provide evidence of the benefits of coupling MCML and MLR based on their complementarity, which means that MCML contributes towards MLR learning a more satisfactory space.

The role of MCML in CML can be understood from the point of view of noise reduction. Low-dimensional space projection by MCML has the effectiveness of data de-noising which coincides with MLR's target.

Traditionally, PCA is widely used to find low-dimensional embedding for de-noising before supervised learning. Although it has a similar functionality as MCML, PCA does not have the discriminative ability as that of MCML, and the space mapping given by it is not guaranteed to improve the performance of MLR.

The objective of MLR is that the distance between the samples in the same class should be smaller than that between the samples from different classes.

In the ideal case, when MCML meets its target perfectly, MLR's objective will also be indirectly and perfectly achieved, because intra-class distances will all be zero and inter-class distances will all be infinite. Hence, in general, MCML works in the same direction as MLR.

Although, in the real world, it is impossible to obtain ideal data for MCML or MLR, it is intuitive that, if samples from the same class become closer together and samples from different classes become farther away, it will be easier for MLR to make intra-class distances smaller than inter-class distances.

Therefore, MCML is able to contribute towards MLR learning a more satisfactory space, namely "MCML+MLR $>_{\text{perf}}$ MLR". If all the contributions from MCML is redundant to MLR which is very unlikely to happen, at least we can get "MCML+MLR $=_{\text{perf}}$ MLR". However, it has never appeared in our experimental results to be presented later.

Comparison with Other Modeling Choices

We will compare MCML+MLR with other modeling choices for CML, including MLR+MLR, MCML+MCML, and MLR+MCML, where the "+" sign denotes a strict "left-to-right" order (i.e., MLR maps the original space to a new linear projective space for further optimization by MCML).

One simple and direct reason to reject MLR+MLR and MCML+MCML is that the Mahalanobis metric learned by MLR has a unique solution given certain training samples, and so does MCML. Therefore, from a performance perspective, "MLR+MLR $=_{\text{perf}}$ MLR" and "MCML+MCML $=_{\text{perf}}$ MCML".

A more in-depth explanation is required for rejecting MLR+MCML, or indeed for not simply implementing MCML.

It is easy to discern the similarity between MCML and MLR. Obviously, both use a convex optimization framework to learn a Mahalanobis metric. The key difference is that, MCML forces intra-class distances to be zero and inter-class distances to be infinite.MCML aims to minimize the sum of the loss function for intra-class distances to be zero and inter-class distances to be infinite, described by Kullback-Leibler divergence. It is a kind of dipolar balance between intra-class distances and inter-class distances. Therefore, MCML may seek to balance the compactness of samples in the same class and the separation between samples from different classes. Undoubtedly, an optimal combination of intra-class compactness and inter-class separation will be the most beneficial for classification.

On the contrary, MLR concerns the relative relationship between samples in the same class and samples from different classes. It forces intra-class distances to be smaller than inter-class distances for all samples, which is different from the dipolar balance addressed by MCML. Optimizing the relative comparison between intra-class distances and inter-class distances will be conducive to ranking, because it matches the requirement of a good ranking: given a query, samples within a corpus from the same class as the query have smaller distances than samples from different classes.

With an ideal MCML metric, all intra-class distances will be zero and all inter-class distances will be infinite. Thus, the inter-class distances will definitely be

larger than the intra-class distances, so the optimization of relative distance comparison can also be achieved. Nevertheless, in real cases, the data distribution is complex due to small between-class but large within-class variations, especially in the case of an extremely small sample class size, as discussed in this chapter. There is the possibility that MCML will sacrifice the relative comparison between intra-class distances and inter-class distances to minimize its own objective function.

To illustrate this, we give an example, in which the dipolar balance addressed by MCML may impair the relative comparison between intra-class distances and inter-class distances. An illustration is shown in Figure 3.1.

Consider the layout of three points in a two-dimensional space; the scaling of axes by the metric matrix is simplified in two orthogonal dimensions, with (u, v) denoting the scaling parameters. The two points i and j are in the same class, marked by the same color, and point k is from a different class, marked by a different color. With i assigned as the query, we can see that the intra-class distance is initially smaller than the inter-class distance.

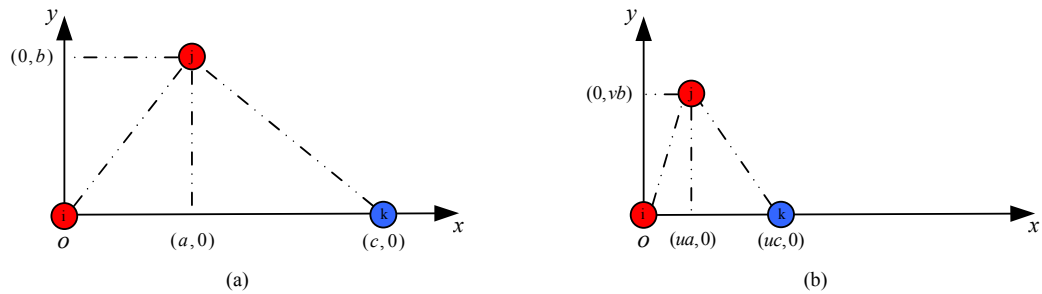


Figure 3.1: Exemplar to show the dipolar balance addressed by MCML may impair the relative comparison between intra-class distances and inter-class distances. Sample classes are distinguished by color.

As mentioned in [13], the objective function of the optimization framework described by Equation 3.4 is convex in matrix M. Equation 3.4 can be re-written as:

$$\arg\min_{M} \sum_{i} P_0(j|i) \log \frac{P_0(j|i)}{P^M(j|i)}. \tag{3.9}$$

In order to seek M to optimize Equation 3.9, the constant part can be ignored and the part containing M should be maintained; $P_0(j|i) \log \frac{P_0(j|i)}{P^M(j|i)} =$

$P_0(j|i)(\log P_0(j|i) - \log P^M(j|i)) = P_0(j|i)\log P_0(j|i) - P_0(j|i)\log P^M(j|i)$, which shows that these terms are an additive constant and a positive multiplicative constant with respect to $-\log P^M(j|i)$ and thus can be ignored. Therefore, minimizing Equation 3.9 is equivalent to minimizing $f(M)$ as below:

$$f(M) = -\sum_{i,j;y_i^{\text{label}}=y_j^{\text{label}}} \ln P^M(j|i) = \sum_{i,j;y_i^{\text{label}}=y_j^{\text{label}}} d_{ij}^M + \sum_i \ln Z_i. \tag{3.10}$$

In each step, the metric scaling parameters will be adjusted to simulate the decrease of Equation 3.10. If the example appears to achieve the dipolar balance at the expense of the relative comparison between intra-class distances and inter-class distances, the following system of inequalities will have a non-empty solution set for u and v.

$$\begin{cases} a^2+b^2+a^2+b^2+\ln(e^{-(a^2+b^2)}+e^{-c^2})+\ln(e^{-(a^2+b^2)}+ \\ \qquad e^{-((c-a)^2+b^2)}) > (ua)^2+(vb)^2+(ua)^2+(vb)^2+ \\ \qquad \ln(e^{-((ua)^2+(vb)^2)}+e^{-((uc-ua)^2+(vb)^2)})+ \\ \qquad \ln(e^{-((ua)^2+(vb)^2)}+e^{-(uc)^2}); \\ a^2+b^2 < c^2; \\ a^2+b^2 < (c-a)^2+b^2; \\ (ua)^2+(vb)^2 \geq (uc)^2; \\ u \geq 0; \\ v \geq 0; \\ v+u > 0. \end{cases} \tag{3.11}$$

In this system, the first inequality means that in each convergence step, after scaling, the value of the objective function is smaller than in the previous step. The second and third inequalities describe the initial distance relationship among the three points, and the fourth describes the impairment of the relative distance comparison after scaling by the metric matrix. The fifth, sixth, and seventh inequalities are the constraints of u and v. u and v cannot be zero simultaneously, because this would imply that the metric matrix of MCML is the zero matrix, at which point MCML loses its meaning.

To prove the existence of a non-empty solution set for u and v, we do not require all of the analytical solutions for the system of inequalities. Instead, we focus on the case $(ua)^2+(vb)^2 = (uc)^2$, because this is the critical condition of the system of inequalities. We can then obtain $v > 0$, and thus

$\left\{(u,v)|0<u<\sqrt{\ln\frac{(1+e^{-c^2+a^2+b^2})(1+e^{-c^2+2ac})-2}{2e^{-c^2+2ac}}}, v=\sqrt{\frac{(uc)^2-(ua)^2}{b^2}}\right\}$. In order to ensure the solution set is not empty, we set $\frac{(1+e^{-c^2+a^2+b^2})(1+e^{-c^2+2ac})-2}{2e^{-c^2+2ac}}>1$, and then acquire $e^{c^2-2ac}+e^{a^2+b^2-2ac}+e^{-c^2+a^2+b^2}>3$. According to the property of inequalities that the arithmetic mean is greater than the geometric mean, and considering other inequalities, we use the amplification and minification method to obtain the conditions for a, b and c as $\sqrt{a^2+b^2}<c<\frac{a^2+b^2}{2a}$, where $0<a<\frac{b}{\sqrt{3}}$. This proves the fact that dipolar balance may be minimized at the expense of relative comparison between intra-class distances and inter-class distances. In other words, MCML may impair the ranking to optimize its own objective function. Therefore, MCML is not equivalent to MLR, and cannot replace MLR. If we force MCML to perform the ranking directly, it is not guaranteed that we will obtain the optimum ranking results, though there is some potential for this. Hence, unlike "MCML+MLR", "MLR+MCML" is likely to be no better than MLR itself, i.e., "MLR+MCML $\leq_{\text{perf}}$ MLR".

Therefore, the proposed MCML+MLR is the most reasonable choice, and cannot be replaced by other model forms such as MLR+MCML, MLR+MLR, and MCML+MCML.

Conclusively, MCML and MLR have their own strengths. MCML optimizes the metric for classification and MLR optimizes the metric for ranking, and one cannot replace the other's role. Nevertheless, in a sense, it is this difference between MCML and MLR that offers the space for their complementarity and cooperation. MCML+MLR is the best choice for exploring this complementarity.

3.2.4 Experiments and Results

Experimental Setup

We demonstrate the effectiveness of our method on on of the most representative public benchmark datasets for single-shot vs. single-shot re-identification: VIPeR [15].

We use all of the people in each dataset. We normalize all images to 48×128 pixels, then randomly halve each dataset into training data and testing data. We repeat this 10 times for cross-validation and average the results for evaluation. The experimental results are illustrated by CMC curves. For each person, we randomly select two images, one as the query image and the other as the corpus image, in both the training and testing stages. Each time, we use the same selected data for comparing the methods.

According to current research, local color statistical descriptors perform remarkably well for human re-identification. As each dataset has its own characteristics, we propose the most suitable feature representation for each set on an individual basis.

VIPeR is distinguished by its large viewpoint, illumination, and pose variations. Hence, it is reasonable to explore the local color statistical descriptors by considering the human-body-structure information. We recommend the concatenation of WHSV [11, 51] and DSCH [8, 51], denoted by "WHSV+DSCH". DSCH are able to deal with illumination variations and occlusion due to their dense sampling of the cells that contain the local color statistical description of human body appearance. WHSV has the merit of dealing with viewpoint and pose changes because it not only makes use of global color information, but also takes symmetric and asymmetric properties of the human body into consideration.

Result Analysis

We compare the proposed CML method with both the related methods and the state-of-the-art to demonstrate the effectiveness and superiority of our model. These methods include SDALF, OMRR, MCML, MLR+MCML, and RDC.

As a representative method that can deal with single-shot vs. single-shot human re-identification, the SDALF compared in our experiments is not exactly the original one. SDALF is generalized to indicate a direct matching by the most suitable feature representation measured with the Bhattacharyya metric. Although SDALF is not a learning-based approach, it uses the same features as those in our method, and thus this pure matching based method is valuable for demonstrating the effectiveness of our learning based algorithm.

Besides OMRR and MCML, the MLR+MCML alternative is also adopted to validate the reasoning behind our proposed CML modeling formulation. RDC is a state-of-the-art method which focuses on learning a reliable metric based on the novel relative distance comparison modeling and has achieved encouraging results compared with other metric learning methods for human re-identification. Here, we use the original code of RDC model provided by the authors for comparison [59].

Although OMRR and RDC can work on single-shot human re-identification, they actually use multiple images per-person in the training and testing stages if possible [59, 51] (except on the VIPeR dataset), whereas the proposed method CML focuses on single-shot vs. single-shot, using only 2 images per person for both training and testing. Fewer training samples will make the problem more

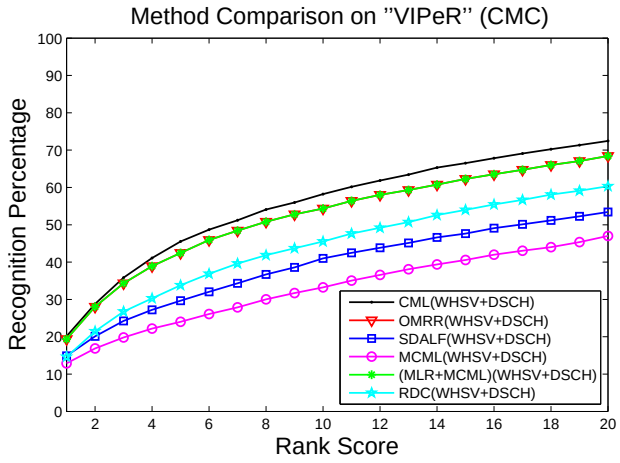


Figure 3.2: CMC performance comparison on the VIPeR dataset.

challenging.

In fact, single-shot vs. single-shot re-identification tends to be sensitive to the selected human image data, and thus, almost all the solutions, including CML, to the single-shot vs. single-shot re-identification issue may suffer from this headaching difficulty. However, even so, this cannot deny the value of these methods. So, more strictly, we can say CML has its statistical advantage in the considered scale, and this advantage has been comparatively more readily confirmed by the representative single-shot vs. single-shot dataset, VIPeR, as shown in Figure 3.2.

In the beginning of this chapter, we have mentioned discriminative metric space can avail PCNNM. We expect to know whether the strengthend metric space can increase the performance of PCNNM, and thus compensate for the insufficiency of class samples during the single-shot vs. single-shot case. Therefore, we conduct related experiments and check the results in Figure 3.3. For this figure, we can see, CML+PCNNM wins CML, but loses to MLR+PCNNM slightly; generally they are well-matched. This means, though CML has provided a more discriminative space, the strengthened metric space doesn't largely enhance the performance of PCNNM. Then, a new problem is brought out: how to effectively use strengthened metric space to tap the potential of PCNNM for the single-shot vs. single-shot case.

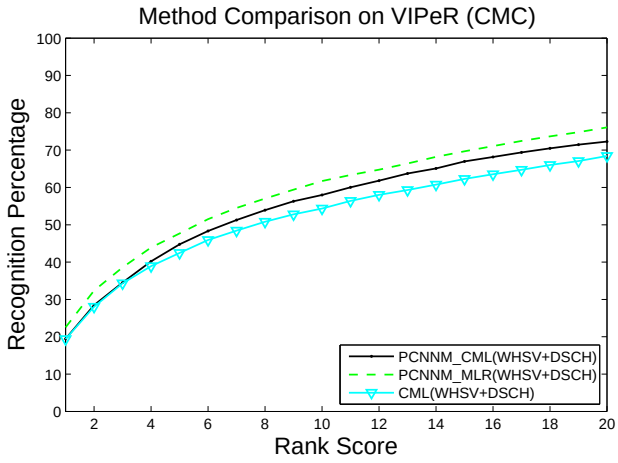


Figure 3.3: CMC performance comparison on the VIPeR dataset.

3.3 Point-level Common-Near-Neighbor Metric Learning

The single-shot vs. single-shot case is specific and challenging for re-identification, CML strengthens the metric space by using MCML to assist OMRR based on their complementarity, however, such strengthening has not leaded up to the substantial performance increment for PCNNM. The bottleneck seems to arise from the pipeline formulation of OMRR and PCNNM. We resort to PCNNM by incorporating the neighborhood information to overcome the vulnerability of Euclidean distance. The objective of OMRR is in essence to improve the Euclidean distance by a learned Mahalanobis metric. Though OMRR can de-noise the list-wise rankings that will ameliorate rank-order-list-based PCNNM, the objective of OMRR is not exactly the same as PCNNM dissimilarity measure, which deviates from the universal faith of consistency between training and testing for learning strategies, and this inconsistency may narrow down the re-identification performance.

Aimed at inheriting the merits of OMRR and PCNNM, and simultaneously overcoming their imperfections, this chapter proposes a novel metric learning method, “Point-level Common-Near-Neighbor Metric Learning (PCNNML)”, which directly optimizes the PCNNM dissimilarity so as to accord the training stage with the testing stage. PCNNML considers the neighborhood-wise comparison relationship between samples, thus breaks the traditional metric learning that solely considers the point-wise comparison relationship. Experiments

on widely-used benchmark dataset exhibit that PCNNML exceeds OMRR, PCNNA, and CML+PCNNM under a unified evaluation setting for the single-shot vs. single-shot case.

3.3.1 Method Elaboration

Conventional Metric Learning

Reviewing the metric learning approaches, we can start with using Mahalanobis metric matrix w to adapt the measurement of Euclidean distance, denoted by

$$D_w(x_i, x_j) = (x_i - x_j)^\top w(x_i - x_j), \tag{3.12}$$

where x_i and x_j are two samples, and w is the Mahalanobis metric matrix. Since the matrix w is PSD, it can be decomposed into $w = L^\top L$. Hence, an equivalent variant can be reformulated as:

$$\begin{aligned} D_L(x_i, x_j) &= (x_i - x_j)^\top L^\top L(x_i - x_j), \\ &= (L(x_i - x_j))^\top L(x_i - x_j), \\ &= (Lx_i - Lx_j)^\top (Lx_i - Lx_j). \end{aligned} \tag{3.13}$$

The decomposed metric matrix L can project the original feature space onto the new one by scaling the coordinates to correct for correlation between sample vectors. In this new space, we can conveniently measure the Euclidean distance between samples still.

Recently, metric learning has become an active research topic for predicting structured outputs, which is a common demand in many real applications. OMRR is a typical exemplar to adjust the space by learned Mahalabobis metric matrix to output the expectative permutations/ ordering of items.

Point-level Common-Near-Neighbor Metric Learning

In OMRR, the ranking y_q^{ranking} that maximizes $w^\top \psi(q, y_q^{\text{ranking}}, \mathcal{X})$ is obained by simply sorting the score $g_w(x_{qi}) = w^\top \phi_{qi}(x_{qi}, q)$ in a descending order [33]. If the ranking ability of g_w is improved, the better performance of OMRR will have. Since $\phi_{qi}(x_{qi}, q)$ encodes the point-to-point relationship between samples, badly-distributed samples, which can be traced back the to noisy data collected from the real scenario, becomes a burden to OMRR. Though OMRR aims at solving these problems by learning a capable metric, the small sample size in the single-shot vs.

single-shot re-identification case will more or less limits its power. This is also the reason CML seeks for a suitable and discriminatory projection by MCML to de-noise the space for MLR.

Neighborhood information provides an effective way for alleviating the sensitivity of metric learning. By delivering the effectiveness of the well-distributed sample points to the badly-distributed ones, PCNNM takes advantage of the neighborhood information and has been proved effective even for the single-shot vs. singe-shot case. This dissimilarity consists of a symmetric term and an asymmetric term combined by the function in Equation 2.4. PCNNM has been fully discussed in Chapter 2, so we are not going to detail it here.

One the one hand, encapsulating the neighborhood information into metric learning may help overcome the sensitivity of point-wise distance metric learning and thus strengthen the learned metric space. One the other hand, when the metric training stage is consistent with the PCNNM dissimilarity testing stage, the learned metric space will especially be suitable for the SCNNM dissimilarity measurement, and thus a substantial performance improvement for SCNNM can be anticipated. By addressing the relative comparison between intra-class SCNNM dissimilarities and inter-class SCNNM dissimilarities when optimizing the Mahalanobis distance, we plan to embed PCNNM into the MLR framework.

However, PCNNM dissimilarity cannot be expressed into the binary operation of two feature vectors straightforwardly. To solve this, we take advantage of the decomposed property of Mahalanobis metric matrix to describe g_w implicitly instead of explicitly, rephrased by

$$g_L^{\text{PCNNM}} = D^{\text{PCNNM}}(La, Lb). \tag{3.14}$$

It is similar to the kernel tricks if g_L^{PCNNM} is viewed as the kernel function. With it, we can learn the metric space based on PCNNM dissimilarity. This metric learning method is named "PCNNML".

For inference, PCNNML iterates the stage of learning the metric matrix, decomposing the metric matrix, mapping the feature space, measuring the PCNNM dissimilarity before convergence.

3.3.2 Experiments and Results

Experiments have been set up following the universal evaluation protocol in [11]. The results are drawn by the CMC curves in Figure 3.4. Comparisons to related competitors OMRR, PCNNA, and CML+PCNNM are provided on the challenging, representative, and popular VIPeR dataset in the same feature space of

WHSV+DSCH, which have clearly demonstrated the superiority of our proposed method PCNNML in this chapter.

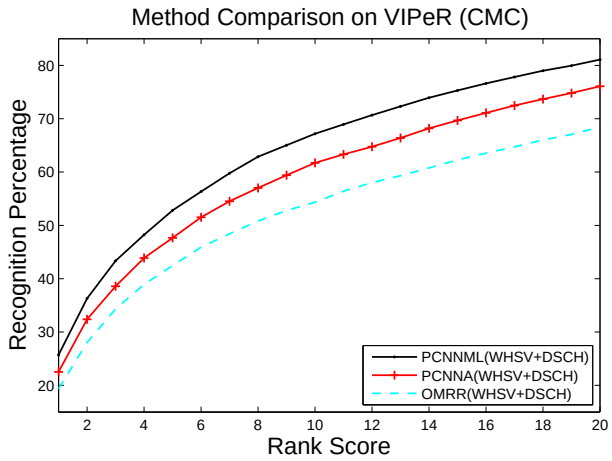


Figure 3.4: Method comparison for PCNNML.

3.4 Summary

In this chapter, we have proposed two novel methods, CML and PCNNML, to obatin the strengthened metric space for single-shot vs. single-shot human re-identification. CML uses MCML to de-noise the feature space for MLR based on their complementarity. Experimental results have verified the superiority of CML to strengthen the metric space. However, CML doesn't seems to largely enhance the performance of PCNNM. PCNNML incorporates the constraints of PCNNM dissimilarity comparison into MLR to ensure a well consistency between training and testing. Experimental results have shown PCNNML outperforms PCNNA by a large margin. Accordingly, we can conclude that suitably strengthened metric space will help increase the performance of PCNNM. Finally, we provide the illustration of CML and PCNNML in Figure 3.5.

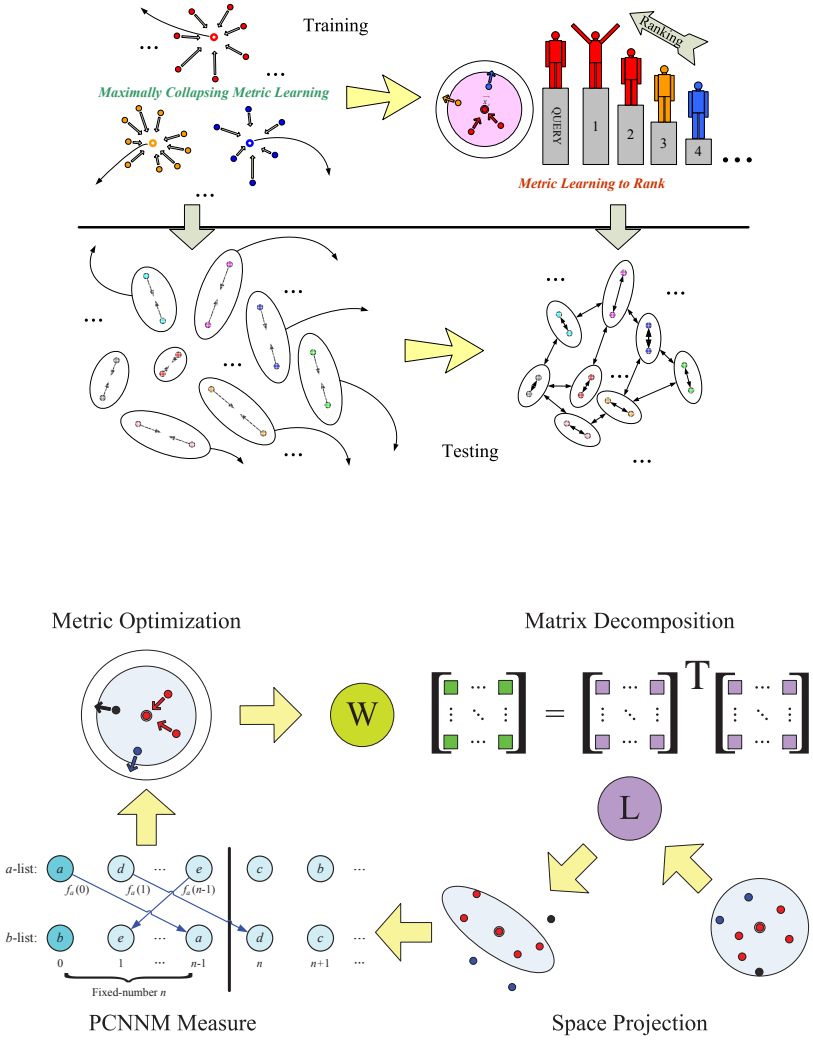


Figure 3.5: The upper part illustrates the framework of CML. The metrics are consecutively learned by MCML and MLR, and then used for projecting the testing samples. The lower part illustrates the training stage of PCNNML. In each iteration before convergence, the learned metric matrix is decomposed to project the training samples into a discriminative space where the intra-class and inter-class distances are re-measured by PCNNM for the next round of iteration.

Chapter 4

Set-level Common-Near-Neighbor Information

4.1 Introduction

Chapter 2 and 3 have studied the direction of single-shot human re-identification. From this chapter, we start to study the direction of multiple-shot human re-identification. In this direction, for each identity, there are multiple images acquired from the same camera collected as one set. The target of multiple-shot re-identification is to match these sets between cameras. Different than the point based matching in the single-shot case, the set based matching has more resources whilst challenges due to the availability and variability of within-set distribution information.

This chapter will propose two novel methods, “Riemannian Set-level Common-Near-Neighbor Modeling (RSCNNM)” and “Bi-level Relative Information Analysis (BRIA)”. In these methods, “Set-level Common-Near-Neighbor Modeling (SCNNM)” is designed to exploit the set-level common-near-neighbor information in the feature spaces projected by MRCG and TPCR, respectively. MRCG can condense human image sets into representative imaginary points in the Riemannian space, and TPCR can compact the human image sets to depress the intruders and outliers in the Hilbert space. Both of them provide the discriminative spaces which are suitable for SCNNM to deliver the effectiveness of the well-distributed sets to the badly-distributed ones.

4.2 Set-level Common-Near-Neighbor Modeling

In this chapter, we recast the human re-identification problem into a set based ranking problem that depends on set-to-set dissimilarity measured by SCNNM. SCNNM is extended from PCNNM. Accordingly, in order to present SCNNM, it is necessary to describe PCNNM beforehand.

In Chapter 2, PCNNM has been proven to be effective to deal with the single-shot human re-identification problem. Based on the expectation using those well-distributed samples to help improve those badly-distributed samples in a discriminative space, PCNNM explores the neighborhood structure comparison information to further make intra-class dissimilarities smaller than inter-class dissimilarities for all samples. As the core part, PCNNM is composed of the symmetric dissimilarity and the asymmetric dissimilarity, as expressed in Equation 2.4.

However, this method operates on the sample level and is designed for the target of single-shot human re-identification, which is much different from the multiple-shot case [11]. Though it's possible to directly apply it to multiple-shot problems, it is undesirable to do so. If we transform the multiple-shot problem into a single-shot problem to solve, both efficiency and effectiveness will be low. From the efficiency perspective, because the dissimilarity between each pair of samples is required to measure in this case, when the sample number increases in each set, the computation will be combinatorial explosion. From the effectiveness perspective, PCNNM explores the point-level common-near-neighbor information for every sample pair, so it can neither model the correlations among the multiple images within the same set which is important for robustness to within-set variations, nor maintain the robustness to the noisy outliers for the set.

Inspired by PCNNM, we propose a new model called "Set-level Common-Near-Neighbor Modeling (SCNNM)" to explore the neighborhood information among sets instead of samples towards the multiple-shot human re-identification problem. By using set based neighborhood structure comparison information SCNNM further ensures inter-class dissimilarities are larger than intra-class dissimilarities for all sets.

Technically, SCNNM incorporates the set-level neighborhood information into a novel dissimilarity of SCNNM, which is composed of a symmetric term and an asymmetric term combined by the following function:

$$H^{\text{SCNNM}}(A,B) = H^{\text{SCNNM}}_{\text{Symmetric}}(A,B) + 2\Lambda N H^{\text{SCNNM}}_{\text{Asymmetric}}(A,B). \tag{4.1}$$

In Equation 4.1, A and B denote two arbitrary sets; λ is the trade-off parameter between $H^{\text{SCNNM}}_{\text{Symmetric}}(A,B)$ and $H^{\text{SCNNM}}_{\text{Asymmetric}}(A,B)$; the "Fixed-number" for the

neighborhood size, denoted by N, is suggested to be half of the average number of sets per class. Considering symmetry, $H^{\text{SCNNM}}_{\text{Symmetric}}(A, B)$ is given by:

$$H^{\text{SCNNM}}_{\text{Symmetric}}(A, B) = H^{\text{Fixed-number}}_{A}(B) + H^{\text{Fixed-number}}_{B}(A), \tag{4.2}$$

where

$$H^{\text{Fixed-number}}_{A}(B) = \sum_{I=0}^{N-1} O_B(F_A(I)). \tag{4.3}$$

In Equation 4.2, $H^{\text{Fixed-number}}_{A}(B)$ sums the rank orders of A's list's top sets in B's rank order list under the setting of N, as shown in Figure 4.1, and $H^{\text{Fixed-number}}_{B}(A)$ is calculated in the similar way. In Equation 4.3, $F_A(I)$ is the I^{th} set in A's rank order list; $O_B(F_A(I))$ returns the rank order of $F_A(I)$ in B's list. Here, the rank order list of an assigned set is formed by the ranking of all the other sets according to their dissimilarities to this set.

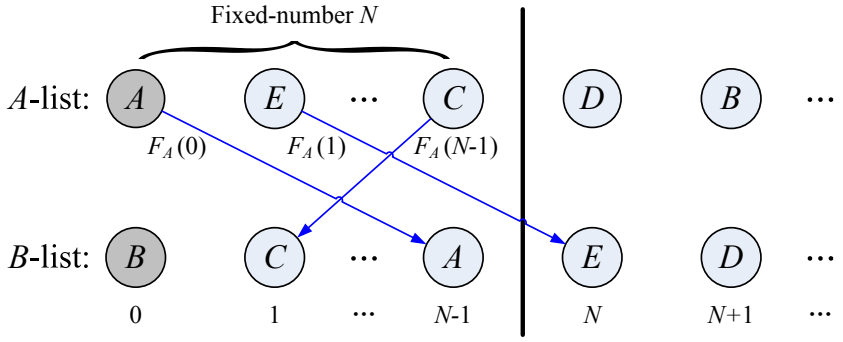


Figure 4.1: Sets are denoted by A, B, C, D, E and so on. $H^{\text{Fixed-number}}_{A}(B)$ is calculated from the 0^{th} to the $(N-1)^{th}$ nearest neighbor in A's rank order list.

As the asymmetric term, $H^{\text{SCNNM}}_{\text{Asymmetric}}(A, B)$ is given by:

$$H^{\text{SCNNM}}_{\text{Asymmetric}}(A, B) \quad = \quad min(O_A(B), O_B(A)). \tag{4.4}$$

In Equation 4.4, $O_A(B)$ is the rank order of B in A's rank order list, and $O_B(A)$ is defined in the similar way.

If SCNNM can continue the strength of PCNNM, it will be a powerful weapon for multiple-shot re-identification. From the efficiency perspective, SCNNM treats the samples in the same class as one whole set, so it is much faster than PCNNM

when there are multiple-shot images in each set. Suppose the average number of samples belonging to the same class (i.e., a set) is k_0, and there are totally C_0 classes, then the computational cost for re-identifying a query set (k_0 samples) in PCNNM will be $o(k_0^3 C_0^2 \log(k_0 C_0))$, given that the fast sorting algorithm is adopted (i.e., $o(n_0 \log n_0)$ complexity for n_0 items). However, re-identifying cost for the same task in SCNNM will be only $o(k_0 C_0 \log k_0)$. Therefore, SCNNM is more than $k_0^2 C_0$ times faster than PCNNM. From the effectiveness perspective, by quantifying the local neighborhood structure of the paired sets in each other's neighborhood structures, SCNNM not only can maintain the robustness to the noisy outliers for the set, but also can deliver the effectiveness from the well-distributed sets to the badly-distributed sets.

Similar to PCNNM of the single-shot case, "Fixed-number" N is an important tunable parameter in SCNNM, which may influence the symmetric term, so it deserves in-depth explanation. N describes the neighborhood size concerned by the symmetric term. If the neighborhood size is too large, set pair from different classes may share many common near neighbors for the top "Fixed-number" sets in both rank order lists, thus, the symmetric term will be reduced for sets from different classes with regard to the sets within the same class; if the neighborhood size is too small, set pair in the same class may share few common near neighbors for the top "Fixed-number" sets in both rank order lists, then, the symmetric term will be enlarged for the sets in the same class with regard to the sets from different classes. Obviously, both cases have negative influence on dissimilarity-based ranking, thus should be avoided. In order to measure by a robust symmetric term, it is reasonable to propose the choice of "Fixed-number" N to be approximate half of the average set number in each class in a compromise, in case the neighborhood size is too large or too small.

The steps of SCNNM are displayed in Algorithm 2.

4.3 SCNNM in Riemannian Space

Deliberating on the comparison between PCNNM and SCNNM, we can find their formulations are identical when exploiting the common-near-neighbor information. The difference is the former works on the points, while the latter operates on the sets.

As the low-level measure, PCNNM is based on the point-to-point distance, while SCNNM is based on set-to-set distance. Different from PCNNM, SCNNM is sensitive to the within-set variation, but fortunately traded off by the available

Algorithm 2 SET-LEVEL COMMON-NEAR-NEIGHBOR MODELING (SCNNM)

Require: Query image sets X_qs and corpus image sets X_cs; feature space $\mathcal{F}$.
Ensure: Ranking Y^{ranking} of all X_cs for each X_q.

1: Project all X_qs and X_cs into $\mathcal{F}$ as $\mathcal{X}_q$s and $\mathcal{X}_c$s, respectively.
2: List all $\mathcal{X}_q$s and $\mathcal{X}_c$s together into $\mathcal{X}$.
3: Sort $\mathcal{X}$ by set-to-set dissimilarity to acquire the rank order lists for each $\mathcal{X}_q$ and $\mathcal{X}_c$.
4: Measure SCNNM dissimilarity H^{SCNNM} for each pair of $\mathcal{X}_q$ and $\mathcal{X}_c$ based on their rank order lists.
5: Re-rank all $\mathcal{X}_c$s according to H^{SCNNM}s calculated in step 4 for each $\mathcal{X}_q$ to return Y^{ranking}.

useful within-set distribution information. So it seems the ambiguity for the unconfirmed SCNNM is whether the set-to-set distance effectively undertakes the low-level measure role.

Chapter 2 has proved the reasonability of PCNNM, thus, normally, if we can measure set-to-set distance just like point-to-point distance, the same rationality will be easily adapted to SCNNM. Along the way, we propose to project the each set into an imaginary point in the Riemannian space using MRCG [40]. Surely, even so, this kind of imaginary point-to-point distance still belongs to the scope of set-to-set distance in nature, because it works on sets.

4.3.1 Mean Riemannian Covariance Grid

MRCG [40] is one state-of-the-art method for the multiple-shot human re-identification problem. It uses the covariance descriptors computed from dense overlapping grids on the images, which effectively capture the discriminative information of appearance details for each person. The dissimilarity measurement for this covariance-based representation is measured in the Riemannian space. By condensing the information of all the samples within the set into a Karcher mean based signature, MRCG can deal with the large intra-class variations caused by pose changing, illumination varying, and occlusions. The steps of MRCG are detailed in Algorithm 3.

Algorithm 3 MEAN RIEMANNIAN COVARIANCE GRID (MRCG)

Require: Query image sets X_qs and corpus image sets X_cs; person number p; image number per person N_0; grid size (h_0, w_0).

Ensure: Ranking Y^{ranking} of all X_cs for each X_q.

1: Tile grids on each human image with the horizontal step and vertical step size both as $min(h_0, w_0)/2$.

2: Calculate Karcher mean μ for each grid within the image set: $\arg\min_{\mu} \sum_{r=1}^{N_0} \rho^2(\mu, C_r)$, where $\{C_1, ..., C_{N_0}\}$ denote a set covariance descriptors; r is the image index in the set; ρ denotes the Riemannian distance.

3: Calculate MRC discriminants σ for each grid: $\sigma^z_{i,k} = \frac{1}{v_0-1} \sum_{i=1, i\neq j}^{v_0} \rho^2(\mu^z_{i,k}, \mu^z_{j,k})$, where z denotes the camera; i and j are the image set indices; v_0 denotes the image set number; k is the grid index.

4: Calculate set-to-set similarity $\mathcal{S}_0$ between cameras: $\mathcal{S}_0(A, B) = \sum_{k=1}^{K} \frac{\sigma^{z_1}_{A,k} + \sigma^{z_2}_{B,k}}{\rho(\mu^{z_1}_{A,k} + \mu^{z_2}_{B,k})} / K$, where K is the total number of grids for each image; z_1 and z_2 denote different cameras; A and B are image sets from z_1 and z_2, respectively. Thus, set-to-set dissimilarity $\mathcal{S}_1$ can be obtained by $\mathcal{S}_1 = 1/\mathcal{S}_0$.

5: Re-rank all X_cs according to $\mathcal{S}_1$s calculated in step 4 for each X_q to return Y^{ranking}.

4.3.2 Collaboration of MRCG and SCNNM

Being impressed by the effectiveness of MRCG on the issue of multiple-shot human re-identification, we embed it into the SCNNM model, resulting in our proposed "RSCNNM" method which makes use of the merits of both methods. The general steps of RSCNNM are in the following: since multiple-shot images for the same person within the same camera are treated to stay in one set, firstly, the covariance-based representation is extracted by MRCG for each set; then, with these representations, the SCNNM dissimilarities between query sets and corpus sets are measured in the Riemannian space; after that, set based ranking is carried out according to these dissimilarities.

4.3.3 Experiments and Results

We demonstrate our proposed method RSCNNM on public benchmark datasets: ETHZ [10], i-LIDS-MA [40], and i-LIDS-AA [40].

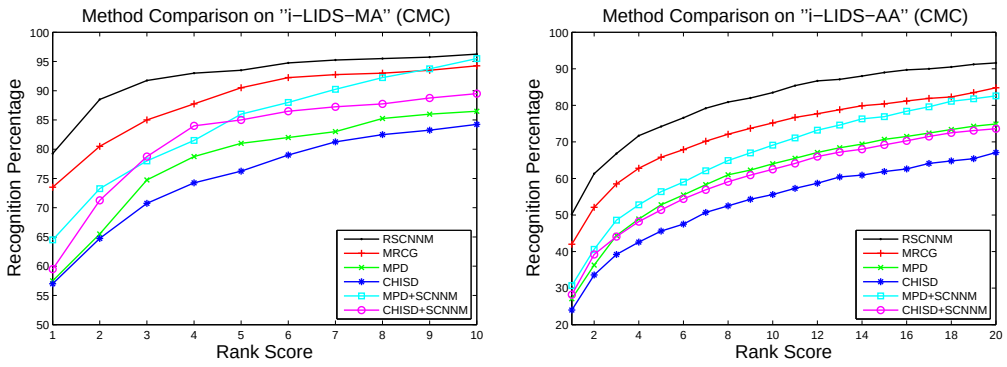


Figure 4.2: CMC performance comparison on the i-LIDS-MA and i-LIDS-AA datasets.

We use all the persons in each mentioned dataset. We normalize all the images into 64×192 pixels. We repeat it for 10 times cross-validation and average the results for evaluation. Experimental results are illustrated by the CMC curves, which represent the expectation of finding the correct match in the top several matches. For each person, we randomly select 10 images each time. And for fairness, in each time, we use the same selected data for method comparison. For each image, the covariance grid size is set to 16×16. The "Fixed-number" n and the trade-off parameter λ of SCNNM are suggested as "$n = 1$" and "$\lambda = 1$", because each person only has two sets (one query set and one corpus set from two different cameras).

We compare our proposed method RSCNNM with MPD [11], CHISD [5], and MRCG [40]. For MPD and CHISD, we adopt the discriminative vector descriptor concatenated of DSCH [8, 51], SFB, and GT [16, 59, 50] in Euclidean space, which has been widely used by state-of-the-art methods. In RSCNNM, we use the same feature representation designed in MRCG to demonstrate the advantage of using SCNNM, which can tackle the non-uniform distribution problem and the asymmetric ranking problem of sets in the Riemannian space. Meanwhile, for conviction, we test the reliability and robustness of SCNNM by its collaboration with MPD and CHISD, denoted by "MPD+SCNNM" and "CHISD+SCNNM", respectively.

As a result, on ETHZ1, RSCNNM obtains 99.04% recognition rate on Rank-1; on ETHZ2 and ETHZ3, RSCNNM gets perfect results on Rank-1. These results are better than any other reported ones. The experimental results on iLIDS-MA and iLIDS-AA are drawn in Figure 4.2. Obviously, the proposed method RSC-

NNM outperforms the state-of-the-art methods. In greater detail, SCNNM has satisfactory collaboration with not only MRCG in the Riemannian space, but also MPD and CHISD in Euclidean space. Although MRCG still performs better than MPD, CHISD, MPD+SCNNM, and CHISD+SCNNM, by performing SCNNM in the Riemannian space, our proposed RSCNNM greatly prevails over MRCG, showing that it enhances the power of covariance-based representation by integrating the set-level neighborhood information.

4.4 SCNNM in Hilbert Space

Multiple-shot human re-identification tackles the problem of judging the reappearance of the person by using image sets acquired from distributed cameras. Despite the value of this direction, how to build the correct correspondence between sets remains challenging due to the real-world complexities. Last section has presented an idea to cope with the multiple-shot problem by condensing human image sets into imaginary points based on covariance descriptors in the Riemannian space. This method has obtained inspiring results, however, delving deep into this approach, we find covariance descriptors are not regular, and there are few learning tools compatible with them, which to a certain degree limits the further development of RSCNNM.

Thereout, this section returns to relying on vectorial features in the familiar Hilbert space to measure the set-to-set distance regarding the within-set distribution information while depressing the negative impact of outliers and intruders. We creatively propose a novel perspective to resolve the multiple-shot re-identification problem, named “BRIA”, which integrates the advantages of two levels of complementary relative information in an effective, efficient, and elegant manner. Concretely, BRIA measures the set-level relative dissimilarity presented by SCNNM, which incorporates the set-level neighborhood structure information, in the discriminative Hilbert space constructed by the sample-level relative feature instantiated by TPCR. The integration of these dual relative information will have generated the significant performance improvement during experiments.

4.4.1 Third-Party Collaborative Representation

TPCR gains large performance enhancement for multiple-shot human re-identification. TPCR relies on the reconstructed coefficients from collaborative representation [55], which evolves from sparse representation [47].

Collaborative representation inclines to use a few words in the dictionary to represent each sample. TPCR uses the third-party data as the dictionary. For each sample, TPCR compacts a kind of relative information referring to the words in the dictionary into a feature vector. Those words with large weights tend to characterizes a kind of neighborhood information on the sample level. The steps of TPCR are detailed in Algorithm 4.

Algorithm 4 THIRD-PARTY COLLABORATIVE REPRESENTATION (TPCR)

Require: Dataset X_0 of corpus samples and query samples; third-party dataset X_{tp} of Z_0 classes; regularization parameter μ.

Ensure: A collaborative representation based description $\hat{\beta}'(x)$ for each sample $x \in X_0$ over X_{tp}.

1: Normalize the columns of X_0 and X_{tp} to have the unit l_2-norm.

2: Solve the collaborative representation problem:
$\hat{\alpha} = \arg\min_\alpha \{\|x - X_{tp}\alpha\|_2^2 + \mu\|\alpha\|_2^2\}$,
with a closed-form solution $\hat{\alpha} = P_{tp}x$, where $P_{tp} = \left(X_{tp}^\top X_{tp} + \mu \cdot I_u\right)^{-1} X_{tp}^\top$, and I_u is the unit matrix. Note that P_{tp} can be pre-computed once X_{tp} is given.

3: Compute the summed coefficients within each class:
$\hat{\beta}_i(x) = \sum_{j=1}^{n_i} \hat{\alpha}_{ij}, \forall i \in \{1, \ldots, Z_0\}$, where n_i is the sample number of class i in the third-party data.

4: Normalize $\hat{\beta}(x)$: $\hat{\beta}'(x) = \hat{\beta}(x) / \sum_{i=1}^{Z_0} \left|\hat{\beta}_i(x)\right|$, and return $\hat{\beta}'(x)$.

TPCR introduces the third-party data to enhance the representative power of the dictionary, and further concatenates the reconstructed coefficients into a feature vector with summing the dimensions that belong to the same class.

The effectiveness of TPCR comes from two aspects. One is using the abundant and diverse third-party data as the dictionary. The prepared third-party data covers enough information of pose variations, illumination changes, viewpoint alterations, occlusions, and so forth. These data can help enhance the power of the dictionary used for collaborative representation. The other is using the intra-class sum to compress the class structured feature dimensions. Such compression glues together the feature vector entries that suffer from intra-class variations, so the intra-class variations of the represented samples will be reduced. Benefitting from the intra-class information compression and the third-party data dictionary,

TPCR is robust to the real-world complexities of human image data. Moreover, as expected, the performance of TPCR largely outstrips original features under traditional set-to-set distances for the re-identification tasks [49].

TPCR feature provides a kind of relative information on the sample level referring to the words represented by the basic features in the dictionary. The subjectiveness in the basic feature representation and dictionary building will inevitably lead to the limited discriminative power of the TPCR feature space, in which intra-class dissimilarities may still be larger than inter-class dissimilarities for some samples. Thus, traditional set-to-set distances such as MPD [11] and CHISD [5] in the TPCR feature space are still far from being perfect. Though being impressive, either MPD or CHISD with TPCR has weakness due to the fact that they rely on the measure between only some local parts of the sets. More concretely, MPD depends on the nearest samples between the sets. In this case, outliers of each class may easily influence the measuring reliability. CHISD tries to improve it by considering the distance between convex hulls for the set pair, however, it is unavoidably influenced by the layout of nearest samples between the sets which support the convex hulls. The illustration of MPD and CHISD is shown in Figure 4.3. Such sensitivity may easily cause the asymmetric ranking, which means, a pair of sets usually don't have the same rank order for each other in their own set-level Rank- Order lists. Thus, it is unfair to judge the rank order only considering one side of them.

Based on the robust CHISD, we propose to use SCNNM to explore the relative information among sets instead of samples towards the multiple-shot human re-identification problem. When most sets of the same class stay closer to each other than those from different classes, the sets within the same class will share more common-near-neighbor sets than those from different classes. SCNNM utilizes such kind of information to further ensure inter-class dissimilarities to be larger than intra-class dissimilarities for all sets instead of samples.

4.4.2 Collaboration of TPCR and SCNNM

In this section, we proposed to take advantage of the within-set distribution information when measuring the low level set-to-set distance upon MPD or CHISD. Meanwhile, we expect to prevent intruders and outliers which may degrade the low-level set-to-set distance. Due to samples in the same set share the one label, we propose to select a discriminative feature space, in which for all samples intra-class distances are smaller than inter-class distances. And TPCR meets such requirement.

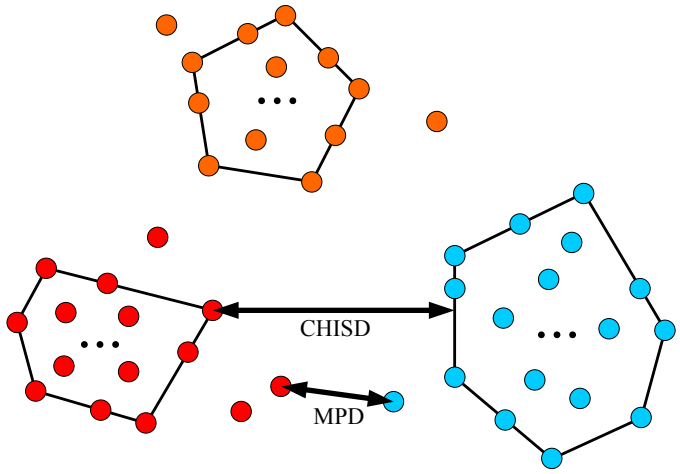


Figure 4.3: Illustration of MPD and CHISD. For the set pair, MPD considers the minimum distance between points, while CHISD concerns the minimum distance between convex hulls.

In the TPCR feature space, SCNNM overcomes the weakness of sensitivity for MPD and CHISD. This dissimilarity incorporates the relative information on the set level, which complements the relative feature TPCR on the sample level. Such complementarity can be understood from two aspects. Firstly, sample-level and set-level information are simultaneously considered; secondly, the dissimilarity measurement is suitable for the feature representation, whereby the subjectiveness of TPCR is indirectly reduced, thus leading up to a remarkable performance for BRIA.

In procedure, firstly, BRIA extracts TPCR feature for each sample over the third-party data; then, since a group of TPCR features from the same class are treated within one set, set-to-set dissimilarities are measured by SCNNM. The steps of BRIA are presented in Algorithm 5, in which, step 1 to 2 belong to TPCR feature mapping, and step 3 to 5 belong to SCNNM dissimilarity measure.

BRIA is different from conventional methods for multiple-shot human re-identification. Technically, BRIA considers for the issue from two levels: sample level and set level, while most of traditional methods focus one aspect. Methodologically, BRIA addresses the relative information, which is rarely concerned by current existing methods. Relative information considers the neighborhood topological information by encoding the relationship between the concerned sam-

Algorithm 5 BI-LEVEL RELATIVE INFORMATION ANALYSIS (BRIA)

Require: The labeled third-party dataset X_{tp}; the testing data of corpus sets X_cs and query sets X_qs.

Ensure: Ranking Y^{ranking} of all X_cs for each X_q.

1: Perform TPCR algorithm using X_{tp}s as the dictionary to acquire the feature function P_s.
2: Project X_qs and X_cs into the TPCR feature space by P_s as $\mathbf{X}_q$s and all $\mathbf{X}_c$s, respectively.
3: List all $\mathbf{X}_c$s and $\mathbf{X}_q$s together into $\mathbf{X}$s.
4: Sort $\mathbf{X}$s by CHISD to acquire the rank order lists for each $\mathbf{X}_c$ and $\mathbf{X}_q$.
5: Measure SCNNM dissimilarity H^{SCNNM} between each pair of $\mathbf{X}_c$ and $\mathbf{X}_q$ based on their rank order lists.
6: Re-rank all $\mathbf{X}_c$s according to H^{SCNNM}s calculated in step 5 for each $\mathbf{X}_q$ to return Y^{ranking}.

ple/set and other several distributed samples/sets. It is robust especially when the size of each class is not large and samples/sets themselves are non-uniform distributed. Taking advantage of the collaboration between feature representation and dissimilarity measurement, BRIA enhances the performance as far as possible, which will also be experimentally demonstrated latter.

The proposed method BRIA has some limitations as well. It requires the third-party data to build a dictionary for TPCR. Thus, the ability of TPCR is inevitably influenced by the quality of the dictionary. Currently, there is no optimization method on dictionary selection to maximize the ability of TPCR. Even so, TPCR has promising effectiveness with the recommended dictionary [49]. Furthermore, the SCNNM dissimilarity in BRIA is based on rank order lists that are formed by low-level set-to-set distance measurements, thus the capability of it highly depends on the robustness of these measurements. If the low-level set-to-set distance is too sensitive to the noises of each set, the reliability of SCNNM dissimilarity might be reduced, which will give rise to the low performance of BRIA.

4.4.3 Experiments and Results

Experimental Setup

We demonstrate the superiority of BRIA on several public benchmark datasets: ETHZ [10], iLIDS [57], iLIDS-MA [41], and iLIDS-AA [41]. All of them have

multiple images of spatial-temporal variations for each person.

We normalize all the images into 48×128 pixels, and then randomly select 10 images per person for each query set and corpus set, respectively (coming from different cameras if possible). For i-LIDS, each person has at most eight images, so we use them all by one half as the query set and the other half as the corpus set. For persuasiveness, we average the results for ten-fold cross validation with random corpus-query data splitting.

According to the current research situation, we suggest the basic feature for TPCR to be concatenated by DSCH, SFB, and GT [39], to capture the color and texture information. By contrast with the relative feature TPCR, this concatenated original feature, denoted by "Ori", can be considered as a kind of absolute feature, which is also valuable for demonstrating and comparing.

We emphasize the flexibility of the third-party data for building a descriptive and representative dictionary. Although the dictionary selection based on optimization seems more mathematically strict, it is not the focus of this chapter. According to [49], even some heuristic selection of the third-party data as the dictionary would not reduce the capability of TPCR. So we just follow the suggestions in [49], and after some trials, we recommend the third-party data for each ETHZ dataset to be the rest two similar ETHZ datasets together with a very different dataset i-LIDS-AA. Taking ETHZ1 for example, we use ETHZ2, ETHZ3, and i-LIDS-AA together, denoted by "ETHZ2+ETHZ3+i-LIDS-AA". We also recommend the third-party data for i-LIDS to be "ETHZ3+i-LIDS-MA+i-LIDS-AA", for i-LIDS-MA to be "i-LIDS-AA", and for i-LIDS-AA to be "i-LIDS-MA". If there are labeled data from the same dataset as the corpus belong to, we may also involve these data together with the third-party data in the dictionary. In the process of dictionary building, we limit the image number to be no larger than 46 for each person, which is the largest class size in i-LIDS-MA. Honestly, in our experiments, the third-party data selection are not guaranteed to be the best, but it will not influence the validation of the effectiveness of BRIA. Moreover, such tolerance to the flexible representation of TPCR feature mirrors the stability and reliability of BRIA to a certain degree.

Parameter Discussion

The Fixed-number N may influence the performance of the symmetric term in SCNNM. To show the robustness of BRIA, we display the results by changing N for both single-set vs. single-set cases and multiple-set vs. single-set cases. In the single-set vs. single-set cases, there is one set on the query side and one

set on the corpus side for each person. In the multiple-set vs. single-set cases, there are multiple sets on the query side and one set on the corpus side for each person. Results are evaluated by MRR scores. Here, p denotes the class number, S denotes the set number in each class, and q denotes the average sample number per set.

Table 4.1: MRR scores (%) with different Ns for BRIA in single-set vs. single-set cases.

i-LIDS	$N=1$	$N=5$	$N=10$	$N=15$	$N=20$
$p=30, 1\leq q\leq 4$	**76.97**	69.32	67.55	64.43	64.78
$p=70, 1\leq q\leq 4$	**66.25**	59.62	54.29	53.78	53.11
$p=119, 1\leq q\leq 4$	**60.07**	54.92	49.33	47.67	47.25
i-LIDS-MA	$N=1$	$N=5$	$N=10$	$N=15$	$N=20$
$p=20, q=10$	**99.25**	90.33	80.84	80.92	75.37
$p=40, q=10$	**98.02**	88.99	83.78	81.21	80.50
i-LIDS-AA	$N=1$	$N=5$	$N=10$	$N=15$	$N=20$
$p=30, q=10$	**71.01**	65.65	59.12	57.10	55.66

Table 4.2: MRR scores (%) generated by tuning Λ for BRIA in single-set vs. single-set cases.

Λ	i-LIDS $p=30$	i-LIDS $p=70$	i-LIDS $p=119$	i-LIDS -MA $p=20$	i-LIDS -MA $p=40$	i-LIDS -AA $p=30$
0	76.97	66.25	60.07	99.25	98.02	71.01
0.001	77.13	66.47	60.27	99.25	98.02	71.21
0.01	77.13	66.47	60.27	99.25	98.02	71.21
0.1	77.13	66.47	60.22	99.25	98.02	71.64
1	**77.75**	66.91	60.69	99.25	**98.04**	**72.40**
10	77.59	**67.54**	**60.95**	**99.50**	97.68	71.80
100	77.58	67.43	60.75	99.50	97.54	71.46
1000	77.58	67.43	60.75	99.50	97.54	71.46
∞	73.31	63.31	56.49	98.42	96.72	70.87

For the single-set vs. single-set cases, from Table 4.1, we can see the overall decreasing trend of the results when N grows, and the symmetric term achieves

its best performance with $N = 1$. With a suitable N, we further study the balancing parameter Λ of the model, as shown in Table 4.2. Generally, joining the symmetric term and asymmetric term for SCNNM can bring a better performance. In most cases, having $\Lambda \in [1, 10]$ is a good choice. Even so, tuning Λ does not substantially change the results, which shows the stability of SCNNM as well.

In order to explain the importance of the Fixed-number N, we further evaluate this parameter on the multiple-set vs. single-set cases with different set numbers in each class. Experiments are carried out on i-LIDS-MA, because this dataset is not only challenging, but also has enough samples to conduct experiments for the multiple-set vs. single-set cases. By contrast, the results on ETHZ are saturated, and the image number for each identity in i-LIDS and i-LIDS-AA are not enough to satisfy the required experimental condition. We simply separate this dataset into several sets without separating the data from different cameras. Results are displayed in Table 4.3. Obviously, when the Fixed-number N is set to approximate half of the set number in each class, the symmetric term achieves its best performance. This phenomenon not only shows the importance of the Fixed-number N, but also supports the recommendation for it. Besides the Fixed-number N, we also test the balancing parameter Λ, as described in Table 4.4. Generally, being coincident with the results in single-set vs. single-set cases, the proposed modeling plays its best performance with $\Lambda \in [1, 10]$ for multiple-set vs. multiple-set cases as well.

Note that, the discussion and recommendation of the "Fixed-number" parameter and the trade-off parameter settings are within the scope of our examined data.

Table 4.3: MRR scores (%) with different Ns for BRIA in multiple-set vs. single-set cases.

i-LIDS-MA	$N = 2$	$N = 4$	$N = 6$	$N = 8$	$N = 10$
$S = 4, q = 5$	**93.38**	92.94	90.69	87.95	86.18
$S = 8, q = 5$	94.64	**94.95**	94.64	94.42	93.17
i-LIDS-MA	$N = 1$	$N = 3$	$N = 5$	$N = 7$	$N = 9$
$S = 2, q = 5$	**93.76**	89.53	85.47	81.46	80.62
$S = 6, q = 5$	93.58	**94.23**	93.92	92.85	91.51
$S = 10, q = 5$	94.44	94.45	**94.55**	94.46	94.18

Table 4.4: MRR scores (%) generated by tuning Λ for BRIA in multiple-set vs. single-set cases.

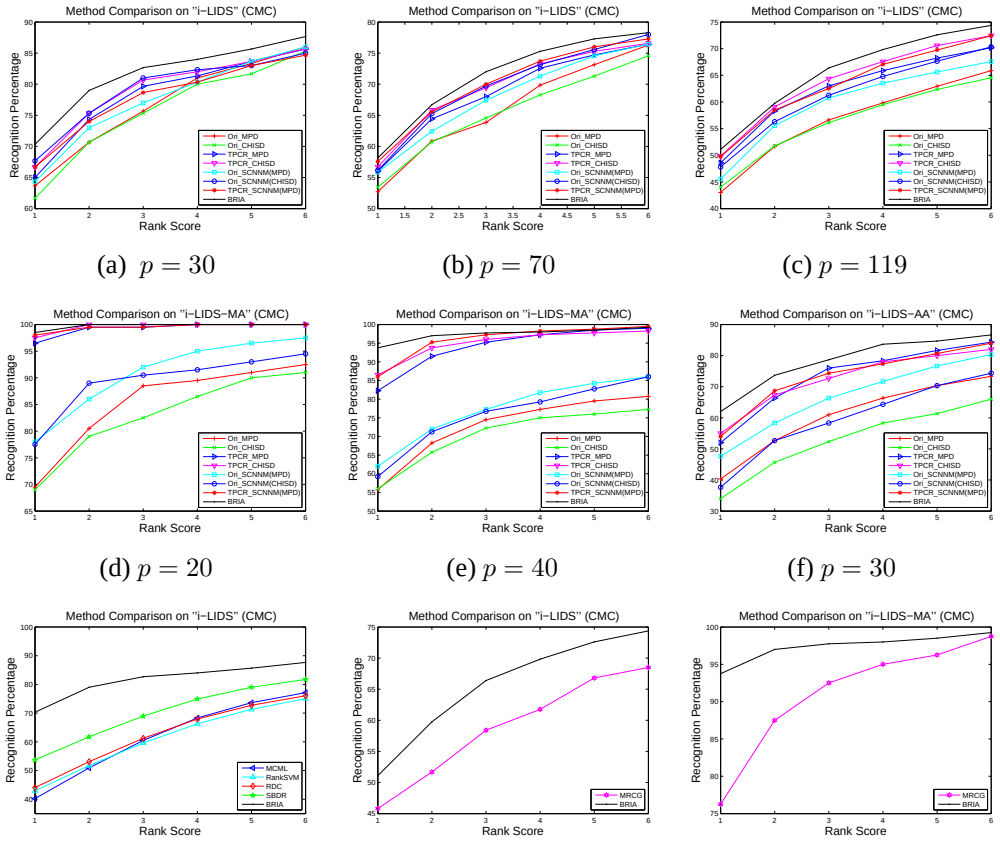
Λ	$S=2$, $q=5$	$S=4$, $q=5$	$S=6$, $q=5$	$S=8$, $q=5$	$S=10$, $q=5$
0	93.76	93.38	94.23	94.95	94.55
0.001	93.66	93.34	94.24	94.96	94.57
0.01	93.66	93.34	94.24	94.96	94.58
0.1	93.66	93.35	94.30	95.01	94.57
1	93.82	**93.55**	**94.48**	**95.18**	**94.88**
10	**94.06**	93.29	93.89	94.52	94.29
100	94.05	93.08	93.47	93.78	93.37
1000	94.05	93.08	93.47	93.77	93.36
∞	92.81	92.10	92.94	93.32	93.04

Method Comparison

To show the advantage of BRIA, which is also denoted by "TPCR_SCNNM(CHISD)", we compare it with typical related methods for human re-identification, including original feature and TPCR feature under different set-to-set distances like MPD and CHISD. In order to further validate the capability of SCNNM dissimilarity itself, we conduct experiments on its cooperation with several possible combinations of features and low-level set-to-set distances, such as "Ori_SCNNM(MPD)", "Ori_SCNNM(CHISD)", and "TPCR_SCNNM(MPD)". Moreover, we demonstrate the capability of our method by comparing with typical state-of-the-art methods as well, including the unsupervised method MRCG, and the supervised methods MCML [13], RankSVM [36], RDC [59], and SBDR [50]. Because each person only has two sets (one query set and one corpus set), the "Fixed-number" N and the balancing parameter Λ for SCNNM dissimilarity are suggested as "$N=1$" and "$\Lambda=1$", respectively. Experimental results are illustrated by the CMC curve, which visualizes the expectation of the correct match at each rank based on the ranking of each of the corpus with regard to the query [36].

Results are illustrated in Figure 4.4 except those on ETHZ, because for ETHZ1, ETHZ2, and ETHZ3, BRIA approaches 100% recognition rate on Rank-1 of CMC for all persons, which are superior to any other state-of-the-art methods. In Figure 4.4, p denotes the number of persons. For i-LIDS, we test on 30 persons, 70 persons, and 119 persons; for i-LIDS-MA, we experiment on 20 persons and 40 persons; for i-LIDS-AA; we demonstrate on 30 persons. Overall, significantly,

BRIA outperforms all the other concerned methods, as the evidence for the effectiveness of complementarity between the sample-level relative feature TPCR and the set-level relative dissimilarity SCNNM.



(a) $p = 30$ (b) $p = 70$ (c) $p = 119$

(d) $p = 20$ (e) $p = 40$ (f) $p = 30$

(g) $p = 30$ (h) $p = 119$ (i) $p = 40$

Figure 4.4: CMC performance comparison on the i-LIDS, i-LIDS-MA, and i-LIDS-AA datasets.

The third-party data based dictionary covers kinds of variations, occlusion, and localization errors for different persons. Benefited from such a dictionary, TPCR can be robust to these similar variations, occlusion, and localization errors. For example, i-LIDS-AA covers the viewpoint variation, illumination variation, pose variation, occlusion, and localization errors, as shown in Figure 1.5. If it is used as a dictionary, it may help to handle such difficulties in another similar dataset i-LIDS-MA. Even if TPCR feature cannot guarantee the enough ability to

satisfactorily discriminate the person image sets, SCNNM will further exploit set-level common-near-neighbor information to make up for TPCR, so as to ensure intra-class dissimilarities are smaller than inter-class dissimilarities for all sets.

Furthermore, we can see TPCR_SCNNM(MPD), which is also the collaboration of two levels of relative information, but joined by a sensitive low-level set-to-set distance MPD, cannot work as well as BRIA which relies on CHISD, though TPCR_SCNNM(MPD) has potentiality to enhance the performance compared with other analogues and competitors. It justifies the argument that SCNNM is not only necessarily dependent on but also inevitably influenced by the robustness of low-level set-to-set distance measure. CHISD is more robust to noisy outliers, so it can provide a better platform for SCNNM.

When the person number p increases, though in i-LIDS, the original feature collaborating with SCNNM may be competitive with the TPCR feature collaborating with SCNNM, our proposed BRIA still have significant advantages, and such advantages are especially remarkable on i-LIDS-MA and i-LIDS-AA. From another perspective, we can clarify the strong adaptability of our proposed SCNNM dissimilarity according to the facts that SCNNM dissimilarity not only works very well with the TPCR feature, but also has good cooperation with the original feature. Analytically, original feature can be seen as a kind of sample-level absolute feature in a sense, potentially complementary to the set-level relative information as well.

We can also see that MPD performs better than CHISD in some results on i-LIDS-MA and i-LIDS-AA, especially in case of the original features. MPD depends on the nearest sample points between the sets. Therefore, outliers of each class may easily influence the measuring reliability. CHISD tries to improve it by considering the distance between convex hulls for the set pair. However, it is unavoidably influenced by the layout of nearest sample points between the sets which support the convex hulls.

Actually, the convex hulls play a role to produce other interpolated points on them. The distribution of these interpolated points is determined by the existing sample points in the feature space. If the feature space is discriminative, existing sample points will be well-distributed. In this situation, the interpolated points on the convex hulls will be reliable. Otherwise, existing sample points will be unsatisfactorily distributed. On this case, the convex hulls will be unreliable, and the interpolated points will bring more noises. Obviously, as demonstrated, TPCR can provide a more discriminative feature space than the original one. Therefore, in the TPCR feature space, CHISD can bring its superiority into play. But in the original feature space, CHISD loses its effect to a certain level, so that it may be

outperformed by MPD. As for i-LIDS, sample points per set are very few and unsatisfactorily distributed in the original feature space. Thus, CHISD cannot either exploit its advantages or interpolate more noisy points, so will stay at a similar capability level to MPD.

In the i-LIDS dataset, there are less than 10 images per person. Because the expected performance of MRCG relies on the enough number of images per person to effectively extract the Karcher mean based covariance descriptors to condense the within-class correlations, image number per person in i-LIDS stays as a bottleneck for MRCG. However, our method remarkably outperforms it. Indeed, adequate image number per person can display the superiority of our method, but even when the image number per person is comparatively small, the usage of relative information can offset the degrading of performance to some extent as well.

MCML, RankSVM, RDC, and SBDR are supervised approaches. The performance of them are unavoidably influenced by whether training samples and testing samples are independently and identically distributed, which cannot be ensured always in fact. Consequently, in i-LIDS, the insufficiency and complexity of person images in each class limit the performance of these methods. In i-LIDS-MA, it is difficult to carry out training and testing because there are merely 40 persons together. If we randomly split it into training data and testing data, too few persons for training will easily cause overfitting, and too few persons for testing will be unconvincing for method comparison as well. By contrast, BRIA doesn't require implementing learning using extensive training samples which need the matched people to be tediously annotated across camera views in the real scene. Taking advantage of the third-party data as the dictionary to represent the relative feature TPCR on the sample level, and making use of rank order lists to model the relative dissimilarity SCNNM on the set level, BRIA has obvious superiority to all the concerned methods.

4.5 Summary

This chapter has proposed two novel methods, RSCNNM and BRIA, to exploit the set-level common-near-neighbor information in discriminative spaces for multiple-shot human re-identification. In RSCNNM and BRIA, SCNNM is designed to deliver the effectiveness of the well-distributed sets to the badly-distributed ones in the Riemannian space constructed by MRCG and the Hilbert space constructed by TPCR, respectively. Extensive experiments on standard datasets have shown their significant superiority to the related state-of-the-art competitors. Ultimately,

we attach the description of RSCNNM and BRIA in Figure 4.5.

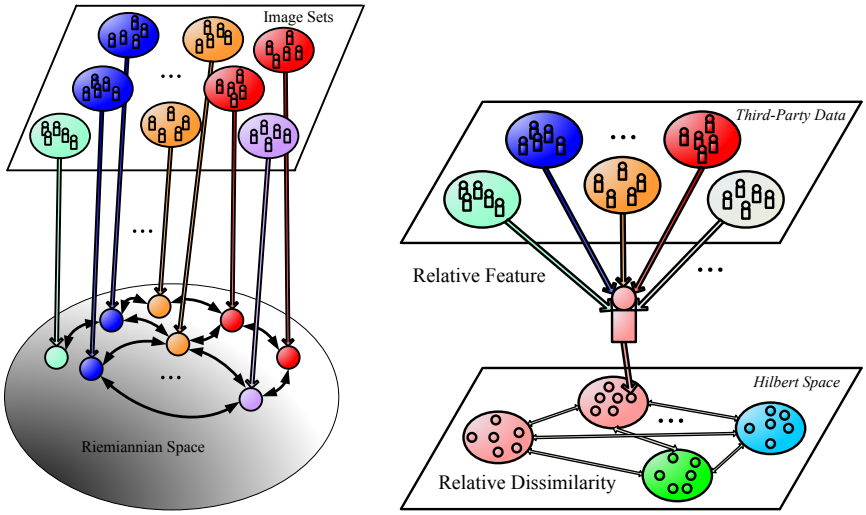


Figure 4.5: Depiction of RSCNNM and BRIA. For RSCNNM (left), image sets are projected into the points in the Riemannian space for SCNNM dissimilarity measure; for BRIA (right), each relative feature TPCR will be represented by the third-party data, and all relative dissimilarities SCNNM will be measured between sets in the Hilbert space.

Chapter 5

Locality Based Discriminative Measure

5.1 Introduction

To address the multiple-shot re-identification problem, Chapter 4 has proposed the RSCNNM and BRIA methods. In these methods, SCNNM is designed to exploit the set-level common-near-neighbor information in two discriminative feature spaces constructed by MRCG and TPCR. However, MRCG requires adequate and typical within-set images to characterize the densely tiled grids on human appearance, and TPCR requires diverse and labeled dictionary data to cover plentiful prototypes and variations of human appearance. Both of them are expensive in fact.

We expect to avoid these expensive requirements for the specific single-set vs. single-set re-identification case. This chapter will propose a novel set based matching model referred to as "Locality Based Discriminative Measure (LBDM)". LBDM comprises three primary steps: set-to-set distance crafting, local metric field constructing, and set based matching. The first step involves designing a novel set-to-set distance, which accumulates the local point based minimum approach distance between human image sets, and then, in the second step, local metric learning is intended to pull closer together the sample points of the same set than those from different sets to ensure a more reliable set-to-set distance in the local metric field space, so that, in the final step, the SCNNM dissimilarity can be fully leveraged for effective matching.

5.2 Locality Based Discriminative Measure

5.2.1 Set-to-set Distance Crafting

To correctly match the sets, an opportune set-to-set distance is important. Most previous methods have spotlighted minority-based distance, while claiming the effectiveness of this strategy.

Minority-based distance takes within-set variation into account by measuring the closest local minorities of each paired sets. Two exemplary methods are MPD [11] and CHISD [5]. MPD measures the minimum point-wise distance between sample points from two arbitrary sets in Euclidean space. It is susceptible because outliers of each set may easily disturb the distance measure and greatly change the results. CHISD is intended to improve upon it by calculating the distance between convex hulls of two sets instead of finding the closest pair of points from them. However, it is unavoidably vulnerable to the layout of point minorities that support the convex hulls.

Minority-based distance focuses on the difference of sample points within the set. On the other hand, majority-based distance, such as Average Point-wise Distance (APD), accounts for the majorities of point-wise distances between sets. This type of global distance pursues robustness to the small number of irregular outliers in each set by averaging. Nevertheless, it is incapable of preserving the information of within-set difference; it therefore cannot perform as well.

We thus propose a simple but effective set-to-set distance for the issue of multiple-shot human re-identification. It inherits the advantages of both minority-based and majority-based distances while overcoming their drawbacks. It is referred to as the "Mean Approach Distance (MAD)", because it collects the mean point-to-set approach into the set-to-set distance measure. As will be shown, such a design can balance the discriminability of minority-based distance and the robustness of majority-based distance.

Let us denote a to be an arbitrary point in set A, b to be an arbitrary point in set B, l to be the point-to-point distance, and D_{p2s} to be the point-to-set distance. Thus, considering symmetry, D^{MAD} is given by:

$$D^{\text{MAD}}(A,B) = \frac{1}{|A|}\sum_{a\in A} D_{p2s}(a,B) + \frac{1}{|B|}\sum_{b\in B} D_{p2s}(A,b). \tag{5.1}$$

where

$$D_{p2s}(a,B) = min\{l(a,b)|b\in B\}, \tag{5.2}$$

$$D_{p2s}(A,b) = min\{l(a,b)|a\in A\}. \tag{5.3}$$

In Equation 5.1, $|\bullet|$ means the cardinality of a set. In Equation 5.2 and 5.3, l is chosen to be the l_2-norm. Equation 5.1 is a global measure covering all of the sample points from two sets, and Equation 5.2 and 5.3 consider the local minority of sample points in each set with regard to samples in another set.

Although the recommended minority based distance might not be the best among all the possibilities, it fits the proposed MAD measure well and has been proved efficacious, without loss of MAD's generality.

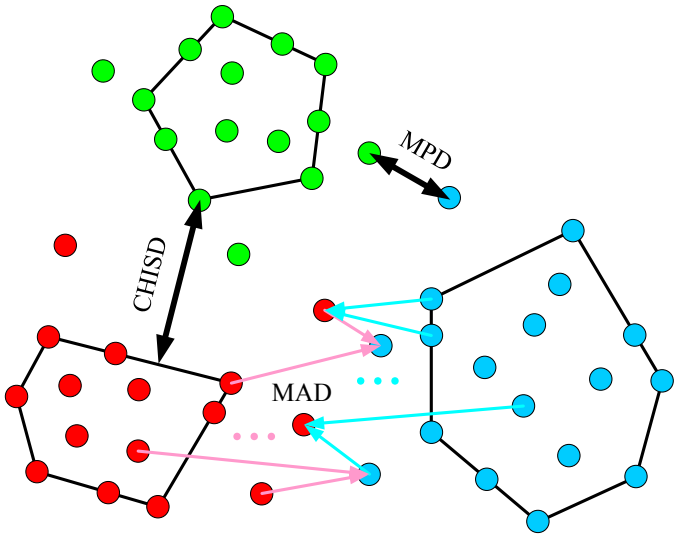


Figure 5.1: MPD, CHISD, and MAD visualization. Set classes are distinguished by colors; convex hulls are shown by polygons; set-to-set distances are denoted by two-way arrows; and point-to-set distances in Equations 5.2 and 5.3 are marked by one-way arrows. It is difficult to directly visualize MAD; we therefore draw some point-to-set distances that MAD consists of to make it understandable.

A visualization of MPD, CHISD, and MAD is provided in Figure 5.1.

5.2.2 Local Metric Field Constructing

MAD provides a proper way to adapt the point based distance to the set level in the feature space. To calculate a discriminative MAD, a suitable metric space is also indispensable. As the baseline, the Euclidean metric suffers from the possible weakness of heuristic feature representation. By contrast, a learned Mahalanobis metric has been proven much superior to it [46, 51, 59, 50].

This chapter discusses a set based matching problem. For each set of images of individuals within the same camera, the class label (human identity) has been fixed, although unknown, for all the sample points belonging to it, and different from those in the other sets. Under this precondition, we propose a new local metric learning model to accommodate the local variations for all sets of samples. We hypothesize that each set remains in a local metric field space, and suggest to learn a possibly different local metric for each set rather than a unique global one. The local metric has been proved very effective because of its case-sensitive discriminative power [37]. To the best of our knowledge, to date few works have applied local metric learning to the issue of human re-identification because of complexity of real-world situations and limitations of sample size.

Our local metric field constructing consists of the following steps: neighborhood determining, metric learning, and distance measuring. In neighborhood determining,we determine the set-level neighborhood in each camera by utilizing the new MAD measure. Because there is only one image set for each identity and the within-set samples are relevant, this neighborhood will decide the irrelevant samples for learning the local metric. In metric learning, we employ the capable metric learning method OMRR [51] to exploit the potentiality of MAD. In distance measuring, we seek for a suitable way to match the sets between cameras, by treating all the learned local metrics as the points on a Riemannian manifold.

We have argued the merits of MAD. Although it is representable for informative inliers and immune to irregular outliers, unexpected set distribution may still weaken its reliability. A desired set based discriminative measure will satisfy the requirement that intra-class set-to-set distances should be smaller than inter-class set-to-set distances. Here, we reconsider this requirement by the sample level, because when the situation of intra-class compactness and inter-class separation is improved for sample points, the set they compose will be easier to discriminate.

Before local metric learning, the neighborhood size needs to be determined for every query set and corpus set, respectively. We do this by selecting the predefined number of nearest neighbor sets by measuring MAD with an Euclidean metric. Here, for each set in the query and corpus sides, we treat samples within it to be relevant, and samples in its neighborhood to be irrelevant, respectively. Point-to-point distance $l(a, b)$ in Equations 5.2 and 5.3, with a learned metric matrix w, can be expressed as $l_w(a, b) = (a - b)^\top w(a - b)$. We expect the learned metric matrix to be capable of improving the discriminability of MAD. Among the existing Mahalanobis metric learning models for this issue, OMRR is a sound method [51]. The framework of OMRR can refer to Equation 2.11

After constructing the local metric fields, the next task is determining how to

measure the distance between paired sets. Because all learned local metrics are positive definite matrices, we can deduce that they are on a differentiable manifold a natural Riemannian structure [3]. The pair of query and corpus sets belonging to the same class will remain close in the feature space; therefore, their local metrics will also tend to remain nearby on the Riemannian manifold. Let w_A and w_B be the learned local metrics of two arbitrary sets A and B, respectively. Then, there exists a unique geodesic joining of A and B. This geodesic has a parametrization as below:

$$\gamma = w_A^{1/2}(w_A^{-1/2}w_B^{1/2}w_A^{-1/2})^t w_A^{1/2}, \tag{5.4}$$

where $0 < t < 1$. In compromise, we use the midpoint ($t = 0.5$) of the geodesic γ joining w_A and w_B on the manifold as a new metric matrix. This metric matrix can be used to isomorphically project the pair of query and corpus sets from the source space to the target space. In the projected space, we can re-measure MAD between the set pairs. To speed up the computation, such a midpoint can also be directly approximated by the algebraic average of the metric matrix pair.

5.2.3 Set Based Matching

Algorithm 6 LOCALITY BASED DISCRIMINATIVE MEASURE (LBDM)

Require: Testing data of corpus sets X_cs and query sets X_qs; local neighborhood size L_n.

Ensure: Ranking Y^{ranking} of all X_cs for each X_q.

1: Measure MAD between each X_q and all the other sets in the query side with a Euclidean metric.
2: Determine local neighboring sets from different classes for each X_q with a predefined L_n.
3: Use OMRR to learn the metric for every query set in its local area to construct the local metric field for all query sets.
4: Construct the local metric field for all corpus sets similarly to step 1-3.
5: Measure SCNNM dissimilarity H^{SCNNM} in the learned local metric field to obtain the dissimilarity between each pair of X_q and X_c.
6: Re-rank all X_cs according to H^{SCNNM}s calculated in step 5 for each X_q to return Y^{ranking}.

The discriminability of MAD is enhanced by studying the local metric field. To further reduce erroneous matches between cameras, this ability can be addi-

tionally improved by studying the common neighborhood information for all sets, which is formulated by SCNNM.

.

SCNNM, given by Equation 4.1, can deliver the effectiveness of the well-distributed sets to the badly-distributed ones in a discriminative space, which has been studied by MAD and LMF in this chapter. The steps of LBDM are described in Algorithm 6.

In sum, the idea of LBDM is to incorporate three layers of locality information for a discriminative measure. In MAD, the average of the collected point-to-set approach distances de-noises the locality information between sets. In LMF (Local Metric Field), the neighboring sets participating in metric learning optimize the locality information for each set. In SCNNM, the "Fixed-number" quantize the locality information for paired sets in their neighborhood structures. Their integration will lead to the anticipated performance improvement, which will have been experimentally demonstrated.

5.3 Experiments and Results

5.3.1 Experimental Setup

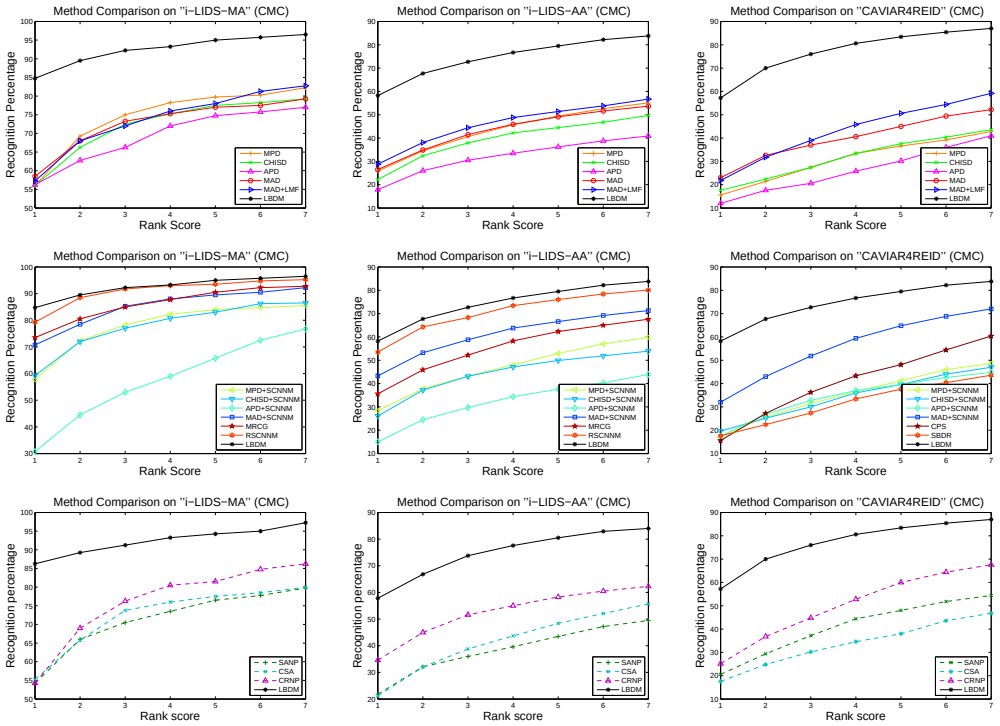
We set up experiments on several public benchmark datasets, including ETHZ1, ETHZ2, ETHZ3 [10], i-LIDS-MA, i-LIDS-AA [40], and CAVIAR4REID [6].

We normalize all the images into 48×128 pixels and randomly select 5 (CAVIAR4REID) or 10 (i-LIDS-MA and i-LIDS-AA) images per person for each query and corpus set, respectively (produced from different cameras, if possible). For evaluation, we perform ten-fold cross-validation. We adopt the reliable feature for our modeling, which simultaneously encodes the color and texture information by concatenating DSCH [8, 51], SFB, and GT [16, 59, 50]. Although feature representation is not the focus of this chapter, we admit introducing a more effective feature to LBDM could likely result in improved performance. The local neighborhood size L_n in the local metric learning model is set to one-fifth of the number of persons in each camera. This heuristic setting is not guaranteed to result in the best LBDM performance. Because different datasets may have different sample distributions, tuning this parameter may improve the performance, which will be discussed later. The SCNNM parameter settings can be referred in Chapter 4; therefore, we will not detail them here.

The state-of-the-art methods compared here include MPD [11], CHISD [5],

APD, MRCG [40], RSCNNM, Custom Pictorial Structures (CPS) [6], SBDR [50], SANP [19], Collaborative Sparse Approximation (CSA), CRNP [48], and BRIA.

5.3.2 Result Analysis



(a) $p = 40$ (b) $p = 100$ (c) $p = 50$

Figure 5.2: CMC performance comparison on the i-LIDS-MA, i-LIDS-AA, and CAVIAR4REID datasets.

All results are illustrated in Figure 5.2 except those on the ETHZ datasets, because our method approaches a 100% recognition rate on each of them, respectively; it is therefore superior to any other method. In Figure 5.2, p denotes the person number. From the reported results, we can observe that the proposed LBDM overall provides a large performance enhancement compared with the original feature it cooperates with; therefore, it consistently exceeds all competitors. In greater details, MAD shows an advantage over the conventional minority- and majority-based set-to-set distances (MPD, CHISD, and APD) on Rank-1, which

is a significant evaluating indicator for matching. This advantage is especially remarkable on CAVIAR4REID. Moreover, the quality of the low-level distance measure will influence the performance of the high-level dissimilarity measure. Otherwise, the former will deteriorate the latter. In contrast with other distance measures, such as MPD, CHISD, and APD, MAD can serve as a compatible partner to work with the SCNNM high-level dissimilarity measure for the desired result. Furthermore, it is clear that the learned LMF is instrumental to MAD by comparing MAD against "MAD+LMF". In addition, we can also confirm the effectiveness of LMF by comparing our model with and without it, denoted by LBDM and "MAD+SCNNM", respectively. A series of experiments clearly show that LMF contributes to a substantial performance improvement. Furthermore, when competing with the recently well-performed methods, SANP, CSA, and CRNP, LBDM shows obvious advantages.

We then evaluate LBDM by changing the local neighborhood size L_n. Since LMF is learned within each camera, L_n determines the number of irrelevant samples when learning the local metric for each set. Results are detailed in Table 5.1 by the MRR scores. We test different size ratios of the local neighborhood from 0.1 to 1 with the step size of 0.1 of the total person number in each camera, respectively.

Table 5.1: LBDM locality evaluation with different neighborhood size ratios in terms of MRR scores (%).

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
i-LIDS-MA	86.08	89.27	89.88	89.94	**90.61**	90.55	89.43	90.00	90.20	90.52
i-LIDS-AA	66.58	70.92	67.96	**73.25**	69.19	68.97	69.38	68.36	69.26	66.52
CAVIAR4REID	65.26	68.98	**69.40**	68.87	68.05	67.57	67.77	67.67	67.75	68.44

As shown in Table 5.1, overall, LBDM has a good performance, but the local neighborhood size L_n has an influence on the performance of it. This shows the importance of local neighborhood size selection to learn LMF in the formulation. In more detail, it is evident that performance seems to vary more drastically in i-LIDS-AA than in i-LIDS-MA and CAVIAR4REID, which means, i-LIDS-AA is more sensitive to the number of irrelevant samples to learn the local metric for each human image set. This phenomenon can be traced to the severe local variation of the individual images inherent with the birth of this dataset from automatic detecting and tracking.

5.4 Summary

This chapter has proposed a novel method, LBDM, for single-set vs. single-set human re-identification. In LBDM, a new set-to-set distance MAD is crafted and its discriminability is exploited by LMF to help enhance the matching ability of SCNNM. Results have confirmed the reliability and superiority of LBDM. In the end, we give the architecture of LBDM in Figure 5.3.

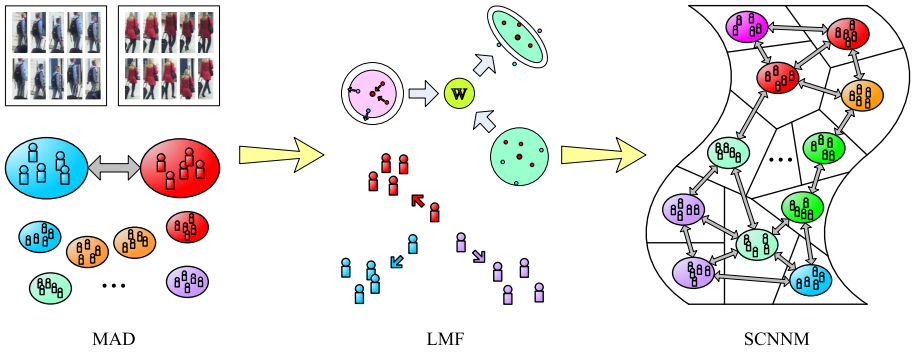


Figure 5.3: LBDM comprises three primary steps: set-to-set distance crafting (left), local metric field constructing (middle), and set based matching (right). In LBDM, a new set-to-set distance MAD is crafted by encoding the local minority distribution information between paired sets, and upon this distance, an effective LMF space is constructed to accommodate the local variation of each set, before the SCNNM dissimilarity is measured among all sets.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis has proposed to exploit the common-near-neighbor information in discriminative spaces for human re-identification.

In Chapter 2, to cope with the problem of large within-class variations and small between-class differences of human image data in single-shot re-identification, we have proposed the PCNNA method. This method uses the PCNNM dissimilarity to study the point-level common-near-neighbor information. We have also paid effort in improving the original feature space by metric learning to make intra-class distances smaller than inter-class distances. Experimental results have confirmed the advantage of PCNNA.

Towards the specific single-shot vs. single-shot re-identification, in Chapter 3, we have suggested using the strengthened metric space to compensate for the insufficiency of within-class distribution information. First, we have presented the CML method, which couples MCML and MLR based on their complementarity. As justified by a serial of experiments, though CML can learn a more discriminative metric space than MLR, it cannot largely increase the performance of the followed PCNNM. Then, we have presented the PCNNML method, which incorporates PCNNM into the metric learning framework, so that the training is consistent with the testing. Results have experimentally clarified PCNNML can bring the substantial performance enhancement for PCNNM, which satisfactorily ends the stories in the single-shot re-identification branch.

In Chapter 4, we have proposed the RSCNNM and BRIA methods to exploit the set-level common-near-neighbor information in discriminative spaces. For

effectively measuring the SCNNM dissimilarity, RSCNNM uses MRCG to condense the human image sets into representative imaginary points in the Riemennian space, and BRIA uses TPCR to compact human image sets to depress the intruders and outliers in the Hilbert space. Remarkable performance has shown the superiority of RSCNNM and BRIA.

For the specific single-set vs. single-set re-identification re-identification, in Chapter 5, we have proposed the LBDM method, which avoids the expensive requirements in RSCNNM and BRIA. This method studies the locality information by learning the local metric field upon the new set-to-set distance MAD to provide a discriminative space for SCNNM. Experiments have confirmed LBDM outstrips state-of-the-arts by a large margin, which draws to a successful end for the multiple-shot re-identification branch.

The intention of this thesis is not only discussing several methods, but, more importantly, is explaining the philosophy of exploiting common-near-neighbor information in discriminative spaces throughout all the stories. If we pay more attention to the philosophy of the thesis, we can draw a serial of conclusions from these chapters as well.

- (a) For the single-shot case, PCNNM is effective in a discriminative metric space (Chapter 2);
- (b) For the specific single-shot vs. single-shot case, suitably strengthened metric space will help increase the performance of PCNNM (Chapter 3);
- (c) For the multiple-shot case, SCNNM is effective when the metric space is discriminative (Chapter 4);
- (d) For the specific single-set vs. single-set case, improving the metric space by effective set-to-set distance and local metric field can help increase the performance of SCNNM (Chapter 5).

6.2 Future Work

End is another beginning. The future work will include, but not limited to, the following two aspects.

Methodological exploration Based on the same philosophy of exploiting the common-near-neighbor information in discriminative spaces, we can explore new

models. For example, we can enrich this research with graph theory by treating the concerned element and its neighbors as the nodes in graph.

Application extension Re-identification can deal with the problem of false alarms and identity switches in cross-camera tracking. And the methods proposed in this thesis may help build the correspondences between the tracklets in the surveillance system, regardless of whether the tracklets contain multiple-shot or single-shot human image data.

.

Appendix

We annotate some terminologies and list some related feature representations and dissimilarity measurements in the appendix for better understanding the contents of this thesis, though they have been detailed in the mainbody of the thesis.

Terminology Annotation in the Thesis

Human Re-identification

Matching people across camera views at different sites, known as human re-identification, is challenging and valuable for both academia and industry. This visual surveillance issue can help determine the reappearance for the person of interest who has been observed in the camera network in public places, such as the shopping mall, hospital, airport, and so forth. However, the issue itself is difficult due to the real-world complexities, including pose variations, illumination changes, viewpoint alterations, clutters, occlusions, and possibly similar body shapes and clothing styles.

Common-Near-Neighbor Information

Given a pair of elements, we intend to quantify the information on the local neighborhood structure of paired elements in each other's neighborhood structures, and thereby conceptualizes this information as "common-near-neighbor information" to measure how common near for the local layout of the paired samples in each other's neighborhood structures.

Discriminative Space

Discriminative space can be understood as a metric space in which the points or the sets can be easily separated by some baseline distance measurement, such as Euclidean distance, MPD, and so forth, due to the intra-class compactness and inter-class separation of the sample space.

Neighborhood Structure

In this thesis, neighborhood structure conceptually indicates the spatial layout relationship between the concerned element and its neighboring elements. In Chapter 2, we compress and characterize this structural relationship into one-dimensional rank-order lists in terms of nearness and farness.

Metric Space

A metric space is a space where a notion of distance between interior elements has been defined. In the thesis, a metric space can be expressed by (M, d), where M denotes the feature representations, and d denotes the distance for measuring them. Here, d is instantiated by the point-to-point distance in the single-shot cases, while the set-to-set distance in the multiple-shot cases.

Local Metric Field

Local metric field is a field space in which possibly different metrics are assigned at local areas. In Chapter 5, local metric field has been constructed for each set to accommodate the local variation of human image data.

Relative Information

Relative Information is a concept opposite to original information. In Chapter 4, relative information has been instantiated by relative feature and relative dissimilarity, both of which, are acquired by encoding some relative relationship between the concerned element and the reference elements.

Feature Representation in the Thesis

Weighted HSV color histograms (WHSV)

WHSV encodes all the chromatic content of each part of the pedestrian in consideration of the distance to the symmetric and asymmetric axes of the body: each pixel is weighted by a one-dimensional Gaussian kernel. In this way, pixel values near the axes count more in the final histogram. More details can refer to [11, 51].

Dense Sampled Color Histogram (DSCH)

DSCH uses color histograms extracted from rectangular regions to represent the images. The rectangular regions are densely collected from a regular grid, providing an overlapping representation. Color histograms uses RGB and HSV spaces. Histograms extracted from an image can be concatenated as feature vector before dimension reduction through PCA. More details can refer to [8, 51].

Schmid-Filter-Bank (SFB)

Human images are divided into six horizontal stripes, and SFB is extracted for each strip. SFB can capture the texture information. It is rotationally invariant and has 13 isotropic, which are robust to viewpoint and pose variations. More details can refer to [16, 59, 50].

Gabor Transform (GT)

Human images are divided into six horizontal stripes, and GT is extracted for each strip. Frequency and orientation representations of Gabor filters are similar to those of the human visual system, and they have been found to be particularly appropriate for texture representation and discrimination. Hence, a set of 8 Gabor filters with different frequencies and orientations are adopted. More details can refer to [16, 59, 50].

Third-Party Collaborative Representation (TPCR)

TPCR is a kind of relative feature. It resorts to the third-party data as the dictionary, and then concatenates the reconstructed coefficients from collaborative

representation for each sample based on this dictionary, with summing the coefficients that correspond to axioms of the same class in the dictionary. The effectiveness of TPCR comes from two aspects. One is using the intra-class sum to compress the class structured feature dimensions. Such compression glues together the entries that have intra-class variations, so the intra-class variations of each represented sample will be reduced. The other is using the diverse third-party data as the dictionary. These third-party data can help describe each image sample representatively to enlarge the inter-class differences. More details can refer to [49].

Mean Riemannian Covariance Grid (MRCG)

MRCG uses grids to tile the human image, and then encodes the mean of covariance descriptors for the images belonging to the same person in Riemannian space before measurement. Basically, “Covariance Grid” can characterize human appearance, and “Mean Riemannian” can condense set of covariance grids into imaginary points. More details can refer to [40].

Dissimilarity Measurement in the Thesis

Minimum Point-wise Distance (MPD)

MPD is a kind of baseline set-to-set distance in this thesis. It measures the minimum point-wise distance between paired sets. More details can refer to [11].

Convex Hull Image Set Distance (CHISD)

CHISD is a kind of set-to-set distance. It measures the distance between the convex hulls constructed for paired sets, in order to depress the negative impact from outliers. More details can refer to [5].

Maximally Collapsing Metric Learning (MCML)

MCML is a metric learning approach. It learns Mahalanobis distance relying on the intuition that all points in the same class should be mapped to a single location in the feature space and all points in other classes should be mapped to other locations. More details can refer to [13].

Metric Learning to Rank (MLR)

MLR is a metric learning approach. It learns Mahalanobis distance based on expectation that rankings of data induced by distance from a query can be optimized against various ranking measures, such as AUC (Area under the Curve of ROC), Precision-at-k, MRR, MAP (Mean Average Precision), or NDCG (Normalized Discounted Cumulative Gain). More details can refer to [33].

Bibliography

[1] Le An, Mehran Kafai, Songfan Yang, and Bir Bhanu. Reference-based person re-identification. In *Proceedings of 10th International Conference on Advanced Video and Signal Based Surveillance, AVSS*, Krakow, Porland, Aug. 2013. IEEE Press.

[2] Loris Bazzani, Marco Cristani, and Vittorio Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117:130–144, 2013.

[3] Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.

[4] Yinghao Cai and Matti Pietikainen. Person re-identification based on global color context. In *Workshop on ACCV*. IEEE Press, Nov. 2010.

[5] Hakan Cevikalp and Bill Triggs. Face recognition based on image sets. In *Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2567–2573. IEEE Press, Jun. 2010.

[6] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *Proceedings of the British Machine Vision Conference, BMVA*, pages 68.1–68.11. BMVA Press, Aug. 2011.

[7] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, Jun. 2007.

[8] Mert Dikmen, Emre Akbas, Thomas S. Huang, and Narendra Ahuja. Pedestrian recognition with a learned metric. In *Proceedings of the 10th Asian*

Conference on Computer Vision, ACCV, pages 501–512. Springer-Verlag, Nov. 2010.

[9] Yonina C. Eldar and Gitta Kutyniok. *Compressed Sensing Theory and Applications*. Cambridge University Press, 2012.

[10] Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and appearance for mobile scene analysis. In *Proceedings of the 11th International Conference on Computer Vision, ICCV*. IEEE Press, Oct. 2007.

[11] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the 23rd Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2360–2367. IEEE Press, Jun. 2010.

[12] Dario Figueira, Loris Bazzani, Ha Quang Minh, Marco Cristani, Alexandre Bernardino, and Vittorio Murino. Semi-supervised multi-feature learning for person re-identification. In *Proceedings of the 10th International Conference on Advanced Video and Signal Based Surveillance, AVSS*, pages 111–116. IEEE, Aug. 2013.

[13] Amir Globerson and Sam Roweis. Metric learning by collapsing classes. *Advances in Neural Information Processing Systems*, 18(13):451–458, 2006.

[14] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person re-identification*. Springer London, 2014.

[15] Doug Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proceedings of the 10th International Workshop on Performance Evaluation for Tracking and Surveillance, PETS*, pages 41–47. IEEE Press, Oct. 2007.

[16] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the 10th European Conference on Computer Vision*, volume 5302, pages 262–275. Springer, Oct. 2008.

[17] Martin Hirzer, Peter M. Roth, and Horst Bischof. Person re-identification by efficient impostor-based metric learning. In *Proceedings of the 19th International Conference on Advanced Video and Signal-Based Surveillance, AVSS*, pages 203–208. IEEE Computer Society, Sep. 2012.

[18] Martin Hirzer, Peter M. Roth, Martin Kostinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In *Proceedings of the 12th European Conference on Computer Vision, ECCV*, volume 7577, pages 780–793. Springer, 2012.

[19] Yiqun Hu, Ajmal S. Mian, and Robyn Owens. Sparse approximated nearest points for image set classification. In *Proceedings of the 24th Conference on Computer Vison and Pattern Recognition, CVPR*, pages 121–128. IEEE Press, Jun. 2011.

[20] Yoshihisa Ijiri, Shihong Lao, Tony X Han, and Hiroshi Murase. Human re-identification through distance metric learning based on jensen-shannon kernel. In *Proceedings of the 9th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 603–612. SciTePress, Feb. 2012.

[21] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. Cutting-plane training of structural svms. *Machine Learning*, 77:27–59, 2009.

[22] Nebojsa Jojic, Alessandro Perina, Marco Cristani, Vittorio Murino, and Brendan Frey4. Stel component analysis: Modeling spatial correlations in image class structure. In *Proceedings of the 22th Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2044–2051. IEEE Press, Jun. 2009.

[23] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischo. Large scale metric learning from equivalence constraints. In *Proceedings of the 25th Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2288–2295. IEEE Press, Jun. 2012.

[24] Cheng-Hao Kuo, Sameh Khamis, and Vinay Shet. Person re-identification using semantic color names and rankboost. In *Workshop on Applications of Computer Vision*, pages 281–287. IEEE Computer Society, Jan. 2013.

[25] Igor Kviatkovsky, Amit Adam, and Ehud Rivlin. Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1622–1634, 2013.

[26] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *Proceedings of the 26th Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Press, Jun. 2013.

[27] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidenti

cation with transferred metric learning. In *Proceedings of the 11th Asian Conference on Computer Vision, ACCV*. Springer-Verlag, Nov. 2012.

[28] Kai Liu, Xin Guo, Zhicheng Zhao, and Anni Cai. Person re-identification using matrix completion. In *Proceedings of the 20th International Conference on Image Processing, ICIP*. IEEE Press, Sep. 2013.

[29] Bazzani Loris, Cristani Marco, Perina Alessandro, Farenzena Michela, and Murino Vittorio. Multiple-shot person re-identification by hpe signature. In *Proceedings of the 20th International Conference on Pattern Recognition, ICPR*, pages 1413–1416. IEEE Press, Aug. 2010.

[30] Chen Change Loy, Chunxiao Liu, and Shaogang Gong. Person re-identification by manifold ranking. In *Proceedings of the 20th International Conference on Image Processing, ICIP*. IEEE Press, Sep. 2013.

[31] Bingpeng Ma, Yu Su, , and Frederic Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *Proceedings of the 1st Workshop on the 12th European Conference on Computer Vision, ECCV Workshop(1)*, pages 413–422. IEEE Press, Oct. 2012.

[32] Bingpeng Ma, Yu Su, and Frederic Jurie. Bicov: a novel image representation for person re-identification and face verification. In *Proceedings of the 23rd British Machine Vision Conference, BMVC*, pages 57.1–57.11. BMVA Press, Sep. 2012.

[33] Brian McFee and Gert Lanckriet. Metric learning to rank. In *Proceedings of the 27th International Conference on Machine Learning, ICML*, pages 775–782. Omnipress, Jun. 2010.

[34] Alexis Mignon and Frederic Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *Proceedings of the 25th Conference on Computer Vison and Pattern Recognition, CVPR*, pages 2666–2672. IEEE Press, Jun. 2012.

[35] Hieu V. Nguyen. and Li Bai. Cosine similarity metric learning for face verification. In *Proceedings of the 10th Asica Conference on Computer Vision, ACCV*, volume 6493 of *Lecture Notes in Computer Science*, pages 709–720. Springer, Nov 2010.

[36] Bryan Prosser, Weishi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by support vector ranking. In *Proceedings of the British Machine Vision Conference*, pages 21.1–21.11. BMVA Press, Aug. 2010.

[37] Deva Ramanan and Simon Baker. Local distance functions: A taxonomy, new algorithms, and an evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):794–806, 2011.

[38] Mohammad Ali Saghafi, Aini Hussain, Halimah Badioze Zaman, and Mohamad Hanif Md. Saad. Color-spatial person re-identification by a voting matching scheme. In *Proceedings of the 3rd International Visual Informatics Conference*, volume 8237, pages 470–482. Springer, Nov. 2013.

[39] William Robson Schwartz and Larry S. Davis. Learning discriminative appearance-based models using partial least squares. In *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing, SIBGRAPI*, pages 322–329. IEEE Press, Oct. 2009.

[40] Bak Slawomir, Corvee Etienne, Bremond Francois, and Thonnat Monique. Multiple-shot human re-identification by mean riemannian covariance grid. In *Proceedings of the 8th International Conference on Advanced Video and Signal-Based Surveillance, AVSS*, pages 179–184. IEEE Press, Aug. 2011.

[41] Bak Slawomir, Corvee Etienne, Bremond Francois, and Thonnat Monique. Boosted human re-identification using riemannian manifolds. *Image Vision Computing*, 30:443–452, Jun. 2012.

[42] Boyd Stephen and Vandenberghe Lieven. *Convex Optimization*. Cambridge University Press, Mar. 2004.

[43] Dapeng Tao, Lianwen Jin, Yongfei Wang, Yuan Yuan, and Xuelong Li. Person re-identification by regularized smoothing kiss metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 23:1675–1685, 2013.

[44] Yimin Wang, Ruimin Hu, Chao Liang, Chunjie Zhang, and Qingming Leng. Person re-identification by smooth metric learning. In *Proceedings of the 14th Pacific-Rim Conference on Multimedia (PCM)*, volume 8294, pages 561–573. Springer, Nov. 2013.

[45] Michael Weber, Martin Bauml, and Rainer Stiefelhagen. Part-based clothing segmentation for person retrieval. In *Proceedings of the 8th International Conference on Advanced Video and Signal-Based Surveillance, AVSS*, pages 361–366. IEEE Computer Society, Aug. 2011.

[46] Kilian Q. Weinberger, John Blitzer, and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, Feb. 2009.

[47] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:210–227, 2009.

[48] Yang Wu, Michihiko Minoh, and Masayuki Mukunoki. Collaboratively regularized nearest points for set based recognition. In *Proceedings of the 24th British Machine Vision Conference, BMVC*. BMVA Press, Sept. 2013.

[49] Yang Wu, Michihiko Minoh, Masayuki Mukunoki, and Shihong Lao. Robust object recognition via third-party collaborative representation. In *Proceedings of the 21st International Conference on Pattern Recognition, ICPR*. IEEE Press, Nov. 2012.

[50] Yang Wu, Michihiko Minoh, Masayuki Mukunoki, and Shihong Lao. Set based discriminative ranking for recognition. In *Proceedings of the 12th European Conference on Computer Vision, ECCV*, pages 497–510. IEEE Press, Oct. 2012.

[51] Yang Wu, Masayuki Mukunoki, Takuya Funatomi, and Michihiko Minoh. Optimizing mean reciprocal rank for person re-identification. In *Proceedings of the 8th International Conference on Advanced Video and Signal-Based Surveillance, AVSS*, pages 408–413. IEEE Press, Aug. 2011.

[52] Yuanlu Xu, Liang Lin, Weishi Zheng, and Xiaobai Liu. Human re-identification by matching compositional template with cluster sampling. In *Proceedings of th 14th International Conference on Computer Vision, ICCV*. IEEE Press, Dec. 2013.

[53] Gang Yuan, Zhaoxiang Zhang, and Yunhong Wang. Enhancing person re-identification by robust structural metric learning. In *Proceedings of the 7th International Conference on Image and Graphics*, pages 453 – 458. Conference Publishing Services, Jul. 2013.

[54] Guanwen Zhang, Yu Wang, Jien Kato, Takafumi Marutani, and Kenji Mase. Local distance comparison for multiple-shot people re-identification. In *Proceedings of the 11th Asian Conference on Computer Vision*, pages 677–690, Nov. 2012.

[55] Lei Zhang, Meng Yang, and Xiangchu Feng. Sparse representation or collaborative representation: Which helps face recognition? In *Proceedings of the 13th International Conference on Computer Vision, ICCV*, pages 471–478. IEEE Press, Nov. 2011.

[56] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *Proceedings of the 26th Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Press, Jun. 2013.

[57] Weishi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *Proceedings of the 20th British Machine Vision Conference, BMVC*, pages 23.1–23.11. BMVA Press, Sep. 2009.

[58] Weishi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *Proceedings of the 24th Conference on Computer Vision and Pattern Recognition*, pages 649–656. IEEE Computer Society, Jun. 2011.

[59] Weishi Zheng, Shaogang Gong, and Tao Xiang. Re-identification by relative distance comparison. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE Press, March 2013.

[60] Chunhui Zhu, Fang Wen, and Jian Sun. A rank-order distance based clustering algorithm for face tagging. In *Proceedings of the 24th Conference on Computer Vision and Pattern Recognition, CVPR*, pages 481–488. IEEE Press, Jun. 2011.

List of Publications

Journal Articles

1. Wei Li, Masayuki Mukunoki, Yinghui Kuang, Yang Wu, Michihiko Minoh, “Person Re-identification by Common-Near-Neighbor Analysis.” Submitted to IEICE Transactions on Information and Systems, 2014.

2. Wei Li, Yang Wu, Masayuki Mukunoki, Michihiko Minoh, “Bi-level Relative Information Analysis for Multiple-Shot Person Re-Identification.” IEICE Transactions on Information and Systems, Vol.E96-D, No.11, pp.2450-2461, 2013.

3. Wei Li, Yang Wu, Masayuki Mukunoki, Michihiko Minoh, “Coupled Metric Learning for Single-shot versus Single-shot Person Reidentification.” Optical Engineering, Vol.52, No.2, pp.027203-1-10, 2013.

Refereed Conference Presentations

1. Yang Wu, Wei Li, Michihiko Minoh, Masayuki Mukunoki, “Can Feature-Based Inductive Transfer Learning Help Person Re-Identification?” In Proceedings of the 20th IEEE International Conference on Image Processing (ICIP), Sep. 2013.

2. Wei Li, Yang Wu, Masayuki Mukunoki, Michihiko Minoh, “Locality Based Discriminative Measure for Multiple-shot Person Re-identification.” In Proceedings of the 10th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), Aug. 2013.

3. Wei Li, Yang Wu, Yasutomo Kawanishi, Masayuki Mukunoki, Michihiko Minoh, “Riemannian Set-level Common-Near-Neighbor Analysis for

Multiple-shot Person Re-identification." In Proceedings of the 13th IAPR Conference on Machine Vision Applications (MVA), May 2013.

4. Wei Li, Yang Wu, Masayuki Mukunoki, Michihiko Minoh, "Common-Near-Neighbor Analysis for Person Re-identification." In Proceedings of the 19th IEEE International Conference on Image Processing (ICIP), Sep. 2012.

5. Yang Wu, Michihiko Minoh, Masayuki Mukunoki, Wei Li, Shihong Lao, "Collaborative Sparse Approximation for Multiple-Shot Across-Camera Person Re-identification." In Proceedings of the 9th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), Sep. 2012.