密级：_____

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

博士学位论文

黎曼度量学习及其在视频人脸识别中的应用研究

作者姓名：黄智武

指导教师：山世光　研究员

中国科学院计算技术研究所

学位类别：工学博士

学科专业：计算机应用技术

培养单位：中国科学院计算技术研究所

2015年05月

Riemannian Metric Learning and Its application to Video Face Recognition

By

Huang Zhiwu

A Dissertation Submitted to

University of Chinese Academy of Sciences

In partial fulfillment of the requirement

For the degree of

Doctor of Engineering

Institute of Computing Technology

Chinese Academy of Sciences

May,2015

声　　明

我声明本论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，本论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

作者签名：　　日期：

论文版权使用授权书

本人授权中国科学院计算技术研究所可以保留并向国家有关部门或机构送交本论文的复印件和电子文档，允许本论文被查阅和借阅，可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本论文。

（保密论文在解密后适用本授权书。）

作者签名：　　导师签名：　　日期：

摘　要

在机器学习领域，越来越多的研究者认识到当数据具有非欧氏结构时，采用开发于欧氏空间的机器学习方法通常会由于忽略数据的这一特殊的几何结构而产生次优的结果。为了克服这一缺陷，目前有一类新的机器学习方法假设用于表示输入数据的黎曼流形是显式已知的，通过开发已知流形的黎曼度量来在训练数据上学习有效的判别模型。本文将这一类机器学习方法称为黎曼度量学习。近年来，基于传统统计模型的黎曼度量学习方法已经被成功应用于诸多计算机视觉与模式识别领域的分类问题。以具体的基于视频的人脸识别任务为例，人脸视频序列通常包含非常丰富的人脸动态时序信息和多视空间信息。为了解决视频人脸识别问题，一些传统的统计模型（如线性子空间、协方差矩阵和高斯概率模型）通常可以用来有效编码视频序列中的人脸变化模式，从而成为一种鲁棒的视频特征。由于这些统计模型通常位于一个特定的黎曼流形上，现有的黎曼度量学习方法可以为这一问题提供一种有效的判别学习策略。因此，本文针对视频人脸识别问题，从对视频序列的统计建模出发，围绕黎曼度量学习研究展开以下四个主要工作：

(1) 基于视频序列的线性子空间建模，提出了一种在格拉斯曼流形上的投影度量学习方法来解决视频-视频人脸识别问题。为了在以线性子空间为基本元素的格拉斯曼流形上进行判别学习，该方法提出一个从原始的格拉斯曼流形到一个新的、更具判别性的格拉斯曼流形的映射学习框架。为了求解这个度量学习问题，该方法采用类Fisher准则来定义相应的目标函数，并开发黎曼共轭梯度优化算法。

(2) 基于视频序列的双阶统计量建模，提出了一种跨欧氏-黎曼度量学习框架来同时解决三种不同的基于视频的人脸识别问题，即视频-图像、图像-视频和视频-视频人脸识别。该方法采用双阶统计量（即均值和协方差）对视频数据进行建模，进而将这三种视频人脸识别问题统一形式化成欧氏数据和黎曼数据的匹配/融合问题。为了解决这一问题，该方法提出了一个统一的基于多视判别学习的异质度量学习框架将异质数据映射到一个公共子空间，从而实现了跨异质空间的距离计算。

(3) 基于视频序列的高斯分布函数建模，提出了一种在对称正定矩阵流形上的对数欧氏度量学习方法来解决视频-视频人脸识别问题。该方法借鉴经典的信息几何理论将高斯概率模型所在的空间嵌入到一个特定的对称正定矩阵流形上，并在此流形上推导出一个基于切空间映射的对数欧氏度量学习框架来学习更具判别性的对称正定矩阵对数。该方法通过优化所提出的基于LogDet散度的目标函数来学习新度量学习问题对应的类马氏矩矩阵。

(4) 基于视频序列的多种统计建模，提出了一种混合欧氏-黎曼度量学习框架来有效融合样本均值、样本协方差和高斯模型这三种统计模型，从而更有效地解决视频-视频

人脸识别问题。为了减少高斯模型所在的空间与其它两种统计模型的空间之间的异质性，该方法同样首先将高斯分布的空间嵌入到一个对称正定矩阵流形上，然后通过设计一个基于LogDet散度的目标函数来学习多个马氏矩阵，从而达到多统计模型的有效融合的目的。

综上所述，本文针对基于统计建模的黎曼度量学习在视频人脸识别上的应用开展了广泛和深入的研究，提出了一系列在特定黎曼流形上的黎曼度量学习方法。大量实验结果表明，本文提出的方法可以有效地提升视频人脸识别的性能。

关键词：黎曼度量学习， 视频人脸识别， 统计建模， 投影度量学习， 跨欧氏-黎曼度量学习， 对数欧氏度量学习， 混合欧氏-黎曼度量学习

Riemannian Metric Learning and Its application to Video Face Recognition

Huang Zhiwu (Computer Application Technology)
Advised by Professor Shan Shiguang

In the field of machine learning, researchers have increasingly realized that if the data is intrinsically non-Euclidean, the machine learning methods developed in Euclidean space commonly yield suboptimal results due to ignoring the special geometrical structure of data. To overcome this limitation, a new family of machine learning methods typically assumes the Riemannian manifold representing the input data is known explicitly, and learns an effective discriminant model on training data by exploiting Riemannian metric on the known manifold. This paper calls this class of machine learning methods as Riemannian metric learning. Recently, Riemannian metric learning on traditional statistical model has been successfully applied to many computer vision and pattern recognition classification tasks. Take the task of video face recognition for example. In this task, video sequences often contain very rich face dynamic temporal information and multi-view spatial information. To solve the problem of video face recognition, traditional statistical models (e.g., liner subspace, covariance matrix and Gaussian probabilistic model) can be employed to effectively encode the pattern variations of faces in video sequence, and thus become a robust video-based feature. Since these statistical models commonly reside on a specific Riemannian manifold, existing Riemannian metric learning can provide an effective discriminant learning scheme for this problem. Therefore, this paper goes from statistical modeling of video sequence for the problem of video face recognition, and launches the following four main works on the research of Riemannian metric learning:

(1) Based on the linear subspace modeling for video sequence, proposes a Projection Metric Learning (PML) on Grassmann manifold to solve the problem of Video-to-Video (V2V) face recognition. To perform discriminative learning on Grassmamm manifold with linear subspace as basic element, this method proposes a transformation learning framework which maps the original Grassmann manifold to a new more discriminant Grassmann manifold. To handle this problem of metric learning, this method defines an objective function by adopting a Fisher-like criterion, and exploits the optimization algorithm of Riemannian Conjugate Gradient.

(2) Based on the double-order statistics modeling for video sequence, proposes a Cross Euclidean-to-Riemannian Metric Leaning (CERML) method to simultaneously solve the three different problems of video face recognition, i.e., Video-to-Still (V2S), Still-to-Video (S2V) and Video-to-Video (V2V) face recognition. This method employs double-order statistics (i.e., mean and covariance) to model video data, and then uniformly formulates the three video face

recognition tasks as a problem of matching/fusing of Euclidean data and Riemannian data. By adopting a multi-view learning analysis based objective function, this method designs a uniformed heterogeneous metric learning framework to transform the heterogeneous data into a common Euclidean subspace to realize the distance computation across heterogeneous spaces.

(3) Based on the Gaussian modeling for video sequence, proposes a Log-Euclidean Metric Learning (LEML) on Symmetric Positive Definite (SPD) manifold to solve the V2V face recognition problem. Inspired by the well-known information geometry theory, this method embeds the space of Gaussian components into a specific SPD manifold, and derives a tangent map based Log-Euclidean metric learning framework on this manifold to learn more discriminant SPD matrix logarithms. By designing a LogDet divergence based objective function, the Mahalanobis-like matrices in this problem of metric learning can be learned.

(4) Based on the multiple statistical modeling for video sequence, proposes a Hybrid Euclidean-and-Riemannian Metric Learning (HERML) to effectively fuse sample mean, sample covariance matrix and Gaussian model for more robust V2V face recognition. To alleviate the heterogeneity with the Euclidean space of mean and the SPD manifold of covariance matrix, this method first embeds the space of Gaussian distribution into another specific SPD manifold. Then, this method designs a LogDet divergence based objective function to learn multiple Mahalanobis matrices to fuse the heterogeneous data.

In conclusion, this paper conducts extensive and deep research on statistical model based Riemannian metric learning with application to video face recognition methods, and proposes a series of Riemannian metric learning methods on specified Riemannian manifolds. Extensive experiments demonstrate the approaches proposed in this paper are all qualified to improve video face recognition.

Keywords: Riemannian metric learning, video face recognition, statistical modeling, projection metric learning, cross Euclidean-to-Riemannian metric learning, Log-Euclidean metric learning, hybrid Euclidean-and-Riemannian metric learning

目　录

图 目 录

表 目 录

第一章　绪论

《吕氏春秋·察今》里有这样的记述：“有道之士，贵以近知远，以今知古，以所见知所不见。故审堂下之阴，而知日月之行，阴阳之变；见瓶水之冰，而知天下之寒，鱼鳖之藏也；尝一脔肉，而知一镬之味，一鼎之调。”这段文字告诉人们可以根据“所见”的事物，把握事物之间的本质联系，探求“所不见”事物的属性、状态和规律。同样，面对错综复杂的现象，科学研究的主要任务正是去探寻隐藏在观察表象背后的本质规律。

1.1　课题研究背景及意义

传统的机器学习和数学分析方法通常要求输入数据能够表示为在欧氏空间上的向量。虽然这一假设在很多应用领域都取得了成功，但是越来越多的研究发现如果输入数据具有非欧氏结构而在机器学习过程中忽视这一几何结构将产生次优的结果。为了解决这一问题，目前主要有两种机器学习策略来研究如何开发位于黎曼流形上的数据几何的问题。第一种是著名的（黎曼）流形学习策略，它主要假设数据位于一个未知的黎曼流形上，通过利用无标签或者有标签的训练数据来挖掘未知流形的几何结构。第二种机器学习策略主要假设用于表示输入数据的黎曼流形是显式已知的，通过开发已知流形的黎曼度量来在训练数据上学习有效的分类模型。根据这一机器学习策略的特点，本文将之统一称为黎曼度量学习。在计算机视觉与模式识别领域，黎曼度量学习已经开始被成功应用于许多实际分类问题上。比如，对称正定矩阵流形和线性子空间的格拉斯曼流形以及相对应的黎曼度量学习方法已经被广泛应用于行人检测、行为识别、纹理分类以及人脸识别等任务上。

以近年来受到广泛关注的视频人脸识别问题为例，其研究的基本对象是视频人脸序列。随着视频监控设备、家用摄像机以及智能手机的日益普及，各种动态人脸视频序列呈现爆炸式地增长。相比静态人脸图像，视频人脸序列通常包含非常丰富的人脸动态时序信息和多视空间信息。然而如何充分有效地挖掘这些视频数据当中的有用的时空信息，则给研究者带来了很大的挑战。从计算机视频与模式识别的角度来看，视频人脸识别问题主要面临两方面的挑战：第一，如何从视频人脸序列的数据中提取有效、紧致的特征表示；第二，如何针对人脸视频特征设计合理有效的分类算法。在统计模型趋于成熟的大背景下，目前一些传统的统计模型（如线性子空间、协方差矩阵和高斯概率模型）可以有效开发视频序列中的人脸变化模式，从而为视频人脸识别问题提供了一种有效的视频特征。由于这些统计模型通常位于一个特定的黎曼流形上，因此传统的黎曼度量学习方法可以为这一问题提供了一种有效的判别学习策略从而达到更加鲁棒的视频人脸识别。

虽然现有的基于统计建模的黎曼度量学习方法在视频人脸识别问题上已经取得了巨大的成功，但是它们还是有以下两方面的问题：第一，在统计建模方面，目前大多数方法提出的统计模型一般只刻画了视频图像集合的某一方面的统计信息（比如协方差统计量）而忽视了其他类型的统计信息（比如均值统计量），从而没有完全编码出视频中的所有有价值信息。因此，研究一种更高阶的统计量或者融合多种统计模型来对视频序列中的人脸多视时空信息进行更加鲁棒地编码成为一个非常有意义的研究问题；第二，在黎曼度量学习方面，目前一些主流方法研究如何将统计模型表示成黎曼流形上的元素，然后通过推导出一些基于黎曼度量的核函数来将传统的核方法适配到黎曼流形上进行有监督学习。然而，这些方法也把传统的核方法的一些固有缺陷也带到黎曼流形判别分析的框架里。比如，核方法一般不能学习到显式的映射，而且随着样本数目的增加核方法的复杂度会呈指数级地增长。因此，如何直接在统计模型所在的黎曼流形上学习有效的黎曼度量进而可以更加有效地在黎曼流形上进行分类是值得深入探索的一个研究方向。

综上所述，本文在开发各种不同统计模型的基础上展开的黎曼度量学习方法及其在视频人脸识别中的应用研究对于推进机器学习、计算机视觉和模式识别等多个领域的研究都具有广泛的意义和价值。本章接下来首先介绍黎曼度量学习的相关研究进展，接着回顾视频人脸识别的国内外研究现状，最后介绍本文在黎曼度量学习及其在视频人脸识别中的应用研究上所作出的相应贡献。

1.2 黎曼度量学习概述

在微分流形和黎曼几何中，黎曼流形是一种具有黎曼度量的微分流形。换句话说，只有配备了有效的黎曼度量的微分流形才能被称为黎曼流形。本节将分两小节分别介绍黎曼流形的一些基本概念以及现有的黎曼度量学习方法。

1.2.1 黎曼流形

黎曼流形是一种局部具有欧氏空间性质的特殊拓扑空间，是欧氏空间中的曲线、曲面等概念的推广。下面给出一些有关黎曼流形的一些基本定义和概念[3–6, 15, 16, 41, 84, 106, 130, 144]。

定义1.1（拓扑流形） 给定一个n维的拓扑空间$\mathcal{M}$，若$\mathcal{M}$上的每一个点都有一个开邻域$U \subseteq \mathcal{M}$，使得U同胚于一个n维的欧氏空间$\mathbb{R}^n$中的一个开子集，则称$\mathcal{M}$是一个n维的拓扑流形（一般简称为流形）。

定义1.2（微分结构） 给定一个n维的拓扑流形$\mathcal{M}$，假定$\mathcal{A} = \{(U_a, \phi_a)|a \in I\}$是$\mathcal{M}$的坐标卡上的一个集合，并满足以下条件：
(1) $\{(U_a, \phi_a)|a \in I\}$形成流形$\mathcal{M}$的一个覆盖；

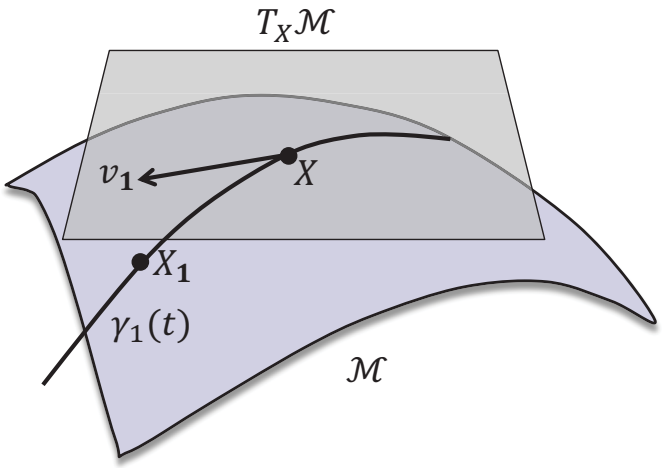


图 1.1. 黎曼流形基本概念: 黎曼流形$\mathcal{M}$，流形上的点$\boldsymbol{X}$和$\boldsymbol{X}_1$，在点$\boldsymbol{X}$上的切空间$T_X\mathcal{M}$，$\boldsymbol{X}_1$在切空间$T_X\mathcal{M}$上的切向量v_1，$\boldsymbol{X}$与$\boldsymbol{X}_1$间的测地线$\gamma_1(t)$。

(2) $\mathcal{A}$中的任意两个坐标卡都是C^r相关的；
(3) $\mathcal{A}$是极大的，即：若$\mathcal{M}$的坐标卡(U,ϕ)与$\mathcal{A}$中的每一个坐标卡是C^r相关的，则有(U,ϕ)属于$\mathcal{A}$。
此时，坐标卡集$\mathcal{A}$是拓扑流形$\mathcal{M}$上的一个C^r微分结构。

定义1.3（微分流形） 给定一个n维的拓扑流形$\mathcal{M}$，若$\mathcal{M}$拥有全局定义的C^r微分结构，则称流形$\mathcal{M}$为一个n维C^r微分流形。

定义1.4（切空间） 设x为n维的微分流形$\mathcal{M}$上的一个点，$\mathcal{M}$在$\boldsymbol{X}$的切向量v如果满足下列条件的一个映射$v: C_X^\infty \to \mathbb{R}$：
(1) $\forall f,g \in C_x^\infty$，有$v(f+g)=v(f)+v(g)$;
(2) $\forall f \in C_x^\infty, \lambda \in \mathbb{R}$，有$v(\lambda f)=\lambda v(f)$；
(3) $\forall f,g \in C_x^\infty$，有$v(f\cdot g)=v(f)\cdot v(g)+v(g)\cdot v(f)$。
那么，由映射v得到的所有切向量张成的n维向量空间$T_X\mathcal{M}$被称为微分流形$\mathcal{M}$在点x处的切空间。

定义1.5（黎曼度量） 给定一个n维的微分流形$\mathcal{M}$，在流形$\mathcal{M}$上的每一个点的切空间上的（正定）内积族称为黎曼度量：

$$g_x = T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}, \quad x \in \mathcal{M}. \tag{1.1}$$

定义1.6（黎曼流形） 给定一个n维的微分流形$\mathcal{M}$，若再给$\mathcal{M}$配备一个特定的黎曼度量，也就是在$\mathcal{M}$上的每一个点的切空间上配备一个切空间内积度量，则称微分流形$\mathcal{M}$为黎曼流形。

如图1.1所示，黎曼流形包含了以下几个基本概念：流形上的点$\boldsymbol{X}$和$\boldsymbol{X}_1$，在点$\boldsymbol{X}$上的切空间$T_X\mathcal{M}$，$\boldsymbol{X}_1$在切空间$T_X\mathcal{M}$上的切向量v_1，$\boldsymbol{X}$与$\boldsymbol{X}_1$间的测地线$\gamma_1(t)$。常见的黎曼流形有球体（Sphere）、正交群（Orthogonal group）、施蒂费尔流形（Stiefel manifold）、格拉斯曼流形（Grassmann manifold）以及对称正定矩阵流形（Symmetric Positive Definite, SPD manifold）。接下来重点介绍一下施蒂费尔流形、格拉斯曼流形和对称正定矩阵流形。

1.2.1.1 施蒂费尔流形与格拉斯曼流形

定义1.7（施蒂费尔流形） 由所有$d \times p$ $(0 < p < d)$的半正交矩阵张成的黎曼流形称为施蒂费尔流形$\mathcal{S}(p,d)$，即$\mathcal{S}(p,d) \triangleq \{\boldsymbol{X} \in \mathbb{R}^{d\times p} : \boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_p\}$，其中$\boldsymbol{I}_p$表示大小为$p \times p$的单位矩阵。

格拉斯曼流形可以定义为施蒂费尔流形的带有等价类操作的商流形（quotient manifold）。这一等价类操作是$\boldsymbol{X}_1 \sim \boldsymbol{X}_2$当且仅当$Span(\boldsymbol{X}_1) = Span(\boldsymbol{X}_2)$，其中$Span(\boldsymbol{X})$表示由$\boldsymbol{X} \in \mathcal{S}(p,d)$张成的线性子空间。

定义1.8（格拉斯曼流形） 由所有在$\mathbb{R}^d$上的p维线性子空间张成的黎曼流形称为格拉斯曼流形$\mathcal{G}(p,d)$。

在格拉斯曼流形$\mathcal{G}(p,d)$上，两个切向量Δ_1, Δ_2在点$Span(\boldsymbol{X}) \in \mathcal{G}(p,d)$上的切空间的内积可以表示为$\langle\Delta_1, \Delta_2\rangle$。这一黎曼结构导出了在格拉斯曼流形的测地距离（也就是流形上两点之间的最短曲线的长度，表示为$\delta_g(\boldsymbol{X}_1, \boldsymbol{X}_2)$=$tr(\Delta_1^T(\boldsymbol{I}_d - \frac{1}{2}\boldsymbol{X}\boldsymbol{X}^T)\Delta_2) = tr(\Delta_1^T\Delta_2)$）。这一测地距离可以看作是一个线性子空间旋转到另一个线性子空间的最小旋转幅度。令$\Theta = [\theta_1, \theta_2, \ldots, \theta_p]$为两个线性子空间$Span(\boldsymbol{X}_1), Span(\boldsymbol{X}_2)$之间的主夹角序列[7]（详见以下主夹角定义），那么它们之间的测地距离为:

$$\delta_g(\boldsymbol{X}_1, \boldsymbol{X}_2) = \|\Theta\|_2. \tag{1.2}$$

定义1.9 (主夹角) 假定$\boldsymbol{X}_1$和$\boldsymbol{X}_2$为两个大小为$d \times p$的正交基矩阵，那么两个线性子空间$Span(\boldsymbol{X}_1), Span(\boldsymbol{X}_2)$之间的主夹角$0 \le \theta_1 \le \ldots \le \theta_p \le \pi/2$可以递归地定义成：

$$\begin{aligned} \cos(\theta_i) = &\max_{\boldsymbol{u}_i \in span(\boldsymbol{X}_1)} \max_{\boldsymbol{v}_i \in span(\boldsymbol{X}_2)} \boldsymbol{u}_i^T\boldsymbol{v}_i \\ s.t. \quad &\boldsymbol{u}_i^T\boldsymbol{u}_i = 1, \boldsymbol{v}_i^T\boldsymbol{v}_i = 1 \\ &\boldsymbol{u}_i^T\boldsymbol{u}_j = 0, \boldsymbol{v}_i^T\boldsymbol{v}_j = 0, (j = 1, \ldots, i-1). \end{aligned} \tag{1.3}$$

其中，$\cos(\theta_i), 0 \le i \le p$被称为典型相关系数。

基于线性子空间的主夹角，文献[41]在投影映射$\Phi(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{X}^T$的框架下提出了另一种表示格拉斯曼流形的基本元素的策略，也就是用投影矩阵$\boldsymbol{X}\boldsymbol{X}^T$来表示格拉斯曼流形元素，即线性子空间$Span(\boldsymbol{X})$。文献[62]证明了这一投影嵌入是一个从格拉斯曼流形到秩为q的幂等对称矩阵空间的微分同胚映射。也就是说，它是一个一对一、连续、可微的映射，并且它的逆映射也是个连续、可微的。因此，在格拉斯曼流形上的每个元素只对应唯一的一个投影矩阵。在这一投映射框架下，文献[41]提出了在格拉斯曼流形上的投影度量（Projection metric）。

定义1.10 (投影度量) 由于投影算子$\Phi(\boldsymbol{X})$是一个$d \times d$维的对称矩阵，因此其对应的内积形式可以定义为$\langle \boldsymbol{X}_1, \boldsymbol{X}_2 \rangle_\Phi = tr(\Phi(\boldsymbol{X}_1)^T \Phi(\boldsymbol{X}_2))$。这一内积形式对线性子空间的具体实现形式具有不变性，由此可以推导出对应的距离形式：

$$\delta_p(\boldsymbol{X}_1, \boldsymbol{X}_2) = \left(\sum_{i=1}^{p} \sin^2 \theta_i \right)^{1/2} = 2^{-1/2} \| \boldsymbol{X}_1 \boldsymbol{X}_1^T - \boldsymbol{X}_2 \boldsymbol{X}_2^T \|_F. \tag{1.4}$$

其中，$\| \cdot \|_F$表示矩阵的弗罗贝尼乌斯范数（Frobenius norm）。文献[55]指出这一距离被证明满足了度量的三个要素条件，即非负性、对称性与三角不等式，因此它又被称为投影度量（Projection metric）。除此之外，文献[62]证明了投影度量可以以最大尺度为$\sqrt{2}$来估计格拉斯曼流形上的真实测地距离。因此，投影度量已经成为了在格拉斯曼流形上最受欢迎的度量之一。

基于线性子空间的主夹角，文献[143]也提出了一种在格拉斯曼流形上的有效度量称为比奈-柯西度量（Binet-Cauchy metric）。

定义1.14（比奈-柯西度量） 在格拉斯曼流形上的两个线性子空间之间的距离可以定义为典型相关系数间的乘积：

$$\delta_b(\boldsymbol{X}_1, \boldsymbol{X}_2) = \left(1 - \prod_{i=1}^{p} \cos^2 \theta_i \right)^{1/2} = \det \left(\boldsymbol{X}_1^T \boldsymbol{X}_1 - \boldsymbol{X}_2^T \boldsymbol{X}_2 \right) \tag{1.5}$$

文献[55]将比奈-柯西度量理解成一种嵌入。假定s为$\{1, \ldots, d\}$的一个子集（表示成$s = \{r_1, \ldots, r_p\}$），$\boldsymbol{X}^{(s)}$为$p \times p$维矩阵，它的每行对应正交基矩阵$\boldsymbol{X}$的第$r_1, \ldots, r_p$行。如果$s_1, \ldots, s_n$是子集s的所有选择，那么比奈-柯西嵌入可以定义为

$$\Psi_{BC} : \mathcal{G}(p, d) \to \mathbb{R}^n, \quad \boldsymbol{X} \mapsto \left(\det \boldsymbol{X}^{(s_1)}, \ldots, \det \boldsymbol{X}^{(s_n)} \right) \tag{1.6}$$

其中，$n = C_d^p$是从d行中选出p行的个数。采用这一比奈-柯西映射框架可以得到一个内积形式$\sum_{r=1}^{n} \det \boldsymbol{X}_1^{(s_i)} \boldsymbol{X}_2^{(s_i)}$。

1.2.1.2 对称正定矩阵流形

定义1.12（对称正定矩阵流形） 当$d \times d$的对称正定矩阵矩阵（*SPD*）张成的空间被赋予一个合适的黎曼度量时，这一空间就可以构成一个特定类型的黎曼流形，也就是所谓的对称正定矩阵（*SPD*）流形$\mathbb{S}_+^d$。

通过采用对数映射$\log_{\boldsymbol{S}_1}: \mathbb{S}_+^d \to T_{\boldsymbol{X}_1}\mathbb{S}_+^d$，$\boldsymbol{X}_1 \in \mathbb{S}_+^d$，在SPD流形上的点$\boldsymbol{X}_1$上的切线都位于对应的切空间上$T_{\boldsymbol{X}_1}\mathbb{S}_+^d$。这一切空间上定义了内积$\langle , \rangle_{\boldsymbol{X}_1}$。在所有切空间上定义的内积族就是所谓的流形上的黎曼度量。如果对SPD流形配备黎曼度量，那么在流形上的两点$\boldsymbol{X}_1, \boldsymbol{X}_2$之间的测地距离一般可计算为：

$$\delta_s(\boldsymbol{X}_1, \boldsymbol{X}_2) = \langle \log_{\boldsymbol{X}_1}(\boldsymbol{X}_2), \log_{\boldsymbol{X}_1}(\boldsymbol{X}_2) \rangle_{\boldsymbol{X}_1}. \tag{1.7}$$

目前，有两个最广为应用的黎曼度量是仿射不变度量（Affine-Invariant metric）[106] 和对数欧氏度量（Log-Euclidean metric）[16]。它们均设计了对称正定矩阵流形上两元素之间的测地距离（公式1.7）的具体形式。

文献[105]在SPD流形$\mathbb{S}_+^n$ 上定义了仿射不变黎曼度量。在仿射不变黎曼度量的框架下，在流形$\mathbb{S}_+^n$ 上的点$\boldsymbol{X}_1$处的切空间上两个切向量的内积可以表示为:

$$\langle \boldsymbol{H}_1, \boldsymbol{H}_2 \rangle_{\boldsymbol{X}_1} = \langle \boldsymbol{X}_1^{-1/2} \boldsymbol{H}_1 \boldsymbol{X}_1^{-1/2}, \boldsymbol{X}_1^{-1/2} \boldsymbol{H}_2 \boldsymbol{X}_1^{-1/2} \rangle. \tag{1.8}$$

在这一黎曼度量框架下, 文献[105]给出了对应的黎曼指数映射与对数映射:

$$\begin{aligned} \exp_{\boldsymbol{X}_1}(\boldsymbol{H}) &= \boldsymbol{X}_1^{1/2} \exp(\boldsymbol{X}_1^{-1/2} \boldsymbol{H} \boldsymbol{X}_1^{-1/2}) \boldsymbol{X}_1^{1/2} \\ \log_{\boldsymbol{X}_1}(\boldsymbol{X}_2) &= \boldsymbol{X}_1^{1/2} \log(\boldsymbol{X}_1^{-1/2} \boldsymbol{X}_2 \boldsymbol{X}_1^{-1/2}) \boldsymbol{X}_1^{1/2} \end{aligned} \tag{1.9}$$

这两个映射均是局部微分同胚（local diffeomorphisms）映射（即在某个领域内是个一对一映射、满射且是连续可微映射）。

定义1.14（仿射不变度量） 根据公式*1.8*和公式*1.9*，在*SPD*流形上的两个*SPD*矩阵之间的测地距离可以表示为以下仿射不变度量形式：

$$\begin{aligned} \delta_a(\boldsymbol{X}_1, \boldsymbol{X}_2) &= \langle \log_{\boldsymbol{X}_1}(\boldsymbol{X}_2), \log_{\boldsymbol{X}_1}(\boldsymbol{X}_2) \rangle_{\boldsymbol{X}_1} \\ &= \| \log(\boldsymbol{X}_1^{-1/2} \boldsymbol{X}_2 \boldsymbol{X}_1^{-1/2}) \|_F. \end{aligned} \tag{1.10}$$

虽然仿射不变度量有着一系列非常优越的属性，但是由文献[16]指出，这一度量在实际应用中的成功常常以巨大的时间开销为代价。为了解决这一问题，文献[16]引进了一种新的在SPD流形上的黎曼度量叫作对数欧氏度量。具体地，他们通过开发SPD流形$\mathbb{S}_+^d$ 的李群结构从而推导出了在SPD流形上的对数欧氏度量。李群结构对应的群操作

是对于任意$\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{S}_+^d$, $\boldsymbol{X}_1 \odot \boldsymbol{X}_2 := \exp(\log(\boldsymbol{X}_1) + \log(\boldsymbol{X}_2))$, 其中$\exp(\cdot)$和$\log(\cdot)$分别表示矩阵的指数操作和对数操作。

在SPD矩阵的李群上定义的对数欧氏度量对应于在SPD矩阵对数域上的欧氏度量。在赋予SPD流形$\mathbb{S}_+^d$以对数欧氏度量的框架下，在点$\boldsymbol{X}$上的切空间$T_{\boldsymbol{X}}\mathbb{S}_+^d$里的两个基本元素$\boldsymbol{T}_1$，$\boldsymbol{T}_2$的内积可计算为：

$$\langle \boldsymbol{T}_1, \boldsymbol{T}_2 \rangle_{\boldsymbol{X}} = \langle D_{\boldsymbol{X}} \log . \boldsymbol{T}_1, D_{\boldsymbol{X}} \log . \boldsymbol{T}_2 \rangle. \tag{1.11}$$

其中，$D_{\boldsymbol{X}} \log . \boldsymbol{T}$表示$\boldsymbol{X}$的矩阵对数沿着$\boldsymbol{T}$的方向导数（directional derivative）。与对数欧氏度量相关联的对数映射和指数映射分别定义为：

$$\begin{aligned} \log_{\boldsymbol{X}_1}(\boldsymbol{X}_2) &= D_{\log(\boldsymbol{X}_1)} \exp .(\log(\boldsymbol{X}_2) - \log(\boldsymbol{X}_1)), \\ \exp_{\boldsymbol{X}_1}(\boldsymbol{T}_2) &= \exp(\log(\boldsymbol{X}_1) + D_{\boldsymbol{X}_1} \log . \boldsymbol{T}_2)). \end{aligned} \tag{1.12}$$

其中，由于$\log \circ \exp = \boldsymbol{I}$ ($\boldsymbol{I}$为单位矩阵)，那么$D_{\log(\boldsymbol{X})} \exp . = (D_{\boldsymbol{X}} \log .)^{-1}$。关于更详细的对公式1.11和公式1.12的推导过程，请读者参考文献[16]。

定义1.14（对数欧氏度量） *根据公式1.11和公式1.12，在SPD流形上的两个SPD矩阵之间的测地距离可以表示为对数欧氏度量形式：*

$$\begin{aligned} \delta_l(\boldsymbol{X}_1, \boldsymbol{X}_2) &= \langle \log_{\boldsymbol{X}_1}(\boldsymbol{X}_2), \log_{\boldsymbol{X}_1}(\boldsymbol{X}_2) \rangle_{\boldsymbol{X}_1} \\ &= \| \log(\boldsymbol{X}_1) - \log(\boldsymbol{X}_2) \|_F. \end{aligned} \tag{1.13}$$

实际上，对数欧氏度量对应的是在矩阵对数域（也就是在单位矩阵上的切空间）上的欧氏距离。换句话说，在对数欧氏度量的框架下，在SPD流形上的两个元素之间的测地距离可以归纳为在单位矩阵处的切空间上的两个矩阵对数之间的欧氏距离。因此，通过采用对数欧氏度量，SPD矩阵流形就转化成了一个平坦的黎曼流形。

1.2.2　黎曼度量学习方法

黎曼度量学习方法主要目标是以黎曼流形上的数据作为基本元素通过开发对应的黎曼度量来学习判别函数从而能够对流形上的元素进行有效地分类。如图1.2所示，现有的黎曼度量学习方法通常采用三种不同的学习策略。第一类方法主要在切空间估计的基础上学习流形-欧氏空间映射，使得从流形上变换到目标欧氏空间上的同类（异类）元素之间的距离最小化（最大化）。第二类方法采用了流形-核空间映射学习策略，首先将流形嵌入到一个高维的可再生核希尔伯特空间（Reproducing Kernel Hilbert Space, RKHS），然后采用传统的核方法对RKHS的元素进行判别分析。第三类方法在流形-流形映射框架下学习判别函数。由于目标空间也是个黎曼流形，那么这样的学习机制保持了原始流形的黎曼几何，从而可以学习到一个有效的判别函数。下面将分别介绍这三种类型的黎曼度量学习方法。

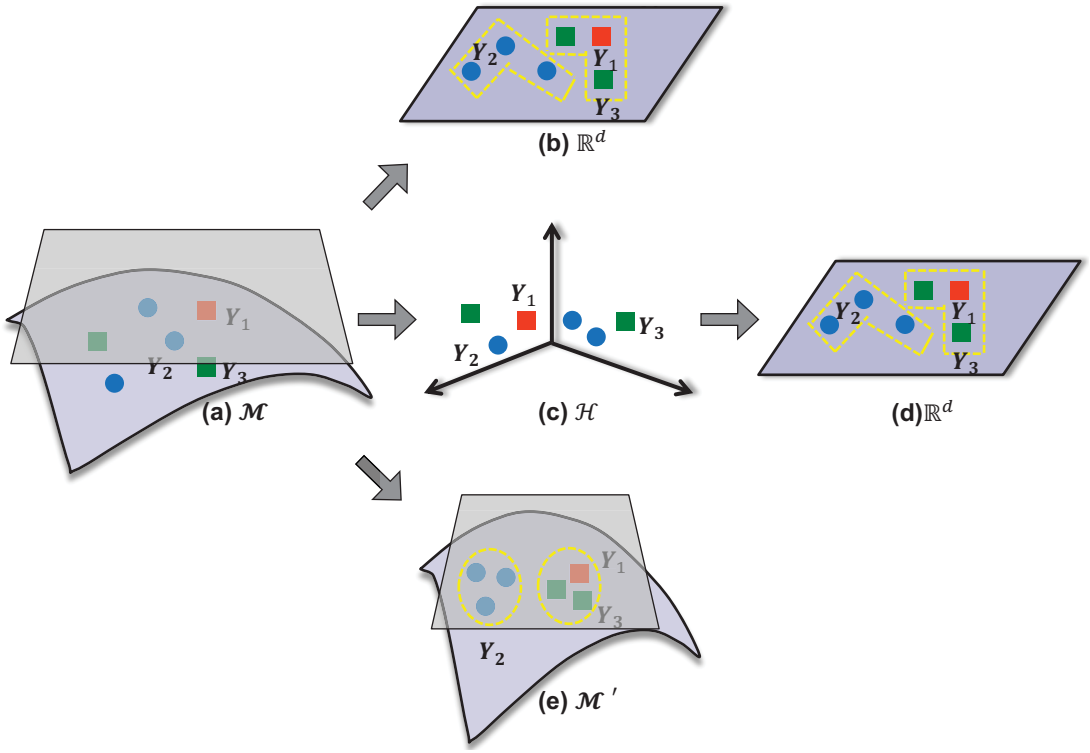


图 1.2. 三种不同的黎曼度量学习策略。(a)-(b): 流形-欧氏空间映射学习策略；(a)-(c)-(d): 流形-核空间映射学习策略；(a)-(e): 流形-流形映射学习策略。

1.2.2.1 流形-欧氏空间映射学习策略

这一类黎曼度量学习方法[28, 29, 119, 128, 132, 134]通常在切空间对流形的黎曼几何进行局部估计的基础上，对切空间上的切向量再学习判别函数，从而学习到一个位于新的欧氏空间上的更有判别性的向量形式（请看图1.2(a)-(b)）。

在早期，文献[132]为了缓解切空间与流形之间的局部微分同胚映射问题，采用了Boosting的框架将流形的点映射到多个不同的切平面上，然后在多个不同的切空间上学习多个分类器并对它们进行有效融合。具体地，这一工作是在对称矩阵流形的仿射不变度量的框架下，采用公式1.9的对数映射$\log_{\boldsymbol{X}_1}(\boldsymbol{X}_2)$将对称矩阵流形上的点都映射到以流形的均值点（Karcher mean）为支点的切平面上。这一均值点$\boldsymbol{\mu}$可由以下公式计算得到：

$$\boldsymbol{\mu} = \arg\min_{\boldsymbol{X}\in\mathcal{M}} \sum_{i=1}^{N} \delta_a(\boldsymbol{X}_i, \boldsymbol{X}). \tag{1.14}$$

其中，$\delta_a(\boldsymbol{X}_i, \boldsymbol{X})$是仿射不变度量（公式1.10），$N$是流形上的点的个数。在这一均值点对应的切空间上，这一工作然后将所有的切向量都进行向量化操作：

$$vec_{\boldsymbol{\mu}}(\boldsymbol{y}) = vec_{\boldsymbol{I}}(\boldsymbol{\mu}^{\frac{1}{2}}\boldsymbol{y}\boldsymbol{\mu}^{\frac{1}{2}}). \tag{1.15}$$

其中，$\boldsymbol{I}$是单位矩阵，这一在单位矩阵上的向量化操作可以表示为

$$vec_{\boldsymbol{I}}(\boldsymbol{y}) = [y_{1,1}, \sqrt{2}y_{1,2}, \sqrt{2}y_{1,3}, \ldots, y_{2,2}, \sqrt{2}y_{2,3}, \ldots, y_{d,d}]^T. \tag{1.16}$$

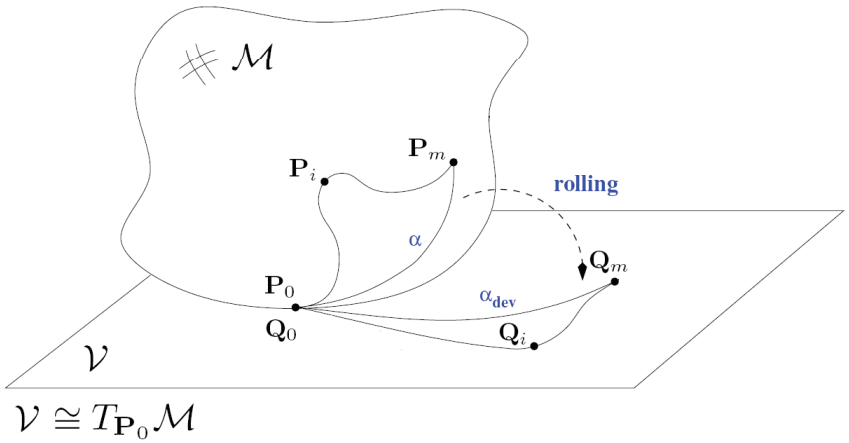


图 1.3. 卷映射框架[29]。它采用卷映射通过学习一个仿射切空间$\mathcal{V}$来全局估计黎曼流形$\mathcal{M}$的几何结构。

经过向量化操作1.15，这一工作最后以这些向量形式为基本元素开发了经典的LogitBoost分类学习算法。由于在boosting的学习过程中增加了那些被错分的样本的权重，那么流形的带权重均值会移向这些样本点从而对它们更准确地分类。然而，在多类分类问题里，不同类的样本点分布在流形上各个不同的地方，很难用流形上的均值点来统一逼近这些点，从而使在难分点的近邻（均值）处寻找切空间的作法失效。为了解决这一问题，文献[128]采用了对数欧氏度量（公式1.13），只在以对称正定矩阵流形上的单位矩阵为支点的切空间上通过开发LogitBoost分类器进行多类分类。但是这一工作同样还是无法完全解决这一类方法的切空间局部估计问题。类似地，文献[134]也采用了对数欧氏度量框架首先将映射到切空间上的所有对数矩阵转化为向量形式，然后在向量形式上利用了经典的信息理论度量学习方法（Information-Theoretic Metric Learning, ITML）[40]学习对应的马氏矩阵，从而生成更具判别力的向量表示。

为了完全克服局部微分同胚映射带来的负面影响，文献[29]引入了一个新的卷映射（Rolling Map）框架来替换对数映射框架。具体地，如图1.3所示，这一框架首先不作任何滑动操作（slipping）及扭曲操作（twisting）地将流形卷起成一个刚体，然后将流形上的黎曼数据都映射到一个仿射切空间上，最后在这一仿射切空间上同样采用LogitBoost算法进行分类。具体地，在流形$\mathcal{M}$上的卷映射是沿着一条平滑的Rolling曲线$\alpha : [0, T] \to \mathcal{M}$的一个平滑映射$h$:

$$\begin{aligned} h : [0, T] &\to SE_n = SO_n \ltimes \mathbb{R}^n \\ t &\mapsto h(t) = (R(t), s(t)). \end{aligned} \tag{1.17}$$

其中SE_n是特殊欧氏群（Special Euclidean group），SO_n是特殊正交群（Special Orthogonal）。特别地，这一映射需要在任意$t \in [0, T]$时都需要满足文献[29]里定义的三个条件，即滚动条件（Rolling condition）、不滑动条件（No-slip condition）和不扭曲条件（No-twist condition）。但是很遗憾的是，由于这一卷映射操作相当复杂导致计算杂度非

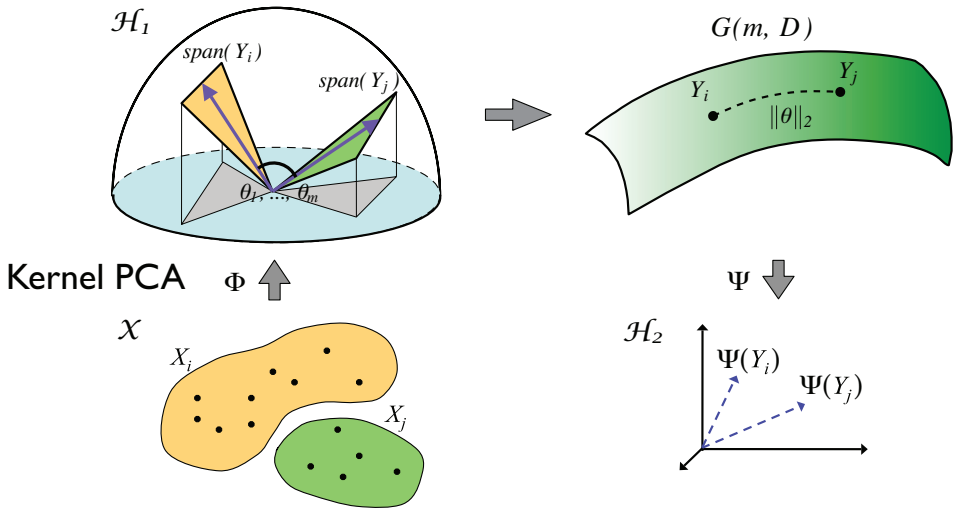


图 1.4. 双核空间嵌入学习框架[54]。它首先对欧氏数据的空间作核嵌入操作得到一些高维的线性子空间，然后再对由这些线性子空间所张成的格拉斯曼流形再作核嵌入操作，最后在这一核空间上学习一个判别函数。

常高，因此在实际应用中非常受限。

1.2.2.2 流形-核空间映射学习策略

另一类黎曼度量学习方法[55, 56, 58, 60, 61, 70, 72, 99, 135, 139, 143]致力于开发在黎曼流形上的正定核函数来将流形嵌入到一个高维的可再生核希尔伯特空间（Reproducing Kernel Hilbert Space, RKHS）上（如图1.2(a)-(c)-(d)所示）。由于希尔伯特空间从广义上来讲是个欧氏空间，所以这一操作也可以看作是从黎曼流形到欧氏空间的转换。在这之后，这些工作可以采用传统在欧氏空间上设计的基于核的分类器或特征提取方法（比如核判别分析方法[20]）对变换之后的元素进行分类或特征学习。这一过程可以表示成从希尔伯特空间到另一个低维欧氏空间的变换。

作为这类方法的一个先驱性的工作，文献[55]通过推导出基于格拉斯曼流形上的投影度量（公式1.4）和比奈-柯西度量（公式1.5）的核函数将原始的格拉斯曼流形作核空间嵌入，最后采用传统的核方法（kernel LDA）来进行判别学习。具体地，基于投影度量的内积核函数和基于比奈-柯西度量的内积核函数分别定义为：

$$k_p(\boldsymbol{X}_1, \boldsymbol{X}_2) = tr(\boldsymbol{X}_1\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{X}_2^T) = \|\boldsymbol{X}_1^T\boldsymbol{X}_2\|_F^2. \tag{1.18}$$

$$k_{bc}(\boldsymbol{X}_1, \boldsymbol{X}_2) = (\det \boldsymbol{X}_1^T\boldsymbol{X}_2)^2 = \det \boldsymbol{X}_1^T\boldsymbol{X}_1\boldsymbol{X}_2^T\boldsymbol{X}_2. \tag{1.19}$$

在这一工作的基础上，另一些工作[56, 58, 60, 72]推导出了更多的在格拉斯曼流形上的正定核函数，比如基于拉普拉斯（Laplace）形式和二项式（Binomial）形式的正定

核函数：

$$k_{l,bc}(\boldsymbol{X}_1, \boldsymbol{X}_2) = exp(-\beta\sqrt{1 - |\det(\boldsymbol{X}_1^T \boldsymbol{X}_2)|}). \tag{1.20}$$

$$k_{l,p}(\boldsymbol{X}_1, \boldsymbol{X}_2) = exp(-\beta\sqrt{p - \|\boldsymbol{X}_1^T \boldsymbol{X}_2\|_F^2}). \tag{1.21}$$

$$k_{bi,bc}(\boldsymbol{X}_1, \boldsymbol{X}_2) = (\beta - |\det(\boldsymbol{X}_1^T \boldsymbol{X}_2)|)^{-\alpha}. \tag{1.22}$$

$$k_{bi,p}(\boldsymbol{X}_1, \boldsymbol{X}_2) = (\beta - \|\boldsymbol{X}_1^T \boldsymbol{X}_2\|_F^2)^{-\alpha}. \tag{1.23}$$

另外一些工作[70, 72, 135, 139]也是在类似的核嵌入的框架下，对对称正定矩阵流形推导出了基于对数欧氏度量（公式1.13）的内积形式和高斯形式的核函数：

$$k_{\log}^{i}(\boldsymbol{X}_1, \boldsymbol{X}_2) = tr(\log(\boldsymbol{X}_1)\log(\boldsymbol{X}_2)). \tag{1.24}$$

$$k_{\log}^{g}(\boldsymbol{X}_1, \boldsymbol{X}_2) = exp(-\|\log(\boldsymbol{X}_1) - \log(\boldsymbol{X}_2)\|_F^2/2\sigma^2). \tag{1.25}$$

如图1.4所示，还有一些方法[54, 61, 99, 143]采用了双核空间嵌入的方法，它们首先对欧氏数据的空间作核嵌入操作得到一些高维的黎曼数据，然后再对由高维的黎曼数据所张成的黎曼流形再作核嵌入操作最后学习一个判别函数。比如，工作[61]首先将原始数据映射到一个高维的希尔伯特空间上，然后在这高维空间上计算对应的协方差矩阵（可正则化成SPD矩阵），最后基于杰弗里斯（Jeffreys）或斯坦（Stein）散度的核函数对这高维的SPD矩阵张成的SPD流形再嵌入到一个高维的核空间上。具体地，基于杰弗里斯和斯坦的核函数可分别形式化成：

$$k_j(\boldsymbol{X}_1, \boldsymbol{X}_2) = exp(-\beta\frac{1}{2}tr(\boldsymbol{X}_1^{-1}\boldsymbol{X}_2 + \boldsymbol{X}_2^{-1}\boldsymbol{X}_1)). \tag{1.26}$$

$$k_s(\boldsymbol{X}_1, \boldsymbol{X}_2) = exp(-\beta(\log\det(\frac{1}{2}\boldsymbol{X}_1 + \frac{1}{2}\boldsymbol{X}_2) - \frac{1}{2}\log\det(\boldsymbol{X}_1\boldsymbol{X}_2))). \tag{1.27}$$

一些研究[55, 60, 61, 72, 99, 139]均验证了这一核嵌入的学习策略取得了比只用切空间估计的学习策略更好的分类性能。由文献[60]所述，这主要是因为切空间只是对流形上的真实黎曼几何作第一阶估计，而一个更高维的RKHS能更好地刻画流形的非线性结构。但是这一类方法还是有挺多缺陷。它们实际上也是首先将位于单位矩阵的切空间上的矩阵对数转成了向量形式，然后再学习更加有判别性的向量表示。然而，由于矩阵对数（即位于单位矩阵的切空间上的基本元素）并不是普通矩阵而是对称矩阵，因此这些方法的向量化操作不可避免地破坏了切空间的内部几何结构。另外，虽然这些工作通过在黎曼流形上推导出有效的核函数从而有效地利用了一些开发在欧氏空间的经典核学习算法，但是这些工作同时引入了传统核方法的一些缺陷。比如，推导出来的核函数需要通过复杂的理论证明来满足Mercer定理才能产生一个有效的可再生核希尔伯特空间。而且，在传统核方法中，变换到希尔伯特特空间上的数据通常是隐式的，只有它们之间的相似性（或内积）可以通过核函数来得到。另外，传统核方法随着数据量的增大，其计算核矩阵的时间复杂度将呈指数级增长。

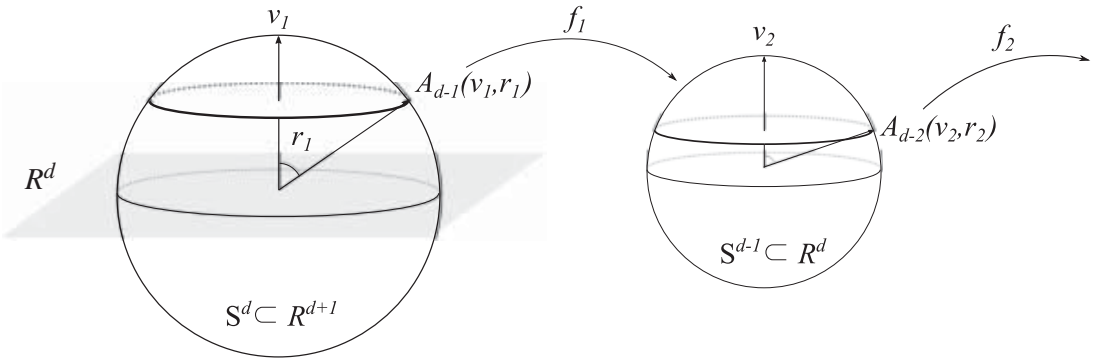


图 1.5. 流形-流形投影学习框架[76]通过变换f_1将在球体S^d里的子球体$A_{d-1}(v_1, r_1)$与另一球体S^{d-1}建立关联。递归地，子球体$A_{d-2}(v_2, r_2)$可以在S^{d-1}里寻找到，并由变换f_2来确定下一个球体S^{d-2}。这里d为球体的维数，$v_1 \in S^d, v_2 \in S^{d-1}, r_1, r_2$为子球体半径。

1.2.2.3 流形-流形映射学习策略

如上所述，不管是流形-欧氏空间映射学习策略还是流形-核空间映射学习策略都是将黎曼流形估计为一个欧氏空间。然而，黎曼流形是一个局部同胚于欧氏空间的微分流形，这两种学习策略显然不可避免地破坏了原始流形的黎曼几何结构。为了解决这一问题，最近两个工作[59, 76]研究了流形-流形映射学习问题。

文献[76]提出了一个从高维的球体到降维后的子流形的投影学习框架。类比于传统的PCA（Principal Component Analysis）降维技术，这一方法主要学习从高维球体到低维子球体的投影矩阵。具体地，如图1.5所示，这一工作定义了一系列从球体到子球体的映射，通过变换f_1将在球体S^d（这里d为球体的维数）里的子球体$A_{d-1}(v_1, r_1)$（这里$v_1 \in S^d, r_1$为子球体半径）与另一球体S^{d-1}建立关联。递归地，子球体$A_{d-2}(v_2, r_2)$可以在S^{d-1}里寻找到，并通过变换f_2来确定下一个球体S^{d-2}。那么对于每个k，子球体A_{d-k}是在不同的空间S^{d-k+1}所定义的。这一工作采用最小化重构残差的方式来计算得到每个球体里的子球体。另外，给定$v_k \in S^m$和$r_k \in (0, \pi/2)$，这一系列的变换可以形式化成：

$$f_k(x) = \frac{1}{\sin(r_k)} R^-(v_k)x, \quad x \in A_{d-k}. \tag{1.28}$$

$$f_k^{-1}(x^\dagger) = R^T(v_k) \begin{bmatrix} \sin(r_k)x^\dagger \\ \cos(r_k) \end{bmatrix}^T, \quad x^\dagger \in S^{d-k}. \tag{1.29}$$

其中，$R(v_k)$是一个大小为$(m+1) \times (m+1)$的旋转矩阵，$R^-(v_k)$是一个包含$R(v_k)$的前m行的大小为$m \times (m+1)$的矩阵。

在求得子球体和相应的变换之后，一个嵌套（nested）球体可以由位于原始空间S^d里的子球体所定义。从形式上来讲，一个$(d-k)$维的嵌套球体$\mathcal{A}_{d-k}$最终可以通过

以下公式计算得到：

$$\mathcal{A}_{d-k} = \begin{cases} f_1^{-1} \circ f_2^{-1} \circ \ldots f_{k-1}^{-1}(A_{d-k}), & (k = 2, \ldots, d-1) \\ A_{d-k}, & (k = 1). \end{cases} \tag{1.30}$$

最近，基于类似的流形-流形映射学习的思想，文献[59]开发了一个从原始对称正定矩阵流形$\mathbb{S}_+^d$到更低维的、更具判别性的另一个对称正定矩阵流形$\mathbb{S}_+^k$之间的映射学习框架。具体地，给定一个矩阵$\boldsymbol{X} \in \mathbb{S}_+^d$，这一流形-流形映射可以形式化成：

$$f(\boldsymbol{X}, \boldsymbol{W}) = \boldsymbol{W}^T \boldsymbol{X} \boldsymbol{W}. \tag{1.31}$$

其中，变换矩阵$\boldsymbol{W}$需要满足满秩约束的条件才能使得新生成的$\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{W}$是正定的，从而张成了一个新的对称正定矩阵流形。为了更好求解参数矩阵$\boldsymbol{W}$，这一工作对它又额外施加了一个正交约束，即$\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_m$，其中$\boldsymbol{I}_m$是单位矩阵。这一约束要求流形配备的度量需要满足仿射不变的属性。因此，这一框架只能包容如仿射不变度量以及斯坦散度等流形上的距离度量。为了让新生成的对称正定矩阵流形上的两个元素之间的度量更具判别性，本工作设计了一个类似图嵌入（Graph-Embedding）框架的目标函数：

$$\boldsymbol{W}^* = \arg \min_{\boldsymbol{W} \in \mathbb{R}^{d \times k}} \sum_{i,j} \boldsymbol{A}_{ij} \delta^2(\boldsymbol{W}^T \boldsymbol{X}_i \boldsymbol{W}, \boldsymbol{W}^T \boldsymbol{X}_j \boldsymbol{W}) \quad s.t. \boldsymbol{W}^T \boldsymbol{W} = \boldsymbol{I}_m. \tag{1.32}$$

其中，$\delta^2(\boldsymbol{W}^T \boldsymbol{X}_i \boldsymbol{W}, \boldsymbol{W}^T \boldsymbol{X}_j \boldsymbol{W})$可由在对称矩阵流形上的仿射不变度量（公式1.10）来计算得到。$\boldsymbol{A}_{ij} = \boldsymbol{G}_w - \boldsymbol{G}_b$, 这里$\boldsymbol{G}_w, \boldsymbol{G}_b$分别是类内图和类间图的关联矩阵：

$$\boldsymbol{G}_w(i,j) = \begin{cases} 1, & \text{if } \boldsymbol{X}_i \in N_w(\boldsymbol{X}_j) \text{ or } \boldsymbol{X}_j \in N_w(\boldsymbol{X}_i) \\ 0, & \text{oherwise} \end{cases} \tag{1.33}$$

$$\boldsymbol{G}_b(i,j) = \begin{cases} 1, & \text{if } \boldsymbol{X}_i \in N_b(\boldsymbol{X}_j) \text{ or } \boldsymbol{X}_j \in N_b(\boldsymbol{X}_i) \\ 0, & \text{oherwise} \end{cases} \tag{1.34}$$

其中，$N_w(\boldsymbol{X}_i)$表示$\boldsymbol{X}_i$的同类v_w-近邻样本，$N_b(\boldsymbol{X}_i)$表示$\boldsymbol{X}_i$的异类v_b-近邻样本。

1.3　视频人脸识别概述

相比静态人脸图像，视频人脸序列可以提供更加丰富的人脸动态时序信息和多视空间信息。如果能充分利用视频中的这些丰富的时空信息那就可以为非限定条件下的人脸识别问题的解决带来机遇。本节首先介绍一下基于视频的人脸识别问题，然后再系统回顾一下现有基于视频的人脸识别方法。

表 1.1. 三种不同的基于视频的人脸识别场景

目标集 \ 查询集	静态图像	动态视频
静态图像	/	视频-图像
动态视频	图像-视频	视频-视频

1.3.1 视频人脸识别问题

基于视频的人脸识别任务需要对出现在视频序列中的人脸图像识别出其对应的身份。具体地，如表1.1所示，基于视频的人脸识别一般研究三种不同的场景，即视频-图像（Video-to-Still, V2S）、图像-视频（Still-to-Video, S2V）和视频-视频（Video-to-Video, V2V）人脸识别。其中，视频-图像人脸识别场景需要将待查询的视频序列与目标数据库里的静态图像（如通缉照片、身份证、驾驶证等）进行匹配从而识别出查询视频中的人脸身份。这一视频人脸识别场景在黑名单监控系统里是非常常见的。相反，图像-视频人脸识别任务需要在由视频序列构成的数据库里查询出输入静态图像中的人脸身份。这一识别场景可以应用于在监控视频数据库里查询给定的身份证照片来定位可疑人物。第三种基于视频的人脸识别场景是视频-视频人脸识别，它主要是在一个由视频序列构成的目标集合里查询在给定视频中出现的人脸身份。比如，它可以应用于通过匹配两段在不同位置拍摄的视频序列来跟踪敏感人物。

在视频监控系统采集条件下，人脸图像质量受人的活动以及环境的影响是巨大的。在通常情况下，视频中的人脸帧一般是些低质量的（如模糊现象严重、低分辨率）、大角度姿态的、真实复杂光照下的图像。基于视频的人脸识别研究的主要任务是关注如何在非常差的可见条件下取得精准的人脸识别效果。具体地，Zhou和Chelleppa [158] 指出了视频数据能够提供以下三种有用属性来帮助更有效的人脸识别：

- 一系列观察对象：由于一个视频序列通常包含来自同一个人的多张人脸图像，所以它提供了这一人脸在不同条件下的表观信息。
- 时间动态：视频序列中包含了人脸图像的时序信息。
- 三维信息：一个视频序列展示了同一个物体在不同角度的表观，比如二维视频隐式地包含了三维的人脸几何信息。

而针对在非可控条件下的视频人脸识别应用，一些不利因素包括：

- 姿态变化：非可控摄像机采集的是处于不同角度的非理想的人脸图像，从而使得在不同图像上，人脸的各个像素之间的对应关系是不同的。

- 光照变化: 拍摄对象通常经历一些极为不同的光照角度及强度，因此人脸表观在不同的时间点通常是不一样的。
- 表情变化: 人脸的表观随着人脸的表情变化而变化。
- 尺度变化: 随着采集对像移向或远离摄像机，其人脸将会在视频帧里占据更大或更小的区域。在这一情况下，当视频帧中的人脸分辨率小到一定程度时将很难被识别。同时，图像的空间分辨率也取决于摄像机的属性，比如其镜头场深度。
- 运动模糊: 如果摄像机的曝光时间设置得太长或者摄像头移动得过快都可能导致拍摄出来的人脸会出现严重模糊现象。
- 遮挡: 在真实世界里，视频中的人脸往往会被某些位于所处环境中的物体所遮挡，从而导致从视频中识别人脸甚至将人脸从背景中区分出来都会变得困难许多。

以上这些因素都可能使得同一个人在不同的场景下拍摄的表观差异要大于不同人在相同场景下拍摄的表观差异。尽管姿态和光照是对传统意义的人脸识别影响最大的两个不利因素，然而文献[109]提供了一些证据表明了以上其它因素同样也会极大影响在非可控场景下的视频人脸识别的性能。

1.3.2　视频人脸识别方法

相比于传统基于图像的人脸识别，基于视频的人脸识别研究还相对比较不成熟。所以直到现在，目前只有三篇完整的关于基于视频的人脸识别的综述文献[17, 137, 150]。除此之外，文献[157] 提供了一个广义的人脸识别综述。他们将人脸识别分为静态图像、多模态和时空三大类，并在关于视频人脸识别的章节讨论了在当时属于最优的人脸跟踪与三维建模方法。还有Zhou和Chelleppa [158]阐述了他们提出的概率身份描述（probabilistic identity characterization）框架用以解决基于视频的人脸识别问题并在文章里综述了视频的一些属性。

参照文献[17]，基于视频的人脸识别方法通常可以分成两大类：基于序列（sequence-based）的方法和基于集合（set-based）的方法。基于序列的方法显式地利用了时序信息和空间信息来提高在采集条件比较差的视频人脸识别的效果。比如，基于时序建模的方法[21, 34, 96, 153]通常只开发人脸的运动信息来进行人脸识别以及基于时空建模的方法[8, 19, 42, 78, 86, 92, 100, 127, 159, 160]通过同时挖掘表观信息和运动线索来获得更加鲁棒的识别结果。与基于序列的方法相比，基于集合的方法把视频序列处理成无序的图像集合进而可以通过融合人脸图像的多视空间信息来进行人脸识别。这类方法主要是通过采用融合视频中人脸的多视图信息或者捕获视频中人脸的变化模式的技术，可以在非限定的采集条件下达到比较鲁棒的识别性能。基于集合的方法一般可以再细分为基于超分辨率、基于三维建模、基于帧选择以及基于统计

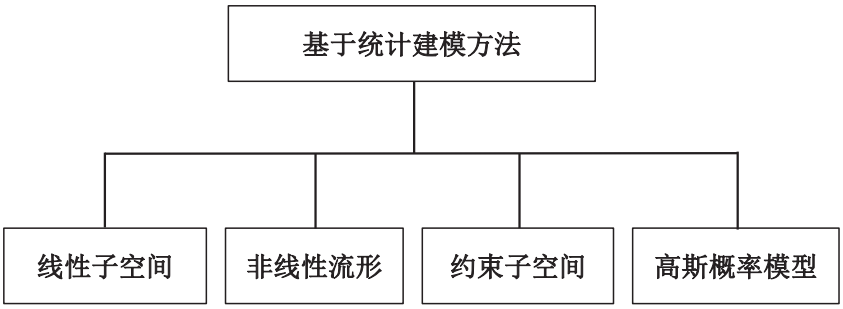


图 1.6. 基于统计建模的视频人脸识别方法

建模四种方法。具体地，基于超分辨率的技术[9, 11, 12, 52, 66, 73, 75, 155, 161, 164]就是在视频人脸数据层面或特征层面上增强视频人脸的分辨率。基于三维建模的技术[91, 103, 104, 125, 148]通过开发在大量的视频集合里的人脸的多视图信息来重构人脸的三维结构，从而可以帮助获得对姿态不变的属性。基于帧选择的方法[23, 43, 53, 98, 122, 145, 156]是在原始视频帧集合里去选择一个包含高质量的视频帧子集或者一个包含表观变化范围比较大的视频帧子集，然后对这些选出来的视频帧在最后的得分层面上进行融合。那目前最主流的基于集合的方法是基于统计建模技术。它们主要是通过采用统计模型来表示每个视频数据的统计量，从而有效地刻画出视频中数据的类内变化模式并以此作为视频识别的特征。更具体地说，基于统计建模的方法通常通过将视频人脸帧集合建模成一个线性子空间或者一个非线性流形或者一个高斯分布来有效地刻画姿态、光照和表情等的变化模式，从而达到对这些变化模式不变的鲁棒性。而且，这一种方法可以采用成熟的统计方法和数学模型来更好地解决这一问题。根据采用的统计模型的类型，本文将基于统计模型的方法又细分为基于线性子空间、基于非线性流形、基于仿射子空间以及基于高斯概率模型四种不同方法（如图1.6 所示）。下面再详细介绍这些基于统计建模的方法。

在传统的人脸识别研究里，研究者普遍认为人脸图像在图像空间可以构成一个低维的线性子空间。在过去十几年里，大量工作[44, 55, 58, 82, 88, 89, 101, 143, 149]采用主成分分析（Principal Component Analysis, PCA）将一个选定人的人脸图像集合表示成一个线性子空间。比如，文献[149] 开创性地提出了互子空间方法（Mutual Subspace Method, MSM），在对人脸图像集合进行线性子空间建模的基础上，采用经典相关性（Canonical Correlation）来度量两个子空间模型的相似度，从而可以对集合进行有效地分类。基于线性子空间主夹角的概念，文献[143]对线性子空间上推导出基于矩阵的正定核函数，然后采用SVM对以线性子空间为基本元素进行分类。除了这一子空间到子空间的距离计算方法，文献[36] 还设计了一种最近特征空间方法（Nearest Feature Space, NFS）计算点到子空间的距离。随着越来越多的图像集合数据被标注，文献[81]提出了判别典型相关方法（Discriminative Canonical Correlation, DCC），采用一种类似LDA的判别学习框架来最大化同类线性子空间的典型相关性同时最小化异类线性子

空间的典型相关性。另一个代表性的工作是文献[55] 提出一种格拉斯曼判别学习方法（Grassmann Discriminant Analysis, GDA）。这一方法借助代数几何的格拉斯曼流形将基于子空间的学习形式转化成了在格拉斯曼流形的分类问题。具体地，它首先定义了各种基于格拉斯曼流形的度量的核函数，然后通过采用Kernel LDA的学习框架在流形上学习判别函数。同样，文献[58]在格拉斯曼流形上采用了一个更加广义的图嵌入判别学习框架（Grasmannian Graph-embedding Disriminant Analysis, GGDA）来学习对应的判别函数。

当视频里的人脸图像只包含在某种维度空间（比如光照空间）上的变化，线性子空间是首选的用于集合建模的统计模型之一。然而在真实的世界里，人脸图像比如视频中的人脸帧往往包含了非常丰富的变化模式（比如分辨率、姿态、表情、光照等变化）。为了解决这一问题，有些基于非线性流形的方法[35, 39, 80, 138, 140] 通过开发非线性流形来估计图像集合里的人脸图像的非线性变化模式。比如，文献[140] 提出了一种流形-流形距离距离（Manifold-Manifold Distance, MMD）计算方法。它首先采用最大线性块（Maximal Linear Patch）准则将一个流形分割成多个局部线性子空间模型，然后通过融合两个流形里的所有线性子空间之间的距离（也就是典型相关性）来度量流形-流形的距离。为了学习基于流形的判别信息，文献[138] 设计了一个流形判别学习（Manifold Discriminant Analysis, MDA）方法来最大化流形间隔（manifold margin）从而达到在流形进行有判别地分类的目的。为了增强参与距离计算的子空间对间的对应关系，文献[39] 提出一种流形对齐方法使得参与匹配的流形都对齐到一个参考流形上。除此之外，文献[35] 提出一种稀疏估计最近子空间方法（Sparse Approximated Nearest Subspaces, SANS），首先将每个流形包含的子空间都建模在格拉斯曼流形上，然后通过最小化在格拉斯曼流形上的联合稀疏重构残差来计算两个非线性流形间的最近局部线性子空间距离。

虽然基于非线性流形可以同时建模集合里的数据的多种变化模式，但是它要求集合里的数据要足够密集。如果集合数据有限的情况下，这一些方法通常会失效。基于约束子空间（比如，仿射包、凸包等）的统计模型方法[32, 33, 67, 136, 152, 162, 163]通过估计集合里的样本的仿射组合或凸组合来解决此问题。比如，文献[136] 在基于对集合作仿射包/凸包建模的基础上提出了K-局部超平面/凸包距离最近邻算法（K-local Hyperplane/Convex Distance Neighbor, HKNN/CKNN）来度量点到仿射包/凸包之间的距离。文献[33] 采用仿射包的交集和超球面边界来表示集合数据，开发了最近超磁盘边界（Nearest Hyperdisk Bounding, NHD）分类器来计算点到仿射包的距离。之后，文献[32] 又开发了基于仿射包/凸包的图像集距离（Affine/Convex Hull based Image Set Distance, AHISD/CHISD）计算方法来度量两个仿射包/凸包之间的距离。在这一工作的基础上，工作[67, 152, 163]在计算仿射包/凸包之间的距离时加入l_1或l_2约束使得仿射包/凸包之间的距离计算更加鲁棒。除了前面几种无监督的方法，文献[162]提出点到集合以及集合到集合的度量学习（Point-to-Set/Set-to-Set Distance Metric Learning, PSDML/SSDML）

框架来学习更加有判别性的点到仿射包以及仿射包到仿射包的距离度量，从而达到更加鲁棒的基于图像集/视频的人脸识别性能。

与前面几种统计模型不同，基于高斯概率模型的方法一般采用高斯概率分布函数来建模图像集合。比如，文献[114] 将图像集人脸的变化模式建模成人脸图像空间中的单一高斯分布函数，然后采用诸如KL散度（Kullback-Leibler Divergence, KLD）之类的度量来计算两个概率分布函数之间的相似度。显然，这一简单的概率建模无法刻画复杂的模式分布；为了解决这一问题，文献[13] 提出了流形密度散度方法。具体地，他们采用了更加符合真实复杂条件的高斯混合模型（Gaussian mixture models, GMM）来取代单高斯模型。最近，有几个工作[94, 135, 139] 采用高斯分布的二阶统计量即协方差矩阵来建模图像集数据。为了能够更加有效地解决集合分类问题，这些方法首先将协方差矩阵表示成特定流形即对称正定矩阵（Symmetric Positive Definite, SPD）流形上的元素，然后将基于视频的人脸识别形式化成在SPD流形上的分类问题。更具体地，文献[139] 推导出一种基于对数欧氏度量（Log-Euclidean metric）的黎曼内积核函数，然后将其输入到核判别分析（Kernel Discriminant Analysis, KDA）或者核偏最小二乘（Kernel Partial Least Squares, KPLS）框架里在SPD流形上进行有监督地分类任务。为了提高后端的判别能力，最近有些工作[94, 135] 提出了多核/多度量学习的框架来进行更加鲁棒地分类。

1.4 本文主要贡献

如前面章节所述，基于统计建模的视频人脸识别方法主要集中关注两个关键性问题：1）如何对视频序列中的人脸图像集合进行更加鲁棒地统计建模；2）如何对统计模型学习更加有判别性的函数从而进行更加有效的分类。在统计建模问题上，目前大多数方法提出的统计模型一般只在某一方面刻画了视频中的人脸图像集合的的统计信息（比如协方差统计量），而忽视了其他有用的统计信息（比如均值统计量），从而导致识别不够鲁棒。在判别学习问题上，目前一些主流方法（也就是黎曼度量学习方法）通常需要推导出一些基于黎曼度量的核函数来将传统的核方法适配到黎曼流形上进行有监督学习。然而，这些方法也把传统的核方法的一些固有缺陷也带到黎曼流形判别分析的框架里。比如，核方法一般不能学习到显式的映射，而且随着样本数目的增加核方法的复杂度会呈指数级地增长。

针对以上两个问题，本文研究了双阶/多阶统计建模方法以及一系列基于统计模型的黎曼度量学习方法来有效解决视频人脸识别问题。具体地，在统计建模问题上，本文主要提出了将一阶统计统计量（均值）与二阶统计统计量（协方差）甚至与完整的高斯概率模型进行融合的统计建模策略。在判别学习问题上，本文针对基于视频的人脸识别问题，主要从视频的不同统计建模出发设计了一系列有效的黎曼度量学习方法来开展研究。本文旨在黎曼度量学习方法在视频人脸识别问题中的应用研究做了一些

有益的尝试。本文的研究内容与主要贡献总结如下：

- 提出基于线性子空间建模的投影度量学习方法（Projection Metric Learning，PML）

基于视频序列的线性子空间建模，本文提出一种在格拉斯曼流形上的投影度量学习方法来解决视频-视频人脸识别问题。该方法主要采用了流形-流形映射（如图1.7(a)-(d)）的黎曼度量学习策略。具体地，该方法在以线性子空间为基本元素的格拉斯曼流形上，提出一个将原始的格拉斯曼流形变换到一个新格拉斯曼流形上的度量学习框架来学习更具判别性的线性子空间之间的投影度量。该方法不仅可以为作为一种格拉斯曼流形上的度量学习方法，而且也可以成为一种施加于格拉斯曼流形的降维技术。通过系统深入的实验表明该方法在四个极具挑战性的数据库上的性能均达到了与当前最优的基于集合的判别式学习方法同等可比的水平。同时实验也表明了该方法作为降维方法与其它方法结合之后的性能可以在一定程度上超过同类型的方法。

- 提出基于双阶统计量建模的跨欧氏-黎曼度量学习方法（Cross Euclidean-to-Riemannian Metric Learning，CERML）

基于视频序列的双阶统计量建模，本文提出一种跨欧氏-黎曼度量学习框架来同时解决三种不同的基于视频的人脸识别问题，即视频-图像、图像-视频和视频-视频人脸识别。该方法主要采用了流形-核空间映射的黎曼度量学习策略（如图1.7(a)-(b)-(c)）。具体地，该方法采用双阶统计量（即均值和协方差）对视频数据进行建模，然后将这三种视频人脸识别问题统一形式化成异质数据（即欧氏和黎曼数据）的匹配/融合问题，最后提出一个新型的用于匹配/融合异质数据的异质度量学习统一框架。这一新的异质度量学习框架可以通用于不同的二阶统计量，如协方差、线性子空间和仿射子空间，因此成为一个可以包容多数统计模型的通用度量学习框架。通过大量的视频-图像/图像-视频和视频-视频人脸识别实验对该方法进行了验证，最后的实验结果表明了该方法在四个极具挑战性的数据库上均达到了明显优于其它方法的性能。

- 提出基于高斯分布函数建模的对数欧氏度量学习方法（Log-Euclidean Metric Learning，LEML）

基于视频序列的高斯分布函数建模，本文借鉴著名的信息几何理论将高斯模型的空间嵌入到一个对称正定矩阵流形上，进而提出一种在对称正定矩阵流形上的对数欧氏度量学习方法。该方法主要采用了流形-流形映射（如图1.7(a)-(d)）的黎曼度量学习策略。具体地，该方法主要是去学习一个切映射将原始的对称正定矩阵流形上的切空间嵌入到另一个流形的切空间上，目标是要让在新的切空间上的对数欧氏度量更有判

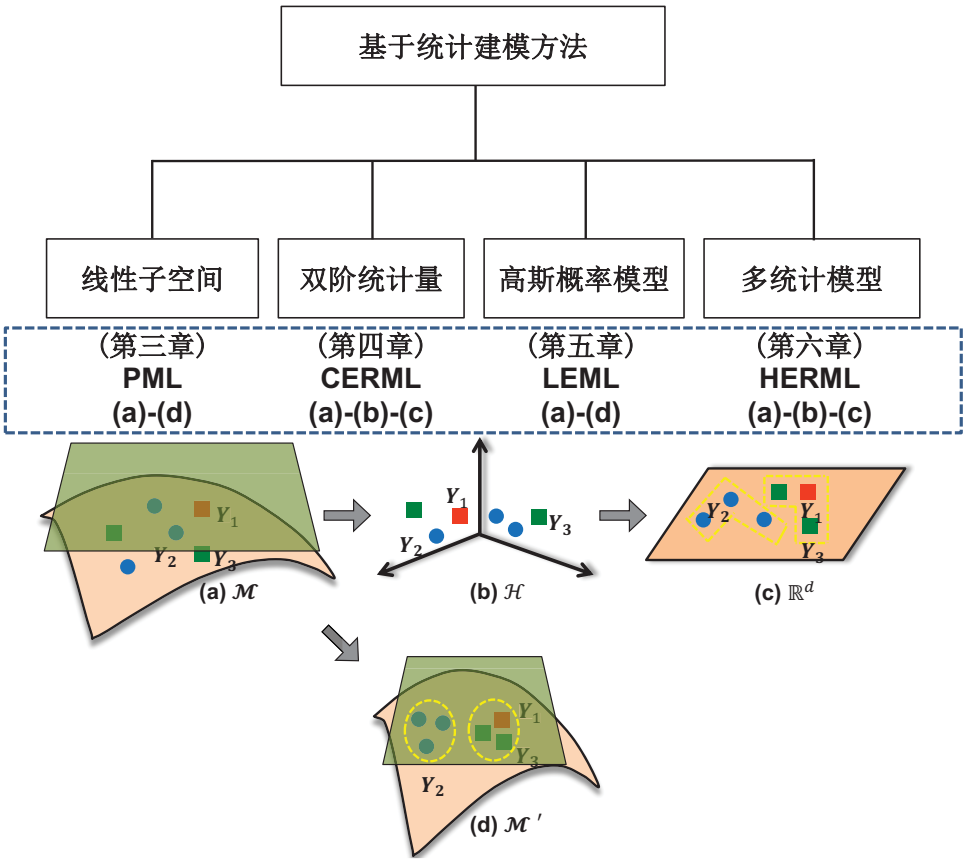


图 1.7. 本文组织结构

别性。新提出方法的有效性通过系统的实验对比进行了验证，在视频-视频人脸识别和视频-视频人脸确认两个任务上，算法均表现出良好的性能。与同类型的基于对称正定矩阵建模的分类方法相比，本文方法取得了与当前最优方法相当甚至更优的性能，从而表明其可以有效地解决视频人脸识别问题。

- 提出基于多统计模型的混合度量学习方法（Hybrid Euclidean-and-Riemannian Metric Learning，HERML）

基于视频序列的多种统计建模，本文提出一种混合欧氏-黎曼度量学习框架来有效融合样本均值、样本协方差和高斯模型这三种统计模型，从而达到更加鲁棒的视频-视频人脸识别。该方法主要采用了流形-核空间映射的黎曼度量学习策略（如图1.7(a)-(b)-(c)）。具体地，为了减少高斯模型所在的空间与其它两种统计模型的空间之间的异质性，该方法首先同样借鉴经典的几何理论将高斯分布的空间嵌入到一个对称正定矩阵流形上，然后设计了一个混合度量学习框架同时在这些统计模型所在的多个异质空间上学习多个马氏矩阵来对它们进行有效融合。实验结果表明该方法在这四个数据库上的两个不同视频人脸识别任务上均达到了当前最好的性能。另外也通过对比不同统计模型的表现可以发现融合这三种统计模型可以提高视频人脸识别的性能。

1.5　本文组织结构

本文的组织结构如下（如图1.7所示）：

第一章为绪论部分。主要介绍了本文的研究背景、黎曼度量学习方法以及视频人脸识别的国内外研究现状。最后给出了本文的主要贡献和组织结构。

第二章为数据准备部分，主要介绍了一些现有的公开视频人脸数据库以及一个新的图像/视频人脸数据库并在其上设计评测协议。

第三章提出了基于视频的线性子空间建模的投影度量学习方法，用于解决视频-视频人脸识别问题，并从理论和实验上进行了分析与验证。

第四章提出了基于视频的双阶统计量建模的跨欧氏-黎曼度量学习方法，用于同时解决视频-图像、图像-视频以及视频-视频人脸识别问题，并与已有的工作进行了全面的对比和讨论。

第五章提出了基于视频的高斯分布建模的对数欧氏度量学习方法，用于解决视频-视频人脸识别问题，并与已有的工作进行了全面的对比和讨论。

第六章提出了基于视频的多种统计建模的混合欧氏-黎曼度量学习方法，用于解决视频-视频人脸识别问题，并与已有方法进行了全面的对比与讨论。

第七章对全文的工作进行了总结。

第二章　数据准备

2.1　引言

为了衡量基于视频人脸的识别研究问题的进展情况，构建一个符合真实场景的视频人脸数据库并在其上设计合适的评测协议是一种必要的手段。在过去的十几年里，国内外的大量研究机构都致力于收集一些基于视频的人脸数据库。参考综述文献[17]，表2.1里通过罗列出一些已经公开的视频人脸数据库来简单描述一下它们的一些关键特性，包括采集人数、视频数以及数据库中视频人脸的变化模式和外界影响因素。从表2.1中可以发现尽管这些数据库都已经被研究者广泛使用，但是大多数数据库还不足以支持基于视频的人脸识别研究。比如说，有相当多的数据库只包含了少量的志愿者或者只采集有限数目的视频。还有一些视频数据库的视频中只包含少量的视频帧。比如，ScFace [50] 数据库的每段视频序列中只抽取了3帧人脸图像。另外，在有些数据库如UT Dallas [37] 中，视频一般是由同一个摄像机采集的。更重要的一个问题是，大部分的数据库只设计了三种基于视频的人脸识别场景（视频-图像、图像-视频和视频-视频）中的一种或两种。比如，YouTube Celebrities [79] 数据库主要是为视频-视频的人脸确认任务而采集的，而PaSC [24] 数据库只设计了视频-图像和视频-视频的人脸确认协议。在数据库规模方面，本文发现只有YouTube Face DB [142] 和Celebrity-1000 [90]数据库包含超过1,000人的视频数据。除此之外，在几乎所有的数据库中，大多数的志愿者都来非亚洲地区，因此也带来严重的肤色上的偏差。

为了弥补现有数据库的不足，本章准备了一个大规模的模拟真实世界的视频监控场景的图像/视频人脸数据库（命名为COX[1]人脸库）来评测三种不同的基于视频的人脸识别任务，即视频-图像、图像-视频和视频-视频人脸识别。与大多数现有的数据库相比，COX人脸库包括了更多的采集对象（1,000人），更多的视频序列（3,000段）以及更多在姿态、表情、光照、模糊和分辨率上呈自然变化的视频帧。除此之外，由于所有的志愿者都来自中国，因此这一数据库可以作为现有数据库的一个很好的补充。更重要的是这一数据库不仅包含了由三个不同摄像机采集的视频集合，而且还收集了由一个数码相机拍摄高质量的静态正面人脸图像集合。

本章内容安排如下：2.2节简要回顾现有的视频人脸数据库；2.3节详细描述新收集的COX人脸数据库；2.4节对本章内容进行总结。

[1]COX的命名来源于它是由中国科学院计算技术研究所（Chinese Academy of Sciences）、欧母龙（OMRON Social Solutions Co. Ltd）和新疆大学（Xinjiang University）一起合作创建的一个数据库。

表 2.1. 现有的视频人脸数据库。其中，“变化模式”包含: 姿态（pose (p)）、光照（illumination (l)）、表情（expression (e)）、分辨率（resolution (r)）、运动模糊（motion blur (b)）和行走（walking (w)），“场景”包括视频-图像（V2S）、图像-视频（S2V）和视频-视频（V2V）三种人脸识别场景。

Datasets	人数	视频数	变化模式	场景
CMU MoBo [51]	25	150	w	V2V
First Honda/UCSD [85]	20	75	p	V2V
Sec. Honda/UCSD [86]	15	30	p	V2V
CMU FIA [46]	214	214	p,l,e	V2V
CamFace [2]	100	1,400	p,l	V2V
Faces96 [1]	152	152	l,r	V2V
VidTIMIT [112]	43	43	p,e	V2V
YouTube Celebrities [79]	47	1,910	p,l,e,r,b,w	V2V
MBGC [108]	821	3,764	p,l,e,r,b,w	V2S/V2V
ND-Flip-QO [18]	90	14	l,e,r,b	V2V
YouTube Faces DB [142]	1,595	3,425	p,l,e,r,b,w	V2V
Chokepoint [145]	29	48	p,l,r,b,w	V2V
ScFace [50]	130	910	p,l,r	V2S/V2V
UT Dallas [102]	284	1,016	p,l,e,r,b,w	V2S/V2V
UMD Comcast10 [37]	16	12	p,l,r,b,w	V2V
PaSC [24]	2,65	2,802	p,l,e,r,b,w	V2S/V2V
Celebrity-1000 [90]	1,000	7,021	p,l,e,r,b,w	V2V

2.2 现有的视频人脸数据库

参考综述文献[17], 本节在表2.1罗列出更多的已经公开的视频人脸数据库。下面简单介绍一下这些数据库的基本信息。

CMU Motion of Body（MoBo）[51] 数据库包含25个人在跑步机上以四种不同的方式行走的共有150段视频序列。行走的方式包括慢速行走、快速行走、斜面行走和拿球行走四种不同的场景。每个人的视频数据是由6个摄像机采集的。每段视频的帧率是30帧每秒。

Honda和UCSD一起收集了两个包含大范围姿态变化的视频人脸数据库。第一个数据库[85]包括了来自20个人的75段视频，第二个数据库[86]收集了来自15个人的30段视频。每段视频的分辨率为640×480，帧率是15帧每秒，持续时间至少是15秒。由于两个数据库都是在室内采集，因此光照变化不是很大。

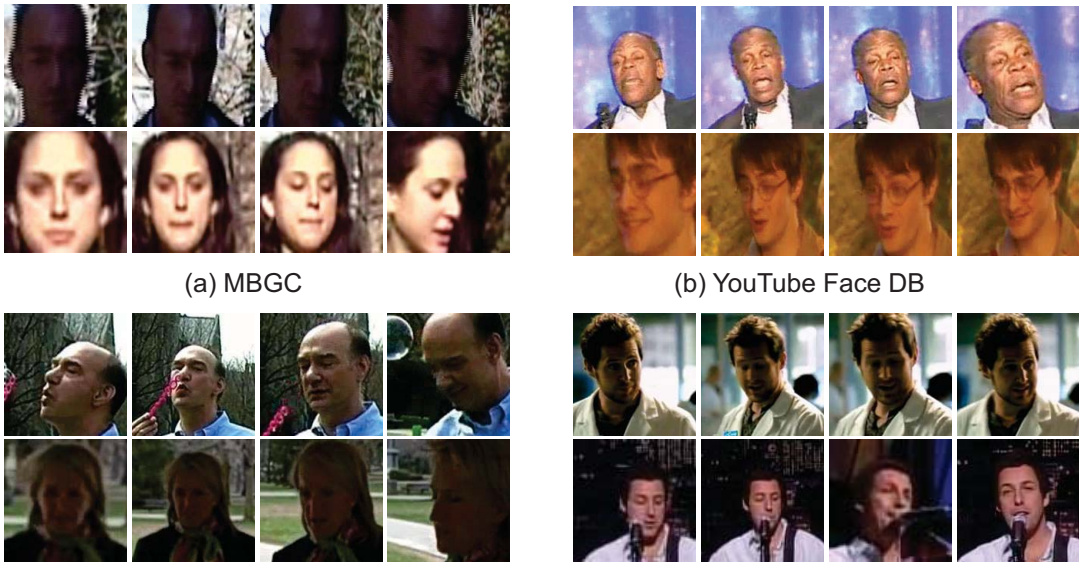


图 2.1. 几个代表性视频人脸数据库示例

CMU Face in Action（FIA）[46] 数据库收集了20段来自214个人，分辨率为640×480的视频。在视频里，拍摄对象随机调整他们的人脸表情和姿态来模拟护照登记时的场景。这些视频序列是分别在室内和室外采集的，帧率是30帧每秒。很多参与拍摄的人都是分三个时间段过来采集数据。

CamFace[2] 数据库里包含了来自不同年龄、不同种族的67个男士和33个女士的视频数据。每个人有14段分辨率为320×240，帧率为10帧每秒的视频。这些视频片断包含了多种光源的不同设置。在视频里，采集对象是自由行走并带有各种不同的姿态，但表情变化相对较小。

Faces96 [1]数据库收集了来自152个人的分辨率为196×196的视频序列。这些视频序列是在室内采集的，包括了比较大的姿态以及光照的变化。每个视频是在同一天采集的，帧率为每秒0.5帧。

VidTIMIT [112]数据库采集了一些来自24个男士和19个女士在办公室里谈话场景的视频。采集对象在可控的情况下旋转他们的头部。该数据库采用了一个broadcast质量的视频摄像机来采集视频序列。

YouTube Celebrities [79] 数据库包括了1,910段从YouTube上收集到的由47个演员和政治家构成的视频序列。在这一数据库里，大部分的视频是低分辨率、高压缩的。除此之外，姿态、光照和表情也是高度非可控的。

Multiple Biometric Grand Challenge（MBGC）[108] 数据库采集了来自821个人的3,764段视频。这一数据库是在室内和室外采集的，包括了多种不同变化的视频序列。这些视频同时采集了志愿者在行走、活动和谈话时的场景下的正面和非正面姿态的人脸序列。因此这一数据库包括了非可控的光照、移动、姿态等变化。

ND-Flip-QO [18]数据库收集了包括由90个人构成的14段人群视频（分辨率为640×480）的一个视频数据库，其中有5个人出现在多个视频里，其余的85人只出现在一个视频里。这些视频由一个Flip摄像机在室内和室外两个不同场景下拍摄的。在拍摄过程中，被采集者允许自由改变人脸表情。

YouTube Faces DB [142]数据库的所有视频均是从YouTube上收集的，即全部是非可控条件下采集的。YouTube数据库包含1,595个人的3,425段视频，平均每个人有2.15段视频，每段视频的平均长度约为181帧。因为是在YouTube采集的，这些视频包括比较大的姿态、光照、分辨率等变化。

Chokepoint [145]数据库是为真实世界监控条件而设计的视频人脸数据库。这一数据库主要拍摄志愿者通过室内门口的场景，在第一个门口采集了25个人的视频序列，在第二个门口采集了29个人的视频片断。在两个门口采集的时间间隔有1个月。每个视频的帧率是30帧每秒，图像分辨率为800×600像素。这一数据库总共有48段视频序列。除此之外，这一数据库还采集了64,204静态人脸图像。

ScFace [50]数据库的视频人脸图像是在一个非可控的室内采集的。这一数据库包含了130人的数据，每个人由7个不同的摄像机拍摄的7小段视频片断，每小段只有3帧人脸图像（每帧人脸图像是志愿者在固定的三个位置上进行采集）。另外，每个人还有11张静态人脸图像。

UT Dallas [102]数据库包含1,016段视频序列，其中510段是志愿者在行走（正面人脸）的视频以及另外506段是志愿者在任意活动（非正面人脸）的视频。视频中的人脸帧的分辨率是720×480像素。

UMD Comcast10 [37]数据库包含来自16个人一起合拍的12段视频。这些视频包括站立没有行走的视频以及志愿者向摄像机走近的视频。其中，每个视频的分辨率为1,920×1,080。

Point-and-Shoot Challenge（PaSC）数据库不仅采集了来自293个人的9,376张静态图像而且也收集了来自265个人（是上面293人的子集）的2,802段视频。在视频采集过程中，每个人都执行了某种简单的动作，比如接电话、吹泡泡、拿书本等，每种动作用手持相像和固定摄像机拍摄。

Celebrity-1000 [90] 是从YouTube和YouKu上收集到的来自1,000个名人的7,021段视频序列。这些视频场景包括采访现场、演唱会和新闻发布会等。数据库中的绝大多数视频包括比较大的姿态、光照、分辨率等变化。

2.3 新采集的COX数据库描述

本节将详细介绍新收集的COX人脸数据库的一些细节，包括数据采集、数据处理以及在数据库上的协议设置。

2.3.1　数据采集

为了采集一个能够评测三种不同的基于视频的人脸识别场景（特别是像视频监控之类的应用场景），需要尽可能地采集符合真实世界条件下的视频序列。这样的视频应该不仅包括了在人脸姿态、人脸表情、人脸分辨率以及环境光照方面的丰富的变化，还包含了各种有关图像质量方面的噪声和模糊现象。为了设计视频-图像和图像-视频的人脸识别场景，也需要采集每个志愿者的静态人脸图像来模拟身份证上的人脸照片。基于以上这些考虑，本节精心设计了以下一系列的图像/视频采集程序。

2.3.1.1　数据采集途径

所有的视频序列都是在一个大型的体育场馆里采集的。这一体育馆一边是透明大玻璃墙，顶上是一个非常高的天花板。这样的环境除了能保护志愿者和采集设备不受意外事件干扰，还能形成一个带有复杂的半室外光照的图像/视频采集环境。另外，为了模拟真实的夜晚光照条件，还有一部分视频是在晚上体育馆里打开日光灯的条件下采集的。因此，在新数据库里，打在人脸上的光照都是非常自然的而且也非常接近于许多实际应用。

2.3.1.2　数据采集设备

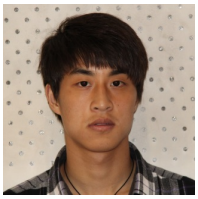
用来采集新数据库的设备包括3台摄像机和1台数码相机。具体地，采用Cannon EOS 500D DC来为志愿者采集高质量的静态图像，使用3台SONY HDR-CX350E DV 并将它们固定在3个约3米高的三角架上采集志愿者的视频序列。关于如何使用这些设备采集静态图像和视频序列的更详细介绍请参考接下来的小节。

2.3.1.3　数据采集设备

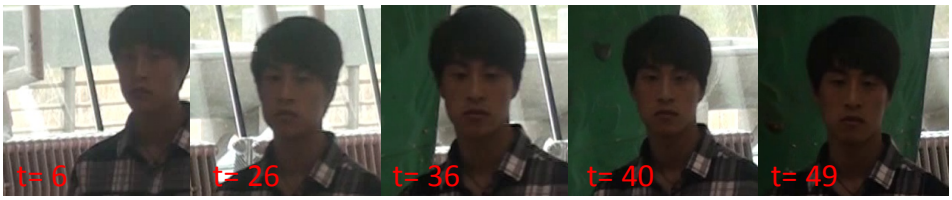
对每个志愿者进行静态图像采集是在一个拥有标准室内光的采集室里进行的。为了能够采集接近身份证照片的人脸图像，数码相机被搭置在一个离志愿者3米远的三角架上。志愿者被要求坐在一张普通的椅子上并保持正面及中立表情。在采集过程中，相机的闪光灯一直处于打开状态来防止一些阴影现象的产生。图2.2（a）中是一张静态图像示例。

2.3.1.4　视频序列采集

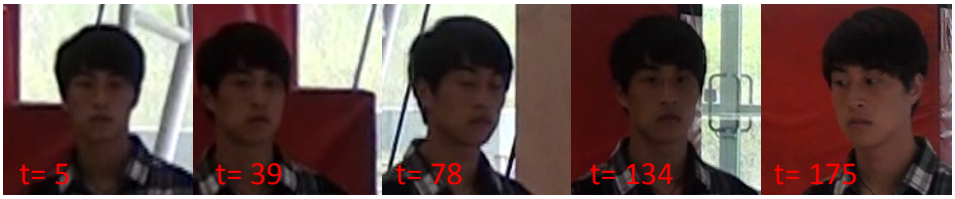
为了模拟视频监控场景，这一数据库采集了志愿者在行走状态下的视频序列。为了采集人脸更多的变化模式，需要事先精心设计一个行走路线以及摄像机的摆放设置。具体地，如图2.3所示，志愿者被要求大致沿着S形的路线从指定的起点自由行走到指定的终点。3台标记为Cam1、Cam2和Cam3的摄像机分别被固定在3个不同的位置，并均被架在约2米高的三角架上。如图2.2（b）（c）（d）所示，这三台不同的摄像机分别采



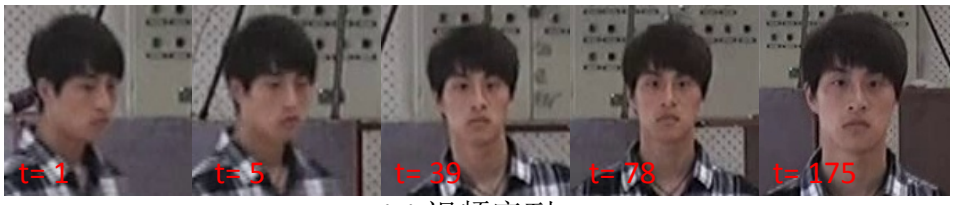
(a) 静态图像



(b) 视频序列1



(c) 视频序列2



(d) 视频序列3

图 2.2. COX人脸数据库示例。(a)由数码相机采集的静态图像示例。(b)(c)(d)分别是由三台不同摄像机采集的视频序列示例。其中，每张视频帧上标红的t值表示当前帧在视频序列中的序号。

集志愿者在S型路线的不同方位（也就是三段分别被标成红色、绿色和蓝色的路线）上行走时的视频序列。

很显然，设计这样的一个S型路线有利于采集志愿者更多的人脸变化模式包括在姿态、光照、模糊状态以及人脸分辨上的变化。实际上，随着志愿者在S形的两个半圆上行走，他会很自然地连续调整人脸的方向，从而导致人脸姿态的改变。除此之外，由于采集室里的大玻璃能透进室外光，因此人脸上的光照变化也是足够大的。

2.3.2 数据处理

在采集完1,000人的视频序列之后，还需要这些数据进行预处理以方便未来的数据库评测研究。之所以要做数据处理主要是因为在视频采集的过程中，摄像机一直处在充电状态并采集了很多志愿者没有出现在屏幕内的视频序列。除此之外，为了达到方便评测的目的，所有的视频帧中要求只包含志愿者的头肩部分甚至只有头部。

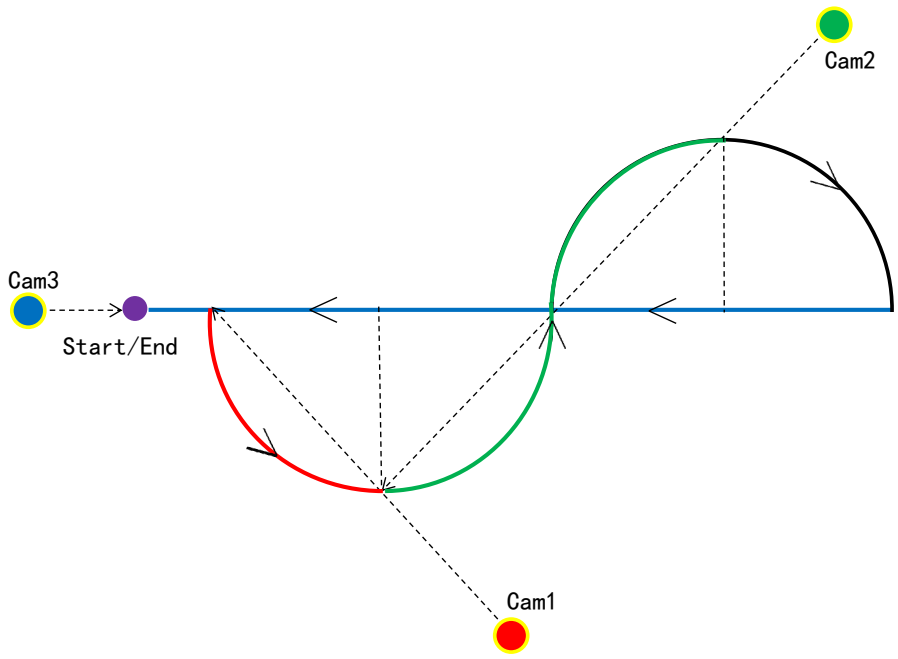


图 2.3. 志愿者的行走路线和相机的摆放设置。每个志愿者被要求沿着一条S形路线从指定的起点自由行走到指定的终点。3台摄像机Cam1、Cam2、Cam3 被分别固定在3个不同的位置，分别采集志愿者行走在标成红色、绿色以及蓝色的路线上时的视频序列。S形中两个半圆的半径均为3米。

为了满足以上要求，需要首先将每段原始长视频序列切割成多段较短的视频片断。其中，每个片断只包含一个志愿者在从起点出发到离开视野的整个视频序列。所有的长视频序列最终被切分成了3,000段视频片断。在这之后，通过采用一个商业人脸检测软件（OKAO[2]）来检测在视频中出现的人脸。然而，由于在这一真实场景下的人脸检测结果通常不是非常理想，因此会产生一些不准确甚至错误的检测结果。为了方便后续的处理，开发了一种类似人脸跟踪的策略来去除一些可能不准确或错误的检测结果。具体地，如果在某帧中检测出来的人脸中心离前几帧的人脸中心太远的话，这一预处理程序会将这一张人脸帧当作异常帧并将它去掉。尽管这一过程不可避免地导致少量视频帧的丢失，但是实际上这并不影响最终的评测。另外，还需要特别指出的是由于不同志愿者的行走速度都非常不同，从而导致了数据中的视频序列的长度都不尽相同。图2.4分别统计了3台摄像机采集的所有视频片断中包含的帧数情况。从图中可以看到大部分的视频序列都包含了超过100张单个人的视频帧，特别是Cam3里的大多数视频序列包含了约有170张视频帧。

接下来，还需要做一些数据处理来减少数据库的最终存储量。由于3台摄像机采集的都是分辨为1,920×1,080的视频序列，这样导致非常大的存储要求从而会影响数据库的广泛传播。而且，由于这些摄像机实际上是以隔行模式来采集视频，因此通过采用

[2]http://www.omron.com/r_d/technavi/vision/okao/detection.html

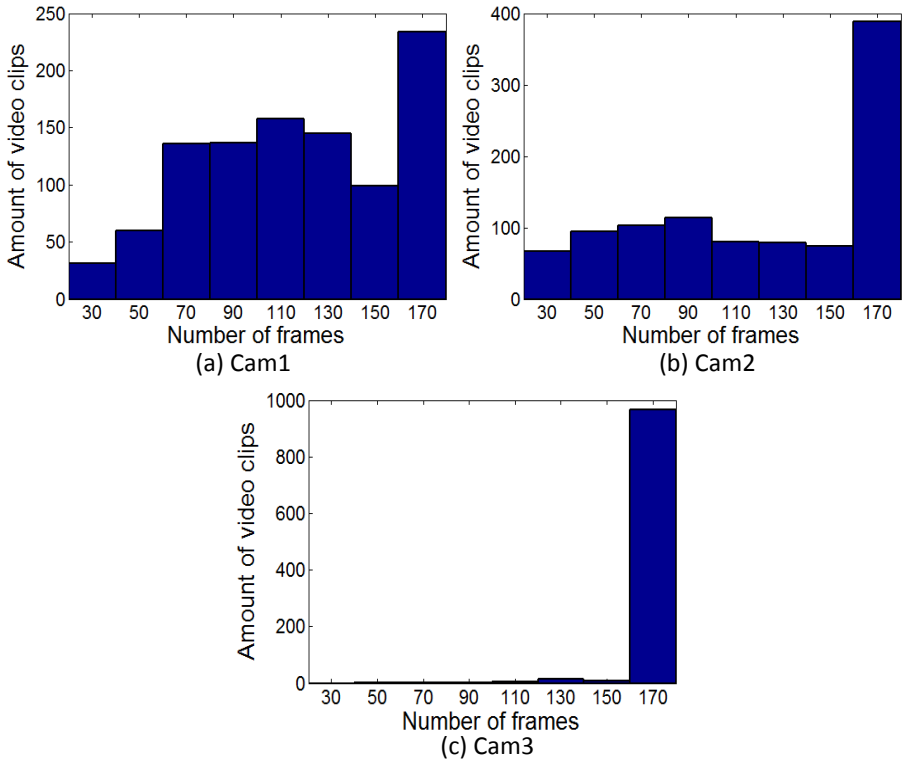


图 2.4. 三台不同摄像机采集的视频序列在帧数上的统计

一个商用工具（Aunsoft Final Mate[3]）来对所有的视频进行去隔行操作。然后，为了进一步减少数据库的大小，在每个能检测到人脸帧里把以志愿者的头部为中心的图像区域裁剪出来。这一处理过程尽可能地避免各种几何变换操作，从而杜绝任何像素插值的可能。换句话说，在原始图像中的所有像素的灰度值都被直接拷贝到新图像上。最后，切割出来的头肩图像被保存成JPG格式的文件。通过上述的处理，切割出来的人脸图像区域大小变化很大，从最小的66×66像素到最大的798×798像素。

为了方便评测，每张静态图像和视频帧都被以特定的格式统一命名。具体地，所有的静态图像以”subjectID-frontal.jpg”的格式命名，视频帧以”subjectID-camNum-frameNum.jpg”的格式命名。以这样的命名方式，数据库里的每张图像/视频帧都被赋予了一个独一无二的文件名（图像文件名只包括了志愿者编号信息，视频帧文件名包含了志愿者编号、摄像机编号以及视频帧序号三部分信息）。比如，以“201103180001-1-0133.jpg”命名的人脸图像是由摄像机Cam1采集的编号为“201103180001”的志愿者的第133张视频帧。所有处理后的数据都被组织成一个简单的文件结构。具体地，所有静态图像都被放置在一个以“still”命名的文件夹里，而所有的视频帧都被放置在一个以“video”命名的文件里。“video”文件夹里有三个以“cam1”、“cam2”和“cam3”命名的子文件夹，每个文件夹里都包含了1,000个以志愿者编号命名的子文件夹。最底层的子文件夹里放置着对应志愿者由对应摄像机采集的所有视频人脸帧。

[3]http://www.aunsoft.com/final-mate/

表 2.2. COX数据库里志愿者的年龄分布

年龄	10-19	20-29	30-39	40-49	50-59	60-69	70-79
人数	19	933	23	15	6	3	1

2.3.3　人员统计

这一数据库是在新疆大学采集的。因此，在这一数据库里，大部分的志愿者来自这一学校的学生、老师以及在这一学校附近的居民。在这1,000个志愿者里，有435人是男性，565人是女性。其中，大约有一半是蒙古人种一半是高加索人种。由于主要是从学校招集来的志愿者，所以大部分被采集对象是青年人，关于更详细的人员统计信息，请参考表2.2。

2.3.4　测试协议

由于设计COX数据库的目的是评测所有基于视频的人脸识别方法，所以这一新数据库分别为三种不同的基于视频的人脸识别场景（也就是视频-图像、图像-视频和视频-视频）设计不同的测试协议。对于大部分的视频监控应用，人脸识别和人脸确认是两种最常用的任务。所以，新数据库同时为人脸识别和人脸确认设计了协议。具体地，人脸识别性能由首选识别率（Rank-1 Recognition Rate）来度量，人脸确认率由接收者操作特征（Receiver Operating Characteristic，ROC）曲线来衡量。与最近的数据库[24]里的确认协议一致，测试协议要求人脸识别算法必须对所有的图像/视频对（即匹配来自target的所有图像/视频与来自query的所有图像/视频）计算一个相似度矩阵。最终得到的相似度矩阵可以用来绘制用户操作特征曲线。

为了确保不同的方法能进行公平地对比，该测试协议对所有评测场景都严格定义对应的训练集和测试集。需要指出的是，这里的“训练集”表示用以训练特征抽取器或者分类模型的图像/视频集合，“测试集”表示用以形成目标集合（即gallery）和查询集合（即probe）。这里，有一个问题就是如何确定训练数据和测试数据的比率：如果用更多的数据进行训练通常会导致更高的性能，而如果用更多的测试数据可以得到更准确的性能评估。而且，对于人脸识别任务来说，我们更期望在目标集里有更多的识别对象。因此，该测试协议使用了更多的数据用以测试。具体地，这一协议经验性地将这一比率设置为3:7，也就是随机选择300个人的数据用以训练模型，用剩余的700个人的数据作测试。对所有的三种场景下的评测，测试协议都预先定义好了10组随机的300/700划分。因此，该测试协议要求未来在这一数据库上的所有测试都要报告10次人脸识别结果的均值和方差。对于人脸验证任务，和数据库[142]的协议类似，该测试协议要求所有测试算法将跑完10次的相似度矩阵合成一个最终的相似度矩阵，然后根据这一矩阵来绘制ROC曲线。除了以上描述的公共部分，下面为三种不同的基于视频

表 2.3. COX数据库上的视频-图像识别场景的训练集与测试集配置

		人数	图像数	视频数	图像/视频来源	备注
训练集		300	300	900	DC, Cam 1, 2, 3	
测试集	目标集	700	700	-	DC	
	查询集1	700	-	700	Cam 1	V1-S
	查询集2	700	-	700	Cam 2	V2-S
	查询集3	700	-	700	Cam 3	V3-S

表 2.4. COX数据库上的图像-视频识别场景的训练集与测试集配置

		人数	图像数	视频数	图像/视频来源	备注
训练集		300	300	900	DC, Cam 1, 2, 3	
测试集	目标集1	700	-	700	Cam 1	S-V1
	目标集2	700	-	700	Cam 2	S-V2
	目标集3	700	-	700	Cam 3	S-V3
	查询集	700	700	-	DC	

的人脸识别场景定义对应协议的特殊部分。

2.3.4.1 视频-图像测试协议

在视频-图像人脸识别场景里，目标集合包含了已知对象身份的静态图像，查询集合是那些需要先与目标静态图像作匹配然后被识别的人脸视频片断。因此，对于这一场景来说，该测试协议设计了如表2.3所示的关于训练数据与测试数据的几种不同的配置。具体地，由3个不同摄像机采集的视频构成了3种不同的实验，也就是V1-S、V2-S和V3-S。用于训练和测试的10次随机300/700划分设置均放在COX发布数据的“V2S partitions”文件夹里。

2.3.4.2 图像-视频测试协议

与视频-图像人脸识别场景相反，图像-视频人脸识别场景里的目标集合包含的是视频序列，而查询集合是静态图像集合。因此，如表2.4所示，该测试协议根据目标集里视频对应的不同采集摄像机同样构建3种不同的实验，也就是S-V1、S-V2和S-V3。同样，10次随机300/700划分均设置在COX发布数据的“S2V partitions”文件夹里。

表 2.5. COX数据库上的视频-视频识别场景的训练集与测试集配置

		人数	视频数	视频来源	备注
训练集		300	900	Cam 1, 2, 3	
测试集	目标集 查询集	700	700 700	Cam 1 Cam 2	V2-V1
	目标集 查询集	700	700 700	Cam 1 Cam 3	V3-V1
	目标集 查询集	700	700 700	Cam 2 Cam 3	V3-V2
	目标集 查询集	700	700 700	Cam 2 Cam 1	V1-V2
	目标集 查询集	700	700 700	Cam 3 Cam 1	V1-V3
	目标集 查询集	700	700 700	Cam 3 Cam 2	V2-V3

2.3.4.3　视频-视频测试协议

为了构建视频-视频人脸识别场景评测，该测试协议将目标集合和查询集合分别设置成来自3个不同摄像机的其中2个摄像机的数据集合。因此，这些不同的组合可以构成6组不同的实验（如表2.5所示）。除此之外，还可以将3种不同的视频数据的1种或2种去构成目标集，把剩余的视频数据作为查询对象。基于减少评测任务的考虑，本协议暂时不考虑这些实验。与前两种评测协议相同，同样将这一场景的10次随机300/700划分设置在COX发布数据的“V2V partitions”文件夹里。

2.4　本章小结

本章收集了一个大规模人脸数据库，这一数据库包括来自1,000人的由数码相机采集的类证件照的静态图像和由三个不同摄像机拍摄的类监控场景的视频。这一数据库以及所设计的评测协议是为了去评测三种不同的基于视频的人脸识别任务，即视频-图像、图像-视频和视频-视频人脸识别。

第三章　基于线性子空间建模的投影度量学习方法

3.1　引言

在过去的十几年里，大量的工作[35, 44, 55, 56, 58, 81, 101, 143, 149]已经成功将线性子空间统计模型应用于基于视频的人脸识别问题。在这些工作里，某个特定人物的视频序列首先被处理成无序的图像集合，然后采用主成分分析（Principal Component Analysis, PCA）技术将这一图像集合表示成一个线性子空间来对人脸图像集合中的数据变化模式进行全局建模，从而在非限定条件的基于视频的人脸识别中对人脸姿态及表情等变化更加鲁棒。除此之外，在匹配规模比较大的图像集合时，这一类方法相比传统逐帧匹配的方法具有更低的计算复杂度。尽管这些优点极大促进了线性子空间在视频人脸识别中的广泛应用，但是由于线性子空间拥有自己特有的非欧氏几何结构，因此如何去准确表示并处理它们是一个非常具有挑战性的研究问题。

大量的研究工作[41, 45, 55, 68, 71, 84, 121, 129, 144]表明具有相同维数的线性子空间是位于一个特定类型的黎曼流形上，这一流形被人们称为格拉斯曼流形（Grassmann manifold）。由于格拉斯曼流形是个非线性结构的空间，所以绝大多数应用于欧氏空间的主流技术都不能直接处理这一格拉斯曼流形上的数据。为了解决这一问题，一些工作[7, 31, 41, 55, 56, 58, 62, 121, 144]研究和开发了在格拉斯曼流形上的特定黎曼几何。其中，有些工作[31, 41, 55, 56, 58, 62, 144]通过引进了著名的投影映射（Projection mapping）框架将格拉斯曼流形上的基本元素（即线性子空间）表示成对应的投影矩阵。投影映射框架推导出了投影距离来计算两个投影矩阵之间的距离。根据文献[62]所述，投影距离可以以非常小的近似尺度接近于格拉斯曼流形上的真实测地距离。投影距离的另一个特点是满足距离函数度量的三个要素条件，即非负性、对称性与三角不等式[55]，所以它又被称为投影度量（Projection metric）。由于投影度量有以上一些良好属性，最近一些工作利用它来估计格拉斯曼流形的黎曼几何进而将应用在欧氏空间的传统算法扩展到格拉斯曼流形上。比如，文献[31]和文献[62]采用投影度量分别在格拉斯曼流形上开发了一种新的聚类算法和新的字典学习方法。

本章主要研究如何在格拉斯曼流形上进行判别分析并应用于基于视频的人脸识别任务上。在投影映射框架下，最近一些工作[55, 56, 58, 60, 135]致力于开发在黎曼流形上的正定核函数来将流形嵌入到一个高维的可再生核希尔伯特空间（Reproducing Kernel Hilbert Space, RKHS）上（如图3.1 (a)-(b)-(d)所示）。由于希尔伯特空间从广义上来讲是个欧氏空间，所以这一操作可以看作是从黎曼流形到欧氏空间的转换。在这之后，这些工作可以采用传统在欧氏空间上设计的基于核的分类器或特征提取方法（比如核判别分析方法[20]）对变换之后的元素进行分类或特征学习。这一过程可以表示成从希尔伯特空间到另一个低维欧氏空间的变换（如图3.1 (d)-(e)所示）。虽然这些工作通

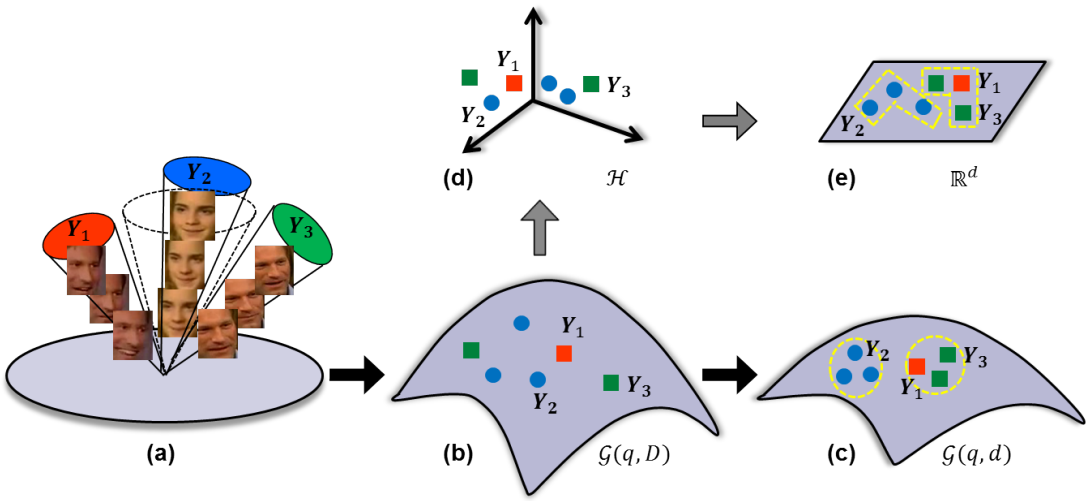


图 3.1. 本章新提出的在格拉斯曼流形上的投影度量学习方法（Projection Metric Learning, PML）概念图。传统格拉斯曼判别分析方法是沿着线路(a)-(b)-(d)-(e)首先将原始的格拉斯曼流形$\mathcal{G}(q,D)$ (b)嵌入到一个高维的希尔伯特空间$\mathcal{H}$，然后再学习一个从希尔伯特空间$\mathcal{H}$到一个更低维、更有判别性的欧氏空间$\mathbb{R}^d$ (e)。与传统方法相比，本章新提出的PML方法沿着另一条路线(a)-(b)-(c)直接学习从原始的格拉斯曼流形$\mathcal{G}(q,D)$ (b)到更低维、更有判别性的新格拉斯曼流形$\mathcal{G}(q,d)$的映射。

过在格拉斯曼流形上推导出有效的核函数从而可以利用一些开发在欧氏空间的经典核学习算法进行判别学习。但是这些工作同时引入了传统核方法的一些缺陷。比如，推导出来的核函数需要通过复杂的理论证明来满足Mercer定理才能产生一个有效的可再生核希尔伯特空间。而且，在传统核方法中，变换到希尔伯特特空间上的数据通常是隐式的，只有它们之间的相似性（或内积）可以通过核函数来得到。另外，传统核方法随着数据量的增大，其计算核矩阵的时间复杂度将呈指数级增长。

为了克服这些现有的格拉斯曼判别分析方法的缺陷，本章通过挖掘投影度量来直接在格拉斯曼流形上学习类马氏矩阵。对比已有的方法，本章新提的方法是通过直接在格拉斯曼流形进行学习并开发其上的黎曼几何来学习到一个新的格拉斯曼流形数据表示，从而能受益于原始流形有用的黎曼几何属性。而且，新方法学习到的类马氏矩阵可以进一步分解成用以降维的变换矩阵。这一变换矩阵是将原始的格拉斯曼流形映射到一个更低维的、更具判别性的格拉斯曼流形（如图3.1 (a)-(b)-(c)所示）。虽然已有一些工作提出了基于线性子空间的降维技术[44, 81, 101]，但是它们是基于典型相关性的距离来学习线性子空间到低维线性子空间的变换。由于基于典型相关性的距离并不是一种结构化度量[55]，所以这些技术并没有开发出线性子空间特有的黎曼流形结构。与它们不同的是，本章的新方法通过开发格拉斯曼流形上的黎曼几何来学习投影度量。据[55]证明，投影度量是可以推导出一个正定核，因而本章提出的方法可以被看作是其它基于格拉斯曼流形上核方法的一个预处理步骤，也就是新提的方法可以将学习得到的更低维、更具判别性的格拉斯曼流形输入到这些方法里进而提高它们的性能。

本章接下来的安排如下：3.2节系统调研相关工作并讨论与本章新工作的关系；3.3节简单描述新工作的背景知识；3.4节详细介绍新提出的投影度量学习工作，重点描述算法的问题形式化，并提出一种有效的优化求解算法；3.5节给出了算法在四个公开视频人脸数据库上与其它方法的对比结果；3.6节对算法进行了总结讨论。

3.2　相关工作

本节将更详细地回顾一些现有的基于线性子空间的降维方法、基于核的格拉斯曼流形判别分析方法以及学习流形-流形映射的算法。

在早期，有一些相关工作是对线性子空间的降维学习方法比如受限互子空间方法（Constrained Mutual Subspace Method, CMSM）[44, 101]和判别典型相关系数方法（Discriminant Canonical Correlations, DCC）。CMSM方法开发一个受限的子空间，使得来自不同类的子空间样本对有比较小的典型相关系数。然而，这一方法对受限子空间的维数十分敏感。DCC方法通过学习一个从原始子空间到低维子空间的变换，同时试图最大化类内子空间对的典型相关系数，最小化类间子空间对典型相关系数。由工作[55]指出，在当所有线性子空间相交于一个点时，CMSM和DCC所使用的基于最大典型相关系数的距离对所有数据都会趋近于0。除此之外，基于最大典型相关系数的距离由于不满足距离度量的三个要素条件，所以不能成为一种有效的度量，从而不能跟一些经典的度量学习方法一起使用。最后，这些工作没有去开发线性子空间特有的数据结构（即格拉斯曼流形的黎曼几何），从而导致学到的判别函数有可能是次优的。

在最近几年里，有些工作[55, 56, 58, 60, 135]开始研究在格拉斯曼流形上学习基于子空间的判别函数的问题。比如，通过推导出基于知名的投影度量的格拉斯曼核函数，文献[55]首先将格拉斯曼流形嵌入到一个高维的希尔伯特空间，然后在核判别分析方法的框架下在希尔伯特空间上进行判别核学习。在这一工作基础上，文献[58] 采用了一个更加有推广性的图嵌入判别学习框架将基于投影度量的核函数与基于典型相关相似度的核函数有效地结合起来进行核空间判别式学习。尽管这些方法都可以进行有效地有监督地分类，然而它们非常受限于Mercer核，导致只能产生隐式的映射以及只能采用基于核的分类器。更重要的是，这些基于核的方法的计算复杂度会随着训练样本的增加而急剧地增长。

为了解决这些问题，本章新提出的新方法不采用希尔伯特空间嵌入，而是通过利用投影度量来开发格拉斯曼流形的黎曼几何并直接在格拉斯曼流形上学习一个类马氏矩阵，也就是对称半正定矩阵（Symmetric Positive Semidefinite, PSD）。在计算机视觉与模式识别领域里，本章发现有两个工作[107, 133]也是在某种特定的黎曼流形上学习一个参数化的矩阵。但是它们的思想和解决的问题都跟本章的工作很不一样。这两个工作主要是在施蒂费尔流形（Stiefel manifold）上优化一个变换矩阵来解决对位于欧氏空间的数据向量作降维的问题。与它们相比，本章的新工作是在一个PSD流形上学习

一个类马氏矩阵来解决以线性子空间为基本元素的格拉斯曼流形上的投影度量学习问题。另外，这一学习到的马氏矩阵可以进一步分解成一个从原始格拉斯曼流形映射到另一个更低维的、更具判别性的格拉斯曼流形的降维变换矩阵。据调研，目前只有两个工作[59, 76]研究从流形到流形的映射学习问题。不过，这两个工作与本工作还是有巨大的不同：工作[76]主要是学习从高维的球体流形到降维之后的子流形的映射，工作[59]学习的是从一个高维的对称正定矩阵（Symmetric Positive Definite, SPD）流形到一个低维的对称正定矩阵流形的变换矩阵。

3.3 背景知识

本节先简单介绍一下格拉斯曼流形的基本黎曼几何来为本章的新方法提供一些背景知识。关于格拉斯曼流形的更详细的知识和相关研究，请读者参阅文献[7, 41, 62, 65, 84, 121, 144]。

一个格拉斯曼流形$\mathcal{G}(q,D)$被定义成在欧氏空间$\mathbb{R}^D$的q维线性子空间的集合，这一集合构成了一个$q(D-q)$维的紧致（compact）黎曼流形。具体地，格拉斯曼流形的基本元素是一个线性子空间，由span($\boldsymbol{Y}$)来表示，是由它的大小为$D \times q$ 的正交基矩阵$\boldsymbol{Y}$张成的，其中$\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{I}_q$，$\boldsymbol{I}_q$是大小为$q \times q$的单位矩阵。

在投影映射$\Phi(\boldsymbol{Y}) = \boldsymbol{Y}\boldsymbol{Y}^T$的框架下，工作[41]提出了另一种表示格拉斯曼流形的基本元素的策略，也就是用投影矩阵$\boldsymbol{Y}\boldsymbol{Y}^T$来表示格拉斯曼流形元素。根据工作[62]所证明，这一投影嵌入是一个从格拉斯曼流形到秩为q的幂等对称矩阵空间的微分同胚映射。也就是说，它是一个一对一、连续、可微的映射，并且它的逆映射也是个连续、可微的。因此，在格拉斯曼流形上的每个元素只对应唯一的一个投影矩阵。

由于投影算子$\Phi(\boldsymbol{Y})$ 是一个$D \times D$维的对称矩阵，因此其对应的内积形式可以定义为$\langle \boldsymbol{Y}_1, \boldsymbol{Y}_2 \rangle_\Phi = tr(\Phi(\boldsymbol{Y}_1)^T\Phi(\boldsymbol{Y}_2))$。这一内积形式对线性子空间的具体实现形式具有不变性，由此可以推导出对应的距离形式：

$$d_p(\boldsymbol{Y}_1\boldsymbol{Y}_1^T, \boldsymbol{Y}_2\boldsymbol{Y}_2^T) = 2^{-1/2}\|\boldsymbol{Y}_1\boldsymbol{Y}_1^T - \boldsymbol{Y}_2\boldsymbol{Y}_2^T\|_F. \tag{3.1}$$

其中，$\|\cdot\|_F$ 表示矩阵的弗罗贝尼乌斯范数（Frobenius norm）。在工作[55]，由于这一距离被证明满足了度量的三个要素条件，即非负性、对称性与三角不等式，因此它又被称为投影度量（Projection metric）。除此之外，工作[62]证明了投影度量可以以最大尺度为$\sqrt{2}$来估计格拉斯曼流形上的真实测地距离。因此，投影度量已经成为了在格拉斯曼流形上最受欢迎的度量之一。

3.4 投影度量学习方法

本节首先形式化新提出的在格拉斯曼流形上的投影度量学习方法（Projection Metric Learning, PML）应用于基于视频的人脸识别的问题，然后介绍一个新的求解算法来优化这一度量学习问题。

3.4.1 问题形式化

给定m个视频人脸序列$\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m\}$，其中$\boldsymbol{X}_i \in \mathbb{R}^{D\times n_i}$表示包括$n_i$张人脸帧的第$i$段视频序列，每张人脸帧被表示成一个$D$维的特征向量。在这些数据里，每个视频属于人脸类别中的某一类，表示成C_i。第i段视频$\boldsymbol{X}_i$然后被建模成一个由正交基矩阵$\boldsymbol{Y}_i \in \mathbb{R}^{D\times q}$张成的$q$维线性子空间span($\boldsymbol{Y}_i$)，即$\boldsymbol{X}_i\boldsymbol{X}_i^T \simeq \boldsymbol{Y}_i\boldsymbol{\Lambda}_i\boldsymbol{Y}_i^T$，其中$\boldsymbol{\Lambda}_i, \boldsymbol{Y}_i$分别对应于前$q$大的特征值以及相应特征向量构成的矩阵。

由3.3小节的背景知识所述，所有q维线性子空间span($\boldsymbol{Y}_i$)均位于一个格拉斯曼流形$\mathcal{G}(q, D)$，并一一对应于一个特定的投影矩阵$\boldsymbol{Y}_i\boldsymbol{Y}_i^T$。在此基础上，本章需要学习一个广义的映射$f : \mathcal{G}(q, D) \to \mathcal{G}(q, d)$，具体定义为：

$$f(\boldsymbol{Y}_i\boldsymbol{Y}_i^T) = \boldsymbol{W}^T\boldsymbol{Y}_i\boldsymbol{Y}_i^T\boldsymbol{W} = (\boldsymbol{W}^T\boldsymbol{Y}_i)(\boldsymbol{W}^T\boldsymbol{Y}_i)^T. \tag{3.2}$$

其中，$\boldsymbol{W} \in \mathbb{R}^{D\times d}$ $(d \leq D)$是一个列满秩的变换矩阵。有了这一映射，原始的格拉斯曼流形$\mathcal{G}(q, D)$可以被变换到一个更低维的格拉斯曼流形$\mathcal{G}(q, d)$。然而，除非$\boldsymbol{W}$为正交矩阵，要不然$\boldsymbol{W}^T\boldsymbol{Y}_i$一般不是一个正交基矩阵。而只有由正交基矩阵张成的子空间才能构成一个有效的格拉斯曼流形。为了克服这一问题，本节先暂时用$\boldsymbol{W}^T\boldsymbol{Y}_i$的正交部分（表示成$\boldsymbol{W}^T\boldsymbol{Y}_i'$）来表示变换后的投影矩阵的正交基矩阵。关于如何去获取$\boldsymbol{W}^T\boldsymbol{Y}_i'$，将会在接下来的一小节里给出更详细的描述。现在集中研究在新格拉斯曼流形上的投影度量以及对应的目标函数。

3.4.1.1 新学到的投影度量

在新格拉斯曼流形$\mathcal{G}(q, d)$上，任意两个变换后的投影矩阵$\boldsymbol{W}^T\boldsymbol{Y}_i'\boldsymbol{Y}_i'^T\boldsymbol{W}, \boldsymbol{W}^T\boldsymbol{Y}_j'\boldsymbol{Y}_j'^T\boldsymbol{W}$间的投影度量可以定义为:

$$\begin{aligned} &d_p^2(\boldsymbol{W}^T\boldsymbol{Y}_i'\boldsymbol{Y}_i'^T\boldsymbol{W}, \boldsymbol{W}^T\boldsymbol{Y}_j'\boldsymbol{Y}_j'^T\boldsymbol{W}) \\ &= 2^{-1/2}\|\boldsymbol{W}^T\boldsymbol{Y}_i'\boldsymbol{Y}_i'^T\boldsymbol{W} - \boldsymbol{W}^T\boldsymbol{Y}_j'\boldsymbol{Y}_j'^T\boldsymbol{W}\|_F^2 \\ &= 2^{-1/2}tr(\boldsymbol{P}\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^T\boldsymbol{P}). \end{aligned} \tag{3.3}$$

其中，$\boldsymbol{A}_{ij} = \boldsymbol{Y}_i'\boldsymbol{Y}_i'^T - \boldsymbol{Y}_j'\boldsymbol{Y}_j'^T$, $\boldsymbol{P} = \boldsymbol{W}\boldsymbol{W}^T$. 由于$\boldsymbol{W}$是一个列满秩的矩阵，所以$\boldsymbol{P}$是一个秩为$d$的$D \times D$维的对称半正定矩阵（Symmmetric Positive Semidefinite matix, PSD）。与传统的马氏度量学习相似，这里的矩阵$\boldsymbol{P}$可以看成是一个类马氏矩阵。

3.4.1.2 判别函数

为了使在新格拉斯曼流形上的同类线性子空间的投影距离尽量地接近，异类线性子空间的投影距离尽量地被拉远，本章方法设计了以下目标函数$J(\boldsymbol{P})$来求解最优的$\boldsymbol{P}$矩阵：

$$\boldsymbol{P}^* = \arg\min_{\boldsymbol{P}} J(\boldsymbol{P}) = \arg\min_{\boldsymbol{P}} (J_w(\boldsymbol{P}) - \alpha J_b(\boldsymbol{P})). \tag{3.4}$$

其中，α反映的是类内紧致项$J_w(\boldsymbol{P})$和类间分散项$J_b(\boldsymbol{P})$之间的平衡关系。这两项分别用以描述类内散度的平均值以及类间散度的平均值：

$$J_w(\boldsymbol{P}) = \frac{1}{N_w}\sum_{i=1}^{m}\sum_{j:C_i=C_j} 2^{-1/2} tr(\boldsymbol{P}\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^T\boldsymbol{P}). \tag{3.5}$$

$$J_b(\boldsymbol{P}) = \frac{1}{N_b}\sum_{i=1}^{m}\sum_{j:C_i\neq C_j} 2^{-1/2} tr(\boldsymbol{P}\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^T\boldsymbol{P}). \tag{3.6}$$

其中，N_w和N_b分别是来自同类样本对的总对数和来自不同类样本对的总对数，$\boldsymbol{A}_{ij} = \boldsymbol{Y}_i'\boldsymbol{Y}_i'^T - \boldsymbol{Y}_j'\boldsymbol{Y}_j'^T$，$\boldsymbol{P}$是要学习优化的对称半正定矩阵。

3.4.2 算法优化

这一最优化问题3.4包含了变量$\boldsymbol{P}$和$\boldsymbol{Y}'$。由于$\boldsymbol{Y}'$并不显式地由$\boldsymbol{P}$来表示，所以很难为$\boldsymbol{P}$找到一个闭解。本节提出了一种迭代求解的策略通过迭代的方式来交替求解$\boldsymbol{P}$和$\boldsymbol{Y}'$。为了使$\boldsymbol{W}^T\boldsymbol{Y}$的列正交，这一迭代优化算法每步都要对$\boldsymbol{Y}$进行正则化。由于$\boldsymbol{P}$是一个秩为$d$的对称半正定矩阵，这一优化算法将在秩为$d$的对称半正定矩阵流形上通过开发非线性黎曼共轭梯度（Riemannian Conjugate Gradient, RCG）算法来求解最优的$\boldsymbol{P}$。

3.4.2.1 正则化矩阵$\boldsymbol{Y}$

对所有的i，固定矩阵$\boldsymbol{P}$，矩阵$\boldsymbol{Y}_i$将被正则化成$\boldsymbol{Y}_i'$，使得$\boldsymbol{W}^T\boldsymbol{Y}_i$的列正交。具体地，新提出的优化算法对$\boldsymbol{W}^T\boldsymbol{Y}_i$采用QR分解，即$\boldsymbol{W}^T\boldsymbol{Y}_i = \boldsymbol{Q}_i\boldsymbol{R}_i$，其中$\boldsymbol{Q}_i \in \mathbb{R}^{D\times q}$是正交矩阵，$\boldsymbol{R}_i \in \mathbb{R}^{q\times q}$是上三角可逆矩阵。由于$\boldsymbol{R}_i$可逆并且$\boldsymbol{Q}_i$正交，所以新提出的求解算法通过如下方式将$\boldsymbol{Y}_i$进行归一化，使得$\boldsymbol{W}^T\boldsymbol{Y}_i'$成为一个正交基矩阵：

$$\boldsymbol{Q}_i = \boldsymbol{W}^T(\boldsymbol{Y}_i\boldsymbol{R}_i^{-1}) \quad \rightarrow \quad \boldsymbol{Y}_i' = \boldsymbol{Y}_i\boldsymbol{R}_i^{-1}. \tag{3.7}$$

3.4.2.2 计算对称半正定矩阵$\boldsymbol{P}$

给定矩阵$\boldsymbol{Y}_i$，新提出的求解算法在秩为d的$D\times D$维的对称半正定矩阵流形上采用非线性黎曼共轭梯度下降算法来求解最优的矩阵$\boldsymbol{P}$。在公式3.5和公式3.6中，如果将$\boldsymbol{P}$置于迹操作的外面，那么在公式3.4的目标函数$J(\boldsymbol{P})$可以被进一步转化成：

$$\boldsymbol{P}^* = \arg\min_{\boldsymbol{P}} tr(\boldsymbol{P}\boldsymbol{S}_w\boldsymbol{P}) - \alpha tr(\boldsymbol{P}\boldsymbol{S}_b\boldsymbol{P}). \tag{3.8}$$

其中，$\boldsymbol{S}_w$ 和$\boldsymbol{S}_b$被分别形式化成：

$$\boldsymbol{S}_w = \frac{1}{N_w}\sum_{i=1}^{m}\sum_{j:C_i=C_j} 2^{-1/2} tr(\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^T). \tag{3.9}$$

算法 1 投影度量学习（PML）算法

输入: 在格拉斯曼流形$\mathcal{G}(q,D)$上的观察数据集$\{span(\boldsymbol{Y}_1), span(\boldsymbol{Y}_2), \ldots, span(\boldsymbol{Y}_m)\}$

1. 初始化：$\boldsymbol{W} \leftarrow \boldsymbol{I}_{D\times d}, \boldsymbol{P} \leftarrow \boldsymbol{I}_D$。
2. **重复以下步骤:**
3. 　对所有的i，利用公式3.7正则化$\boldsymbol{Y}_i$。
4. 　利用公式3.9和公式3.10计算$\boldsymbol{S_w}$ and $\boldsymbol{S_b}$。
5. 　利用算法2优化在公式3.8里的$\boldsymbol{P}$。
6. 　通过计算$\boldsymbol{P}$的矩阵平方根来更新$\boldsymbol{W}$。
7. **算法终止。**

输出: 半正定矩阵$\boldsymbol{P}$。

算法 2 黎曼共轭梯度（RCG）算法

输入: 初始的半正定矩阵$\boldsymbol{P}_0$

1. 初始化：$\boldsymbol{H}_0 \leftarrow 0, \boldsymbol{P} \leftarrow \boldsymbol{P}_0$。
2. **重复以下步骤:**
3. 　$\boldsymbol{H}_k \leftarrow -\nabla_{\boldsymbol{P}} J(\boldsymbol{P}_k) + \eta\tau(\boldsymbol{H}_{k-1}, \boldsymbol{P}_{k-1}, \boldsymbol{P}_k)$。
4. 　沿着测地线γ在方向$\boldsymbol{H}_k$上从$\boldsymbol{P}_{k-1} = \gamma(k-1)$ 线性搜索$\boldsymbol{P}_k = \arg\min_{\boldsymbol{P}} J(\boldsymbol{P})$。
5. 　$\boldsymbol{H}_{k-1} \leftarrow \boldsymbol{H}_k, \boldsymbol{P}_{k-1} \leftarrow \boldsymbol{P}_k$.
8. **直至收敛。**

输出: 半正定矩阵$\boldsymbol{P}$

$$\boldsymbol{S}_b = \frac{1}{N_b}\sum_{i=1}^{m}\sum_{j:C_i\neq C_j} 2^{-1/2} tr(\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^T). \tag{3.10}$$

与传统开发在欧氏空间的共轭梯度算法相同，在流形上的共轭梯度算法同样是一个迭代优化过程（算法2）。算法的迭代过程如下：在第k步，先沿着方向为$\boldsymbol{H}_{k-1}$的测地线γ将目标函数的最小值$\boldsymbol{P}_{k-1}$移到$\boldsymbol{P}_k$，然后计算在这一点的黎曼梯度，接着利用旧的搜索方向$\boldsymbol{H}_{k-1}$和新的梯度求解新的搜索方向$\boldsymbol{H}_k \leftarrow -\nabla_{\boldsymbol{P}} J(\boldsymbol{P}_k) + \eta\tau(\boldsymbol{H}_{k-1}, \boldsymbol{P}_{k-1}, \boldsymbol{P}_k)$，最后迭代直到收敛。这里黎曼梯度是由对应的欧氏梯度$D_{\boldsymbol{P}}J(\boldsymbol{P}_k)$计算得到，即$\nabla_{\boldsymbol{P}} J(\boldsymbol{P}_k) = D_{\boldsymbol{P}}J(\boldsymbol{P}_k) - \boldsymbol{P}_k\boldsymbol{P}_k^T D_{\boldsymbol{P}}J(\boldsymbol{P}_k)$，$\tau(\boldsymbol{H}_{k-1}, \boldsymbol{P}_{k-1}, \boldsymbol{P}_k)$ 表示切向量$\boldsymbol{H}_{k-1}$从$\boldsymbol{P}_{k-1}$移到$\boldsymbol{P}_k$的并行转移（parallel transport）。关于对这一算法更详细的描述，请参考文献[7, 41]。现在只需要计算公式3.8 的欧氏梯度：

$$D_{\boldsymbol{P}}J(\boldsymbol{P}_k) = 2(\boldsymbol{S_w} - \alpha\boldsymbol{S_b})\boldsymbol{P}_k. \tag{3.11}$$

综合以上两个步骤，求解本节PML的目标函数的程序在算法1中给出。一旦这一算法求解出最优的半正定矩阵$\boldsymbol{P}$，任意两个线性子空间的比较结果可以由公式3.3来计算。

尽管很难对这一优化求解算法的收敛性给出理论证明，但是在本章的实验部分发现它可以让目标函数3.4能够在有限的迭代次数下收敛到一个稳定的最优解。

3.5 实验验证

本节通过具体的基于视频的人脸识别与人脸验证两个应用问题来评测所提出的PML方法。关于基于视频的人脸识别任务，本节采用了本文在第二章新收集的COX和公开的YouTube Celebrities (YTC)[79]两个数据库。关于基于视频的人脸确认任务，本节采用了YouTube Face DB (YTF) [142]和Point-and-Shoot Challenge (PaSC)[24] 两个公开数库。

在本节所有的实验中，每个视频序列数据都先被处理成图像集合，其数据可以表示为$\boldsymbol{X}_i = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_i}]$，其中$\boldsymbol{x}_j \in \mathbb{R}^D$表示第$i$张视频帧的$D$维的向量化描述子。通过对$\boldsymbol{X}_i$进行奇异值分解（Signular Value Decomposition, SVD），每个图像集合可以被表示成一个线性子空间。具体地，本节里的实验使用了前q个主特征向量构成一个正交基矩阵$\boldsymbol{Y}_i$来表示对应的q维线性子空间，从而形成了一个格拉斯曼流形$\mathcal{G}(q, D)$。下面的几个实验均采用了交叉验证的方式来给q设定合适的值。

为了验证本章新提出的PML方法的有效性，本节实验选取了以下几种类型的代表性方法进行对比：

1. 基于线性子空间的无监督方法：

   Mutual Subspace Method (MSM) [149], Projection Metric (PM) [41]；

2. 基于受限子空间的无监督方法:

   Affine Hull based Image Set Distance (AHISD) [32], Convex Hull based Image Set Distance (CHISD) [32]；

3. 基于线性子空间的判别学习方法：

   Constrained Mutual Subspace Method (CMSM) [44], Discriminative Canonical Correlations (DCC) [81]；

4. 基于受限子空间的判别学习方法:

   Set-to-set distance metric learning (SSDML) [162]；

5. 基于格拉斯曼流形的判别学习方法：

   Grassmann Discriminant Analysis (GDA) [55], Grassmannian Graph-Embedding Discriminant Analysis (GGDA) [58]；

6. 本章新提出的PML方法。

图 3.2. YTC数据库示例

为了保证对比实验的公正性，实验中对所有的方法的关键参数的调整都是根据原始工作的建议进行的。具体地，针对MSM和AHISD方法，实验里均对它们设置成只用第一个典型相关系数或第一个主成分来计算子空间相似度。针对CMSM和DCC两个方法学习得到的判别子空间，其维数在1到10之间进行调整。在基于主夹角/典型相关系数计算子空间相似度阶段，CMSM与DCC均采用了全部相关系数的累加结果。针对SSDML方法，其两个重要参数被实验性地调成$\lambda_1 = 0.001, \lambda_2 = 0.5$。针对GDA和GGDA方法，它们的最终输出数据维数均设置成训练类别数减1。GGDA方法的另一个参数β在$\{1e^2, 1e^3, 1e^4, 1e^5, 1e^6\}$范围内进行调整。针对本章新提出的PML方法，参数$\alpha$ 被设置成0.2。

3.5.1　视频-视频人脸识别评测

视频-视频人脸识别的主要任务是给定一个查询视频，在一个由大量的视频构成的数据库（即目标集）中进行匹配，将最匹配的视频人脸身份赋予查询视频。这一任务通常以首选识别率（Rank-1 Recognition Rate）来作其评测指标。为保证对比实验的充分性与广泛性，本节实验选用了两个具有不同特点的数据集COX和YouTube Celebrities (YTC) [79] 来评测本章新提出的方法在执行视频-视频的人脸识别任务时的有效性。

由于上一章已经详细介绍了本文新收集的COX视频人脸数据库，本节将不再具体描述此数据库。在这一数据库上，本节实验将视频序列中的每张人脸帧都归一化成一张大小为24×30的灰度图像，然后采用了直方图均衡化技术将归一化的图像进行光照预处理。在评测中，严格参照第二章在COX上所设置的协议，报告对比方法在10次测试中的平均性能和标准方差。关于YouTube Celebrities (YTC)数据库，这一数据库由文献[79]收集，用于现实场景中的视频人脸跟踪与识别任务。该数据库包含47个人

表 3.1. PML对比方法在YTC和COX数据库上的视频-视频人脸识别结果（%）

Methods	YTC	COX					
		V2-V1	V3-V1	V3-V2	V1-V2	V1-V3	V2-V3
MSM [149]	60.25±3.05	45.53±0.46	21.47±1.87	11.00±0.85	39.83±0.67	19.36±0.67	9.50±0.67
PM [41]	62.17±3.65	51.57±1.38	46.17±1.33	23.81±0.97	43.36±1.04	34.71±1.08	20.09±0.74
AHISD [32]	63.70±2.89	53.03±2.05	36.13±1.00	17.50±0.81	43.51±0.66	34.99±0.83	18.80±0.69
CHISD [32]	66.62±2.79	56.90±0.64	30.13±0.79	15.03±0.77	44.36±0.51	26.40±0.72	13.69±0.76
CMSM [44]	63.81±3.70	56.90±0.64	30.13±0.79	15.03±0.77	44.36±0.51	26.40±0.72	13.69±0.76
SSDML [162]	68.85±2.32	60.13±0.23	53.14±0.72	28.73±0.53	47.91±0.38	44.42±0.74	27.34±0.85
DCC [81]	65.48±3.51	62.53±6.13	66.10±3.52	50.56±4.21	56.09±11.27	53.84±11.37	45.19±9.81
GDA [55]	65.02±2.91	68.61±1.96	77.70±1.58	71.59±1.11	65.93±1.88	76.11±0.98	74.83±1.80
GGDA [58]	66.37±3.52	70.80±1.24	76.23±1.25	71.99±1.05	69.17±1.01	76.77±1.57	77.43±1.41
PML	**66.69±3.54**	**71.61±1.40**	**66.23±1.38**	**51.56±1.43**	**61.93±1.53**	**54.07±1.73**	**44.53±0.92**
PML-GDA	**68.08±3.78**	**80.63±1.28**	**87.27±1.18**	**81.13±1.59**	**81.86±1.60**	**87.41±0.97**	**79.00±1.26**
PML-GGDA	**70.32±3.69**	**81.91±0.70**	**88.01±1.02**	**83.04±1.31**	**83.47±1.39**	**89.01±1.09**	**80.50±1.17**

的1,910段视频，均来自于YouTube网站上的演艺明星和政治人物的真实视频片段。其中，每个人的视频片段可以分为3个小节（session），每一小节对应于不同的采集时间与场景，各小节中的视频片段数量不等。这些视频普遍具有较低的分辨率与较高的压缩率，且受到不同程度的噪声影响。每段视频包含的图像数一般为数百幅。图3.2给出了2段视频片段中的部分图像，每段视频来自其一个小节。该数据库里视频数据变化模式非常复杂，包括大幅度的姿态变化、复杂的光照、多变的表情以及其它一些变化等。参照前人工作[94, 139]，实验中将该数据库的所有人脸都归一化到20×20大小的图像，并且通过直方图均衡化来减少受光照的影响，然后对每张人脸图像提取灰度特征。按照前人工作[94, 139]对这一数据库设置的协议，本实验将在这一数据库上进行10组交叉验证实验。在每组实验中从每个人的3个小节中随机选用9段视频进行实验，其中每个小节选取3段视频。这9段视频中的3段作为训练集同时也为目标集，6段作为查询集。在COX和YTC两个数据库上，每个视频序列均被表示成10维的线性子空间。

表3.1给出了实验对比的不同方法在YTC和COX两个数据库上的结果，其中首选识别率的均值与标准差是10次实验统计所得。从该表的对比情况分析，可以发现如下一些现象和结论：

（1）基于线性子空间的无监督方法MSM和PM。这两个方法通常会比其它基于子空间的方法在识别性能上要差一些，特别是在COX数据库上的对比更为明显。这主要是因为虽然YTC数据库的视频来自YouTube，但大多数视频的变化模式（姿态、表情及光照等变化）相对比较单一。相比之下，COX数据库主要是模拟非受控条件下的真实监控场景，其中各个视频所包含的变化相对比较复杂。另外，COX数据库里也有一部

分视频是比较短的序列。因此只用线性子空间进行统计建模对一些有可能是短视频序列而且是比较复杂的非线性的变化模式进行全局建模是远远不够的。

（2）基于受限子空间的无监督方法AHISD和CHISD。相比MSM和PM方法，在实验中表现出更高的识别性能。其主要原因是这两个方法采用了受限子空间进行建模。为了解决线性子空间方法不能很好建模时长很短、人脸模式变化却很大的视频序列的这一问题，受限子空间建模方式是通过采用集合里的仿射或凸组合来表示集合的样例，用样例来表示整个集合。如文献[32]所述，由于这一类集合建模方法只是对集合里样本数据计算一个相对松弛的估计，因此它们对集合里的样本点的位置不敏感。

（3）基于线性子空间的判别学习方法CMSM和DCC。相比第一类方法，这一方法的性能有较大幅度的提升。这主要是因为这两种方法都是基于线性子空间建模的基础上再进行了判别式的有监督学习。通过加入判别信息，这三种方法减少了受一些噪声的影响。

（4）基于受限子空间的判别学习方法SSDML方法。由于这一方法是在基本的仿射包的距离度量基础上，开发了一种新的度量学习方法来学习可具判别性的仿射包的距离度量。因此，相比第二类的方法，这一方法对识别性能有进一步的提升。

（5）基于格拉斯曼流形的判别学习方法GDA和GGDA。相比同样是基于线性子空间建模的方法CMSM和DCC，这两个方法在两个数据库特别是在COX数据库上都取得了更高的识别性能。这主要是因为CMSM和DCC没有挖掘线性子空间特殊的非线性流形结构从而只学习到了相对比较次优的判别信息。相比之下，GDA和GGDA充分挖掘了研究比较成熟的格拉斯曼流形上的黎曼几何来帮助基于线性子空间的降维和分类。

（6）本章新提出的PML方法。通过在格拉斯曼流形上学习投影度量，PML方法提高了基准方法PM的识别性能，从而验证了度量学习的有效性。而与CMSM和DCC相比，PML通过利用格拉斯曼流形上的有效黎曼几何来学习更加有判别的投影度量，从而在两个数据库上都取得了更优的识别性能。与GDA和GGDA相比，PML在YTC数据库上达到了与它们相当的性能，而在COX库上的性能比它们差一些。不过，值得一提的是，由于PML方法学到的类马氏矩阵可以进一步分解成降维矩阵，所以这一方法可以作为格拉斯曼流形的降维方法。因此，PML可以作为GDA和GGDA的前端预处理方法，我们在实验中也测试了这一作法也就是PML-GDA和PML-GGDA。从在YTC和COX两个数据库上的实验结果来看，PML-GDA和PML-GGDA都在很大程度上提高了原始方法GDA和GGDA的性能（在两个数据库上均达到了最高的识别性能），从而验证了新提出的PML方法在格拉斯曼流形的降维方面也是非常有效的。

3.5.2　视频-视频人脸确认评测

视频-视频人脸确认实验主要是要判断给定的一对视频序列中的人脸是来自同一个人还是来不同人。这一任务通常以在某一特定的错误接收率（Face Accept Rate,

图 3.3. YTF数据库示例

FAR）下的验证正确率（Verification Rate）来作为其评测指标。针对这一任务，本节在两个极具挑战性的公开视频人脸数据库YouTube Face DB (YTF)[142]和Point-and-Shoot Challenge(PaSC)[24]上进行实验。

YTF数据库是由Wolf等人[142]收集，用于现实场景中的视频人脸确认任务。该数据库包含1,595个人的3,425段视频，每段视频序列平均约有181帧。与YTC数据库类似，这一数据库的视频均来自于YouTube网站上的真实视频片断。因此，这些视频普遍具分辨率低和压缩率高的特点，且包含了各种不同程度的噪声。除此之外，如图3.3所示，该数据库里的视频人脸图像变化模式同样非常复杂，包括大幅度的姿态变化、复杂的光照、多变的表情等。参照前人工作[38]的设置，本节实验首先将该数据库的所有人脸都归一化到24×40大小的图像，然后对每张人脸图像提取灰度特征。按照这一数据库设置的标准协议[142]，本实验需要用给定的5,000视频对来评测人脸确认性能。具体地，这些视频对平均分成10折，每折有250对来自同一个人的视频和250对来自不同人的视频。在实验中，需要随机测试10次，每次将其中的1折作为测试集，另外的9折作为训练集。在这一数据库里，每个视频序列均被表示成10维的线性子空间模型。

PaSC数据库是由Beveridge等人[24]收集，用于现实场景中的视频-视频人脸验证任务。该数据库收集了265个人在执行一些简单动作的2,802段视频。每种动作由两个不同的摄像机采集：一台高质量、分辨率为1,920×1,080架在三角架上的可控（control）摄像机以及五台手持（handheld）摄像机（分辨率从640×480到1,280×7,20变化不等）中的一台摄像机。在这一数据库里，视频数据是在4个不同的小节（session）里采集，而且采集的位置也不同（包括室内和室外），同时离摄像机的距离、姿态等采集设置都会随着小节的变化而变化。除此之外，这一数据库所采集的视频帧一般存在一些复杂的光照变化、严重的运动模糊以及一些散焦状态。因此，这一数据库是目前最具挑战性

图 3.4. PaSC数据库示例

的视频人脸数据库之一。本节将严格按照此数据库设定的协议进行视频-视频人脸验证实验。具体地，实验将执行可控（Control）视频人脸验证和手持（Handheld）视频人脸验证。在两个实验里，目标集和查询集包含了相同的视频。也就是说，在control实验里，目标集和查询集都包含了相同的可控摄像机采集的视频。同样，在handheld实验里也是如此。另外，针对这两个实验，协议还给出了一个包括来自170个人的280段视频的训练集。最后，实验里会计算两个1,401×1,401的相似度矩阵来计算人脸验证率。本节实验首先根据给定的标定数据将该数据库的所有人脸都归一化到256×256大小的图像。为了与YTF数据库保持一致，本节实验同样也对每张人脸帧抽取灰度特征，但是从实验发现，由于这一数据库极具挑战性，如果只用灰度特征在这一数据库上作识别，所有对比方法的性能都极其低（最高的性能大约有10%）而且比较相当。因此，为了体现对比方法的区分性，本节实验采用了深度卷积神经网络特征（Deep Covolutional Neural Network, DCNN）来提高底层图像的特征表示能力。具体地，本节实验利用了Caffe[74]工具对数据库里的人脸图像统一抽取深度卷积神经网络特征。在这一特征的基础上，实验对每个视频对应的线性子空间维数均设置成10。

表6.2给出了实验对比的不同方法在YTF和PaSC两个数据库上的结果。其中，在YTF上的验证率均值与标准差是10次实验统计所得，在PaSC上的两列结果是Control实验和Handheld实验在错误接收率为0.01时的验证率。在YTF上，由于只给定成对的视频样本信息，DCC/GDA/GGDA均不能在这一场景下工作。因此，参照前人工作[77, 123]，本节实验将它们的LDA实现修改成基于样本对（pair-wise）的LDA实验。从表6.2的对比情况分析可以发现与上一节的视频-视频人脸识别实验几乎相同的现象和结论。

从实验结果来看，本章新提出的PML方法可以达到与最优方法GDA和GGDA相当的性能。当将PML学习到的低维格拉斯曼流形输入到GDA和GGDA，这两个方法最后

表 3.2. PML对比方法在YTF和PaSC数据库上的视频-视频人脸确认结果(%)

Methods	YTF	PaSC	
		Control	Handheld
MSM [149]	65.20±1.97	35.80	34.56
PM [41]	65.12±2.00	35.65	33.60
AHISD [32]	64.80±1.54	21.96	14.29
CHISD [32]	66.30±1.21	26.12	20.97
CMSM [44]	66.46±1.54	36.67	36.22
SSDML [162]	65.38±1.86	29.19	22.89
DCC [81]	68.28±2.21	38.87	37.53
GDA [55]	67.00±1.62	41.88	43.25
GGDA [58]	66.56±2.07	43.35	43.09
PML	**67.30±1.76**	**37.25**	**37.23**
PML-GDA	**70.88±1.69**	**42.93**	**43.64**
PML-GGDA	**70.04±2.19**	**43.63**	**43.95**

在这两个数据库上的性能均得到了不同程度的提升（在YTF数据上较为明显，而由于PaSC是一个极具挑战性的数据库，因此在其上的性能提升比较轻微）。

最后，图3.5，图3.6以及表3.3分别给出了本章新提出的PML方法的收敛性分析、参数影响讨论以及与其它方法的时间上的对比，其具体分析内容如下：

（1）收敛性。尽管本章无法对新提出的PML方法的收敛性给出严格的理论证明，但是通过实验发现它在大多数情况下是可以收敛到一个稳定的最优解。图3.5(a)给出了一些PML迭代学习过程的例子。这一例子使用YTF的4折数据来作实验。具体地，给定初始值$\boldsymbol{P} = \boldsymbol{I}$，经过迭代有限次之后，算法的目标函数值都可以收敛到一个稳定的值。在其中的1折数据上，迭代前10次以及迭代到100次时的目标函数值分别为7.69, 8.47, 8.57, 8.58, 8.59, 8.59, 8.58, 8.58, 8.59, 8.58, 8.58。这些结果验证了PML算法可以迭代求解更多次之后仍然能够收敛到一个稳定的解。除此之外，本节还实现了算法在迭代1次之后的性能：64.27% (YTC)；66.34% (YTF)；34.47%, 35.98% (YTF)。这些结果表明了PML算法需要迭代多次才能取得最佳的性能。而且，如图3.5(b)所示，随机对$\boldsymbol{P}$给定不同的初始化，PML求解算法最终得到的目标函数值也是比较接近的。

（2）参数影响。如本章前面讨论，PML方法可以被当作是一种在格拉斯曼流形上的降维技术。因此，其降维后的黎曼流形维数对方法最终性能的影响是一个需要重点讨论的关键问题。同样，本节通过实验来展开对这一问题的讨论。具体地，如图3.6所示，本节在YTC、YTF和PaSC三个数据库上比较了PML方法在不同降维维数时的性能变化情况。从这一图中的三个数据库上的性能变化情况可以发现当降维维数变化到足够大的时候，PML方法的性能趋于稳定。所以，在前面的实验结果报告里，本章均采用了这一图中的最后一个维数设置的性能。

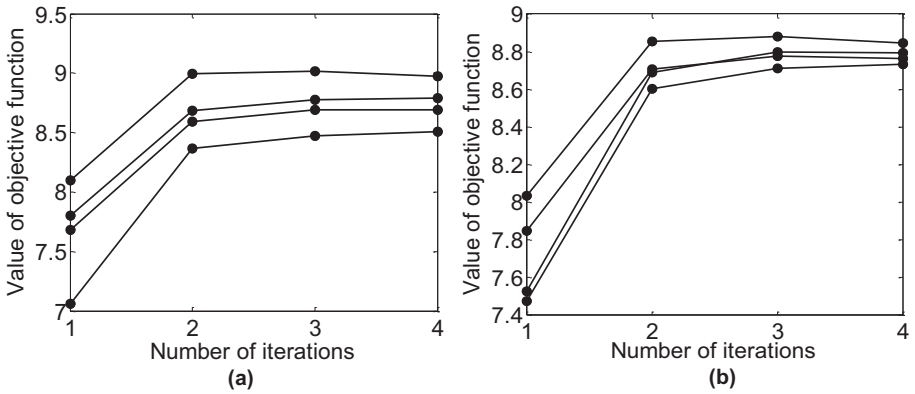


图 3.5. PML方法在优化求解时的收敛特性。(a) 图绘制PML在不同的迭代次数时的目标函数值变化情况，其数据来自YTF实验的其中4折。(b) 图绘制PML在4种不同的初始化时的目标函数值在迭代过程中的变化情况，其数据来自YTF实验的1折。

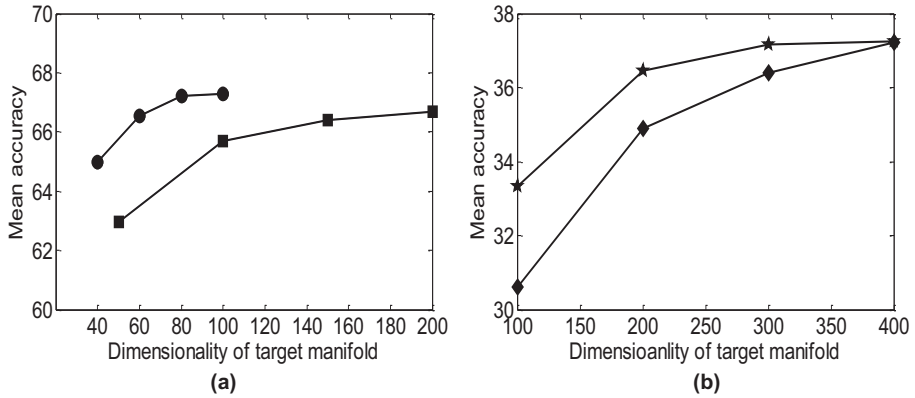


图 3.6. PML方法在输出不同维度的格拉斯曼流形时的平均性能。(a) 图中，带正方形标志的曲线和圆形标志的曲线分别代表在YTC和YTF上的性能变化情况。(b) 图中，带星形标志的曲线和带菱形标志的曲线分别代表PaSC的Control和Handheld实验上的性能变化情况。

（3）运行时间。在YTF数据库上，本节比较了几个代表性方法的运行时间。表3.3罗列出了它们在3.40GHZ PC上的训练时间和测试时间。从表3.3的结果可以发现新提出的PML方法和CMSM/SSDML/DCC在训练和测试时都要远快于GDA和GGDA。这是因为GDA和GGDA均采用了核学习方法，在计算核矩阵时随着数据的增大呈指数级增长因此它们的计算复杂度是极其高的。除此之外，如表3.3所示，本节还计算出了PML-GDA和PML-GGDA在YTF上降维维数为100时的计算时间。可以看出，PML通过降维的方式在一定程度上加速了GDA和GGDA。如果在降维维数设得更低的情况下，相信GDA和GGDA的速度会提得更快。

3.6　本章小结

本章提出一种新的在格拉斯曼流形上的判别学习方法来解决视频-视频人脸识别问题。具体地，该方法为了在以线性子空间为基本元素的格拉斯曼流形上进行判别学习，

表 3.3. PML对比方法在YTF数据库上的计算时间对比（单位：秒）

Methods	CMSM	SSDML	DCC	GDA	GGDA	**PML**	**PML-GDA**	**PML-GGDA**
Train	22.25	30.03	23.15	974.14	1306.13	**33.71**	**932.04**	**1076.33**
Test	9.25	23.15	8.87	278.71	420.17	**9.44**	**195.10**	**226.58**

通过将原始的格拉斯曼流形变换到一个更具判别能力的新格拉斯曼流形上的方式来学习更具判别性的投影度量。新提的方法不仅可以为作为一种格拉斯曼流形上的度量学习方法，而且也可以成为一种施加于格拉斯曼流形的降维技术。以视频人脸识别和视频人脸确认为两个基本任务，本章通过系统深入的实验对提出的方法进行了验证。结果表明，本章方法在四个极具挑战性的人脸视频数据库上的性能均达到了与当前最优的基于格拉斯曼流形的判别式学习方法同等可比的水平。同时，本章的实验也表明了本章的降维方法与其它方法结合之后的性能可以在很大程度上超过原始方法的性能。

本章的工作是对基于格拉斯曼流形的判别学习方法的一种全新的尝试，研究了在格拉斯曼流形上进行降维和度量学习时开发格拉斯曼流形上对应的黎曼几何的重要性。在后续研究工作中将考虑将本工作的投影度量学习框架应用于其它格拉斯曼流形上的度量如比奈-柯西度量（Binet-Cauchy metric）。此外，为了更好地求解本章的度量学习问题，如何开发一种更加有效的可以在理论上能证明收敛的求解算法是值得深入研究的方向。

第四章　基于双阶统计量建模的跨欧氏-黎曼度量学习方法

4.1　引言

针对基于视频的人脸识别问题，上一章提出的基于线性子空间建模的方法主要是利用线性子空间来建模视频人脸数据的变化模式。显然这一统计建模方式只是采用了二阶统计量（等同于协方差），而忽略了一阶统计量，即数据的均值信息。因此，本章研究同时采用双阶统计量来对视频中的集合数据进行建模，在这基础上设计了一种新型的异质度量学习框架，同时解决三种不同的基于视频的人脸识别问题。

如图4.1 (a)所示，传统的异质度量学习方法[26, 95, 97, 110, 146, 147, 154]一般是在多个欧氏空间上学习多个马氏度量或者多个变换到一个公共的欧氏子空间上。比如，文献[110]提出一种度量学习方法对位于不同欧氏空间上的多视数据寻求多个带有领域保持约束的投影变换，从而将它们映射到一个公共子空间上进行判别学习。McFee和Lanckriet[97]采用了多核学习技术将位于多个欧氏空间上的多种异质数据统一映射到一个公共子空间进行融合。文献[154]开发了一种基于联合图正则化的异质度量学习方法对位于两种不同维度的欧氏空间上的异质数据同时学习两种不同的变换使得它们在公共子空间上的距离度量更具判别性。最近，文献[162]为了解决图像-视频（图像集合）分类问题，首先将图像集合建模成仿射包，然后再学习点到仿射包的距离度量。具体地，这一工作采用仿射包的建模方式将图像集合表示为一个虚拟的样本点（即集合中的所有点的一种仿射组合），然后对原始图像和图像集合生成的虚拟样本统一学习同一种线性变换。因此，他们提出的度量学习方法实际是在同一个欧氏空间上学习马氏度量。

相比之下，如图4.1 (b)所示，本章要解决的是欧氏元素与黎曼元素之间的异质度量学习问题：在视频-图像/图像-视频人脸识别场景下，图像一般表示成某种特征向量位于欧氏空间上，而用于视频建模的均值和协方差分别位于欧氏空间和黎曼流形上（一些研究[29, 59, 135, 139]表明协方差通常位于对称正定矩阵流形上），因此这一识别任务可以形式化成欧氏点与黎曼元素之间的匹配问题。在视频-视频人脸识别场景下，由于用于建模视频的均值和协方差在不同方面描述视频序列中的数据，对这两个统计量进行融合显然能够有效提高识别性能。因此，这一对双阶统计量进行融合的任务同样也可以形式化成一种欧氏点与黎曼元素的异质度量学习问题。

针对上述两种异质匹配/融合问题，本章提出一种统一的异质度量学习框架，称为跨欧氏-黎曼度量学习(Cross Euclidean-to-Riemannian Metric Learning, CERML)。在CERML框架里，图像特征向量表示为欧氏点，视频里的图像集合被建模为均值和协方差为欧氏点与黎曼元素，从而将视频-图像/图像-视频的匹配分类问题形式化为欧氏

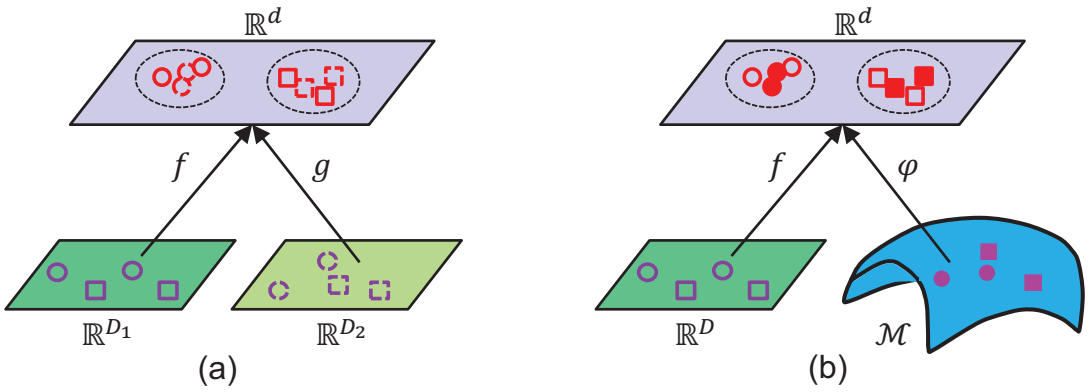


图 4.1. 传统异质度量学习(a)与本章新异质度量学习(b)对比

元素与黎曼元素之间的距离度量计算问题，将视频-视频的分类问题形式化为欧氏元素与黎曼元素之间的距离融合问题。基于欧氏空间与黎曼流形存在的巨大异质性问题，CERML框架采用希尔伯特空间嵌入的方法将原始的异质空间都分别映射到一个高维的希尔伯特空间上。由于希尔伯特空间在广义上遵从的是欧氏几何，这样极大减少了原始欧氏空间和黎曼空间的异质性。另外，通过推导出基于流形上的黎曼度量的核函数，这一希尔伯特空间嵌入可以有效编码原始流形上的黎曼几何。与前人工作相比，本章提出的CERML框架主要有如下几点贡献：

（1）本章通过将视频-图像/图像-视频人脸识别问题形式化成异质数据的匹配问题，将视频-视频人脸识别问题形式化成异质数据的融合问题，首次把这三种不同的基于视频的人脸识别问题放在同一个异质度量学习框架里解决。

（2）本章首次提出了一种用于匹配/融合欧氏数据和非欧氏数据（即黎曼数据）的异质度量学习框架。相比之下，现有的异质度量学习只解决不同欧氏数据的匹配问题（图4.1 (a)）。因此，这一问题更具有挑战性，研究这一新型的异质度量学习问题（图4.1 (b)）具有非常大的理论价值。

（3）本章新提出的CERML框架除了为基于协方差的统计模型而设计，同样也可以适配于其它统计模型，如线性子空间和仿射子空间。因此，CERML可以成为一个包容其它统计模型的通用异质度量学习框架。

（4）本章新提出的CERML框架在三个不同的基于视频的人脸识别场景上进行了系统的实验对比及验证。实验结果表明，CERML作为一种通用的度量学习框架在三种不同的视频人脸识别任务里要一致地优于所对比的当前最优方法。

本章接下来的安排如下：4.2节详细介绍新提出的异质度量学习框架，包括问题形式化、目标函数和优化算法；4.3节重点讨论新提出的异质度量学习框架在视频-图像/图像-视频和视频-视频人脸识别场景下的实例化；4.4节讨论新提出的异质度量学习框架是如何通用于三种不同的统计模型；4.5节给出新提出方法在四个公开人脸数据库上与其它方法的对比结果；4.6节对本章新提出的异质度量学习方法进行了总结与讨论。

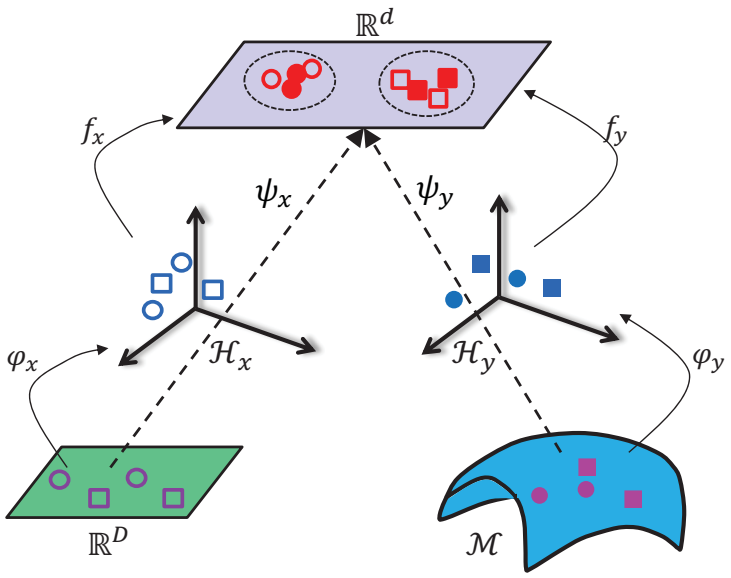


图 4.2. 本章新提出的跨欧氏-黎曼度量学习框架（Cross Euclidean-to-Riemannian Metric Learning, CERML）

4.2　跨欧氏-黎曼度量学习框架

本节将分三个小节来阐述新提出的跨欧氏-黎曼度量学习框架CERML。首先形式化跨欧氏-黎曼异质度量学习问题，然后再提出异质度量学习框架的目标函数，最后设计一种迭代优化算法来解决这一目标函数。

4.2.1　问题形式化

给定一个位于欧氏空间$\mathbb{R}^D$上的数据集合$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$，其类别标签分别为$\{l_1^x, l_2^x, \ldots, l_m^x\}$以及一个位于黎曼流形$\mathcal{M}$上的数据集合$\boldsymbol{y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\}$，其类别标签分别为$\{l_1^y, l_2^y, \ldots, l_n^y\}$。分别在欧氏点集合和黎曼元素集合取出$\boldsymbol{x}_i, \boldsymbol{y}_j$构成一对异质样本，并令$d(\boldsymbol{x}_i, \boldsymbol{y}_j)$表示它们之间的距离。为了计算出这两个异质样本之间的距离度量，如图4.2所示，本章采用了通过学习两个不同映射ψ_x, ψ_y将它们变换到一个公共的欧氏子空间$\mathbb{R}^d$上的策略。在这一公共子空间上，学习得到的距离度量可以转化成经典的欧氏度量：

$$d(\boldsymbol{x}_i, \boldsymbol{y}_j) = \sqrt{(\psi_x(\boldsymbol{x}_i) - \psi_y(\boldsymbol{y}_j))^T(\psi_x(\boldsymbol{x}_i) - \psi_y(\boldsymbol{y}_j))}. \tag{4.1}$$

然而，由于原始的欧氏空间$\mathbb{R}^D$和黎曼流形$\mathcal{M}$遵从的几何结构和其所配置的度量均不同从而导致两者之间存在着非常大的异质性，如果直接用一个线性映射把它们映射到一个公共的欧氏子空间$\mathbb{R}^d$是比较困难的。因此，需要考虑先把黎曼流形转化为一个平坦的欧氏空间来缩小与原始欧氏空间的异质性。目前有两种方式可以将黎曼流形嵌入为一个平坦的欧氏空间。一种是切空间嵌入[119, 128]，另一种是高维的可再生核希尔伯特空间（Reproducing Kernel Hilbert Space，RKHS）嵌

入[28, 70, 135, 139]。切空间嵌入策略是在位于黎曼流形上的某个特定点对应的切空间上估计流形的局部几何结构，而希尔伯特空间嵌入策略通过推导出基于黎曼度量的核函数来编码黎曼流形上的黎曼几何。由工作[70]指出，相比切空间嵌入，希尔伯特空间嵌入可以得到更加丰富的特征表示，从而更加有利于分类任务。因此，本章提出的异质度量学习框架首先通过推导出基于典型的欧氏度量和黎曼度量的核函数来定义隐式的非线性变换的内积形式，以两个非线性变换φ_x and φ_y将欧氏空间$\mathbb{R}^D$和黎曼流形$\mathcal{M}$分别嵌入到一个高维的希尔伯特空间$\mathcal{H}_x, \mathcal{H}_y$上，然后分别从这两个不同的希尔伯特空间上学习两个不同的线性映射f_x, f_y到目标公共子空间$\mathbb{R}^d$上，最终目标是用两个不同的映射$\psi_x = f_x \circ \varphi_x$, $\psi_y = f_y \circ \varphi_y$将分别位于原始的欧氏空间和黎曼流形上的异质数据都映射到一个公共欧氏子空间上，从而使得两两异质数据之间的距离度量可以用经典的欧氏距离即公式4.1来计算。具体地，两个线性映射可以分别表示为$f_x(\boldsymbol{x}_i) = \boldsymbol{V}_x^T \boldsymbol{x}_i$, $f_y(\boldsymbol{y}_j) = \boldsymbol{V}_y^T \boldsymbol{y}_j$，其中$\boldsymbol{V}_x^T, \boldsymbol{V}_y^T$均为线性变换矩阵。借助经典的核技术，可以采用推导核函数的方式将两个非线性映射的内积分别表示为$\langle \varphi_x(\boldsymbol{x}_i), \varphi_x(\boldsymbol{x}_j) \rangle = \boldsymbol{K}_x(\boldsymbol{x}_i, \boldsymbol{x}_j), \langle \varphi_y(\boldsymbol{y}_i), \varphi_y(\boldsymbol{y}_j) \rangle = \boldsymbol{K}_y(\boldsymbol{y}_i, \boldsymbol{y}_j)$，其中$\boldsymbol{K}_x, \boldsymbol{K}_y$分别为两个核函数对应的核矩阵。最后再通过在希尔伯特空间的内积参数化，可以得到最终映射函数ψ_x和ψ_y的具体形式分别为$\psi(\boldsymbol{x}_i) = \boldsymbol{W}_x^T \boldsymbol{K}_{x_i}$，$\psi(\boldsymbol{y}_j) = \boldsymbol{W}_y^T \boldsymbol{K}_{y_j}$。由此，分别位于欧氏空间和黎曼流形上的一对异质数据的距离度量4.1可以进一步具体化为：

$$d(\boldsymbol{x}_i, \boldsymbol{y}_j) = \sqrt{(\boldsymbol{W}_x^T \boldsymbol{K}_{x\_i} - \boldsymbol{W}_y^T \boldsymbol{K}_{y\_j})^T (\boldsymbol{W}_x^T \boldsymbol{K}_{x\_i} - \boldsymbol{W}_y^T \boldsymbol{K}_{y\_j})}. \tag{4.2}$$

另外，根据以上映射模式还可以得到位于相同空间上的一对同质样本变换到目标公共子空间后的距离度量：

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{W}_x^T \boldsymbol{K}_{x\_i} - \boldsymbol{W}_x^T \boldsymbol{K}_{x\_j})^T (\boldsymbol{W}_x^T \boldsymbol{K}_{x\_i} - \boldsymbol{W}_x^T \boldsymbol{K}_{x\_j})}. \tag{4.3}$$

$$d(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sqrt{(\boldsymbol{W}_y^T \boldsymbol{K}_{y\_i} - \boldsymbol{W}_y^T \boldsymbol{K}_{y\_j})^T (\boldsymbol{W}_y^T \boldsymbol{K}_{y\_i} - \boldsymbol{W}_y^T \boldsymbol{K}_{y\_j})}. \tag{4.4}$$

其中$\boldsymbol{K}_x$和$\boldsymbol{K}_y$的具体形式将会在4.4节详细介绍。

4.2.2 目标函数

从公式4.2, 4.3, 4.4中可以看出，这一异质度量学习的框架包含了两个参数化的变换矩阵$\boldsymbol{W}_x, \boldsymbol{W}_y$。为了学习更加有判别性的异质数据的度量，需要设计一个目标函数$J(\boldsymbol{W}_x, \boldsymbol{W}_y)$来优化学习$\boldsymbol{W}_x, \boldsymbol{W}_y$：

$$\min_{\boldsymbol{W}_x, \boldsymbol{W}_y} J(\boldsymbol{W}_x, \boldsymbol{W}_y) = \min_{\boldsymbol{W}_x, \boldsymbol{W}_y} \{D(\boldsymbol{W}_x, \boldsymbol{W}_y) + \lambda_1 G(\boldsymbol{W}_x, \boldsymbol{W}_y) + \lambda_2 T(\boldsymbol{W}_x, \boldsymbol{W}_y)\} \tag{4.5}$$

其中，$D(\boldsymbol{W}_x, \boldsymbol{W}_y)$是定义在相似样本对和不相似样本对集合上的距离约束项，$G(\boldsymbol{W}_x, \boldsymbol{W}_y)$, $T(\boldsymbol{W}_x, \boldsymbol{W}_y)$分别是几何约束项和变换约束项，后两项均是定义在参数矩阵$\boldsymbol{W}_x, \boldsymbol{W}_y$上的正则约束项，$\lambda_1 > 0, \lambda_2 > 0$是平衡参数。

距离约束项。这一项主要是为了约束同类的异质样本对的距离最小化，异类的异质样本对的距离最大化。具体地，采用了经典的距离平方和表达方式来定义这一项：

$$\begin{aligned} D(\boldsymbol{W}_x, \boldsymbol{W}_y) &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{A}(i,j)d^2(\boldsymbol{x}_i, \boldsymbol{y}_j), \\ \boldsymbol{A}(i,j) &= \begin{cases} 1, & \text{if } l_i^x = l_j^y, \\ -1, & \text{if } l_i^x \neq l_j^y. \end{cases} \end{aligned} \tag{4.6}$$

其中，$\boldsymbol{A}(i,j)$用来标记异质样本$d(\boldsymbol{x}_i, \boldsymbol{y}_j)$是否来自同类（异类）。为了消除由于相似样本对和不相似样本对之间的不平衡带来的负面影响，这里通过对矩阵$\boldsymbol{A}$里的每个元素进行求平均操作（即对其除以相似样本/不相似样本的总对数）来归一化$\boldsymbol{A}$。

几何约束项。 这一项主要是分别为欧氏数据和黎曼数据保持相应的欧氏几何和黎曼几何。因此，它具体表示成$G(\boldsymbol{W}_x, \boldsymbol{W}_y) = G_x(\boldsymbol{W}_x)+G_y(\boldsymbol{W}_y)$，其中$G_x(\boldsymbol{W}_x)$和$G_y(\boldsymbol{W}_y)$分别对应欧氏几何保持项和黎曼几何保持项：

$$\begin{aligned} G_x(\boldsymbol{W}_x) &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\boldsymbol{A}_x(i,j)d_x^2(\boldsymbol{x}_i, \boldsymbol{x}_j), \\ \boldsymbol{A}_x(i,j) &= \begin{cases} d_{ij}, & \text{if } l_i^x = l_j^x \text{ and } k_1(i,j), \\ -d_{ij}, & \text{if } l_i^x \neq l_j^x \text{ and } k_2(i,j), \\ 0, & \text{else.} \end{cases} \end{aligned} \tag{4.7}$$

$$\begin{aligned} G_y(\boldsymbol{W}_y) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{A}_y(i,j)d_y^2(\boldsymbol{y}_i, \boldsymbol{x}_j), \\ \boldsymbol{A}_y(i,j) &= \begin{cases} d_{ij}, & \text{if } l_i^y = l_j^y \text{ and } k_1(i,j), \\ -d_{ij}, & \text{if } l_i^y \neq l_j^y \text{ and } k_2(i,j), \\ 0, & \text{else.} \end{cases} \end{aligned} \tag{4.8}$$

其中，$d_{ij} = exp(-\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2/\sigma^2)$, $\boldsymbol{z}$表示数据$\boldsymbol{x}$ 或$\boldsymbol{y}$。$k_1(i,j)$表示样本i是样本j的k_1近邻或者样本j是样本i的k_1近邻。同理，$k_2(i,j)$表示样本i是样本j的k_2近邻或者样本j是样本i的k_2近邻。

变换约束项。 由于在目标公共子空间上计算距离时所有在不同的维度上的值是被相同对待的（即特征向量满足各向同性分布），因此这里的变换约束项可以形式化成：

$$T(\boldsymbol{W}_x, \boldsymbol{W}_y) = \frac{1}{2}(\|\boldsymbol{W}_x^T\boldsymbol{K}_x\|_F^2 + \|\boldsymbol{W}_y^T\boldsymbol{K}_y\|_F^2). \tag{4.9}$$

4.2.3　优化算法

为了优化目标函数$J(\boldsymbol{W}_x, \boldsymbol{W}_y)$（即公式4.5），本节开发了一种迭代优化的算法。首先采用传统Fisher判别分析（Fisher Discriminant Analysis, FDA）的Fisher准则来初始化

参数矩阵$\boldsymbol{W}_x, \boldsymbol{W}_y$，然后采用交替优化的策略进行求解。在介绍具体求解算法之前，先把公式4.6, 4.7, 4.8的矩阵形式重写出来：

$$\begin{aligned} D(\boldsymbol{W}_x, \boldsymbol{W}_y) = \frac{1}{2}(&\boldsymbol{W}_x^T \boldsymbol{K}_x \boldsymbol{B}_x^{'} \boldsymbol{K}_x^T \boldsymbol{W}_x + \boldsymbol{W}_y^T \boldsymbol{K}_y \boldsymbol{B}_y^{'} \boldsymbol{K}_y^T \boldsymbol{W}_y \\ &- 2\boldsymbol{W}_x^T \boldsymbol{K}_x \boldsymbol{A} \boldsymbol{K}_y^T \boldsymbol{W}_y). \end{aligned} \tag{4.10}$$

$$\begin{aligned} G_x(\boldsymbol{W}_x) &= \boldsymbol{W}_x^T \boldsymbol{K}_x \boldsymbol{B}_x \boldsymbol{K}_x^T \boldsymbol{W}_x - \boldsymbol{W}_x^T \boldsymbol{K}_x \boldsymbol{A}_x \boldsymbol{K}_x^T \boldsymbol{W}_x \\ &= \boldsymbol{W}_x^T \boldsymbol{K}_x \boldsymbol{L}_x \boldsymbol{K}_x^T \boldsymbol{W}_x. \end{aligned} \tag{4.11}$$

$$\begin{aligned} G_y(\boldsymbol{W}_y) &= \boldsymbol{W}_y^T \boldsymbol{K}_y \boldsymbol{B}_y \boldsymbol{K}_y^T \boldsymbol{W}_y - \boldsymbol{W}_y^T \boldsymbol{K}_y \boldsymbol{A}_y \boldsymbol{K}_y^T \boldsymbol{W}_y \\ &= \boldsymbol{W}_y^T \boldsymbol{K}_y \boldsymbol{L}_y \boldsymbol{K}_y^T \boldsymbol{W}_y. \end{aligned} \tag{4.12}$$

其中，$\boldsymbol{B}_x^{'}$, $\boldsymbol{B}_y^{'}$, $\boldsymbol{B}_x$均是对角矩阵并分别表示为:$\boldsymbol{B}_x^{'}(i,i) = \sum_{j=1}^{n} \boldsymbol{A}(i,j)$, $\boldsymbol{B}_y^{'}(j,j) = \sum_{i=1}^{m} \boldsymbol{A}(i,j)$, $\boldsymbol{B}_x(i,i) = \sum_{j=1}^{m} \boldsymbol{A}_x(i,j)$, $\boldsymbol{B}_y(i,i) = \sum_{j=1}^{n} \boldsymbol{A}_y(i,j)$。

（1）初始化。首先为三个矩阵$\boldsymbol{A}, \boldsymbol{A}_x, \boldsymbol{A}_y$分别定义对应的类内模板和类间模板$\boldsymbol{A}^w/\boldsymbol{A}^b, \boldsymbol{A}_x^w/\boldsymbol{A}_x^b, \boldsymbol{A}_y^w/\boldsymbol{A}_y^b$，分别表示成：

$$\boldsymbol{A}^w(i,j) = \begin{cases} 1, & \text{if } l_i^x = l_j^y, \\ 0, & \text{if } l_i^x \neq l_j^y. \end{cases} \tag{4.13}$$

$$\boldsymbol{A}^b(i,j) = \begin{cases} 0, & \text{if } l_i^x = l_j^y, \\ 1, & \text{if } l_i^x \neq l_j^y. \end{cases} \tag{4.14}$$

$$\boldsymbol{A}_x^w(i,j) = \begin{cases} d_{ij}, & \text{if } l_i^x = l_j^x \text{ and } k_1(i,j), \\ 0, & \text{else.} \end{cases} \tag{4.15}$$

$$\boldsymbol{A}_x^b(i,j) = \begin{cases} d_{ij}, & \text{if } l_i^x \neq l_j^x \text{ and } k_2(i,j), \\ 0, & \text{else.} \end{cases} \tag{4.16}$$

$$\boldsymbol{A}_y^w(i,j) = \begin{cases} d_{ij}, & \text{if } l_i^y = l_j^y \text{ and } k_1(i,j), \\ 0, & \text{else.} \end{cases} \tag{4.17}$$

$$\boldsymbol{A}_y^b(i,j) = \begin{cases} d_{ij}, & \text{if } l_i^y \neq l_j^y \text{ and } k_2(i,j), \\ 0, & \text{else.} \end{cases} \tag{4.18}$$

将公式4.13,4.14的$\boldsymbol{A}^w, \boldsymbol{A}^b$分别代入公式4.10中，可以分别得到关于$D(\boldsymbol{W}_x, \boldsymbol{W}_y)$对应的类内模板$D^w(\boldsymbol{W}_x, \boldsymbol{W}_y)$和类间模板$D^b(\boldsymbol{W}_x, \boldsymbol{W}_y)$：

$$\begin{aligned} D^w(\boldsymbol{W}_x, \boldsymbol{W}_y) = \frac{1}{2}(&\boldsymbol{W}_x^T \boldsymbol{K}_x \boldsymbol{B}_x^{'w} \boldsymbol{K}_x^T \boldsymbol{W}_x \\ &+ \boldsymbol{W}_y^T \boldsymbol{K}_y \boldsymbol{B}_y^{'w} \boldsymbol{K}_y^T \boldsymbol{W}_y - 2\boldsymbol{W}_x^T \boldsymbol{K}_x \boldsymbol{Z}^w \boldsymbol{K}_y \boldsymbol{W}_y^T). \end{aligned} \tag{4.19}$$

$$\begin{aligned} D^b(\boldsymbol{W}_x,\boldsymbol{W}_y) = \frac{1}{2}(&\boldsymbol{W}_x^T\boldsymbol{K}_x\boldsymbol{B}_x^{'b}\boldsymbol{K}_x^T\boldsymbol{W}_x \\ &+\boldsymbol{W}_y^T\boldsymbol{K}_y\boldsymbol{B}_y^{'b}\boldsymbol{K}_y^T\boldsymbol{W}_y - 2\boldsymbol{W}_x^T\boldsymbol{K}_x\boldsymbol{Z}^b\boldsymbol{K}_y\boldsymbol{W}_y^T). \end{aligned} \tag{4.20}$$

同理，4.15,4.16与公式4.11得到关于$G_x(\boldsymbol{W}_x)$对应的类内模板$G_x^w(\boldsymbol{W}_x)$和类间模板$G_x^b(\boldsymbol{W}_x)$，4.17,4.18与公式4.12得到关于$G_y(\boldsymbol{W}_y)$对应的类内模板$G_y^w(\boldsymbol{W}_y)$和类间模板$G_y^b(\boldsymbol{W}_y)$:

$$\begin{aligned} G_x^w(\boldsymbol{W}_x) &= \boldsymbol{W}_x^T\boldsymbol{K}_x\boldsymbol{B}_x^w\boldsymbol{K}_x^T\boldsymbol{W}_x - \boldsymbol{W}_x^T\boldsymbol{K}_x\boldsymbol{Z}_x^w\boldsymbol{K}_x^T\boldsymbol{W}_x \\ &= \boldsymbol{W}_x^T\boldsymbol{K}_x\boldsymbol{L}_x^w\boldsymbol{K}_x^T\boldsymbol{W}_x. \end{aligned} \tag{4.21}$$

$$\begin{aligned} G_x^b(\boldsymbol{W}_x) &= \boldsymbol{W}_x^T\boldsymbol{K}_x\boldsymbol{B}_x^b\boldsymbol{K}_x^T\boldsymbol{W}_x - \boldsymbol{W}_x^T\boldsymbol{K}_x\boldsymbol{Z}_x^b\boldsymbol{K}_x^T\boldsymbol{W}_x \\ &= \boldsymbol{W}_x^T\boldsymbol{K}_x\boldsymbol{L}_x^b\boldsymbol{K}_x^T\boldsymbol{W}_x. \end{aligned} \tag{4.22}$$

$$\begin{aligned} G_y^w(\boldsymbol{W}_y) &= \boldsymbol{W}_y^T\boldsymbol{K}_y\boldsymbol{B}_y^w\boldsymbol{K}_y^T\boldsymbol{W}_y - \boldsymbol{W}_y^T\boldsymbol{K}_y\boldsymbol{Z}_y^w\boldsymbol{K}_y^T\boldsymbol{W}_y \\ &= \boldsymbol{W}_y^T\boldsymbol{K}_y\boldsymbol{L}_y^w\boldsymbol{K}_y^T\boldsymbol{W}_y. \end{aligned} \tag{4.23}$$

$$\begin{aligned} G_y^b(\boldsymbol{W}_y) &= \boldsymbol{W}_y^T\boldsymbol{K}_y\boldsymbol{B}_y^b\boldsymbol{K}_y^T\boldsymbol{W}_y - \boldsymbol{W}_y^T\boldsymbol{K}_y\boldsymbol{Z}_y^b\boldsymbol{K}_y^T\boldsymbol{W}_y \\ &= \boldsymbol{W}_y^T\boldsymbol{K}_y\boldsymbol{L}_y^b\boldsymbol{K}_y^T\boldsymbol{W}_y. \end{aligned} \tag{4.24}$$

接下来，可以采用传统Fisher判别分析的Fisher准则通过最大化所有类间模板之和同时最小化所有类内模板之和来初始化$\boldsymbol{W}_x,\boldsymbol{W}_y$:

$$\begin{aligned} &\max_{\boldsymbol{W}_x,\boldsymbol{W}_y}\{\boldsymbol{D}^b(\boldsymbol{W}_x,\boldsymbol{W}_y)+\lambda_1 G^b(\boldsymbol{W}_x,\boldsymbol{W}_y)\}, \\ &s.t.\quad \boldsymbol{D}^w(\boldsymbol{W}_x,\boldsymbol{W}_y)+\lambda_1 G^w(\boldsymbol{W}_x,\boldsymbol{W}_y)=1. \end{aligned} \tag{4.25}$$

其中，$G^b(\boldsymbol{W}_x,\boldsymbol{W}_y) = G_x^b(\boldsymbol{W}_x,\boldsymbol{W}_y)+G_y^b(\boldsymbol{W}_x,\boldsymbol{W}_y)$，$G^w(\boldsymbol{W}_x,\boldsymbol{W}_y) = G_x^w(\boldsymbol{W}_x,\boldsymbol{W}_y)+G_y^w(\boldsymbol{W}_x,\boldsymbol{W}_y)$。把公式4.19-4.24代入公式4.25得：

$$\begin{aligned} &\max \begin{bmatrix}\boldsymbol{W}_x\\ \boldsymbol{W}_y\end{bmatrix}^T \begin{bmatrix}\boldsymbol{K}_x\boldsymbol{R}_x^b\boldsymbol{K}_x^T & -\boldsymbol{K}_x\boldsymbol{A}^b\boldsymbol{K}_y^T\\ -\boldsymbol{K}_y(\boldsymbol{A}^b)^T\boldsymbol{K}_x^T & \boldsymbol{K}_y\boldsymbol{R}_y^b\boldsymbol{K}_y^T\end{bmatrix}\begin{bmatrix}\boldsymbol{W}_x\\ \boldsymbol{W}_y\end{bmatrix} \\ &s.t. \begin{bmatrix}\boldsymbol{W}_x\\ \boldsymbol{W}_y\end{bmatrix}^T \begin{bmatrix}\boldsymbol{K}_x\boldsymbol{R}_x^w\boldsymbol{K}_x^T & -\boldsymbol{K}_x\boldsymbol{A}^w\boldsymbol{K}_y^T\\ -\boldsymbol{K}_y(\boldsymbol{A}^w)^T\boldsymbol{K}_x^T & \boldsymbol{K}_y\boldsymbol{R}_y^w\boldsymbol{K}_y^T\end{bmatrix}\begin{bmatrix}\boldsymbol{W}_x\\ \boldsymbol{W}_y\end{bmatrix} = 1. \end{aligned} \tag{4.26}$$

其中，$\boldsymbol{R}_x^b = \boldsymbol{B}_x^b + 2\lambda_1\boldsymbol{L}_x^b$, $\boldsymbol{R}_y^b = \boldsymbol{B}_y^b + 2\lambda_1\boldsymbol{L}_y^b$, $\boldsymbol{R}_x^w = \boldsymbol{B}_x^w + 2\lambda_1\boldsymbol{L}_x^w$，$\boldsymbol{R}_y^w = \boldsymbol{B}_y^w + 2\lambda_1\boldsymbol{L}_y^w$。以上公式可以再进一步简化为：

$$\begin{aligned} &\max \boldsymbol{W}^T\boldsymbol{M}^b\boldsymbol{W},\quad s.t.\boldsymbol{W}^T\boldsymbol{M}^w\boldsymbol{W}=1. \\ &\Rightarrow \boldsymbol{M}^b\boldsymbol{W}=\lambda\boldsymbol{M}^w\boldsymbol{W}. \end{aligned} \tag{4.27}$$

其中，矩阵$\boldsymbol{M}^b = \begin{bmatrix} \boldsymbol{K}_x\boldsymbol{R}_x^b\boldsymbol{K}_x^T & -\boldsymbol{K}_x\boldsymbol{Z}^b\boldsymbol{K}_y^T \\ -\boldsymbol{K}_y(\boldsymbol{Z}^b)^T\boldsymbol{K}_x^T & \boldsymbol{y}\boldsymbol{R}_y^b\boldsymbol{K}_y^T \end{bmatrix}, \boldsymbol{M}^w = \begin{bmatrix} \boldsymbol{K}_x\boldsymbol{R}_x^w\boldsymbol{K}_x^T & -\boldsymbol{K}_x\boldsymbol{Z}^w\boldsymbol{K}_y^T \\ -\boldsymbol{K}_y(\boldsymbol{Z}^w)^T\boldsymbol{K}_x^T & \boldsymbol{y}\boldsymbol{R}_y^w\boldsymbol{K}_y^T \end{bmatrix}$，$\boldsymbol{W}^T = [\boldsymbol{W}_x^T, \boldsymbol{W}_y^T]$。很显然，这一目标函数是个广义特征值问题，可以简单用解析方法来求解。

（2）交替优化。先固定$\boldsymbol{W}_y$来更新$\boldsymbol{W}_x$。结合公式4.10, 4.11, 4.12与公式4.5，对目标函数$J(\boldsymbol{W}_x, \boldsymbol{W}_y)$在$\boldsymbol{W}_x$上进行求导并令其等于0，从而得到下面的等式：

$$\begin{aligned} \frac{\partial Q(\boldsymbol{W}_x, \boldsymbol{W}_y)}{\partial \boldsymbol{W}_x} &= \boldsymbol{K}_x\boldsymbol{B}_x^{'}\boldsymbol{K}_x^T\boldsymbol{W}_x - \boldsymbol{K}_x\boldsymbol{Z}\boldsymbol{K}_y^T\boldsymbol{W}_y \\ &\quad + 2\lambda_1\boldsymbol{K}_x\boldsymbol{L}_x\boldsymbol{K}_x^T\boldsymbol{W}_x + 2\lambda_2\boldsymbol{K}_x\boldsymbol{K}_x^T\boldsymbol{W}_x = 0. \end{aligned} \tag{4.28}$$

然后可以得到$\boldsymbol{W}_x$在$\boldsymbol{W}_y$固定时的解:

$$\boldsymbol{W}_x = (\boldsymbol{K}_x(\boldsymbol{B}_x^{'} + 2\lambda_1\boldsymbol{L}_x + 2\lambda_2\boldsymbol{I})\boldsymbol{K}_x^T)^{-1}\boldsymbol{K}_x\boldsymbol{A}\boldsymbol{K}_y^T\boldsymbol{W}_y. \tag{4.29}$$

同理可以得到$\boldsymbol{W}_y$在$\boldsymbol{W}_x$固定时的解：

$$\boldsymbol{W}_y = (\boldsymbol{K}_y(\boldsymbol{B}_y^{'} + 2\lambda_1\boldsymbol{L}_y + 2\lambda_2\boldsymbol{I})\boldsymbol{K}_y^T)^{-1}\boldsymbol{K}_y\boldsymbol{A}\boldsymbol{K}_x^T\boldsymbol{W}_x. \tag{4.30}$$

可以通过重复以上交替优化策略来求解$\boldsymbol{W}_x$和$\boldsymbol{W}_y$。虽然很难对提出的优化算法的收敛性进行理论证明，但是可以通过实验部分的验证说明了这一算法通过有限次的迭代之后可以收敛到一个稳定的最优解上。关于更详细的实验定量分析结果，请参考实验部分。

4.3 实例化

本节将分两小节分别介绍本章新提出的异质度量学习框架在视频-图像/图像-视频人脸识别以及在视频-视频人脸识别场景下的应用。

4.3.1 视频-图像/图像-视频人脸识别

根据本章引言所述，本章同时采用一阶统计量（即均值）和二阶统计量（即协方差）来对视频中的集合数据进行建模。因此，视频-图像/图像-视频人脸识别任务可以形式化成用于表示图像的欧氏数据与用于建模视频的欧氏数据（均值）和黎曼数据（协方差）进行匹配的问题。形式化上，图像的欧氏数据表示为$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}, \boldsymbol{x}_i \in \mathbb{R}^{D_1}$,其类别标签分别为$\{l_1^x, l_2^x, \ldots, l_m^x\}$。视频的欧氏数据表示为$\boldsymbol{y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\}, \boldsymbol{y}_j \in \mathbb{R}^{D_2}$，其类别标签分别为$\{l_1^y, l_2^y, \ldots, l_n^y\}$。相应地，视频的黎曼数据为$\boldsymbol{z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}, \boldsymbol{z}_j \in \mathcal{M}$，其类别标签分别为$\{l_1^z, l_2^z, \ldots, l_n^z\}$。

距离度量。在4.2.1小节里的距离度量（公式4.2）可以实例化为：

$$\begin{aligned} d(\boldsymbol{x}_i, \boldsymbol{y}_j) + d(\boldsymbol{x}_i, \boldsymbol{z}_j) &= \sqrt{(\boldsymbol{W}_x^T\boldsymbol{K}_{x\_i} - \boldsymbol{W}_y^T\boldsymbol{K}_{y\_j})^T(\boldsymbol{W}_x^T\boldsymbol{K}_{x\_i} - \boldsymbol{W}_y^T\boldsymbol{K}_{y\_j})} \\ &\quad + \sqrt{(\boldsymbol{W}_x^T\boldsymbol{K}_{x\_i} - \boldsymbol{W}_z^T\boldsymbol{K}_{z\_j})^T(\boldsymbol{W}_x^T\boldsymbol{K}_{x\_i} - \boldsymbol{W}_z^T\boldsymbol{K}_{z\_j})}. \end{aligned} \tag{4.31}$$

目标函数。在4.2.2小节里的目标函数（公式4.5）实例化为：

$$\begin{aligned}&\min_{\boldsymbol{W}_x,\boldsymbol{W}_y,\boldsymbol{W}_z} J(\boldsymbol{W}_x,\boldsymbol{W}_y,\boldsymbol{W}_z)\\&=\min_{\boldsymbol{W}_x,\boldsymbol{W}_y,\boldsymbol{W}_z}\{D(\boldsymbol{W}_x,\boldsymbol{W}_y,\boldsymbol{W}_z)+\lambda_1 G(\boldsymbol{W}_x,\boldsymbol{W}_y,\boldsymbol{W}_z)+\lambda_2 T(\boldsymbol{W}_x,\boldsymbol{W}_y,\boldsymbol{W}_z)\}\end{aligned} \tag{4.32}$$

其中，距离约束项$D(\boldsymbol{W}_x,\boldsymbol{W}_y,\boldsymbol{W}_z)=D(\boldsymbol{W}_x,\boldsymbol{W}_y)+D(\boldsymbol{W}_x,\boldsymbol{W}_z)$，几何约束项$G(\boldsymbol{W}_x,\boldsymbol{W}_y,\boldsymbol{W}_z)=G_x(\boldsymbol{W}_x)+G_y(\boldsymbol{W}_y)+G_z(\boldsymbol{W}_z)$，变换约束项$T(\boldsymbol{W}_x,\boldsymbol{W}_y,\boldsymbol{W}_z)=\frac{1}{2}(\|\boldsymbol{W}_x^T\boldsymbol{K}_{\boldsymbol{X}}\|_F^2+\|\boldsymbol{W}_y^T\boldsymbol{K}_{\boldsymbol{Y}}\|_F^2+\|\boldsymbol{W}_z^T\boldsymbol{K}_{\boldsymbol{Z}}\|_F^2)$。

初始化。在4.2.3小节里的优化算法中的初始化目标函数（公式4.26）实例化为：

$$\begin{aligned}\max\ &\begin{bmatrix}\boldsymbol{W}_x\\\boldsymbol{W}_y\\\boldsymbol{W}_z\end{bmatrix}^T\begin{bmatrix}\boldsymbol{K}_x\boldsymbol{R}_x^b\boldsymbol{K}_x^T & -\boldsymbol{K}_x\boldsymbol{A}_{xy}^b\boldsymbol{K}_y^T & -\boldsymbol{K}_x\boldsymbol{A}_{xz}^b\boldsymbol{K}_z^T\\-\boldsymbol{K}_y(\boldsymbol{A}_{xy}^b)^T\boldsymbol{K}_x^T & \boldsymbol{K}_y\boldsymbol{R}_y^b\boldsymbol{K}_y^T & 0\\-\boldsymbol{K}_z(\boldsymbol{A}_{xz}^b)^T\boldsymbol{K}_x^T & 0 & \boldsymbol{K}_z\boldsymbol{R}_z^b\boldsymbol{K}_z^T\end{bmatrix}\begin{bmatrix}\boldsymbol{W}_x\\\boldsymbol{W}_y\\\boldsymbol{W}_z\end{bmatrix}\\s.t.\ &\begin{bmatrix}\boldsymbol{W}_x\\\boldsymbol{W}_y\\\boldsymbol{W}_z\end{bmatrix}^T\begin{bmatrix}\boldsymbol{K}_x\boldsymbol{R}_x^w\boldsymbol{K}_x^T & -\boldsymbol{K}_x\boldsymbol{A}_{xy}^w\boldsymbol{K}_y^T & -\boldsymbol{K}_x\boldsymbol{A}_{xz}^w\boldsymbol{K}_z^T\\-\boldsymbol{K}_y(\boldsymbol{A}_{xy}^w)^T\boldsymbol{K}_x^T & \boldsymbol{K}_y\boldsymbol{R}_y^w\boldsymbol{K}_y^T & 0\\-\boldsymbol{K}_z(\boldsymbol{A}_{xz}^w)^T\boldsymbol{K}_x^T & 0 & \boldsymbol{K}_z\boldsymbol{R}_z^w\boldsymbol{K}_z^T\end{bmatrix}\begin{bmatrix}\boldsymbol{W}_x\\\boldsymbol{W}_y\\\boldsymbol{W}_z\end{bmatrix}=1.\end{aligned} \tag{4.33}$$

交替求解。在4.2.3小节里的优化算法中的交替求解过程对应的解析解分别表示为：

$$\begin{aligned}\boldsymbol{W}_x&=(\boldsymbol{K}_x(2\boldsymbol{B}_x^{'}+2\lambda_1\boldsymbol{L}_x+2\lambda_2\boldsymbol{I})\boldsymbol{K}_x^T)^{-1}(\boldsymbol{K}_x\boldsymbol{A}_{xy}\boldsymbol{K}_y^T\boldsymbol{W}_y+\boldsymbol{K}_x\boldsymbol{A}_{xz}\boldsymbol{K}_z^T\boldsymbol{W}_z).\\\boldsymbol{W}_y&=(\boldsymbol{K}_y(\boldsymbol{B}_y^{'}+2\lambda_1\boldsymbol{L}_y+2\lambda_2\boldsymbol{I})\boldsymbol{K}_y^T)^{-1}\boldsymbol{K}_y\boldsymbol{A}_{xy}\boldsymbol{K}_x^T\boldsymbol{W}_x.\\\boldsymbol{W}_z&=(\boldsymbol{K}_z(\boldsymbol{B}_z^{'}+2\lambda_1\boldsymbol{L}_z+2\lambda_2\boldsymbol{I})\boldsymbol{K}_z^T)^{-1}\boldsymbol{K}_z\boldsymbol{A}_{xz}\boldsymbol{K}_x^T\boldsymbol{W}_x.\end{aligned} \tag{4.34}$$

4.3.2　视频-视频人脸识别

与视频-图像/图像-视频人脸识别相同，视频-视频人脸识别同样同时采用一阶统计量（即均值）和二阶统计量（即协方差）来对视频中的集合数据进行建模。因此，视频-视频人脸识别任务可以形式化成用于建模视频的欧氏数据（均值）和黎曼数据（协方差）距离度量的融合问题。形式化上，视频的欧氏数据表示为$\boldsymbol{y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_n\},\boldsymbol{y}_j\in\mathbb{R}^{D_2}$，其类别标签分别为$\{l_1^y,l_2^y,\ldots,l_n^y\}$。相应地，视频的黎曼数据为$\boldsymbol{z}=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_n\},\boldsymbol{z}_j\in\mathcal{M}$，其类别标签分别为$\{l_1^z,l_2^z,\ldots,l_n^z\}$。

距离度量。在4.2.1小节里的距离度量（公式4.3，4.4）可以实例化为：

$$\begin{aligned}d(\boldsymbol{y}_i,\boldsymbol{y}_j)+d(\boldsymbol{z}_i,\boldsymbol{z}_j)=&\sqrt{(\boldsymbol{W}_y^T\boldsymbol{K}_{y_i}-\boldsymbol{W}_y^T\boldsymbol{K}_{y_j})^T(\boldsymbol{W}_y^T\boldsymbol{K}_{y_i}-\boldsymbol{W}_y^T\boldsymbol{K}_{y_j})}\\&+\sqrt{(\boldsymbol{W}_z^T\boldsymbol{K}_{z_i}-\boldsymbol{W}_z^T\boldsymbol{K}_{z_j})^T(\boldsymbol{W}_z^T\boldsymbol{K}_{z_i}-\boldsymbol{W}_z^T\boldsymbol{K}_{z_j})}.\end{aligned} \tag{4.35}$$

目标函数。在4.2.2小节里的目标函数（公式4.5）实例化为：

$$\min_{\boldsymbol{W}_y,\boldsymbol{W}_z} J(\boldsymbol{W}_y,\boldsymbol{W}_z)=\min_{\boldsymbol{W}_y,\boldsymbol{W}_z}\{D(\boldsymbol{W}_y,\boldsymbol{W}_z)+\lambda_1 G(\boldsymbol{W}_y,\boldsymbol{W}_z)+\lambda_2 T(\boldsymbol{W}_y,\boldsymbol{W}_z)\} \tag{4.36}$$

初始化。在4.2.3小节里的优化算法中的初始化目标函数（公式4.26）实例化为：

$$\begin{aligned}\max & \begin{bmatrix}\boldsymbol{W}_y\\ \boldsymbol{W}_z\end{bmatrix}^T \begin{bmatrix}\boldsymbol{K}_y\boldsymbol{R}_y^b\boldsymbol{K}_y^T & -\boldsymbol{K}_y\boldsymbol{A}^b\boldsymbol{K}_z^T\\ -\boldsymbol{K}_z(\boldsymbol{A}^b)^T\boldsymbol{K}_y^T & \boldsymbol{K}_z\boldsymbol{R}_z^b\boldsymbol{K}_z^T\end{bmatrix} \begin{bmatrix}\boldsymbol{W}_y\\ \boldsymbol{W}_z\end{bmatrix}\\ s.t. & \begin{bmatrix}\boldsymbol{W}_y\\ \boldsymbol{W}_z\end{bmatrix}^T \begin{bmatrix}\boldsymbol{K}_y\boldsymbol{R}_y^w\boldsymbol{K}_y^T & -\boldsymbol{K}_y\boldsymbol{A}^w\boldsymbol{K}_z^T\\ -\boldsymbol{K}_z(\boldsymbol{A}^w)^T\boldsymbol{K}_y^T & \boldsymbol{K}_z\boldsymbol{R}_z^w\boldsymbol{K}_z^T\end{bmatrix} \begin{bmatrix}\boldsymbol{W}_y\\ \boldsymbol{W}_z\end{bmatrix} = 1.\end{aligned} \tag{4.37}$$

交替求解。在4.2.3小节里的优化算法中的交替求解过程对应的解析解分别表示为：

$$\begin{aligned}\boldsymbol{W}_y &= (\boldsymbol{K}_y(\boldsymbol{B}_y^{'} + 2\lambda_1\boldsymbol{L}_y + 2\lambda_2\boldsymbol{I})\boldsymbol{K}_y^T)^{-1}\boldsymbol{K}_y\boldsymbol{A}\boldsymbol{K}_z^T\boldsymbol{W}_z.\\ \boldsymbol{W}_z &= (\boldsymbol{K}_z(\boldsymbol{B}_z^{'} + 2\lambda_1\boldsymbol{L}_z + 2\lambda_2\boldsymbol{I})\boldsymbol{K}_z^T)^{-1}\boldsymbol{K}_z\boldsymbol{A}\boldsymbol{K}_y^T\boldsymbol{W}_y.\end{aligned} \tag{4.38}$$

4.4 通用化

本节将分三小节分别介绍本章新提出的异质度量学习框架在不同的统计模型下的适配。对于欧氏数据，采用著名的正定核函数——径向基函数（Radial Basis Function, RBF）。形式化上，给定在欧氏空间上的一对点数据$\boldsymbol{x}_i, \boldsymbol{x}_j$，这一核函数可以定义为：

$$\boldsymbol{K}_x(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma_x^2) \tag{4.39}$$

其中，$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$是两个欧氏点$\boldsymbol{x}_i, \boldsymbol{x}_j$的经典欧氏距离。

针对黎曼数据，根据一些文献[55, 56, 58, 135, 139]所述，集合常用的统计模型一般位于一个特定的非欧氏空间上（即黎曼流形）上，公式4.39的核函数并不适用于统计模型数据。因此，需要将这一核函数推广到黎曼流形上。为了达到这一目的，给定在黎曼流形上的两个基本元素$\boldsymbol{Y}_i$和$\boldsymbol{Y}_j$，本节方法定义了在黎曼流形上的核函数：

$$\boldsymbol{K}_Y(\boldsymbol{y}_i, \boldsymbol{y}_j) = \exp(-d^2(\boldsymbol{y}_i, \boldsymbol{y}_j)/2\sigma_y^2) \tag{4.40}$$

很显然，这一核函数实际上是采用了公式4.39里的高斯核函数的形式。不同的是，这一核函数的最核心的一项$d^2(\boldsymbol{Y}_i, \boldsymbol{Y}_j)$定义了在黎曼流形上两个元素之间的距离。具体地，对于三种典型的统计模型（即线性子空间、仿射子空间和协方差矩阵），这一距离将分别在以下三个小节进行讨论。

4.4.1 线性子空间模型

根据大量研究[55, 56, 58, 60, 135]，在欧氏空间$\mathbb{R}^D$上的d维线性子空间是位于一个格拉斯曼（Grassmann）流形$\mathcal{G}(d, D)$上。也就是说，格拉斯曼流形$\mathcal{G}(d, D)$上的每个元素是一个线性子空间。由于每个线性子空间是由对应的正交基矩阵张成的，所以每个线性子空间可以对应一个正交基矩阵$\boldsymbol{U}$。具体地，这一正交基矩阵$\boldsymbol{U}$是由集合数据的协方

差矩阵对应的前d个主特征向量构成。因此对于格拉斯曼流形上的两个基本元素$\boldsymbol{Y}_i, \boldsymbol{Y}_j$，可以采用经典的投影度量[55, 56]来计算它们之间的距离：

$$d(\boldsymbol{y}_i, \boldsymbol{y}_j) = 2^{-1/2}\|\boldsymbol{U}_i\boldsymbol{U}_i^T - \boldsymbol{U}_j\boldsymbol{U}_j^T\|_F. \tag{4.41}$$

其中，$\|\cdot\|_F$ 表示矩阵弗罗贝尼乌斯范数（Frobenius norm）。

4.4.2　仿射子空间模型

与线性子空间相比，仿射子空间实际是一个带有位移量的线性子空间，也就是说它比线性子空间多考虑了均值的信息。根据研究[56]表明，仿射子空间是位于一个仿射格拉斯曼流形$\mathcal{AG}(d, D)$上。本章方法采用了文献[56]对两个位于$\mathcal{AG}(d, D)$上的元素的距离计算：

$$\begin{aligned} d(\boldsymbol{y}_i, \boldsymbol{y}_j) = 2^{-1/2}(&\|\boldsymbol{U}_i\boldsymbol{U}_i^T - \boldsymbol{U}_j\boldsymbol{U}_j^T\|_F \\ &+ \|(\boldsymbol{I} - \boldsymbol{U}_i\boldsymbol{U}_i^T)\boldsymbol{\mu}_i - (\boldsymbol{I} - \boldsymbol{U}_j\boldsymbol{U}_j^T)\boldsymbol{\mu}_j\|_F), \end{aligned} \tag{4.42}$$

其中，$\boldsymbol{I} \in \mathbb{R}^{D\times D}$ 是一个单位矩阵。

4.4.3　协方差矩阵模型

一些研究[29, 59, 135, 139]表明当协方差矩阵是非奇异矩阵时（通常是这一情况），它们其实是对称正定矩阵，因此通常位于一个特定的对称正定矩阵（Symmetric Positive Definite, SPD）流形上。为了定义公式4.40里的距离，可以采用经典的对数欧氏度量（Log-Euclidean Metric, LEM）[16]。形式化上，LEM可以表示成在矩阵对数域上的欧氏计算：

$$d(\boldsymbol{y}_i, \boldsymbol{y}_j) = \|\log(\boldsymbol{C}_i) - \log(\boldsymbol{C}_j)\|_F. \tag{4.43}$$

其中，$\log(\boldsymbol{C}) = \boldsymbol{U}\log(\boldsymbol{\Sigma})\boldsymbol{U}^T, \boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T$ 是协方差矩阵$\boldsymbol{C}$的特征值分解形式。

基于上述的三个基本黎曼度量（即公式4.41, 4.42, 4.43），公式4.40就可以生成一个在特定黎曼流形上的高斯核函数。然而，根据Mercer定理，只有满足正定条件的核函数才能生成有效的可再核希尔伯特空间。通过采用文献[70]里的证明方法（如下所示），这些定义在不同的黎曼流形上的核函数都可以很容易被证明是正定的。

定义1. *给定非空集合$\boldsymbol{\mathcal{Y}}$，一个函数$f : (\boldsymbol{\mathcal{Y}} \times \boldsymbol{\mathcal{Y}}) \to \mathbb{R}$被称为一个负定核当且当$f$是对称且*

$$\sum_{i,j=1}^{n} \gamma_i\gamma_j f(\boldsymbol{Y}_i, \boldsymbol{Y}_j) \leq 0 \tag{4.44}$$

对于所有$n \in \mathbb{N}, \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\} \in \boldsymbol{\mathcal{Y}}, \{\gamma_1, \ldots, \gamma_n\} \subseteq \mathbb{R}, \sum_{i,j=1}^{n} \gamma_i = 0.$

给定以上的定义，可以采用以下来自文献[113]的重要定理来证明所提出来的核函数是正定的。

定理1. 给定一个非空集合$\mathcal{Y}$和一个函数$f:(\mathcal{Y}\times\mathcal{Y})\to\mathbb{R}$。当且仅当$f$是负定的，那么核函数$exp(-tf(\boldsymbol{Y}_i,\boldsymbol{Y}_j))$对于所有$t>0$是正定的。

证明. 关于详细的证明，请读者参考文献[22]的第3章的定理2.2。

令$\mathcal{Y}$为非空集合, $\mathcal{H}$为内积空间。引进函数$\psi:\mathcal{Y}\to\mathcal{H}$使得$f(\boldsymbol{Y}_i,\boldsymbol{Y}_j)=\|\psi(\boldsymbol{Y}_i)-\psi(\boldsymbol{Y}_j)\|_F^2$, 只需要证明$f(\boldsymbol{Y}_i,\boldsymbol{Y}_j)$对**定理1**是负定的:

$$\begin{aligned}\sum_{i,j=1}^{n}\gamma_i\gamma_j f(\boldsymbol{Y}_i,\boldsymbol{Y}_j)&=\sum_{i,j=1}^{n}\gamma_i\gamma_j\|\psi(\boldsymbol{Y}_i)-\psi(\boldsymbol{Y}_j)\|_{\mathcal{H}}^2\\&=\sum_{j=1}^{n}\gamma_j\sum_{i=1}^{n}\gamma_i\langle\psi(\boldsymbol{Y}_i),\psi(\boldsymbol{Y}_i)\rangle_{\mathcal{H}}\\&\quad-2\sum_{i,j=1}^{n}\gamma_i\gamma_j\langle\psi(\boldsymbol{Y}_i),\psi(\boldsymbol{Y}_j)\rangle_{\mathcal{H}}\\&\quad+\sum_{i=1}^{n}\gamma_i\sum_{j=1}^{n}\gamma_j\langle\psi(\boldsymbol{Y}_j),\psi(\boldsymbol{Y}_j)\rangle_{\mathcal{H}}\\&=-2\sum_{i,j=1}^{n}\gamma_i\beta_j\langle\psi(\boldsymbol{Y}_i),\psi(\boldsymbol{Y}_j)\rangle_{\mathcal{H}}\\&=-2\|\sum_{i=1}^{n}\gamma_i\psi(\boldsymbol{Y}_i)\|_{\mathcal{H}}^2\leq 0\end{aligned}\tag{4.45}$$

根据以上证明，由于公式4.41, 4.42, 4.43均可以表示成公式4.45中$f(\boldsymbol{Y}_i,\boldsymbol{Y}_j)=\|\psi(\boldsymbol{Y}_i)-\psi(\boldsymbol{Y}_j)\|_F^2$的形式，所以这些定义在不同的黎曼流形上的核函数（公式4.40）都可以证明是正定的。关于更详细的证明过程，请读者参考文献[70]。

根据以上三种不同的统计模型对应的黎曼流形类型，本章新提出的CERML方法实际是一个能同时解决三种异质数据匹配实例：欧氏-格拉斯曼异质匹配（Euclidean-to-Grassmann, EG）、欧氏-仿射格拉斯曼异质匹配（Euclidean-to-AffineGrassmann, EA）以及欧氏-对称正定异质匹配（Euclidean-to-SPD, ES）。因此，CERML方法在这三种匹配实例下分别命名为：CERML-EG, CERML-EA和CERML-ES。

4.5 实验验证

本节分别通过三种不同的基于视频的人脸识别实验任务（即视频-图像、图像-视频和视频-视频的人脸识别）来评测所提出的CERML方法。

4.5.1 视频-图像/图像-视频人脸识别评测

关于视频-图像/图像-视频人脸识别任务，本节采用了本文在第二章新收集的COX和公开的YouTube Celebrities (YTC)[79]两个数据库。关于COX数据库，本节实

验严格遵从第二章在COX数据库上所设置的基于视频-图像/图像-视频两个场景下的评测协议，报告6组相关测试结果（每组在给定的10次随机测试数据集中进行评测，最终给出10次结果的均值和方差）。参照第三章对图像与视频帧的预处理操作，本节将所有的人脸图像和人脸视频帧统一归一化成48×60并对其作直方图均衡化，最后提取灰度特征。关于YTC数据库，由于原始数据库是为视频-视频的人脸识别任务而设计的，为了在它之上设计视频-图像与图像-视频的人脸识别场景，本节实验从这一数据库的训练视频里随机抽取视频帧当作静态图像。参照前人工作[135, 138, 139]在YTC上设计的视频-视频人脸识别评测协议，本节实验同样设计了10次交叉验证实验。在每次随机实验中，对每个人随机抽取3个视频或在其里面随机选中的静态图像作为目标集（target set），另外6个视频或其里面的被选中的静态图像作为查询集（query set）。与前人工作[135, 138, 139]对这一数据库所作的预处理相同，本节实验将所有人脸图像都归一化成20×20的灰度图像，然后作直方图像均衡化，最后提取灰度特征。

为了研究本章新提出的CERML方法在视频-图像/图像-视频人脸识别任务下的有效性，本节实验选取了以下几种代表性方法进行对比：

1. 基准方法：

   Nearest Neighborhood Classifier (NNC), Nearest Feature Subspace (NFS) [36], K-local Hyperplane (Convex) Distance Nearest Neighbor (HKNN, CKNN) [136]；

2. 同质度量学习方法:

   Neighbourhood Components Analysis (NCA) [47], Information-Theoretic Metric Learning (ITML) [40], Local Fisher Discriminant Analysis (LFDA) [124], Large Margin Nearest Neighbor (LMNN) [141]；

3. 异质度量/多视学习方法：

   Point-to-Set Distance Metric Learning (PSDML) [162], Kernel Partial Least Squares (KPLS) [116], Kernel Canonical Correlation Analysis (KCCA) [64] and Kernel Generalized Multiview Linear Discriminant Analysis (KGMA) [117];

4. 本章新提出的CERML方法。

在第一类方法中，NNC方法是最基本的最近邻分类器。具体地，在这一视频人脸识别场景下，NNC可以先将视频序列里的每一张人脸帧分别与静态人脸图像进行匹配，然后再在分数层进行融合。NFS/HKNN/CKNN方法是基于集合建模的最近邻分类器。具体地，在这一视频人脸识别的场景下，这些方法首先对视频序列进行建模，然后直接进行视频-图像/图像-视频分类。与NNC方法的处理方式相同，第二类方法也是先单独匹配视频人脸帧与静态人脸图像，然后再进行融合。这类经典的度量学习方

表 4.1. CERML对比方法在YTC和COX数据库上的视频-图像/图像-视频人脸识别结果（%）

Methods	YTC		COX					
	Video-Still	Still-Video	V1-S	V2-S	V3-S	S-V1	S-V2	S-V3
NNC	52.88±2.33	69.15±2.90	9.96±0.61	7.14±0.68	17.37±6.16	7.60±0.69	5.81±0.42	26.37±9.32
NFS [36]	53.27±2.75	60.21±5.99	9.99±1.17	5.90±0.92	22.23±1.60	11.64±0.81	6.51±0.57	31.67±1.48
HKNN [136]	36.94±2.71	48.01±4.80	4.70±0.33	3.70±0.44	12.70±1.00	6.34±0.50	4.64±0.52	20.41±0.80
CKNN [136]	54.45±3.30	68.94±2.19	7.93±0.56	5.47±0.52	14.83±1.06	8.89±0.75	5.67±0.67	26.24±0.82
NCA [47]	51.74±3.11	60.35±3.09	39.14±1.33	31.57±1.56	57.57±2.03	37.71±1.45	32.14±2.20	58.86±1.87
ITML [40]	47.62±1.73	59.72±4.27	19.83±1.62	18.20±1.80	36.63±2.69	26.66±1.74	25.21±2.10	47.57±2.87
LFDA [124]	57.83±2.67	73.12±2.87	21.41±1.77	22.17±1.77	43.99±1.74	40.54±1.32	33.90±1.40	61.40±1.45
LMNN [141]	55.02±2.71	70.99±3.25	34.44±1.02	30.03±1.36	58.06±1.35	37.84±1.79	35.77±1.37	63.33±2.06
PSDML [162]	55.30±1.90	61.21±5.24	12.14±1.04	9.43±1.41	25.43±1.29	7.04±0.60	4.14±0.52	29.86±1.69
KPLS[116]-**EG**	55.16± 2.92	65.32±3.38	21.83±1.58	18.50±1.53	30.89±2.70	15.01±1.17	12.41±0.86	25.63±1.69
KPLS[116]-**EA**	54.66±3.14	64.40±3.15	21.54±0.93	19.19±1.23	29.41±2.09	15.73±1.11	12.51±1.27	24.54±1.69
KPLS[116]-**ES**	54.02±2.61	64.89±5.18	20.21±2.03	16.21±1.45	27.23±1.80	14.83±1.93	11.61±1.10	23.99±2.43
KCCA[64]-**EG**	57.83±2.62	65.25±2.86	32.51±0.97	28.87±1.70	48.43±1.41	30.16±1.00	27.34±1.44	44.91±1.03
KCCA[64]-**EA**	57.40±2.92	64.82±2.86	30.33±1.17	28.39±1.30	47.74±1.92	28.49±1.12	26.49±1.01	45.21±1.40
KCCA[64]-**ES**	55.80±3.12	67.80±3.71	38.60±1.39	33.20±1.77	53.26±0.80	36.39±1.61	30.87±1.77	50.96±1.44
KGMA[117]-**EG**	65.77± 3.45	95.32±2.42	32.41±0.99	28.96±1.51	48.37±1.52	30.06±1.01	27.57±1.48	44.99±1.14
KGMA[117]-**EA**	64.77±3.03	95.11±2.25	30.60±1.17	28.34±1.40	47.74±1.65	28.54±1.30	26.20±1.03	45.27±1.41
KGMA[117]-**ES**	58.19±4.00	89.36±6.88	41.89±1.54	38.29±1.90	52.87±1.61	38.03±2.42	33.29±1.67	50.06±1.29
CERML-EG	**68.86±2.25**	**98.51±2.25**	**32.63±2.05**	**33.89±1.65**	**49.33±1.82**	**43.29±2.29**	**41.19±1.01**	**58.71±1.32**
CERML-EA	**67.30±3.50**	**96.81±1.81**	**38.77±1.72**	**37.57±1.12**	**53.93±1.56**	**43.93±2.16**	**41.56±1.22**	**57.34±1.27**
CERML-ES	**70.57±0.32**	**97.23±2.25**	**51.41±1.87**	**49.81±1.56**	**64.01±1.01**	**52.39±2.55**	**49.39±1.68**	**65.19±1.54**

法由于学习了在图像空间（或者欧氏空间）上的有效度量，从而可以在图像层面上进行有效地分类。在第三类方法中，除了PSDML方法，其它方法如经典的KCCA, KPLS，KGMA并不是特别为这一视频-图像/图像-视频人脸识别场景而设计的。为了适配这一场景，对于KPLS/KCCA/KGMA，实验将采用本章4.4小节里提出的基于不同统计模型的核函数（线性子空间：EG, 仿射子空间：EA，协方差：ES）。

为了保证对比实验的公正性，实验中对所有对比方法的关键参数的调整都是根据原始工作的建议进行的。具体地，对于HKNN方法，正则化参数λ设置为50。对于ITML方法，上界/下界距离阈值为训练数据的距离均值加上/减去其标准差，其它参数设置为默认值。对于LFDA方法，最近邻参数k设置成7。对于LMNN方法，近邻参数设置为5，最大迭代次数设为500，在训练中用于检验的比率为30%。对于PSDML方法，参数$\nu = 1, \lambda = 0.8$。这几个方法的参数均调整为让它们达到最好的性能。对于本章新提出的CERML方法，实验报告了它在三种不同统计模型（线性子空间：CERML-EG,

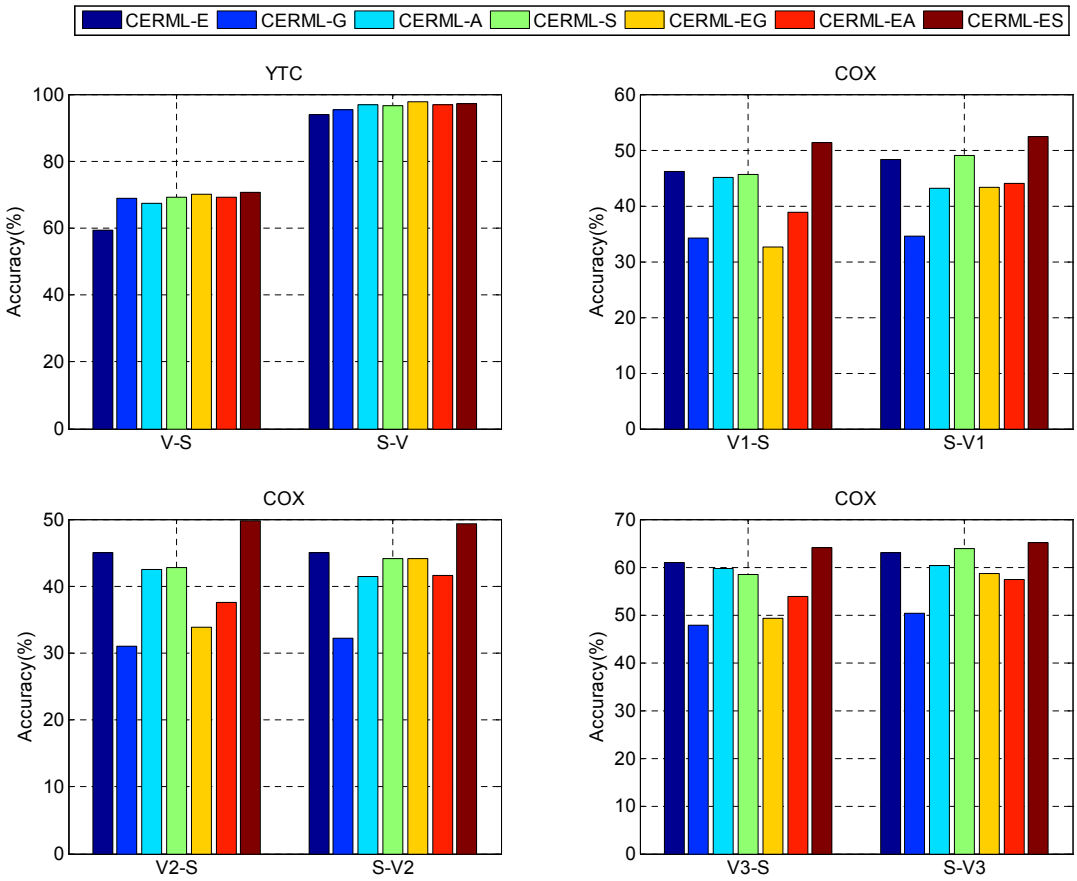


图 4.3. CERML方法在不同异质匹配实例下实现视频-图像/图像-视频人脸识别的性能对比。其中，7种不同的异质匹配实例分别为：CERML-E：图像-视频（均值）；CERML-G: 图像-视频（线性子空间）；CERML-A: 图像-视频（仿射子空间）；CERML-S: 图像-视频（协方差矩阵）；CERML-EG: 图像-视频（均值+线性子空间）；CERML-EA: 图像-视频（均值+仿射子空间）；CERML-ES: 图像-视频（均值+协方差）。

仿射子空间：CERML-EA，协方差：CERML-ES）的性能。CERML的几个主要参数分别设置为：$\lambda_1 = 0.01, \lambda_2 = 0.1, k_1 = 1, k_2 = 20$, 核宽度$\sigma$设置为训练数据的距离均值，算法求解迭代次数设置为30。

表4.1给出了实验对比的不同方法在YTC和COX两个数据库上视频-图像/图像-视频人脸识别的结果，其中首选识别率的均值与标准差是10次实验统计所得。从该表的对比情况分析，可以发现如下一些现象和结论：

（1）基准方法。除了HKNN方法，几个基准方法在两个数据库的性能相当。由于这些方法只是作为基本分类器，并没有通过已知训练集进行有监督学习，因此这类方法的性能要远差于其它几种对比方法。

（2）同质度量学习方法。从结果可以看出，这些方法中特别是LFDA和LMNN在大部分情况下均有效地提升了基准方法的性能。不过也有不尽如人意的地方，比如NCA和ITML方法在YTC上的性能比基准方法还要差一些。这其中可能的原因

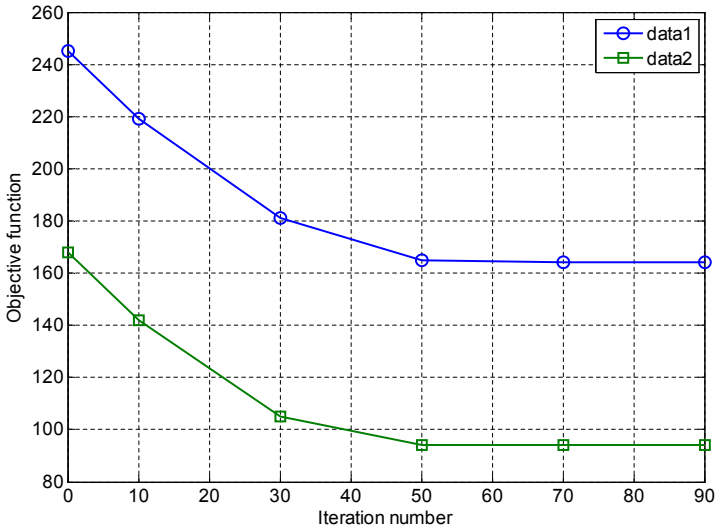


图 4.4. CERML方法在验证数据上迭代不同次数对应的目标函数值

是YTC里的视频一般包含了较大的姿态和场景变化，在这样复杂的场景下这一基于单样本的方法可能会显得更加敏感。

（3）异质度量/多视学习方法。与第二类的大部分方法相比，这一类方法特别是KGMA-ES方法在性能上有一定幅度的提升。这主要是因为这些方法都是基于子空间建模的基础上再进行了判别式的多视图学习。换句话说，这类方法有效挖掘和利用视频（图像集合）提供的整体信息，因此比较适用于这一识别任务。

（4）本章新提出的CERML方法。从表4.1中可以发现，CERML的三种不同统计建模实现在绝大多数情况均取得了明显优于对比方法的识别率。相比NCA和LMNN方法，CERML在COX的V3-S和S-V3上的提升幅度之所以相对比较小一些，是因为在V3视频集里人脸一般是正面人脸图像而且分辨率较高（详见第二章介绍），在这一相对比较可控的场景下，CERML就不能充分发挥它基于统计建模的度量学习的优势。

另外，除了表4.1给出了CERML方法与当前最优方法之间的对比，图4.3还展示了CERML方法在不同异质匹配实例下实现视频-图像/图像-视频人脸识别的性能对比。从图中结果可以看出，同时考虑一阶统计量（即均值）和二阶统计量（即线性子空间、协方差）的CERML实现要比只考虑单阶统计量的CERML实现的性能要高一些。而仿射子空间的实现之所以没有表现出与其它两种统计建模相同的规律主要是因为仿射子空间的计算本身已考虑了数据的均值信息。总的来说，这一实验结果验证了本章提出基于双阶统计量的异质度量学习的动机。

尽管本章无法对新提出的CERML方法的收敛性给出严格的理论证明，但是通过实验发现它在大多数情况是可以收敛到一个稳定的最优解。图4.4给出了CERML求解算法的两个迭代学习过程的例子。这一例子使用了两折的验证数据作实验。通过实验我们发现CERML的目标函数在迭代50次之后最终可以收敛到一个稳定的解上。

4.5.2　视频-视频人脸识别评测

本节采用了视频-视频人脸识别任务和视频-视频人脸确认任务来验证本章新提出的CERML方法的有效性。视频-视频的人脸识别的主要任务是给定来自查询集的一个查询视频，在一个由大量的视频构成的数据库（即目标集）中进行匹配，将最匹配的视频人脸身份赋予查询视频。这一任务通常以首选识别率（Rank-1 Recognition Rate）来作其评测指标。视频-视频的人脸确认实验主要是要判断给定的一对视频序列中的人脸是来自同一个人还是来不同人。这一任务通常以在某一特定的虚警率（Face Accept Rate, FAR）下的验证正确率（Verification Rate）来作为其评测指标。

为保证对比实验的充分性与广泛性，本节实验选用了两个具有不同特点的数据集COX和YouTube Celebrities (YTC) [79]来评测本章新提出的CERML方法在执行视频-视频的人脸识别任务时的有效性。针对视频-视频人脸确认任务，本节在两个极具挑战性的公开视频人脸数据库YouTube Face DB (YTF)[142]和Point-and-Shoot Challenge(PaSC)[24]上进行实验。关于COX，参照第三章的设置，本节同样将视频中的每张人脸帧都归一化成24×30大小的灰度图像，然后采用了直方图均衡化技术将归一化的图像进行光照预处理。在评测中，严格参照第二章在COX上所设置的协议，报告对比方法在10次测试中的平均性能和标准方差。关于YTC数据库，同样参照前人工作[94, 139]将所有视频人脸帧都归一化到20×20大小的图像，并且对其进行直方图均衡化然后提取灰度特征。按照前人工作[94, 139]对这一数据库设置的协议，本实验将在这一数据库上进行10组交叉验证实验。关于TTF数据库，参照前人工作[38]的设置，本节实验首先将该数据库的所有人脸都归一化到24×40大小的图像，然后对每张人脸图像提取灰度特征。按照这一数据库设置的标准协议[142]，本实验需要用给定的5,000视频对来评测人脸确认性能。关于PaSC数据库，与上一章相同，本节实验同样将执行Control和Handheld两种视频人脸验证，将图像归一化成256×256大小的图像，然后同样采用了最近比较优秀的深度卷积神经网络特征（Deep Covolutional Neural Network, DCNN）来提高底层图像的特征表示能力。具体地，本节实验利用了Caffe[74]工具对数据库里的人脸图像统一提取深度卷积神经网络特征。关于这四个数据库的更详细的评测设置，请参考第三章的介绍。

为了验证CERML在视频-视频人脸识别任务下的有效性，本节实验选取了以下几种代表性方法进行对比：

1. 基于线性子空间建模的方法：

   Discriminative Canonical Correlations (DCC) [81], Grassmann Discriminant Analysis (GDA) [55], Grassmannian Graph-Embedding Discriminant Analysis (GGDA) [58]；

2. 基于约束子空间建模的方法:

表 4.2. CERML对比方法在YTC和COX数据库上的视频-视频人脸识别结果（%）

Methods	YTC	COX					
		V2-V1	V3-V1	V3-V2	V1-V2	V1-V3	V2-V3
DCC [81]	68.85±2.32	62.53±6.13	66.10±3.52	50.56±4.21	56.09±11.27	53.84±11.37	45.19±9.81
GDA [55]	65.02±2.91	68.61±1.96	77.70±1.58	71.59±1.11	65.93±1.88	76.11±0.98	74.83±1.80
GGDA [58]	66.37±3.52	70.80±1.24	76.23±1.25	71.99±1.05	69.17±1.01	76.77±1.57	77.43±1.41
AHISD [32]	63.70±2.89	53.03±2.05	36.13±1.00	17.50±0.81	43.51±0.66	34.99±0.83	18.80±0.69
CHISD [32]	66.62±2.79	56.90±0.64	30.13±0.79	15.03±0.77	44.36±0.51	26.40±0.72	13.69±0.76
SSDML [162]	68.85±2.32	60.13±0.23	53.14±0.72	28.73±0.53	47.91±0.38	44.42±0.74	27.34±0.85
LMKML [94]	70.31±2.52	56.14±1.78	44.26±3.54	33.14±4.63	55.37±3.06	39.83±3.35	29.54±3.94
CDL [139]	69.72±2.92	78.43±1.01	85.31±0.97	79.71±1.47	75.56±1.95	85.84±0.86	81.87±1.14
CERML-EG	**68.08±3.78**	**87.59±1.07**	**92.41±0.98**	**88.54±1.30**	**83.21±1.05**	**92.09±0.55**	**91.16±1.19**
CERML-EA	**69.57±2.65**	**87.14±1.06**	**91.94±0.98**	**88.30±1.40**	**82.81±1.66**	**92.03±0.75**	**91.16±1.17**
CERML-ES	**72.38±2.48**	**90.31±0.62**	**94.83±0.41**	**91.51±0.83**	**87.06±0.98**	**95.13±0.79**	**93.89±0.70**

Affine Hull based Image Set Distance (AHISD) [32], Convex Hull based Image Set Distance (CHISD) [32]， Set-to-set distance metric learning (SSDML) [162]；

3. 基于协方差建模的方法：

   Localized Multi-Kernel Metric Learning (LMKML) [94], Covariance Discriminative Learning (CDL) [139];

4. 本章新提出的CERML方法。

由于第一类和第二类方法均在第三章作为对比方法进行过相关实验并且本节的实验同样采用了与第三章相同的设置，因此这里就不再具体重复介绍它们的设置情况。在第三类方法中，对于CDL，由于基于PLS的版本不能在COX的严格协议下工作，所以实验评测了CDL的LDA版本。为了处理CDL的协方差奇异问题，参照前人工作[135, 139] 的设置，实验对每个计算出来的协方差矩阵$\boldsymbol{C}$采用正则化技术：$\boldsymbol{C}^* = \boldsymbol{C} + \lambda \boldsymbol{I}$, 其中$\boldsymbol{I}$是单位矩阵，$\lambda = 10^{-3} \times trace(\boldsymbol{C})$。对于LMKML方法，实验采用了距离均值作为启发式信息来调整核宽度。

表4.2给出了实验对比的不同方法在YTC和COX两个数据库上的结果，其中首选识别率的均值与标准差是10次实验统计所得。另外，表4.3对比了这些方法在YTF和PaSC两个数据库的人脸确认率，其中在YTF上的是10次实验得到的平均人脸确认率和方差，在PaSC上报告的是FAR=0.01时的人脸确认率。对这两个表里的情况进行对比分析，可以发现如下一些现象和结论：

（1）基于线性子空间建模的方法。通过对比DCC，GDA和GGDA方法可以发现GDA和GGDA在YTC和YTF上要稍差于DCC，而在COX和PaSC上却明显优于DCC。

表 4.3. CERML对比方法在YTF和PaSC数据库上的视频-视频人脸确认结果(%)

Methods	YTF	PaSC	
		Control	Handheld
DCC [81]	68.28±2.21	38.87	37.53
GDA [55]	67.00±1.62	41.88	43.25
GGDA [58]	66.56±2.07	43.35	43.09
AHISD [32]	64.80±1.54	21.96	14.29
CHISD [32]	66.30±1.21	26.12	20.97
SSDML [162]	65.38±1.86	29.19	22.89
CDL [139]	64.94±2.38	42.62	42.97
CERML-EG	**69.42±3.25**	**46.00**	**45.49**
CERML-EA	**68.89±1.12**	**45.13**	**44.24**
CERML-ES	**68.36±1.97**	**45.40**	**44.47**

造成这一现象不仅有数据库层面的原因也有方法层面的原因。在数据库层面，这四个数据库里的视频拍摄条件既有区别也有联系：YTC和YTF都是在YouTube上收集的多种不同场景下拍摄的高压缩率低质量的视频序列，而COX和PaSC主要是在模拟某个特定监控场景下采集的一些低质量的视频序列。在方法层面，GDA和GGDA两个均将集合模型表示成某个特定黎曼流形上的元素并且将特征提取和分类两个任务放在同一个空间（即黎曼流形）上，而DCC方法主要采用不一致的策略：在欧氏空间上进行特征提取而最后分类却采用非欧氏距离。因此，GDA和GGDA通过训练更容易拟合某种具体的场景（即COX和PaSC），而DCC这一类方法对多种不同的场景（即YTC和YTF）更具有泛化能力。

（2）基于受限子空间建模的方法。相比AHISD和CHISD两种无监督的方法，SSDML方法由于学习出一个有效的度量，因此在实验中的大多情况表现出更高的识别性能。与DCC方法类似，SSDML同样是在欧氏空间上进行度量学习而最后分类是采用了一个非欧氏的基于仿射子空间的分类器，因此它同样在YTC和YTF上表现出比较好的性能，而在COX和PaSC上要明显差于其它有监督的方法。

（3）基于协方差建模的方法。由于LMKML度量学习方法时间复杂度非常高，本节只在规模相对比较小的YTC和COX上报告它的性能。在这两个库上，相比同样是基于协方差建模的CDL方法，这一方法在YTC上取得了稍好的性能，在COX却远逊于CDL。与上面的分析相同，LMKML采用了与DCC和SSDML相同的判别学习策略而CDL采用了与GDA和GGDA相同的策略，因此会有比较类似的结果。

（4）本章新提出的CERML方法。CERML通过采用异质数据（视频数据的一阶统计量和二阶统计量）的融合策略在几个数据库上一致地取得了要明显好于当前最优方法CDL的性能。同样，图4.5给出了CERML在使用单阶统计量和融合双阶统计量时的性

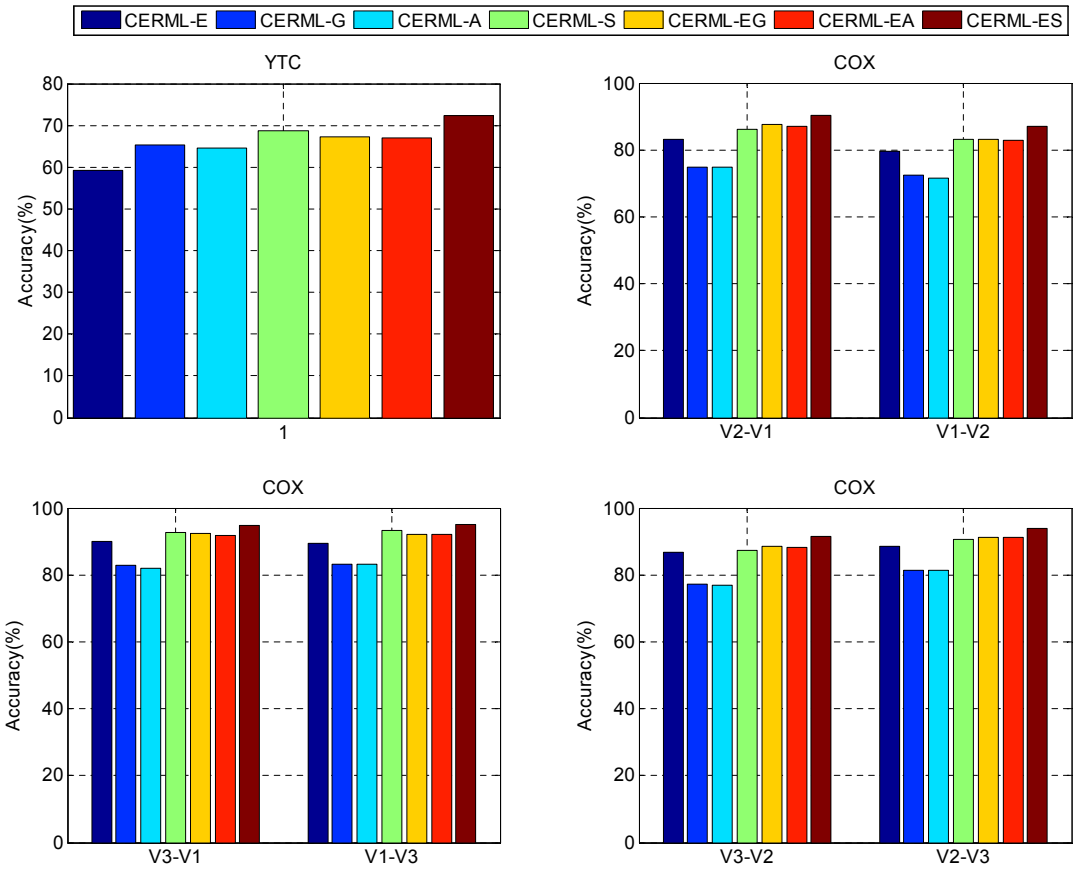


图 4.5. CERML方法在不同异质融合实例下实现视频-视频人脸识别的性能对比。其中，7种不同的异质融合实例分别为：CERML-E：视频（均值）；CERML-G: 视频（线性子空间）；CERML-A: 视频（仿射子空间）；CERML-S: 视频（协方差矩阵）；CERML-EG: 视频（均值+线性子空间）；CERML-EA:视频（均值+仿射子空间）；CERML-ES: 视频（均值+协方差）。

能对比。从图中可以观察到使用双阶统计量时的性能在绝大多数情况下要高于使用单阶统计量时的性能。

4.6 本章小结

本章提出一种基于双阶统计量建模的跨欧氏-黎曼度量学习框架来同时解决三种不同的基于视频的人脸识别问题，即视频-图像、图像-视频和视频-视频人脸识别。具体地，本章采用双阶统计量对视频数据进行建模，然后将这三种视频人脸识别问题统一形式化成异质数据的匹配/融合问题，最后提出一个新型的用于匹配/融合异质数据的异质度量学习统一框架。这一新的异质度量学习框架可以通用于不同的二阶统计量，如协方差、线性子空间和仿射子空间，因此成为一个可以包容多数统计模型的通用度量学习框架。本章通过大量的视频-图像/图像-视频和视频-视频人脸识别实验对提出的方法进行了验证。实验结果表明本章方法在四个极具挑战性的数据库上均取得了明显优于当前最好方法的性能。

由于本章设计了一个用于求解新度量学习目标函数的优化算法只能通过实验的定量分析来证明它可以收敛到一个稳定的解，因此在后续研究工作中可以考虑设计一个可理论证明收敛的求解算法。另外，本章的工作是基于双阶统计量建模而设计的一种度量学习方法，再探索更高阶的统计量或者将双阶统计量合成一个高斯分布模型是未来值得深入研究的方向。

第五章　基于高斯分布函数建模的对数欧氏度量学习方法

5.1　引言

针对基于视频的人脸识别问题，上一章提出的基于双阶统计量建模的方法主要是利用均值和协方差分别对每段视频的人脸数据进行一阶统计和二阶统计。如果要对数据进行一个完整的概率建模，高斯分布函数无疑是一个最自然的选择。根据著名的信息几何（Information geometry）理论[10, 93]，高斯分布函数的空间可以嵌入到一个由对称正定矩阵（Symmetric Positive Definite, SPD）张成的流形上。也就是说，可以把高斯模型转化成一个SPD矩阵。因此，本章主要从高斯分布函数建模出发，研究SPD流形上的黎曼度量学习方法。

如前面章节所述，大量的研究[15, 16, 29, 61, 63, 106, 120, 130–132, 135, 139]表明SPD矩阵通常不是位于向量空间上而是位于一个特定的黎曼流形上。由于没有考虑到对称正定矩阵所在流形特有的黎曼几何，先前在欧氏空间上定义或学习到的度量无法适用于这一类型的矩阵。而且，这些度量甚至还会引起一些不良的影响，比如扩散张量的膨胀效应（the swelling of diffusion tensors）和SPD矩阵求逆之后不满足矩阵对称属性。

为了解决这些问题，一些研究工作[15, 16, 106, 120]提出了为SPD流形而设计的黎曼度量。最为经典的当属由Pennec等人[106]提出的仿射不变度量（Affine-invariant metric, AIM）。对黎曼流形上的SPD矩阵施加以仿射不变度量可以使膨胀效应消失，还能保持SPD矩阵求逆后的对称属性。虽然仿射不变度量有着一系列非常优越的属性，但是由文献[16]指出，这一度量在实际应用中的成功常常以巨大的时间开销为代价。为了解决这一问题，Arsigny等人[16]引进了一种新的在SPD流形上的黎曼度量叫做对数欧氏度量（Log-Euclidean metric, LEM）。这一度量通过赋予SPD流形以一个李群结构（Lie group）并对其施加一个双不变度量（bi-invariant metic）从而将SPD流形归纳为一个平坦的黎曼空间（即在SPD流形的单位矩阵上的切空间）。在这一空间上，只需要计算SPD矩阵的矩阵对数的经典欧氏距离来度量在流形上的两个SPD矩阵的距离。因此，这一对数欧氏黎曼度量克服了仿射不变度量的高计算复杂度问题，而且在SPD流形上同时保持了良好的理论属性[16]。

基于这一对数欧氏度量，目前有相当多的工作[28, 70, 87, 99, 119, 128, 134, 135, 139]致力于研究在SPD流形上学习判别函数问题。比如，有些工作[28, 119, 128, 134]采用对数欧氏度量先将位于单位矩阵的切空间上的$d \times d$大小的SPD矩阵的矩阵对数转成了$\frac{d(d+1)}{2}$大小的向量形式，然后再对这一向量形式学习更加有判别性的l维的向量表示（请看图5.1(a)-(b1)-(c)）。另外一些方法主要是通过推导基于对数欧氏度量的核函数来

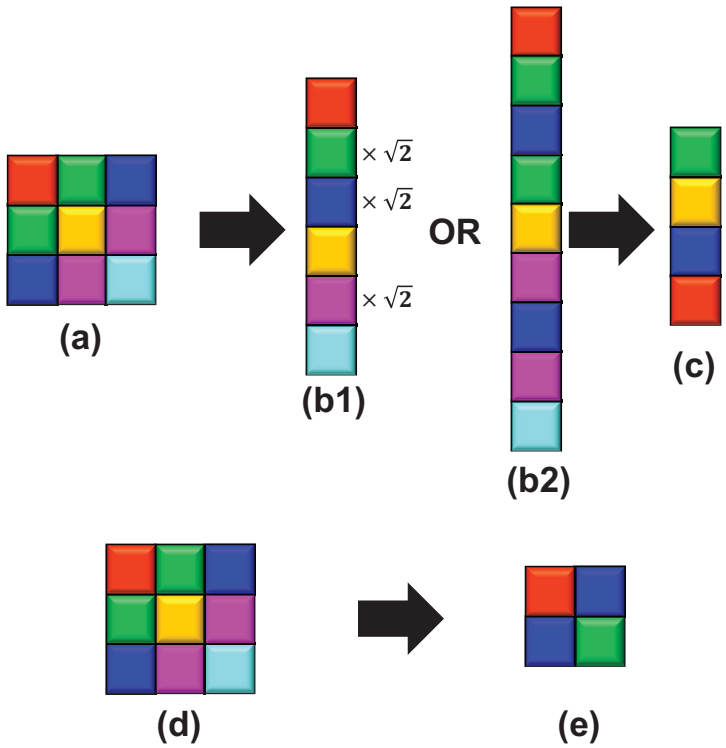


图 5.1. 在对数欧氏度量的框架下，两种处理位于单位矩阵处的切空间上的SPD矩阵对数的不同方案。绝大多数的传统方法一般要先将原始的$d \times d$的SPD矩阵对数(a)转化为一个$\frac{d(d+1)}{2}$维的向量(b1)或者一个d^2维的向量(b2)，最后在这一向量上学习一个l维的向量表示。相比之下，本章新提出的方法直接在原始的$d \times d$维的SPD矩阵对数(d)上学习一个$k \times k$维的SPD矩阵对数(e)。

将原始的SPD黎曼流形嵌入到一个高维的希尔伯特空间上[70, 135, 139]。实际上，这一类方法也是首先将位于单位矩阵的切空间上的$d \times d$大小的SPD矩阵的矩阵对数转成了d^2大小的向量形式，然后再学习更加有判别能力的l维的向量表示。总之，这两类方法所采用的相同技术是首先向量化SPD矩阵的矩阵对数然后再学习更有判别性的向量表示。然而，由于SPD矩阵对数并不是普通矩阵而是对称矩阵，因此这些方法的向量化操作不可避免地破坏了切空间的内部几何结构。

为了克服这一缺陷，本章提出了一种新的在SPD流形上的对数欧氏度量学习方法。如图5.1(d)-(e)所示，不像传统方法需要将SPD矩阵对数先转成向量形式，新提出的方法在对数欧氏度量的框架下直接对SPD矩阵对数学习更有判别能力的$k \times k$大小的对称矩阵（即SPD矩阵对数）。相比传统方法，新提出的方法采用这样的一个学习框架有以下两个主要的优点：1）新方法学到的从原始切空间到另一个更有判别性的SPD矩阵对数空间的映射可以更加如实地遵从原始切空间的几何结构，从而能保持原始切空间的优良的属性。除此之外，如果保持了SPD矩阵对数的对称属性就可以通过矩阵对数的逆操作（即指数操作，两者具有可微同胚映射属性[16]）将其从新学到的切空间反映射到一个有效的SPD流形上，从而可以得到一个更有判别性的SPD流形。2）直接在SPD矩

阵对数上学习判别函数比在SPD矩阵对数的向量形式上去学习判别函数更高效。为了更好地理解这一观点，请读者以PCA和2DPCA[151]作为对照。在传统的基于SPD的判别学习方法里，向量化的SPD矩阵对数的（类内和类间）散度矩阵的维数是$d^2 \times d^2$或者$\frac{d(d+1)}{2} \times \frac{d(d+1)}{2}$，而新方法所求的散度矩阵要小很多，只有$d \times d$的维数。因此，新方法能够更容易去准确估计在SPD矩阵对数上的散度矩阵，从而能够更高效地学习在其上的判别函数。基于这一动机，本章新提出的方法将在SPD矩阵对数上学习新度量任务形式化成了一个学习类马氏矩阵的问题，从而可以继承开发在向量空间上的传统度量学习方法的一些良好特性。

本章接下来的安排如下：5.1节简单介绍相关背景知识；5.3详细介绍新提出的对数欧氏度量学习框架，包括问题形式化、目标函数和优化算法；5.4节给出了新提出度量学习方法在四个具有挑战性的人脸数据库上与其它方法的对比结果；5.5节对本章新提出的对数欧氏度量学习方法进行了总结并分析未来的研究方向。

5.2　背景知识

本节首先介绍高斯分布函数的SPD矩阵形式，然后再对SPD矩阵流形进行描述，最后再简单回顾在SPD流形上的对数欧氏度量。

5.2.1　基于高斯分布函数的集合建模及其SPD矩阵形式

作为一个完整的概率模型，高斯分布函数同时包含了一阶统计量（即均值）和二阶统计量（即协方差）。因此，一个高斯分布函数可以同时刻画一个视频序列（也就是图像集合）数据的位置信息和变化模式。目前，已有一些工作[13, 114]采用高斯分布函数来建模视频序列或图像集合。从形式化上讲，给定一个视频序列/图像集合数据$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$，其中$\boldsymbol{x}_i \in \mathbb{R}^d$是第$i$个图像样本的$d$维特征向量表示，那么这一图像集合对应的高斯分布函数可以定义为$\mathcal{N}(\boldsymbol{m}, \tilde{\boldsymbol{C}})$，其中样本均值可计算成$\boldsymbol{m} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$，观察协方差$\tilde{\boldsymbol{C}}$对应的样本协方差可以由$\boldsymbol{C} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{m})(\boldsymbol{x}_i - \boldsymbol{m})^T$计算得到。

根据信息几何理论[10, 93]所述，高斯分布函数的空间可以嵌入到一个特定的SPD黎曼流形上。具体地，如果随机向量$\boldsymbol{x}$服从正态分布$\mathcal{N}(0, \boldsymbol{I})$，那么它的仿射变换$\boldsymbol{Q}\boldsymbol{x} + \tilde{\boldsymbol{m}}$就服从正态分布$\mathcal{N}(\boldsymbol{m}, \tilde{\boldsymbol{C}})$，其中观察协方差矩阵$\tilde{\boldsymbol{C}}$可以分解成$\tilde{\boldsymbol{C}} = \boldsymbol{Q}\boldsymbol{Q}^T, |\boldsymbol{Q}| > 0$（这里$|\boldsymbol{Q}|$表示矩阵$\boldsymbol{Q}$的行列式）。因此，一个高斯分斯布函数$\mathcal{N}(\boldsymbol{m}, \tilde{\boldsymbol{C}})$可以由一个仿射变换$(\tilde{\boldsymbol{m}}, \boldsymbol{Q})$来表示。根据工作[93]中提到的信息理论，通过一些推导，一个d维的高斯分布函数可以嵌入到一个特定的$(d+1)$维的SPD黎曼流形上，由此可以由唯一一个$(d+1) \times (d+1)$维的SPD矩阵$\boldsymbol{S}$表示：

$$\mathcal{N}(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{C}}) \sim \boldsymbol{S} = |\boldsymbol{Q}|^{-\frac{2}{d+1}} \begin{bmatrix} \boldsymbol{Q}\boldsymbol{Q}^T + \tilde{\boldsymbol{m}}\tilde{\boldsymbol{m}}^T & \tilde{\boldsymbol{m}} \\ \tilde{\boldsymbol{m}}^T & 1 \end{bmatrix} \tag{5.1}$$

关于高斯分布函数空间嵌入到SPD流形的更详细推导过程，请读者参考文献[93]。

5.2.2 SPD矩阵流形

大量研究工作[15, 16, 29, 61, 63, 106, 120, 130–132, 135, 139]表明，当$d \times d$的SPD矩阵张成的空间被赋予一个合适的黎曼度量时，这一空间就可以构成一个特定类型的黎曼流形，也就是所谓的SPD流形$\mathbb{S}_+^d$。SPD流形是一个局部同胚于欧氏空间而全局具有可微结构的拓扑空间。这一全局可微属性使得在此类型的黎曼流形上定义导数曲线成为可能。通过采用对数映射$\log_{\boldsymbol{S}_1} : \mathbb{S}_+^d \to T_{\boldsymbol{S}_1}\mathbb{S}_+^d, \boldsymbol{S}_1 \in \mathbb{S}_+^d$，在SPD流形上的点$\boldsymbol{S}_1$上的切线都位于对应的切空间上$T_{\boldsymbol{S}_1}\mathbb{S}_+^d$。这一切空间上定义了内积$\langle , \rangle_{\boldsymbol{S}_1}$。在所有切空间上定义的内积族就是所谓的流形上的黎曼度量。在配备了黎曼度量之后，在流形上的两点$\boldsymbol{S}_1, \boldsymbol{S}_2$之间的测地距离一般可计算为$\langle \log_{\boldsymbol{S}_1}(\boldsymbol{S}_2), \log_{\boldsymbol{S}_1}(\boldsymbol{S}_2) \rangle_{\boldsymbol{S}_1}$。

目前，有两个最广为应用的黎曼度量是仿射不变度量（Affine-Invariant Metric, AIM）和对数欧氏度量（Log-Euclidean Metric, LEM）。这两个度量之所以被研究者青睐主要是因为它们不仅定义了一种平滑变化的内积，而且还能够推导出SPD黎曼流形上的真实测地线。不过，由于SPD流形不是一个平坦空间，在实际应用中，仿射不变度量的计算复杂度是相当高的。相比之下，对数欧氏度量由于只需要在平坦的SPD矩阵对数空间进行欧氏计算而导致它的计算效率是相当高的。因此，本章主要研究在SPD流形上学习对数欧氏度量的问题。

5.2.3 SPD矩阵流形的对数欧氏度量

Arsigny等人[16]通过开发SPD流形$\mathbb{S}_+^d$ 的李群结构从而推导出了在SPD流形上的对数欧氏度量。李群结构对应的群操作是对于任意$\boldsymbol{S}_1, \boldsymbol{S}_2 \in \mathbb{S}_+^d$, $\boldsymbol{S}_1 \odot \boldsymbol{S}_2 := \exp(\log(\boldsymbol{S}_1) + \log(\boldsymbol{S}_2))$, 其中$\exp(\cdot)$和$\log(\cdot)$分别表示矩阵的指数操作和对数操作。

在SPD矩阵的李群上定义的对数欧氏度量对应于在SPD矩阵对数域上的欧氏度量。在赋予SPD流形$\mathbb{S}_+^d$以对数欧氏度量的框架下，在点$\boldsymbol{S}$上的切空间$T_{\boldsymbol{S}_1}\mathbb{S}_+^d$里的两个基本元素$\boldsymbol{T}_1, \boldsymbol{T}_2$的内积可计算为：

$$\langle \boldsymbol{T}_1, \boldsymbol{T}_2 \rangle_{\boldsymbol{S}} = \langle D_{\boldsymbol{S}} \log . \boldsymbol{T}_1, D_{\boldsymbol{S}} \log . \boldsymbol{T}_2 \rangle. \tag{5.2}$$

其中，$D_{\boldsymbol{S}} \log . \boldsymbol{T}$表示$\boldsymbol{S}$的矩阵对数沿着$\boldsymbol{T}$的方向导数（directional derivative）。与对数欧氏度量相关联的对数映射和指数映射分别定义为：

$$\begin{aligned} \log_{\boldsymbol{S}_1}(\boldsymbol{S}_2) &= D_{\log(\boldsymbol{S}_1)} \exp .(\log(\boldsymbol{S}_2) - \log(\boldsymbol{S}_1)), \\ \exp_{\boldsymbol{S}_1}(\boldsymbol{T}_2) &= \exp(\log(\boldsymbol{S}_1) + D_{\boldsymbol{S}_1} \log . \boldsymbol{T}_2)). \end{aligned} \tag{5.3}$$

其中，由于$\log \circ \exp = \boldsymbol{I}$ ($\boldsymbol{I}$为单位矩阵)，那么$D_{\log(\boldsymbol{S})} \exp . = (D_{\boldsymbol{S}} \log .)^{-1}$。关于更详细的对公式5.2和公式5.3的推导过程，请读者参考文献[16]。

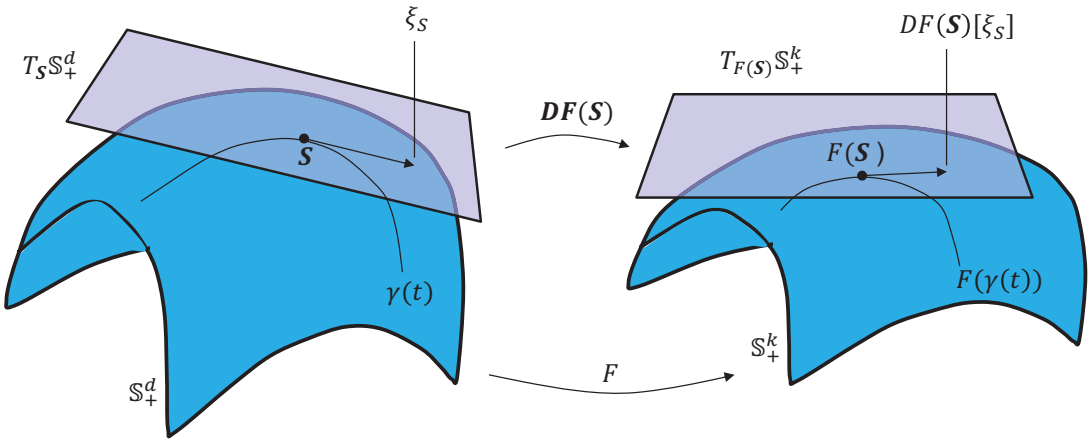


图 5.2. 本章新提出的在SPD矩阵流形上的对数欧氏度量学习方法（Log-Euclidean Metric Learning, LEML）的概念图。为了学习对数欧氏度量，LEML学习一个切射$DF(\boldsymbol{S}) : T_{\boldsymbol{S}}\mathbb{S}_+^d \rightarrow T_{F(\boldsymbol{S})}\mathbb{S}_+^k$将原始切空间$T_{\boldsymbol{S}}\mathbb{S}_+^d$变换到一个更有判别性的新切空间$T_{F(\boldsymbol{S})}\mathbb{S}_+^k$。通过这一映射，对应的映射$F : \mathbb{S}_+^d \rightarrow \mathbb{S}_+^k$是将原始SPD流形$\mathbb{S}_+^d$变换到一个目标SPD流形$\mathbb{S}_+^k$。这里，$\boldsymbol{\xi}_{\boldsymbol{S}}$表示在流形$\mathbb{S}_+^d$上的$\boldsymbol{S}$点的切向量，$\gamma$是$\boldsymbol{\xi}_{\boldsymbol{S}}$对应的测地线。

根据公式5.2和公式5.3，在SPD流形上的两个SPD矩阵之间的测地距离（即对数欧氏距离）可以计算为：

$$\begin{aligned}\mathcal{D}_{le}(\boldsymbol{S}_1, \boldsymbol{S}_2) &= \langle \log_{\boldsymbol{S}_1}(\boldsymbol{S}_2), \log_{\boldsymbol{S}_1}(\boldsymbol{S}_2) \rangle_{\boldsymbol{S}_1} \\ &= \| \log(\boldsymbol{S}_1) - \log(\boldsymbol{S}_2) \|_F^2.\end{aligned} \tag{5.4}$$

实际上，这一度量对应的是在矩阵对数域（也就是在单位矩阵上的切空间）上的欧氏距离。换句话说，在对数欧氏度量的框架下，在SPD流形上的两个元素之间的距离可以由在单位矩阵的切空间的内积计算得到。因此，通过采用对数欧氏度量，SPD矩阵流形就转化成了一个平坦的黎曼流形。

5.3　对数欧氏度量学习方法

本节将分三个小节来阐述新提出的对数欧氏度量学习（Log-Euclidean Metric Learning, LEML）方法。首先介绍切映射是如何将原始的切空间映射到另一个更有判别性的切空间上，然后再具体描述新提出的对数欧氏度量学习框架，最后提出一种优化算法来解决这一新的度量学习问题。

5.3.1　切映射

假设一组由5.2.1小节计算得到的基于SPD矩阵的特征表示为$\mathcal{S} = \{\boldsymbol{S}_1, \ldots, \boldsymbol{S}_n\}$，其中$\boldsymbol{S}_i \in \mathbb{S}_+^d$且其类别标签为$l_i$。将$F : \mathbb{S}_+^d \rightarrow \mathbb{S}_+^k$表示成从原始SPD流形$\mathbb{S}_+^d$到一个新的SPD流形$\mathbb{S}_+^k$ $(k \leq d)$的映射，$\boldsymbol{\xi}_{\boldsymbol{S}}$表示在流形$\mathbb{S}_+^d$上的$\boldsymbol{S}$点的切向量。在新的SPD流形$\mathbb{S}_+^k$上，可以由$F \circ \gamma$（其中，$\gamma$是$\boldsymbol{\xi}_{\boldsymbol{S}}$对应的测地线）计算得到切向量$DF(\boldsymbol{S})[\boldsymbol{\xi}_{\boldsymbol{S}}]$。如

图5.2所示，根据文献[7]所述，以下定义的映射：

$$DF(\boldsymbol{S}) : T_S\mathbb{S}_d^+ \rightarrow T_{F(\boldsymbol{S})}\mathbb{S}_k^+ : \boldsymbol{\xi} \mapsto DF(\boldsymbol{S})[\boldsymbol{\xi}_S]. \tag{5.5}$$

是一个线性变换，被称为F在点$\boldsymbol{S}$切空间$T_{\boldsymbol{S}}\mathbb{S}_+^d$上的切映射（或者可微映射）。由文献[7]指出，当且仅当切映射$DF(\boldsymbol{S}) : T_S\mathbb{S}_d^+ \rightarrow T_{F(\boldsymbol{S})}\mathbb{S}_k^+$对于每个点$\boldsymbol{S} \in \mathbb{S}_+^d$是一个单射（即一对一的映射），那么从流形到流形的映射$F : \mathbb{S}_+^d \rightarrow \mathbb{S}_+^k$是一个沉浸（immersion），也就是一个平滑可微分的单射。

公式5.4表明当给SPD流形配备对数欧氏度量，在流形上的两个SPD矩阵$\boldsymbol{S}_i, \boldsymbol{S}_j$之间的距离可以归纳为它们对应的位于单位矩阵切空间上的矩阵对数之间的欧氏距离。因此，对数欧氏度量框架跟切映射的数据属性是很相容的。基于这一思想，本章主要集中于学习具有单射属性的切映射$DF(\boldsymbol{S})$（由此也可以推导出对应的原始流形到新流形的映射$F(\boldsymbol{S})$）。具体地，首先对位于单位矩阵切空间上的矩阵对数施加以一个变换矩阵$\boldsymbol{W} \in \mathbb{R}^{d\times k}$，从而定义出了切映射的形式化：

$$f(\log(\boldsymbol{S})) = \boldsymbol{W}^T \log(\boldsymbol{S})\boldsymbol{W}. \tag{5.6}$$

为了确保由此切映射生成出的空间是在新流形$\mathbb{S}_+^k$上的切空间，变换$\boldsymbol{W}$需要满足列满秩的条件（即$rank(\boldsymbol{W}) = k$使得$f(\log(\boldsymbol{S}))$是个实对称矩阵从而可以构成一个有效的切空间（也就是SPD矩阵对数的空间））。显然，这一切映射$DF(\boldsymbol{S})$是个单射，从而也可以导出对应有效的流形-流形映射$F : \mathbb{S}_+^d \rightarrow \mathbb{S}_+^k$。

与这一直接对SPD矩阵对数进行操作的切映射相比，大部分基于对数欧氏度量的传统方法[28, 87, 119, 134, 139]需要首先将SPD矩阵对数转化成维数为d^2或者$\frac{d(d+1)}{2}$的向量形式，然后学习一个从这一向量空间到另一个向量空间来得到更有判别性的l维的向量表示。事实上，仅当l能分解成$\frac{k(k+1)}{2}$的形式（比如，$l = 3$, 那么$k = 2$）时，这些方法的学习程序才能生成一个有效的对称矩阵空间。因此，这些方法并总能推导出一个有效的切空间，从而因为不能获得切空间的的有用属性往往学到的是不符合原始数据结构的映射。

据调研，目前只有两个相关的工作[59, 76]研究流形-流形映射学习问题。然而，其中一个工作[76]主要学习的是一个从高维的球体到降维后的子流形的映射。而另一个工作[59] 采用的是两种不同类型的度量（即仿射不变度量和Stein散度度量）来学习从高维SPD矩阵流形到低维SPD矩阵流形的嵌入。在这一工作里，仿射不变度量由于其复杂度相当高而导致整个算法在实际应用中也是相当消耗时间，而Stein散度度量由于不能定义在SPD流形上的测地线而导致学到的散度并是最优的[70]。与这两个工作不同的是，本章新提出的方法目标并不只为了降维而是为了去学习一个类马氏矩阵（虽然类马氏矩阵可以被分解成变换矩阵），因此这一学习方式可以继承传统度量学习的一些良好属性。

5.3.2 度量学习

如前面章节所述，如果学到一个具有单射属性的切映射，一个新的SPD流形$\mathbb{S}_+^k$可以通过这一切映射推导出来。在对数欧氏度量的框架下，在新导出的SPD流形$\mathbb{S}_+^k$上的测地距离可以计算为：（这里为了下文更简便地描述，本节开始用$\boldsymbol{T}$表示$\log(\boldsymbol{S})$，另外还利用了矩阵迹在循环置换时具有的不变属性）：

$$\begin{aligned}\mathcal{D}_{\ell e}^{W}(\boldsymbol{T}_i,\boldsymbol{T}_j) &= \|\boldsymbol{W}^T\boldsymbol{T}_i\boldsymbol{W}-\boldsymbol{W}^T\boldsymbol{T}_j\boldsymbol{W}\|_F^2 \\ &= tr((\boldsymbol{T}_i-\boldsymbol{T}_j)^T\boldsymbol{P}(\boldsymbol{T}_i-\boldsymbol{T}_j)\boldsymbol{P}).\end{aligned} \tag{5.7}$$

其中，$\boldsymbol{P}=\boldsymbol{W}\boldsymbol{W}^T$是一个秩为$k$、维数为$d\times d$的对称半正定（Symmetric Positive Semidefininte, PSD）矩阵。由于$(\boldsymbol{T}_i-\boldsymbol{T}_j)$和$\boldsymbol{P}$都是对称的，那么它们在矩阵迹里进行任意的置换都是允许的[1]，因此我们将公式5.7重写为：

$$\mathcal{D}_{\ell e}^{Q}(\boldsymbol{T}_i,\boldsymbol{T}_j) = tr(\boldsymbol{Q}(\boldsymbol{T}_i-\boldsymbol{T}_j)(\boldsymbol{T}_i-\boldsymbol{T}_j)). \tag{5.8}$$

其中，$\boldsymbol{Q}=\boldsymbol{P}\boldsymbol{P}$。如果对$\boldsymbol{P}$进行奇异值分解$\boldsymbol{P}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$，那么$\boldsymbol{Q}=\boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^T$也是一个秩为$k$、维数为$d\times d$的PSD矩阵（类似于传统度量学习的马氏矩阵）。在实际应用中，通常设置$k=d$将矩阵$\boldsymbol{Q}$初始化为单位矩阵。

如果将公式5.8里的形式化$(\boldsymbol{T}_i-\boldsymbol{T}_j)(\boldsymbol{T}_i-\boldsymbol{T}_j)$看成是对应散度矩阵的逐对计算方式，那么可以发现这一散度矩阵只有$d\times d$大小。与之相比，传统方法[28, 87, 119, 134, 139]一般先将矩阵对数$\boldsymbol{T}_i$转化成大小为d^2或者$\frac{d(d+1)}{2}$的向量形式，然后在学习过程中计算出大小为$\frac{d(d+1)}{2}\times\frac{d(d+1)}{2}$或者$d^2\times d^2$的散度矩阵。由于这一散度矩阵要明显大于公式5.8里的散度矩阵$(\boldsymbol{T}_i-\boldsymbol{T}_j)(\boldsymbol{T}_i-\boldsymbol{T}_j)$，因此这些方法在更高维的空间里学习会低效很多。

由于本章主要研究如何在SPD流形上进行更有效的分类任务，与传统的度量学习方式类似，本节假设在新SPD流形上的样本对之间的距离是先验。具体地，先考虑关于样本对间的相似性或非相似性的约束：在新SPD流形$\mathbb{S}_+^k$上，如果两个样本对之间的测地距离小于一个给定的相对较小的值u即$\mathcal{D}_{\ell e}^{Q}(\boldsymbol{T}_i,\boldsymbol{T}_j)\leq u$，那么可以认为这两个样本是相似的。同理，如果两个样本对之间的测地距离大于一个给定的相对较大的值l即$\mathcal{D}_{\ell e}^{Q}(\boldsymbol{T}_i,\boldsymbol{T}_j)\geq l$，那么可以认为这两个样本是不相似的。

给定上述的距离约束集合，新度量学习问题就是要求解一个PSD矩阵$\boldsymbol{Q}$来参数化在新流形$\mathbb{S}_+^k$上对应的对数欧氏距离。借鉴工作[40]，本章采用了一种具有能够约束并能保持矩阵的秩的良好能力的LogDet散度函数来优化学习这一PSD矩阵$\boldsymbol{Q}$。具体地，本章在SPD流形上设计了一个基于LogDet散度（也称为Burg散度）的目标函数：

$$\begin{aligned}\min_{\boldsymbol{Q}\succeq 0,\boldsymbol{\xi}} \quad & \mathcal{D}_{\ell d}(\boldsymbol{Q},\boldsymbol{Q}_0)+\eta\mathcal{D}_{\ell d}(diag(\boldsymbol{\xi}),diag(\boldsymbol{\xi}_0)), \\ s.t. \quad & \boldsymbol{\delta}_{ij}\mathcal{D}_{\ell e}^{Q}(\boldsymbol{T}_i,\boldsymbol{T}_j)\leq\boldsymbol{\xi}_{ij},\forall c(i,j).\end{aligned} \tag{5.9}$$

[1]需要附加另一条件：$(\boldsymbol{T}_i-\boldsymbol{T}_j)\boldsymbol{P}$为对称矩阵

其中，$\boldsymbol{Q}_0 \in \mathbb{R}^{d\times d}$是矩阵$\boldsymbol{Q}$的初始化，$\mathcal{D}_{\ell d}(\boldsymbol{Q},\boldsymbol{Q}_0) = tr(\boldsymbol{Q}\boldsymbol{Q}_0^{-1}) - logdet(\boldsymbol{Q}\boldsymbol{Q}_0^{-1}) - d$，$d$为在原始SPD流形$\mathbb{S}_+^d$上的数据样本的维数。$\mathcal{D}_{\ell e}^{Q}(\boldsymbol{T}_i,\boldsymbol{T}_j)$表示在新生成的SPD流形$\mathbb{S}_+^k$上的切空间上的两个样本之间的距离，可以由公式5.8计算得到。$c(i,j)$表示关于在切空间上的两个样本$\boldsymbol{T}_i,\boldsymbol{T}_j$的约束。$\boldsymbol{\delta}_{ij} = 1$表示样本约束$c(i,j)$的样本$\boldsymbol{T}_i,\boldsymbol{T}_j$来自同一个类别，$\boldsymbol{\delta}_{ij} = -1$则反之。$\boldsymbol{\xi}_{ij}$是一个包含松驰变量的向量，初始化为$\boldsymbol{\delta}_{ij}\rho - \zeta\tau$，其中$\rho$ 是一个关于距离的约束，τ 是一个边界值，ζ 是一个用于调整边界的尺度。在实际应用中，ρ和τ分别设置为所有训练样本对的距离均值和方差。因此，在这一目标函数里只有两个参数η和ζ需要在实验中进行小范围的调整。

5.3.3 优化算法

为了求解公式5.9里的优化问题，本节采用了循环的Bregman投影算法[25, 30]。这一算法在每次迭代中选择一个距离约束并对其进行投影操作，使得当前解满足所选择的约束。在这一不等约束的实例里，需要对$\boldsymbol{\xi}_{ij}$和$\boldsymbol{Q}$进行调整。这一过程是由循环遍历给定的约束来重复进行。具体地，通过对公式5.9引入对偶变量$\boldsymbol{\alpha}_{ij}$以及采用公式5.8的结果，可以构成一个拉格朗日等式（Lagrangian equation）：$\mathcal{L} = \mathcal{D}_{\ell d}(\boldsymbol{Q}^{t+1},\boldsymbol{Q}^t) + \eta\mathcal{D}_{\ell d}(diag(\boldsymbol{\xi}^{t+1}),diag(\boldsymbol{\xi}^t)) + \boldsymbol{\alpha}_{ij}(\boldsymbol{\delta}_{ij}tr(\boldsymbol{Q}^{t+1}\boldsymbol{A}) - \boldsymbol{\xi}_{ij}^{t+1})$。对这一拉格朗日等式在$\boldsymbol{Q},\boldsymbol{\xi}_{ij}$和$\boldsymbol{\alpha}_{ij}$ 上求完梯度并设置为0可以得到以下三个等式：

$$\boldsymbol{Q}^{t+1} = \boldsymbol{V}_t((\boldsymbol{V}_t^T\boldsymbol{Q}^t\boldsymbol{V}_t)^{-1} - \boldsymbol{\delta}_{ij}\boldsymbol{\alpha}_{ij}(\boldsymbol{V}_t^T\boldsymbol{A}\boldsymbol{V}_t))^{-1}\boldsymbol{V}_t^T. \tag{5.10}$$

$$\boldsymbol{\xi}_{ij}^{t+1} = \frac{\eta\boldsymbol{\xi}_{ij}^t}{\eta + \boldsymbol{\delta}_{ij}\boldsymbol{\alpha}_{ij}\boldsymbol{\xi}_{ij}^t}. \tag{5.11}$$

$$0 = \boldsymbol{\delta}_{ij}tr(\boldsymbol{Q}^{t+1}\boldsymbol{A}) - \boldsymbol{\xi}_{ij}^{t+1}. \tag{5.12}$$

其中，$\boldsymbol{V}_t$是矩阵$\boldsymbol{Q}^t$的特征值分解后的特征向量，$\boldsymbol{T}_i = \log(\boldsymbol{S}_i),\boldsymbol{T}_j = \log(\boldsymbol{S}_j) \in T_I\mathbb{S}_+^d$是在单位矩阵切空间上的约束数据。

在公式5.10中，在PSD矩阵$\boldsymbol{Q}$上的投影更新规则是施加在约束矩阵$\boldsymbol{A} = (\boldsymbol{T}_i - \boldsymbol{T}_j)(\boldsymbol{T}_i - \boldsymbol{T}_j)$，其中$\boldsymbol{T}_i,\boldsymbol{T}_j$为对称矩阵。相比之下，工作[40, 83]里的在PSD矩阵上的投影更新规则是在一个秩为1的约束矩阵，也就是$\boldsymbol{A} = (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T$，其中$\boldsymbol{x}_i,\boldsymbol{x}_j$是数据向量。然而，借鉴工作[83]，这里同样可以采用Shermann-Morrison求逆公式[48, 118]$(B + uv^T)^{-1} = B^{-1} - \frac{B^{-1}uv^TB^{-1}}{1+v^TB^{-1}u}$ 对公式5.10再作进一步推导出在$\boldsymbol{Q}$上的最终更新准则：

$$\boldsymbol{Q}^{t+1} = \boldsymbol{Q}^t + \frac{\boldsymbol{\delta}_{ij}\boldsymbol{\alpha}_{ij}\boldsymbol{Q}^t\boldsymbol{A}\boldsymbol{Q}^t}{1 - \boldsymbol{\alpha}_{ij}tr(\boldsymbol{Q}^t\boldsymbol{A})}. \tag{5.13}$$

将公式5.13和公式5.11代入公式5.12中可以计算出$\boldsymbol{\alpha}_{ij}$的闭解：

$$\boldsymbol{\alpha}_{ij} = \frac{\boldsymbol{\delta}_{ij}\eta}{\eta+1}(\frac{1}{tr(\boldsymbol{Q}^t\boldsymbol{A})} - \frac{1}{\boldsymbol{\xi}_{ij}^t}). \tag{5.14}$$

算法 3 对数欧氏度量学习（LEML）算法

输入: 训练数据$\{\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_n\}, \boldsymbol{T}_i \in T_I\mathbb{S}_+^d$，约束$c(i,j)$及对应的$\boldsymbol{\delta}_{ij} \in \pm 1$初始半正定矩阵$\boldsymbol{Q}_0$，松弛变量$\eta$，距离阈值$\rho$，边界参数$\tau$以及尺度参数$\zeta$。

1. 初始化：$t \leftarrow 1, \boldsymbol{Q}^1 \leftarrow \boldsymbol{Q}_0, \boldsymbol{\xi}_{ij} \leftarrow \boldsymbol{\delta}_{ij}\rho - \zeta\tau, \boldsymbol{\lambda}_{ij} \leftarrow 0, \forall c(i,j)$。
2. **重复以下步骤:**
3. 利用公式5.14和公式5.15更新对偶变量$\boldsymbol{\lambda}_{ij}$。
4. 利用公式5.11更新松弛变量$\boldsymbol{\xi}_{ij}^{t+1}$。
5. 利用公式5.13更新矩阵$\boldsymbol{Q}^{t+1}$.
6. **直至收敛。**

输出: 半正定矩阵$\boldsymbol{Q}$。

由于公式5.9中包含了不等式约束，这里采用了$\boldsymbol{\lambda}_{ij} \geq 0$作为对应的对偶变量。为了保持对偶变量的非负性（这对于满足Karush-Kuhn-Tucker(KKT)条件是必要的），这里通过继续更新$\boldsymbol{\alpha}_{ij}$来求解最终的$\boldsymbol{\alpha}_{ij}$:

$$\boldsymbol{\alpha}_{ij} \leftarrow min(\boldsymbol{\alpha}_{ij}, \boldsymbol{\lambda}_{ij}), \quad \boldsymbol{\lambda}_{ij} \leftarrow \boldsymbol{\lambda}_{ij} - \boldsymbol{\alpha}_{ij}. \tag{5.15}$$

最终的优化程序形式化成算法3。算法的输入是约束数据，初始PSD矩阵$\boldsymbol{Q}_0$，松弛变量η，距离阈值ρ，边界参数τ以及尺度参数ζ。由工作[40, 83]指出，这一循环的Bregman投影算法是可以保证收敛到全局最优解。这一算法的主要时间成本在于公式5.13更新PSD矩阵$\boldsymbol{Q}^{t+1}$上，在每个约束上更新的时间复杂度为$O(d^2)$ (d是原始SPD流形的维数)。因此，主要时间复杂度为$O(Ld^2)$ ，其中L是这一算法在第5步更新$\boldsymbol{Q}^{t+1}$的次数。

5.4 实验验证

与前几章相同，本节将采用同样的一些视频人脸数据库来评测本章新提出的LEML方法。具体地，本节分三小节分别介绍数据库及其设置，评测方法及其设置以及最后的评测结果与实验分析。

5.4.1 实验数据库

本实验选取了与前几章对比实验中相同的数据库，即四个富有挑战性的数据数据库：本文新提出的COX, YouTube Celebrities (YTC) [79], YouTube Face DB (YTF) [142]和Point-and-Shoot Challenge (PaSC) [24]。由于四个视频人脸数据库在前几章已有详细介绍，本节将不再详细介绍这四个数据库。

实验中针对各数据库的训练集/测试集的划分方案与前几章的设置基本相同。针对COX/YTC/YTF数据库，分别进行十组交叉验证，每组实验对应于一组随机选择的训

练集与测试集标号组合。针对PaSC数据库，实验严格遵守原始设计的协议来执行可控（Control）实验和手持（Handheld）实验。

5.4.2 对比方法与参数设置

在这四个数据库上，本节根据5.2.1小节介绍的基于高斯分布函数的统计建模方法来表示视频序列中的人脸图像集合。为了防止高斯模型里的二阶统计量即协方差矩阵$\boldsymbol{C}$奇异，本节统一对每个协方差矩阵$\boldsymbol{C}$加上一个小的扰动$\kappa\boldsymbol{I}$，其中$\kappa = e^{-3} \times \boldsymbol{C}$，$\boldsymbol{C}$是单位矩阵，最后根据公式5.1将高斯分布函数转成一个SPD矩阵。具体地，可以拿YTC数据库作举例，在这一数据库中的每个图像归一成20×20大小的灰度图像，从而可以对每个视频序列的人脸图像集合计算一个400维的样本均值和一个大小为400×400的协方差矩阵，最后可以构成一个401×401维的SPD矩阵。

由于本章提出一个新的在SPD流形上的对数欧氏度量学习方法LEML，所以本节实验中将选取以下三类基于SPD流形度量的基准方法和判别学习方法进行对比：

1. SPD流形上的基准度量：

   Affine-Invariant Metric (AIM) [106], Stein divergence [120], Log-Euclidean Metric (LEM) [16]；

2. 基于AIM和Stein的判别学习方法:

   SPD Manifold Leaning (SPDML-AIM, SPDML-Stein) [59], Riemannian Sparse Representation (RSR) [57]；

3. 基于LEM的判别学习方法：

   Log-Euclidean Kernel (LEK) [87], Covariance Discriminative Learning (CDL) [139], Information-Theoretic Metric Learning with LEM (ITML-LEM) [134]；

4. 本章新提出的LEML方法。

为了保证对比实验的公正性，本节的实验根据这些方法的原始工作的建议对它们的相关参数进行如下调整和设置：针对SPDML方法，实验根据原始工作[59]将类内最近邻个数v_w设置为每类的样本数，通过交叉验证的方式来设置目标流形的维数和类间最近邻个数v_b的值。针对RSR方法，实验在训练样本对间的距离均值范围内调整方法的高斯核宽度参数β，在$\{0.0001, 0.001, 0.01, 0.1\}$范围内调整稀疏表示分类器的正则化参数$\lambda$。针对LEK方法，实验实现了原始工作里的三个版本，也就是多项式核、指数核和高斯核（分别命名为LEK-K_p, LEK-K_e, LEK-K_g）。在LEK-K_p, LEK-K_e里的参数n在1到50的范围内调整，在LEK-K_g里的高斯核宽度参数β与方法RSR的调参策略相同。在三个LEK版本里，稀疏表示分类器的正则化参数λ均

表 5.1. LEML对比方法在YTC和COX数据库上的视频-视频人脸识别结果（%）

Methods	YTC	COX					
		V2-V1	V3-V1	V3-V2	V1-V2	V1-V3	V2-V3
AIM [106]	62.85±3.46	45.53±0.46	21.47±1.87	11.00±0.85	39.83±0.67	19.36±0.67	9.50±0.67
Stein [120]	61.46±3.53	50.17±1.06	37.30±1.12	19.01±0.81	34.06±1.84	29.99±0.96	18.76±0.76
LEM [16]	63.91±3.25	51.33±1.07	42.69±1.68	18.71±1.01	40.04±1.55	32.04±1.05	16.77±0.47
SPDML-AIM [59]	64.66±2.92	N/A	N/A	N/A	N/A	N/A	N/A
SPDML-Stein [59]	61.57±3.43	N/A	N/A	N/A	N/A	N/A	N/A
RSR [57]	72.77±2.69	N/A	N/A	N/A	N/A	N/A	N/A
LEK-K_p [87]	61.85±3.24	N/A	N/A	N/A	N/A	N/A	N/A
LEK-K_e [87]	62.17±3.52	N/A	N/A	N/A	N/A	N/A	N/A
LEK-K_g [87]	56.30±3.62	N/A	N/A	N/A	N/A	N/A	N/A
CDL-LDA [139]	72.70±2.93	83.27±1.08	90.14±0.73	83.51±1.41	83.31±1.07	90.86±1.08	84.47±0.72
CDL-PLS [139]	72.67±2.47	N/A	N/A	N/A	N/A	N/A	N/A
ITML-LEM [134]	66.51±3.67	57.29±2.00	66.81±2.69	59.67±1.01	54.63±1.29	61.84±3.10	59.41±1.83
LEML	**70.53±2.95**	**62.85±3.46**	**61.90±1.96**	**54.73±1.57**	**54.34±1.37**	**67.56±2.57**	**61.34±1.90**
LEML-CDL-LDA	**72.63±2.49**	**84.11±1.28**	**90.84±0.98**	**84.09±1.14**	**84.87±1.31**	**89.70±1.08**	**85.46±1.41**
LEML-CDL-PLS	**73.31±2.49**	N/A	N/A	N/A	N/A	N/A	N/A

在$\{0.0001, 0.001, 0.01, 0.1\}$范围内调整。针对CDL方法，实验采用了与原始工作[139]相同的设置实现了CDL的LDA版本和PLS版本。针对ITML-LEM，实验根据原始ITML工作[40]对其进行调参。对于本章提出的方法LEML，正则化参数η在$\{0.1, 1, 10\}$进行调整，尺度参数ζ在$0.1:0.1:0.5$范围内进行调整，要学习的类马氏矩阵$\boldsymbol{Q}_0$初始化为单位矩阵。

5.4.3　视频-视频人脸识别与确认评测

表5.1和表5.2给出了实验对比的不同方法在四个数据库上的视频-视频人脸识别以及视频-视频人脸确认结果，在YTC/COX/YTF中的识别率/确认率的均值与标准差是通过10次实验统计所得。对于在YTF和PaSC上的视频人脸确认任务，由于RSR和LEK方法采用的是基于稀疏表示的分类器，CDL-PLS采用的是基于回归的分类器，这三种方法都无法在这两个数据库上工作。另外，表5.3给出了对比方法在一台Inter(R) Core(TM) i7-37700M(3.40GHZ)的PC上运行时的时间对比。如表5.3所示，由于SPDML方法的计算复杂度非常之高所以本实验没有报告它在COX和PaSC两个大规模数据库上的性能。从这些表中报告的结果可以得出如下的一些分析和结论：

（1）从表5.1和表5.2中发现，在几个基于SPD矩阵的基准度量方法中，LEM在大部分情况下要好于其它两个基准方法AIM和Stein。这一结果表明了LEM在实际应用中要优于其它两种度量。从表5.3中可以发现基于LEM的CDL/LEK/ITML-LEM在测试时要

表 5.2. LEML对比方法在YTF和PaSC数据库上的视频-视频人脸确认结果(%)

Methods	YTF	PaSC	
		Control	Handheld
AIM [106]	59.28±2.25	28.23	21.13
Stein [120]	58.70±1.97	26.91	20.18
LEM [16]	61.48±2.27	31.04	25.67
SPDML-AIM [59]	62.16±2.16	N/A	N/A
SPDML-Stein [59]	62.56±2.49	N/A	N/A
RSR [57]	N/A	N/A	N/A
LEK-K_p [87]	N/A	N/A	N/A
LEK-K_e [87]	N/A	N/A	N/A
LEK-K_g [87]	N/A	N/A	N/A
CDL-LDA [139]	66.76±1.89	45.24	45.56
CDL-PLS [139]	N/A	N/A	N/A
ITML-LEM [134]	60.02±1.84	36.25	35.19
LEML	**65.12±1.54**	**37.71**	**36.45**
LEML-CDL-LDA	**72.34±2.07**	**47.25**	**46.64**
LEML-CDL-PLS	N/A	N/A	N/A

远快于基于AIM/Stein的SPDML/RSR方法。总之，这些结果表明了度量LEM在实际应用中不仅在性能上而且上效率上都要优于AIM和Stein。

（2）相比于SPD流形的基准黎曼度量LEM，表5.1和表5.2的实验结果验证了新提出的度量学习方法LEML在基于LEM的框架下在四个视频数据库上都可以极大提升LEM的性能。

（3）相比于其它基于SPD流形的判别学习方法，LEML取得与它们可比甚至在有些情况要好于它们的性能。在这些对比方法中，SPDML和ITML-LEM与新提出的LEML方法最相关。同样作为一种学习流形-流形映射的方法，SPDML方法在四个数据库的表现要逊于LEML。这主要是因为SPDML方法是基于AIM和Stein框架设计的流形学习方法，而LEML是在基于LEM框架下学习度量。正如（1）中反映的结论，由于LEM要优于AIM和Stein，所以LEML可以取得更好的结果。另外，同样作为以LEM为基准的度量学习方法，ITML-LEM在向量空间中学习度量，而LEML在矩阵空间中学习度量。从实验结果发现，LEML在很多情况下表现要好于ITML-LEM。这里罗列了两个可能的原因：1）LEML能够更加如实地遵守原始切空间的几何，从而能够获得来自切空间上的良好属性。2）与ITML-LEM相比，LEML在一个更小的散度矩阵上学习判别函数，从而能在SPD流形上学习到更加满意的距离度量。

（4）由于LEML最终学到的类马氏矩阵可以转化成变换矩阵从而可以对原始SPD流形（或其上的切空间）进行降维，因此它可以作为其它基于SPD流形的判

表 5.3. LEML对比方法在YTC数据库上的计算时间对比（单位：秒）

Methods	SPDML-AIM	SPDML-Stein	RSR	LEK-K_p	LEK-K_e	LEK-K_g	CDL-LDA	CDL-PLS	ITML-LEM	**LEML**
Train	15072.56	108.50	10.91	0.92	0.90	0.85	5.30	12.08	92007.13	56.30
Test	9.35	0.04	0.08	0.01	0.01	0.01	0.03	0.05	0.02	0.02

别学习方法的一个预处理操作（如PCA在向量数据分析上的作用相当）。从实验结果发现LEML与CDL进行耦合（即LEML-CDL-LDA/LEML-CDL-PLS）能在一定程度上（特别是在YTF数据库上）提高原始CDL的性能。这一结果表明了LEML可以成为一种有利于后端分类器的SPD流形降维方法。

（5）从表5.3中的时间对比发现，与两个最相关的方法SPDML和ITML-LEM相比，LEML方法在YTC数据库上的训练和测试速度都要快很多。导致这一结果的主要原因是SPDML采用的两种度量框架AIM和Stein相比LEML使用的LEM计算时间开销上要大很多，而ITML-LEM由于在高维的SPD矩阵对数的向量形式上进行学习时间复杂度明显要高于LEML直接在低维的SPD矩阵对数进行度量学习时的时间复杂度。不过，从表中发现相比其它方法，LEML在训练效率方面要相对低一些。这主要原因是, CDL/LEK等方法主要采用计算复杂度较低的解析求解算法，而LEML所采用的循环迭代优化算法通常要迭代10,000次以上，因此在实际应用中其计算复杂度相对比较高。

5.5　本章小结

本章采用高斯分布函数模型对视频序列的人脸图像集合进行建模，借鉴著名的信息几何理论将高斯分布函数模型的空间嵌入到一个SPD流形上，从而将每个高斯模型转化成一个SPD矩阵。为了能够对高斯模型进行有效地分类，本章提出了一种在SPD矩阵流形上的对数欧氏度量学习方法。不同于大多数基于SPD流形的判别学习方法，新提出的方法本质上是试图学习一个从SPD流形的原始切空间到新切空间的映射来直接对SPD矩阵对数进行判别特征学习。本章新提出的方法的有效性通过系统的实验对比进行了验证，在视频人脸识别和视频人脸确认两个任务上，本章新方法均表现出良好的性能。与同类型的基于SPD流形的分类方法相比，本章新方法取得了与当前最优方法相当甚至更优的性能，从而表明其可以有效地解决视频人脸识别问题。

针对算法的进一步研究，其中一个可能的改进方案是可以将目前用于求解最优类马氏矩阵（即PSD矩阵）的基于LogDet散度的目标函数改成如第三章中所采用的在PSD流形上求解最优PSD矩阵的黎曼共轭梯度求解函数。针对算法中采用的SPD流形上的距离度量LEM，将其它两种经典的度量AIM和Stein包容在这一度量学习的框架内也是值得尝试的一个改进方向。

第六章　基于多统计模型的混合欧氏-黎曼度量学习方法

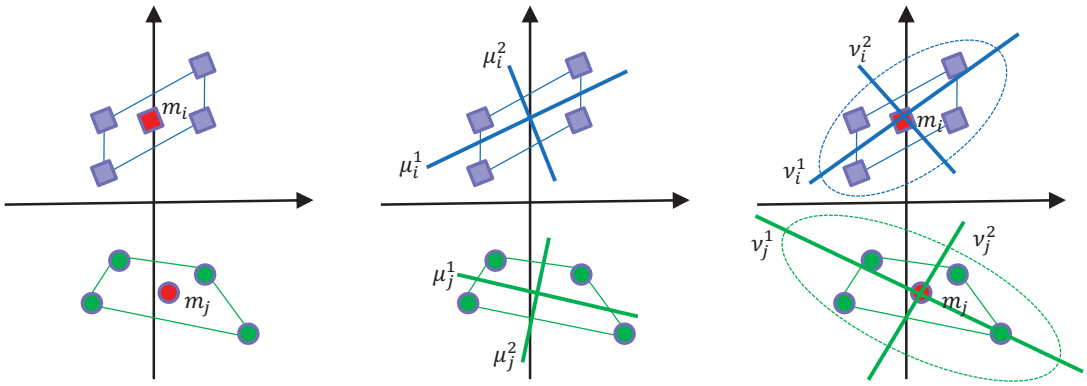
6.1　引言

在真实世界里，由于摄像机的拍摄角度和外界光照条件等非可控因素的变化，人们所采集的视频序列中的人脸表观通常包含了极其复杂的变化模式。因此，在真实场景下的视频序列中的数据通常是服从任意分布的，从而导致了使用单一的统计模型是无法准确刻画出视频数据的分布与结构。如图6.1所示，样本均值通过对集合中的数据求平均可以描述数据的位置信息，而样本协方差通过计算实例样本与样本均值之间的方差刻画了集合中数据的变化模式。很显然，如果单独采用这两种统计模型中的一种只能描述集合数据的某一方面的信息。作为完整的概率分布模型，高斯分布函数通过估计样本均值和观察协方差可以同时统计集合数据的位置和变化信息。值得注意的是，在这一统计模型里，如文献[14, 115, 126]所述，观察协方差是样本协方差的一个最大似然估计，因此可以较松弛地刻画集合中样本的变化模式。虽然高斯分布函数模型有这一良好的特征，但是它总是假设数据是服从高斯分布的。而在现实情况下，有一大部分的数据是不服从这一特定假设的。相比之下，样本均值和样本协方差没有作任何分布的假设。因此，当数据服从高斯分布时，单独采用高斯模型可以描述数据的完整信息；当数据是非高斯分布时，融合样本均值和样本协方差无疑是更好的选择。为了同时考虑这些情况，本章采用了样本均值、样本协方差和高斯分布函数三种统计模型的融合来提高对视频数据的建模能力。

为了更好地融合以上三种统计模型，本章提出一种混合欧氏-黎曼度量学习框架（Hybrid Euclidean-and-Riemannian Metric Learning, HERML）。在HERML框架里，样本均值表示为一维向量位于欧氏空间上，样本协方差通常可以表示成二维SPD矩阵位于SPD流形上（详见第四章内容），高斯分布函数模型可以转化为二维的SPD矩阵位于另一个高一维的SPD流形上（详见第五章内容）。借鉴第四章的异质度量学习机制，本章首先推导出三种异质空间上的核函数将这三种异质数据分别映射到一个高维的可再生核希尔伯特空间上，然后在这三个不同的希尔伯特空间上进行马氏度量学习。与前人工作相比，本章提出的HERML框架主要有如下几点贡献：

（1）本章通过开发样本均值、样本协方差和高斯分布函数模型的互补性将三者同时用于视频序列的统计建模，从而可以更加鲁棒地识别视频中的人脸。

（2）为了减少高斯分布函数模型所在的空间与其它两种统计模型的空间（均值空间为欧氏空间，协方差空间为SPD流形）之间的异质性，与第五章相同，本章借鉴经典的信息几何（Information geometry）理论[10, 93]，将高斯分布的空间也嵌入到一个特定维的SPD流形上。



(a) 均值　　(b) 协方差　　(c) 高斯

图 6.1. 样本均值、样本协方差和高斯分布函数表示集合数据。(a) 样本均值$\boldsymbol{m}_i/\boldsymbol{m}_j$刻画集合数据的位置信息；(b) 样本协方差不同的方向$\boldsymbol{\mu}_i^1, \boldsymbol{\mu}_i^2/\boldsymbol{\mu}_j^1, \boldsymbol{\mu}_j^2$编码集合数据的变化模式；(c) 高斯分布函数模型同时考虑观察均值和观察协方差，其中观察协方差的主要方向$\boldsymbol{v}_i^1, \boldsymbol{v}_i^2/\boldsymbol{v}_j^1, \boldsymbol{v}_j^2$较松弛地刻画集合数据的变化模式。

（3）本章提出一个混合度量学习框架通过在高维的希尔伯特空间上学习三个马氏矩阵将这三种虽互补但异质的统计模型数据进行融合。为了生成有效的高维希尔伯特空间，本章采用了与第四章推导的高斯核函数相对应的线性核函数来定义在均值和协方差上的核空间，而对于高斯模型（如混合高斯模型），本章通过借鉴基于KL散度的核估计理论以及SPD流形的黎曼几何提出了一种新的核函数从而在其上生成一个有效的核空间。

（4）本章HERML框架在四个挑战性的基于视频的人脸数据库上进行了系统的实验对比及验证。实验结果表明，HERML在这四个数据库上均达到了当前最好的性能。

本章接下来的安排如下：6.2小节简单介绍本章新方法的相关工作；6.3小节详细介绍本章新提出的混合欧氏-黎曼度量学习框架，包括问题概述、多统计模型计算与空间嵌入、多统计模型混合度量学习以及所采用的优化算法；6.4小节给出新提出方法在四个公开人脸数据库上与其它方法的对比结果；6.5小节对本章新提出的方法进行了总结并对其未来工作作进一步的讨论。

6.2 相关工作

为了学习混合异质度量，本章的新方法开发了基于LogDet散度（也称为Burg散度）的目标函数。由于著名的信息理论度量学习方法（Information-Theoretic Metric Learning, ITML）[40]同样采用了LogDet散度，因此本节先介绍一下这一方法。另外，本章的新方法为了融合多种统计模型数据提出了一个混合度量学习框架，因此本节也将回顾一下传统的多核学习方法（Multiple Kernel Learning）[38, 69, 94, 97, 111, 135, 147]。

6.2.1 信息理论度量学习方法

经典的信息理论度量学习方法（Information-Theoretic Metric Learning, ITML）[40]将度量学习问题形式化成了一个基于LogDet散度的优化函数来最小化基于线性约束的LogDet散度：

$$\begin{aligned} \underset{\boldsymbol{A}\succeq 0,\boldsymbol{\xi}}{min} \quad & D_{\ell d}(\boldsymbol{A},\boldsymbol{A}_0)+\gamma D_{\ell d}(diag(\boldsymbol{\xi}),diag(\boldsymbol{\xi}_0)) \\ s.t. \quad & tr(\boldsymbol{A}(\boldsymbol{x}_i-x_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T)\le \boldsymbol{\xi}_{ij}, \quad c(i,j)\in S \\ & tr(\boldsymbol{A}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T)\ge \boldsymbol{\xi}_{ij}, \quad c(i,j)\in D \end{aligned} \tag{6.1}$$

其中，$\boldsymbol{A},\boldsymbol{A}_0\in\mathbb{R}^{d\times d}$, $D_{\ell d}(\boldsymbol{A},\boldsymbol{A}_0)=tr(\boldsymbol{A}\boldsymbol{A}_0^{-1})-logdet(\boldsymbol{A}\boldsymbol{A}_0^{-1})-d$, d是数据的维数。$c(i,j)\in S(\in D)$表示样本$\boldsymbol{x}_i,\boldsymbol{x}_j$属于同一类（不同类）的约束。$\boldsymbol{\xi}$是一个由松弛变量构成的向量并初始化为$\boldsymbol{\xi}_0$。$\boldsymbol{\xi}_{ij_0}$当样本对$(i,j)$是同类约束时被设置为给定的距离下限，当样本对$(i,j)$是非同类约束时被设置为给定的距离上限。

ITML同时也可以扩展成一个核学习框架。令$\boldsymbol{K}_0$为初始的核矩阵，也就是$\boldsymbol{K}_0(i,j)=\phi(\boldsymbol{x}_i)^T\boldsymbol{A}_0\phi(\boldsymbol{x}_j)$，其中$\phi$ 是一个从原始空间到高维希尔伯特空间的隐式变换。在核空间上，两点之间的欧氏距离为$\boldsymbol{K}(i,i)+\boldsymbol{K}(j,j)-2\boldsymbol{K}(i,j)=tr(\boldsymbol{K}(\boldsymbol{e}_i-\boldsymbol{e}_j)(\boldsymbol{e}_i-\boldsymbol{e}_j)^T)$，其中$\boldsymbol{K}(i,j)=\phi(\boldsymbol{x}_i)^T\boldsymbol{A}\phi(\boldsymbol{x}_j)$是学习到的核矩阵，$\boldsymbol{A}$ 表示在希尔伯特空间的操作子，$\boldsymbol{e}_i$是第i个典型基向量。基于这些，ITML的核版本可以形式化为：

$$\begin{aligned} \underset{\boldsymbol{K}\succeq 0,\boldsymbol{\xi}}{min} \quad & D_{\ell d}(\boldsymbol{K},\boldsymbol{K}_0)+\gamma D_{\ell d}(diag(\boldsymbol{\xi}),diag(\boldsymbol{\xi}_0)) \\ s.t. \quad & tr(\boldsymbol{K}(\boldsymbol{e}_i-\boldsymbol{e}_j)(\boldsymbol{e}_i-\boldsymbol{e}_j)^T)\le \boldsymbol{\xi}_{ij}, \quad c(i,j)\in S \\ & tr(\boldsymbol{K}(\boldsymbol{e}_i-\boldsymbol{e}_j)(\boldsymbol{e}_i-\boldsymbol{e}_j)^T)\ge \boldsymbol{\xi}_{ij}, \quad c(i,j)\in D \end{aligned} \tag{6.2}$$

6.2.2 多核学习方法

多核学习方法（Multiple Kernel Learning）是一个学习由多个基本核函数组合而成的核机器（kernel machine）的过程。换句话说，现有的大多数的多核学习方法采用了不同的学习技术来确定基本核函数的组合系数。假设有一个基本核函数集合$\{\kappa_r\}_{r=1}^R$，其中R是核函数的个数。那么一个集成的核函数可以表示成：

$$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sum_{r=1}^{R}\boldsymbol{\beta}_r\kappa_r(\boldsymbol{x}_i,\boldsymbol{x}_j), \quad \boldsymbol{\beta}_r\ge 0 \tag{6.3}$$

因此，一个基于二类数据$\{(\boldsymbol{x}_i,\boldsymbol{y}_i\in\pm 1)\}_{i=1}^N$ 的常用MKL模型可以形式化为：

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{i=1}^{N}\boldsymbol{\alpha}_i\boldsymbol{y}_i\kappa(\boldsymbol{x}_i,\boldsymbol{x})+\boldsymbol{b} \\ &= \sum_{i=1}^{N}\boldsymbol{\alpha}_i\boldsymbol{y}_i\sum_{r=1}^{R}\boldsymbol{\beta}_r\kappa_r(\boldsymbol{x}_i,\boldsymbol{x})+\boldsymbol{b} \end{aligned} \tag{6.4}$$

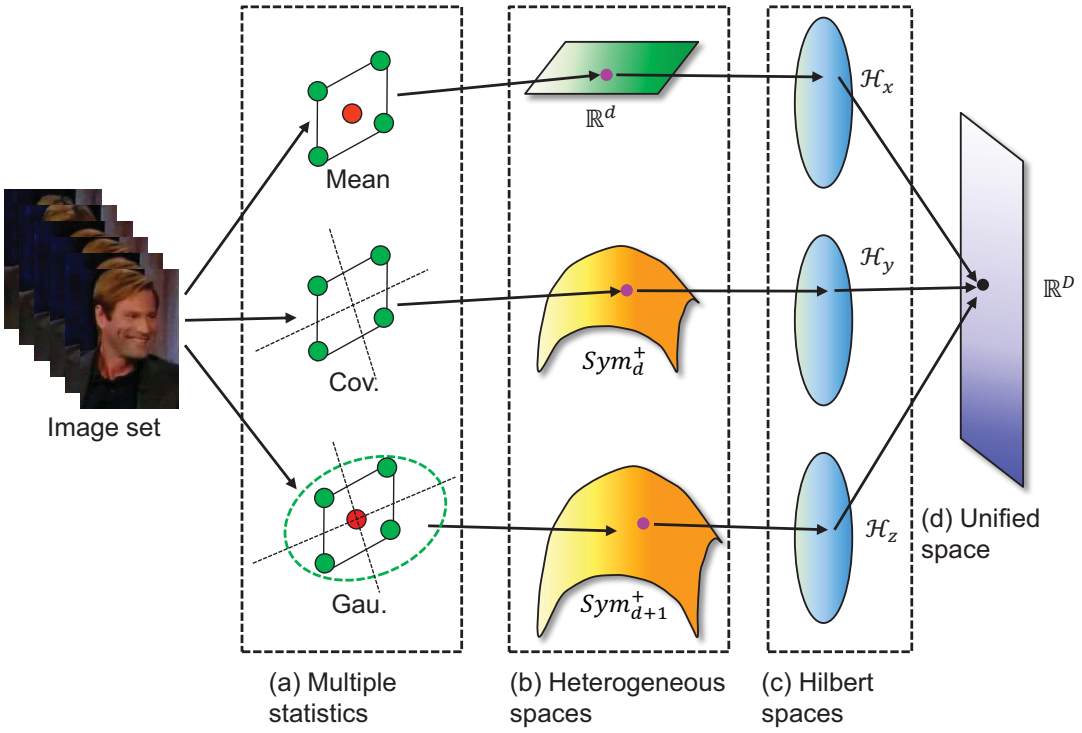


图 6.2. 本章新提出的混合欧氏-黎曼度量学习框架（Hybrid Euclidean-and-Riemannian Metric Learning, HERML）。(a)将来自视频序列的图像集合同时建模成样本均值、样本协方差和高斯模型。(b) 三种统计统计模型分别位于三种异质空间上，即欧氏空间$\mathbb{R}^d$, d维SPD流形$\mathbb{S}_+^d$和$d+1$维SPD流形$\mathbb{S}_+^{d+1}$。(c) 通过学习多个类马氏矩阵将这三种异质数据分别从各自对应的希尔伯特空间$\mathcal{H}_x, \mathcal{H}_y, \mathcal{H}_z$变换到一个公共的欧氏子空间$\mathbb{R}^D$ (d)，从而达到了异质融合的目的。

在两个参数变量$\{\boldsymbol{\alpha}_i\}_{i=1}^N$，$\{\boldsymbol{\beta}_r\}_{r=1}^R$上进行优化学习是MKL问题的一般形式。最近在MKL问题上的一些研究[38, 69, 94, 97, 111, 135, 147] 表明在多个核函数上学习组合系数不仅可以提升分类任务的性能而且也可以增强其解释性。据调研，现有大多数的MKL方法是学习来自多个同质空间（即多个同质的欧氏空间或多个同质的黎曼流形）上的核函数的组合系数。

6.3 混合欧氏-黎曼度量学习框架

本节将分四个小节来阐述新提出的混合度量学习框架。首先简单描述一下这一框架的基本流程，然后再介绍多统计模型的计算与空间嵌入，接着再详细描述基于多统计模型的混合度量学习，最后提出一种循环Bregman投影算法来求解这一问题。

6.3.1 概述

为了更加鲁棒地对视频中的人脸进行识别，本章新提出了一种全新的混合欧氏-黎曼度量学习方法（Hybrid Euclidean-and-Riemannian Metric Learning, HERML）来融合多

种统计模型，即样本均值、样本协方差和高斯分布函数模型。针对这些不同的统计模型，如图6.2 (a)(b)所示，本节首先研究了它们分别张成的三种异质空间：欧氏空间$\mathbb{R}^d$、d维SPD流形$\mathbb{S}_+^d$和$d+1$维SPD流形$\mathbb{S}_+^{d+1}$。因此，本章把混合度量学习形式化成了一个融合多种异质（欧氏与黎曼）数据的问题。为了融合异质数据，本章采用了第四章的异质度量学习机制推导出在这三种异质空间上的核函数将这三种异质数据分别映射到一个高维的可再核希尔伯特空间上（参见图6.2 (b)-(c)）。最后，本章新提出的混合度量学习框架通过同时在这些希尔伯特空间上学习多个马氏矩阵（也可以认为是线性变换矩阵）将异质数据都映射到一个公共欧氏子空间上从而达到多种异质数据融合的目的（参见图6.2 (c)-(d)）。

6.3.2　多统计模型计算与空间嵌入

这一小节首先介绍本章采用的多种统计模型的计算方法，然后再描述这些统计模型是如何嵌入到一个高维的希尔伯特空间上。

6.3.2.1　多统计模型计算

给定一个视频序列的人脸图像集合数据$\boldsymbol{X}_i = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$，其中$\boldsymbol{x}_j \in \mathbb{R}^d$是第$j$个图像样本的$d$维特征向量表示。那么在该数据集合上的均值统计模型可以计算为$\boldsymbol{m} = \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{x}_j$，协方差统计模型可以计算为$\boldsymbol{C} = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{m})(\boldsymbol{x}_j - \boldsymbol{m})^T$。

作为一个完整的概率模型，高斯分布函数同时包含了一阶统计量（即均值）和二阶统计量（即协方差）。从形式化上讲，这一图像集合对应的高斯分布函数可以定义为$\mathcal{N}(\boldsymbol{m}, \tilde{\boldsymbol{C}})$，其中观察协方差$\tilde{\boldsymbol{C}}$是样本协方差$\boldsymbol{C}$对应的最大似然估计。具体地，本章采用了经典的Expectation-Maximization (EM) 算法来计算一个集合数据的高斯混合模型（Gaussian Mixture Model, GMM）：$G = \sum_{i=1}^{M}\omega_i\mathcal{N}(\boldsymbol{m}_i, \tilde{\boldsymbol{C}}_i)$，其中$\mathcal{N}(\boldsymbol{m}_i, \tilde{\boldsymbol{C}}_i)$表示第$i$个高斯组件（Gaussian component），ω_i表示第i个高斯组件的先验。当$M=1$时，GMM就退化成单高斯模型（Single Gaussian Model, SGM）。

6.3.2.2　多统计模型空间嵌入

根据第四章的介绍，均值和协方差分别位于欧氏空间$\mathbb{R}^d$和d维SPD流形$\mathbb{S}_+^d$上。为了减少与前面两个统计模型的嵌入空间的异质性，本节根据第五章借鉴经典的信息几何理论[10, 93]的做法，将每个高斯组件转化成一个$d+1$维的SPD矩阵:

$$\mathcal{N}(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{C}}) \sim \boldsymbol{P} = |\boldsymbol{Q}|^{-\frac{2}{d+1}}\begin{bmatrix}\boldsymbol{Q}\boldsymbol{Q}^T + \tilde{\boldsymbol{m}}\tilde{\boldsymbol{m}}^T & \tilde{\boldsymbol{m}} \\ \tilde{\boldsymbol{m}}^T & 1\end{bmatrix} \tag{6.5}$$

关于这一嵌入过程更详细的推导，请读者参考文献[93]。

为了融合这三种异质数据，借鉴第四章异质数据融合的思想，首先需要将这些异质空间分别嵌入到一个高维的希尔伯特空间上。根据Mercer定理，只有满足正定的条

件的核函数才能生成有效的可再核希尔伯特空间，因此需要为这三种异质空间设计对应的正定核函数。

作为一种经典的正定核，欧氏线性核函数已经成功应用于大量的核学习方法上，因此本章对位于欧氏空间上的均值也同样采用了线性核函数：

$$\kappa_m(\boldsymbol{m}_i, \boldsymbol{m}_j) = \boldsymbol{m}_i^T \boldsymbol{m}_j \tag{6.6}$$

其中，$\boldsymbol{m}_i, \boldsymbol{m}_j$表示两个均值向量。

参照工作[139]，在SPD流形$\mathbb{S}_+^d$上可以通过在其切空间上计算对应的内积来推导相应的黎曼内积核函数。因此，为了将协方差位于的SPD流形$\mathbb{S}_+^d$嵌入到高维的希尔伯特空间上，本章采用了这一类型的黎曼内积核函数：

$$\kappa_C(\boldsymbol{C}_i, \boldsymbol{C}_j) = tr(log(\boldsymbol{C}_i) \cdot log(\boldsymbol{C}_j)) \tag{6.7}$$

其中，$\boldsymbol{C}_i, \boldsymbol{C}_j$表示两个协方差矩阵。由工作[139]所证明，由于这一核函数遵从了矩阵的弗罗贝尼乌斯范数(Frobenius norm)属性，可以很容易证明其具有正定性。

对于高斯模型，为了解决基于KL散度的核函数非正定问题，工作[27]提出了一种具有正定属性的基于KL散度的核估计：

$$\begin{aligned}
\Psi_{KL}(G_i \| G_j) &\leq \sum_{a=1}^{M_a} \sum_{b=1}^{M_b} \Psi_{KL}(p_i^a \| p_j^b) \\
&= \sum_{a=1}^{M_a} \sum_{b=1}^{M_b} \Psi_{KL}(\omega_a f(\tilde{\boldsymbol{m}}_i^a, \tilde{\boldsymbol{C}}_i^a) \| \omega_b f(\tilde{\boldsymbol{m}}_j^b, \tilde{\boldsymbol{C}}_j^b)) \\
&= \sum_{a=1}^{M_a} \sum_{b=1}^{M_b} \omega_a \omega_b \Psi_{KL}(f(\tilde{\boldsymbol{m}}_i^a, \tilde{\boldsymbol{C}}_i^a) \| f(\tilde{\boldsymbol{m}}_j^b, \tilde{\boldsymbol{C}}_j^b))
\end{aligned} \tag{6.8}$$

其中，$\Psi_{KL}(p_i^a \| p_j^b)$是经典的KL散度计算，$\boldsymbol{G}_i, \boldsymbol{G}_j$表示两个高斯模型，$p_i^a, p_j^b$是对应的概率分布，$M_a, M_b$是它们的组件个数，$\omega_a, \omega_b$表示对应高斯组件的先验。由于本章将每个高斯组件转化成了SPD矩阵，因此可以用对应的SPD流形度量来替换公式6.8中的KL散度计算，从而推导出了一种新的核函数:

$$\kappa_G(\boldsymbol{G}_i, \boldsymbol{G}_j) = \sum_{a=1}^{M_a} \sum_{b=1}^{M_b} \omega_a \omega_b tr(log(\boldsymbol{P}_i^a) \cdot log(\boldsymbol{P}_j^a)) \tag{6.9}$$

其中，$\boldsymbol{P}_i^a, \boldsymbol{P}_j^a$是由公式6.5对原始高斯组件计算得到的$d+1$维SPD矩阵。

6.3.3 多统计模型混合度量学习

给定一个由N个视频（图像集合）构成的训练集$\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N]$，其中$\boldsymbol{X}_i = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_i}] \in \mathbb{R}^{n_i \times d}$表示第$i$个图像集合，$n_i$表示这一集合的样本数。上一节定义

的核函数通常可以理解成将原始数据映射到一个高维的希尔伯特空间上，也就是$\phi:\mathbb{R}^d\rightarrow\mathcal{F}$ (或者$\mathbb{S}_+\rightarrow\mathcal{F}$)，然后在新的空间上计算这些高维特征$\boldsymbol{\Phi}_i$，$\boldsymbol{\Phi}_j$ 的内积。虽然这一映射ϕ通常是隐式的，不过为了更简单地分析新提出的度量学习框架，可以先把它当作显式变换来处理。具体地，用$\boldsymbol{\Phi}_i^r$表示图像集合$\boldsymbol{X}_i$的第r个统计模型在高维希尔伯特空间上的特征表示。现在给定一对训练样本$\boldsymbol{X}_i,\boldsymbol{X}_j$，可以计算出它们的第$r$个统计模型在高维希尔伯特空间上的距离度量：

$$d_{A_r}(\boldsymbol{\Phi}_i^r,\boldsymbol{\Phi}_j^r)=tr(\boldsymbol{A}_r(\boldsymbol{\Phi}_i^r-\boldsymbol{\Phi}_j^r)(\boldsymbol{\Phi}_i^r-\boldsymbol{\Phi}_j^r)^T) \tag{6.10}$$

其中，$\boldsymbol{A}_r$表示在高维希尔伯特空间上学习到的马氏矩阵。

如图6.2(c)-(d)所示，假设这些在不同高维希尔伯特空间上的多种统计数据可以统一变换到一个公共的子空间上，本节就可以同时优化学习多个马氏矩阵$\boldsymbol{A}_r,(r=1,\ldots,R)$（这里，$R=3$，这些马氏矩阵都可以转化为变换矩阵）。为了学习这些马氏矩阵，可以设计一个基于LogDet散度的优化函数来最大化类间方差同时最小化类内方差。具体地，此种度量学习的目标函数可以形式化为：

$$\begin{aligned}\underset{\boldsymbol{A_1}\succeq 0,\ldots,\boldsymbol{A_R}\succeq 0,\boldsymbol{\xi}}{min}\quad & \frac{1}{R}\sum_{r=1}^{R}D_{\ell d}(\boldsymbol{A}_r,\boldsymbol{A}_0)+\gamma D_{\ell d}(diag(\boldsymbol{\xi}),diag(\boldsymbol{\xi}_0)),\\ s.t.\quad & \frac{\boldsymbol{\delta}_{ij}}{R}\sum_{r=1}^{R}d_{A_r}(\boldsymbol{\Phi}_i^r,\boldsymbol{\Phi}_j^r)\leq\boldsymbol{\xi}_{ij},\forall(i,j).\end{aligned} \tag{6.11}$$

其中，$d_{A_r}(\boldsymbol{\Phi}_i^r,\boldsymbol{\Phi}_j^r)$可以由公式6.10计算得到，$\boldsymbol{\xi}_{ij}$是一个包含松驰变量的向量，初始化为$\boldsymbol{\delta}_{ij}\rho-\zeta\tau$，其中$\rho$ 是一个关于距离的约束，τ是一个边界值，ζ 是一个用于调整边界的尺度。$\boldsymbol{\delta}_{ij}=1$表示样本约束$c(i,j)$的样本来自同一个类别，$\boldsymbol{\delta}_{ij}=-1$则反之。由于每个马氏矩阵$\boldsymbol{A}_r$都是对称半正定矩阵，因此均可以通过$\boldsymbol{A}_r=\boldsymbol{W}_r\boldsymbol{W}_r^T$得到对应的变换矩阵$\boldsymbol{W}_r=[\boldsymbol{w}_1^r,\ldots,\boldsymbol{w}_{d_r}^r]$。

正如前面所述，由于ϕ^r 通常是隐式的，所以很难或者不可能通过公式6.10计算在希尔伯特空间上的距离度量$d_{A_r}(\boldsymbol{\Phi}_i^r,\boldsymbol{\Phi}_j^r)$。这里，通过采用经典的核技术[20]来将基矩阵$\boldsymbol{w}_k^r$表示为在映射空间上的训练样本的线性组合：

$$\boldsymbol{w}_k^r=\sum_{j=1}^{N}\boldsymbol{u}_j^k\boldsymbol{\Phi}_j^r \tag{6.12}$$

其中，$\boldsymbol{u}_j^k$是扩展系数（expansion coefficients），从而可以得到：

$$\sum_{r=1}^{R}(\boldsymbol{w}_k^r)^T\boldsymbol{\Phi}_i^r=\sum_{r=1}^{R}\sum_{j=1}^{N}\boldsymbol{u}_j^k(\boldsymbol{\Phi}_j^r)^T\boldsymbol{\Phi}_i^r=\sum_{r=1}^{R}(\boldsymbol{u}^k)^T\boldsymbol{K}_{\cdot i}^r \tag{6.13}$$

其中，$\boldsymbol{u}^k$ 是个$N\times 1$维的列向量，它的每个元素是$\boldsymbol{u}_j^k$，$\boldsymbol{K}_{\cdot i}^r$ 是第r个核矩阵$\boldsymbol{K}^r$的第i列向量，$\boldsymbol{K}^r,r=1,\ldots,R$ 是$N\times N$核矩阵分别由公式6.6, 6.7，6.9计算得到。如果对

于$1 \leq r \leq R$，将新马氏矩阵表示成$\boldsymbol{B}_r = \boldsymbol{U}_r\boldsymbol{U}_r^T$，那么目标函数6.11可以改写成：

$$\begin{aligned} \underset{\boldsymbol{B_1}\succeq 0,\ldots,\boldsymbol{B_R}\succeq 0,\boldsymbol{\xi}}{min} \quad & \frac{1}{R}\sum_{r=1}^{R} D_{\ell d}(\boldsymbol{B}_r,\boldsymbol{B}_0) + \gamma D_{\ell d}(diag(\boldsymbol{\xi}), diag(\boldsymbol{\xi}_0)), \\ s.t. \quad & \frac{\boldsymbol{\delta}_{ij}}{R}\sum_{r=1}^{R} d_{B_r}(\boldsymbol{K}_{.i}^r, \boldsymbol{K}_{.j}^r) \leq \boldsymbol{\xi}_{ij}, \forall (i,j). \end{aligned} \tag{6.14}$$

其中，$d_{B_r}(\boldsymbol{K}_{.i}^r, \boldsymbol{K}_{.j}^r)$表示为第$i$个和第$j$个样本的第$r$种统计模型在希尔伯特空间上的距离度量：

$$d_{B_r}(\boldsymbol{K}_{.i}^r, \boldsymbol{K}_{.j}^r) = tr(\boldsymbol{B}_r(\boldsymbol{K}_{.i}^r - \boldsymbol{K}_{.j}^r)(\boldsymbol{K}_{.i}^r - \boldsymbol{K}_{.j}^r)^T) \tag{6.15}$$

一旦最优的马氏矩阵$\boldsymbol{B}_1, \ldots, \boldsymbol{B}_R$被求出，每对参与匹配的视频样本在第$r$种统计模型的距离可以通过公式6.15计算得到，然后对在三种统计模型上的距离求平均作为此样本对的最终距离。

6.3.4 优化算法

为了求解目标函数6.14，本节采用经典的循环Bregman投影算法[25, 30, 83]。Bregman投影算法在每次迭代中只选择一个样本对约束，并执行一个映射使得当前解要满足所选择的约束。在不等式约束情况下，适当的更正也是需要的。这一过程是遍历所有的约束来重复进行。由文献[25, 30, 83]所述，这一循环Bregman投影算法可以收敛到最优解。关于更详细的说明，请读者参考[25, 30, 83]。针对目标函数6.14，本节的优化算法的更新准则如下命题所示。

命题1. *给定在第t次迭代的解$\boldsymbol{B}_r^t$ $(r = 1, \ldots, R)$ 和$\boldsymbol{\xi}_{ij}^t$, 更新第$t+1$次的解$\boldsymbol{B}_r^t$和对应的$\boldsymbol{\xi}_{ij}^t$:*

$$\begin{cases} \boldsymbol{B}_r^{t+1} = \boldsymbol{B}_r^t + \beta_r \boldsymbol{B}_r(\boldsymbol{K}_{.i}^r - \boldsymbol{K}_{.j}^r)(\boldsymbol{K}_{.i}^r - \boldsymbol{K}_{.j}^r)^T\boldsymbol{B}_r, & (6.16) \\ \boldsymbol{\xi}_{ij}^{t+1} = \dfrac{\gamma\boldsymbol{\xi}_{ij}^t}{\gamma + \boldsymbol{\delta}_{ij}\alpha\boldsymbol{\xi}_{ij}^t}, & (6.17) \end{cases}$$

其中，$\beta_r = \boldsymbol{\delta}_{ij}\alpha/(1 - \boldsymbol{\delta}_{ij}\alpha d_{B_r^t}(\boldsymbol{K}_{.i}^r, \boldsymbol{K}_{.j}^r))$，$\alpha$ 可以由以下公式求解:

$$\frac{\boldsymbol{\delta}_{ij}}{R}\sum_{r=1}^{R}\frac{d_{B_r^t}(\boldsymbol{K}_{.i}^r, \boldsymbol{K}_{.j}^r)}{1 - \boldsymbol{\delta}_{ij}\alpha d_{B_r^t}(\boldsymbol{K}_{.i}^r, \boldsymbol{K}_{.j}^r)} - \frac{\gamma\boldsymbol{\xi}_{ij}^t}{\gamma + \boldsymbol{\delta}_{ij}\alpha\boldsymbol{\xi}_{ij}^t} = 0. \tag{6.18}$$

证明. 基于循环Bregman投影算法[25, 30, 83], 可以将公式6.14形式化成拉格朗日形式, 然后在$\boldsymbol{B}_r^{t+1}$, $\boldsymbol{\xi}_{ij}^{t+1}$和$\alpha$之上的梯度设置为零从而可以得到以下几个更新等式:

$$\begin{cases} \nabla D(\boldsymbol{B}_r^{t+1}) = \nabla D(\boldsymbol{B}_r^t) + \boldsymbol{\delta_{ij}}\alpha(\boldsymbol{K}_{.i}^r - \boldsymbol{K}_{.j}^r)(\boldsymbol{K}_{.i}^r - \boldsymbol{K}_{.j}^r)^T, & (6.19) \\ \nabla D(\boldsymbol{\xi}_{ij}^{t+1}) = \nabla D(\boldsymbol{\xi}_{ij}^t) - \dfrac{\boldsymbol{\delta}_{ij}\alpha}{\gamma}, & (6.20) \\ \dfrac{\boldsymbol{\delta}_{ij}}{R}\displaystyle\sum_{r=1}^{R} tr(\boldsymbol{B}_r^{t+1}(\boldsymbol{K}_{.i}^r - \boldsymbol{K}_{.j}^r)(\boldsymbol{K}_{.i}^r - \boldsymbol{K}_{.j}^r)^T) = \boldsymbol{\xi}_{ij}^{t+1}. & (6.21) \end{cases}$$

算法 4 混合欧氏-黎曼度量学习（HERML）算法

输入: 训练样本对$\{(\boldsymbol{K}_{.i}^r, \boldsymbol{K}_{.j}^r), \boldsymbol{\delta}_{ij}\}$, 松弛参数$\gamma$, 初始马氏矩阵$\boldsymbol{B}_0$, 距离阈值$\rho$, 边界参数$\tau$以及调整尺度$\zeta$。

1. 初始化：$t \leftarrow 1, \boldsymbol{B}_r^1 \leftarrow \boldsymbol{B}_0$ for $r = 1, \ldots, R, \boldsymbol{\eta}_{ij} \leftarrow 0, \boldsymbol{\xi}_{ij} \leftarrow \boldsymbol{\delta}_{ij}\rho - \zeta\tau, \forall(i, j)$
2. **重复以下步骤:**
3. 选择约束(i, j)并利用公式6.15计算距离$d_{B_r^t}(\boldsymbol{K}_{.i}^r, \boldsymbol{K}_{.j}^r))$ for $r = 1, \ldots, R$。
4. 利用公式6.18和公式6.22更新α。
5. 利用公式6.16更新$\boldsymbol{B}_r^{t+1}$ for $r = 1, \ldots, R$。
6. 利用公式6.17更新$\boldsymbol{\xi}_{ij}^{t+1}$。
7. **直至收敛。**

输出: 马氏矩阵$\boldsymbol{B}_1, \ldots, \boldsymbol{B}_R$。

由公式6.19和公式6.20可以分别得到公式6.16和公式6.17。然后将公式6.16和公式6.17代入公式6.21可以求得与α相关的更新准则6.18。由于目标函数6.14包含了不等式约束项，这里还需要对α进行更正。具体地，这里需要再增加一个对偶变量$\boldsymbol{\eta}_{ij} \geq 0$。为了保持$\boldsymbol{\eta}_{ij}$非负性约束（这是满足Karush-Kuhn-Tucker(KKT)条件所必须的），参照工作[83]，对α的更正准则如下：

$$\alpha \leftarrow \min(\alpha, \boldsymbol{\eta}_{ij}), \quad \boldsymbol{\eta}_{ij} \leftarrow \boldsymbol{\eta}_{ij} - \alpha \tag{6.22}$$

具体的求解算法由算法4给出。算法的输入是初始化的马氏矩阵$\boldsymbol{B}_1, \ldots, \boldsymbol{B}_R$，样本对约束$\{(\boldsymbol{K}_{.i}^r, \boldsymbol{K}_{.j}^r), \boldsymbol{\delta}_{ij}\}$，松弛参数$\gamma$，距离阈值$\rho$, 边界参数$\tau$以及调整尺度$\zeta$。算法的主要时间复杂度主要在第5步的$\boldsymbol{B}_r^{t+1}$更新，每次迭代复杂度为$O(RN^2)$（其中，$N$是训练样本的数目）。因此，算法的总复杂度为$O(LRN^2)$，其中$L$是执行$\boldsymbol{B}_r^{t+1}$更新的总次数。

6.4 实验验证

本节实验将在视频人脸识别与视频人脸确认两个任务上来验证本章新提出的HERML方法的有效性。接下来将具体描述实验结果和分析。

6.4.1 实验数据库

与前面几章相同，本实验选取了与前几章对比实验中相同的数据库，即四个富有挑战性的数据数据库：COX , YouTube Celebrities (YTC) [79], YouTube Face DB (YTF) [142] 和Point-and-Shoot Challenge (PaSC) [24]。由于四个视频人脸数据库在前几章已有详细介绍，本节将不再详细介绍这四个数据库。

实验中针对各数据库的训练集/测试集的划分方案与前几章的设置基本相同。针对COX/YTC/YTF数据库，分别进行十组交叉验证，每组实验对应于一组随机选择的

训练集与测试集标号组合。针对PaSC数据库，实验严格遵从原始设计的协议执行可控（Control）实验和手持（Handheld）实验。

6.4.2 对比方法与参数设置

本节实验将选取以下四大类方法来对比新提出的HERML方法（其中ITML-Gau方法是采用其kernel版本，并将本章针对高斯模型而提出的核函数作为输入）：

1. 基于样本均值的方法：

   Maximum Mean Discrepancy (MaxMD)[49];

2. 基于线性子空间建模的方法：

   Mutual Subspace Method (MSM) [149], Discriminant Canonical Correlations (DCC)[81]，Grassmann Discriminant Analysis (GDA) [55]；

3. 基于约束子空间建模的方法：

   Affine (Convex) Hull based Image Set Distance (AHISD, CHISD)[32], Set-to-Set Distance Metric Learning (SSDML) [162]；

4. 基于协方差建模的方法:

   Covariance Discriminative Learning (CDL)[139]，Localized Multi-Kernel Metric Learning (LMKML)[94]；

5. 基于高斯分布函数建模的方法:

   Single Gaussian Models (SGM) [115], Gaussian Mixture Models (GMM) [14]，kernelized ITML [40] with the input of kernels for Gaussian (ITML-Gau);

6. 本章新提出的HERML方法。

除了SGM和GMM方法，这里所有的对比方法的源代码都来自原作者。为了公平对比，每个方法的重要参数都是根据原始工作的推荐而调整的。对于MaxMD方法，采用的是Bootstrap版本，将其重要的两个参数分别设置为：$\alpha = 0.1, \sigma = -1$。针对ITML方法，采用了源代码的默认参数。对于MSM方法，在最终进行相互子空间匹配时采用最大的典型相关系数。对于AHISD/CHISD/DCC方法，最后采用10个典型相关系数。对于GDA方法，将子空间的维数设置为10。对于CDL方法，本章实现的是它的LDA版本。对于SSDML方法，其两个关键参数分别设置为：$\lambda_1 = 0.001, \lambda_2 = 0.5$。对于LMKML方法，实验在训练数据间的距离均值范围来调整这一方法中的高斯核宽。对于本章的HERML方法，平衡参数$\gamma = 1$, 距离约束参数ρ 设置为训练数据间的距离均值, 边界参数τ设置为训练数据间的距离标准差，尺度参数ζ在$[0.1, 0.5]$范围内进行调整。

表 6.1. HERML对比方法在YTC和COX数据库上的视频-视频人脸识别结果（%）

Methods	YTC	COX					
		V2-V1	V3-V1	V3-V2	V1-V2	V1-V3	V2-V3
MaxMD [49]	52.6	36.4	19.6	8.90	27.6	19.1	9.60
MSM [149]	60.25±3.05	45.53±0.46	21.47±1.87	11.00±0.85	39.83±0.67	19.36±0.67	9.50±0.67
DCC [81]	68.85±2.32	62.53±6.13	66.10±3.52	50.56±4.21	56.09±11.27	53.84±11.37	45.19±9.81
GDA [55]	65.02±2.91	68.61±1.96	77.70±1.58	71.59±1.11	65.93±1.88	76.11±0.98	74.83±1.80
AHISD [32]	63.70±2.89	53.03±2.05	36.13±1.00	17.50±0.81	43.51±0.66	34.99±0.83	18.80±0.69
CHISD [32]	66.62±2.79	56.90±0.64	30.13±0.79	15.03±0.77	44.36±0.51	26.40±0.72	13.69±0.76
SSDML [162]	68.85±2.32	60.13±0.23	53.14±0.72	28.73±0.53	47.91±0.38	44.42±0.74	27.34±0.85
LMKML [94]	70.31±2.52	56.14±1.78	44.26±3.54	33.14±4.63	55.37±3.06	39.83±3.35	29.54±3.94
CDL [139]	69.72±2.92	78.43±1.01	85.31±0.97	79.71±1.47	75.56±1.95	85.84±0.86	81.87±1.14
SGM [115]	52.03±3.18	26.65±1.40	14.32±2.74	12.38±0.23	26.04±2.52	19.01±2.12	10.34±1.57
GMM [14]	61.95±2.25	30.12±2.16	24.58±1.23	13.03±2.21	28.93±1.34	31.69±1.72	18.93±2.01
ITML-Gau [40]	68.37±2.94	47.89±1.81	48.91±2.10	36.13±1.72	43.12±1.16	35.61±1.82	33.61±1.92
HERML-SGM	**74.57±2.03**	**94.87±1.27**	**96.85±2.12**	**93.32±1.83**	**92.02±1.62**	**96.38±1.26**	**95.32±2.01**
HERML-GMM	**73.34±1.75**	**95.13±1.78**	**96.32±1.72**	**94.22±1.68**	**92.32±1.72**	**95.36±1.82**	**94.53±1.38**

6.4.3　视频-视频人脸识别与确认评测

表6.1和6.2给出了实验对比的不同方法在YTC/COX/YTF/PaSC四个数据库上的结果，其中表6.1首选识别率的均值与标准差是10次实验统计所得，而表中在YTF上的性能是10次实验统计的平均确认率和方差，在PaSC上的是在FAR=0.01时的人脸确认率。从这两个表的对比情况分析，可以发现本章新提出的HERML方法要明显优于对比方法，其中包括ITML-Gau。这主要归功于新方法在统计建模方面具有更强的建模能力：现有的方法大多数是基于单一统计模型对视频中的图像集合进行建模，而本章同时采用了样本均值、样本协方差和高斯模型进行建模，有效地提升了统计建模方面的能力。与现有的多核学习方法LMKML相比，HERML也表现出明显的优越性。这主要是因为LMKML只是将协方差和统计张量统计都转化为了向量，而忽视了协方差和张量特有的矩阵属性（一般位于非欧氏空间上），这样学习的一般是非最优度量。相比之下，HERML在学习马氏度量时，通过推导出每种异质数据特有的嵌入空间所对应的核函数来编码对应的数据几何结构，从而可以学习到更佳的度量。

另外，图6.3和图6.4也对比了HERML框架里不同统计模型在四个数据库上的性能，从结果中可以发现高斯模型SGM/GMM的性能都要好于均值和协方差的性能。而融合均值和协方差后的性能在COX/YTF上要好于高斯模型SGM/GMM，在YTC/PaSC上却表现地相对比较差。可能的原因是在YTC/PaSC上数据相对比较服从高斯分布，用SGM/GMM可以很好拟合数据了，而在COX/YTF上数据并不严格服从高斯分布，样本均值和样本协方差由于没有对数据分布进行假设从而融合之后的性能更佳。从图

表 6.2. HERML对比方法在YTF和PaSC数据库上的视频-视频人脸确认结果(%)

Methods	YTF	PaSC	
		Control	Handheld
MSM [149]	60.25±3.05	35.80	34.56
DCC [81]	68.85±2.32	38.87	37.53
GDA [55]	65.02±2.91	41.88	43.25
AHISD [32]	63.70±2.89	21.96	14.29
CHISD [32]	66.62±2.79	26.12	20.97
SSDML [162]	68.85±2.32	29.19	22.89
CDL [139]	64.94±2.38	42.62	42.97
HERML-SGM	**75.16±0.84**	**45.40**	**45.46**
HERML-GMM	**74.36±1.53**	**46.61**	**46.23**

表 6.3. HERML对比方法在YTC数据库上的计算时间对比（单位：秒）

Method	MaxMD	SSDML	DCC	CDL	LMKML	SGM	GMM	**HERML**
Train	N/A	433.3	11.9	4.3	17511.2	N/A	N/A	27.3
Test	0.1	2.6	0.1	0.1	247.1	0.4	1.9	0.1

中还可以发现将所有的统计模型进行融合后的性能还能有更进一步的提升。不过，HERML-SGM和HERML-GMM在大部分情况下性能相当，甚至有时还要好。发生这一现象可能的原因是GMM里有更多的参数（如高斯组件个数）需要调整。即使本章采用经典的最小描述长度（Minimum Description Length）来估计高斯组件个数也很难找到最佳的参数设置来拟合服从任意分布下的真实数据。然而，由于GMM相比SGM是个更广义的模型，在实际应用中相信会有更强的描述数据的能力。

最后，表6.3也对比了不同方法在一台Intel(R) Core(TM) i7-3770(3.40GHz)的机器上运行的效率。表中仅对监督方法罗列出其训练时间，测试时间是识别每段视频的平均时间。从表中的对比可以发现，除了DCC和CDL方法，HERML要明显快于其它方法特别是同样是多核学习的LMKML方法。这主要是因为LMKML需要将二维的协方差矩阵和三维的张量都转成非常高维的向量形式，而在非常高维的向量空间上进行度量学习的效率自然会相当低效。相比之下，HERML由于直接在维度要低很多的协方差矩阵和高斯模型转化后的SPD矩阵上进行度量学习，其效率要高很多。

6.4.4 与前几章所提方法的对比

由于前几章提出的方法（PML, CERML, LEML）都在这几个数据库的视频-视频人脸识别任务上测试过性能，因此本实验通过绘制图6.5来对比HERML和这几个方法的性

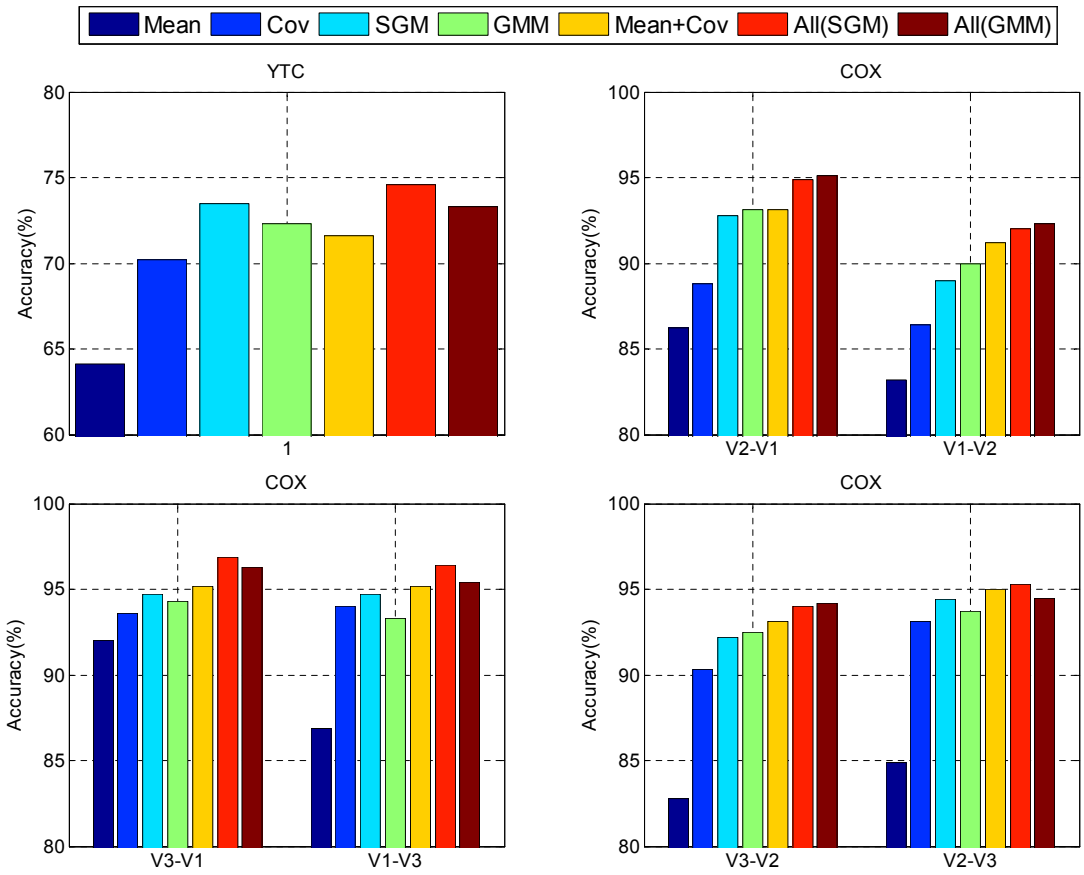


图 6.3. HERML框架里的不同统计模型在YTC和COX数据库上的性能对比。其中，HERML-all-SGM表示融合均值、协方差和单高斯模型，HERML-all-GMM表示融合均值、协方差和混合高斯模型。

能。从图6.5中可以发现，相比于其它方法，HERML由于采用了多统计模型对视频序列中的人脸图像集合进行建模，在四个视频人脸数据库上均达到了目前最高的性能。以下是对本文所提出的方法之间存在的区别与联系作出的详细分析：

从方法前端对视频人脸数据的统计建模角度来说，本文主要提出了四种不同的统计模型来建模视频序列中的人脸变化模式：第三章的PML方法将视频序列中的人脸图像集合表示成一个线性子空间来对人脸图像集合中的数据变化模式进行全局建模，从而在非限定条件的基于视频的人脸识别中对人脸姿态及表情等变化更加鲁棒。由于PML主要是利用线性子空间来建模视频人脸数据的变化模式。显然这一统计建模方式只是采用了二阶统计量，而忽略了一阶统计量，即数据的均值信息。因此，第四章的CERML方法研究同时采用双阶统计量来对视频中的集合数据的一阶信息及二阶信息的刻画，从而更全面地刻画了视频序列的丰富的多视空间信息。为了对数据进行一个完整的概率建模（即分别用模型中的均值和协方差对每段视频的人脸数据进行一阶统计和二阶统计），第五章的LEML开发了高斯分布函数来对视频序列中的人脸空间模式及变化模式进行有效建模；由于在视频序列里的人脸图像数据通常包含了极其复杂的

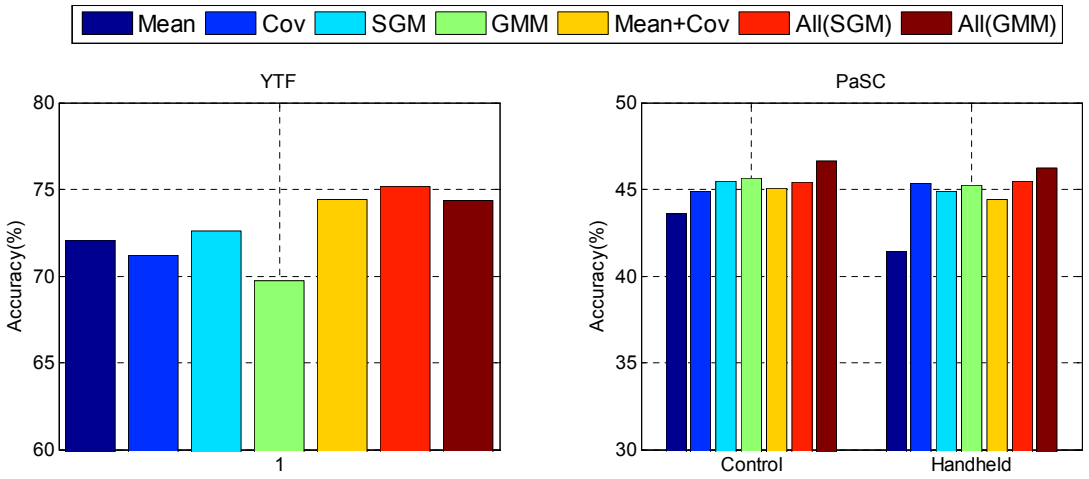


图 6.4. HERML框架里的不同统计模型在YTF和PaSC数据库上的性能对比。其中，HERML-all-SGM表示融合均值、协方差和单高斯模型，HERML-all-GMM表示融合均值、协方差和混合高斯模型。

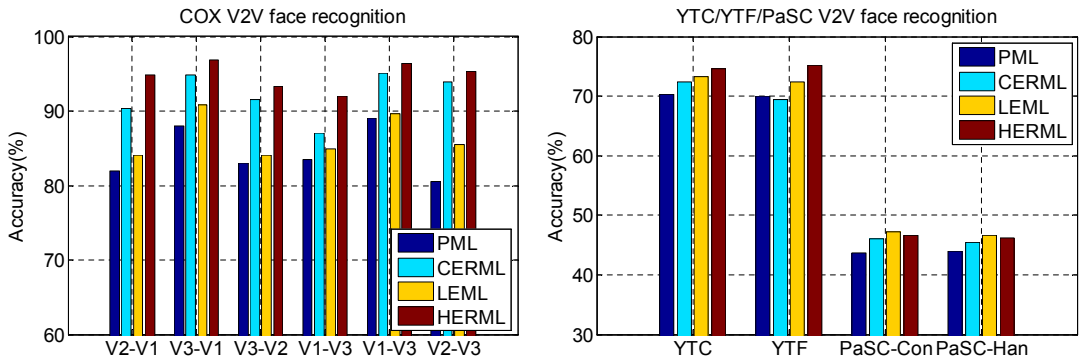


图 6.5. HERML与前几章提出的PML, CERML和LEML在COX, YTC, YTF和PaSC数据库上的性能对比。

变化模式（视频序列中的数据通常是服从任意分布的），那么如果使用单一的统计模型是无法准确刻画出视频数据的分布与结构。基于这一考虑，本章提出的HERML方法通过融合多种统计模型，即均值、协方差和高斯概率模型，来对视频序列的信息进行更加鲁棒地建模。实验结果也表明了本章提出的建模方法在大多数情况下要好于其它三种方法。

从方法后端对统计模型（即黎曼数据）进行黎曼度量学习角度来说，本文所提出的四种不同黎曼度量学习方法共采用了两种不同的学习策略：第三章提出的投影度量学习方法PML和第五章提出的对数欧氏欧氏度量学习方法LEML均在流形-流形映射学习的框架下对黎曼数据学习判别函数。这一学习策略的优点在于它学习到的目标空间跟原始黎曼流形是具有同种拓扑结构以及配备同种黎曼度量，从而使学习到的黎曼度量能受益于原始流形有用的黎曼几何属性。相比之下，第四章提出的跨欧氏-黎曼度量学习方法CERML和第六章提出的混合欧氏-黎曼度量学习方法HERML采用的是

流形-核空间映射学习框架。这一学习策略的优点是可以在将流形嵌入到核空间的基础上直接采用传统开发在欧氏空间的经典核方法从而实现对转换后的黎曼数据进行有效分类的目的。由于LEML和CERML本质上均采用了一阶和二阶统计量的信息，因此为了公平对比这两种学习策略可以考虑LEML和CERML两个方法在实验上的比较。从在COX数据库（库里的测试数据比较接近于训练数据）上的实验结果来看，采用流形-流形映射学习策略的LEML的性能要低于采用流形-核空间映射学习策略的CERML。而在YTC, YTF和PaSC三个数据库（库里的测试数据均与训练数据相差比较大）上，LEML在视频人脸识别任务上的表现要好于CERML。从这两种不同的对比结果可以表明CERML和HERML由于采用了传统的核学习方法拟合能力比较好，而PML和LEML由于在流形上直接采用了线性的分类器泛化能力比较强。另外，还需要指出来的是，CERML和HERML在把传统核方法引进黎曼度量学习框架的同时，这一类方法同样将核方法的一些固有缺陷引了进来，比如不能学习到显式的映射，而且随着样本数目的增加核方法的复杂度也会增大。相比之下，PML和LEML由于是直接在流形上进行度量学习，因此算法复杂度相对较低。

6.5　本章小结

本章通过开发样本均值、样本协方差和高斯模型的互补性将三者同时用于视频序列的统计建模。为了减少高斯模型所在的空间与其它两种统计模型的空间之间的异质性，本章借鉴经典的几何理论将高斯分布的空间也嵌入到一个SPD流形上，最后提出一种混合度量学习框架来有效融合样本均值、样本协方差和高斯模型这三种统计模型。本章在四个挑战性的基于视频的人脸数据库上进行了系统的实验对比及验证。实验结果表明，本章新提出的HERML方法在这四个数据库上的两个不同视频人脸识别任务上均达到了当前最好的性能。另外也通过对比不同统计模型的表现发现融合这三种统计模型可以在一定程度上提高视频人脸识别的性能。

实质上，高斯模型中的观察协方差是样本协方差的一个最大似然估计，可以看作是样本协方差的松驰版本。因此，将高斯模型与样本均值和样本协方差进行融合提升的幅度比较有限。在后续研究工作中可以再继续探索一种更高阶的统计量来真正弥补样本均值和样本协方差不能刻画其它有价值的数据信息的不足。

第七章　结束语

7.1　本文工作总结

作为一种新型的机器学习方法，黎曼度量学习方法已经开始被成功应用于诸多计算机视觉与模式识别领域的分类任务，比如基于视频的人脸识别任务。为了有效编码人脸视频序列中丰富的人脸多视空间信息，一些传统的统计模型通过有效刻画视频序列中的人脸变化模式从而成为一种鲁棒的用于视频人脸识别的特征。由于一般的统计模型通常位于一个特定类型的黎曼流形上，现有的黎曼度量学习方法通过开发对应的黎曼几何的强大数学理论为视频人脸识别问题提供了一种非常有效的判别学习策略。然而通过本文调研，现有的基于统计模型的黎曼度量学习方法通常存在以下两个问题：第一，大多数方法所采用的统计模型只在某一方面刻画了视频中的人脸图像集合数据的统计信息（比如协方差统计量），而忽视了其他有用的统计信息（比如均值统计量），从而导致对视频信息刻画不全的问题；第二，大多数方法通常是采用流形到欧氏空间嵌入或者流形到核空间嵌入的度量学习框架而很少去研究如何直接在黎曼流形上进行度量学习从而更有效地进行视频人脸识别任务。

针对以上两个问题，本文研究了一些双阶/多阶统计建模方法以及一系列基于统计模型的黎曼度量学习方法来更有效地从视频中识别人脸。具体地，在统计建模问题上，本文主要提出了一阶统计统计量（均值）与二阶统计统计量（协方差）融合的策略以及它们与完整的高斯概率模型进行融合的策略。在判别学习问题上，本文以视频人脸识别问题为应用背景，主要从视频序列的不同统计建模出发设计了一系列有效的黎曼度量学习方法。本文旨在推动黎曼度量学习方法理论与应用研究，特别是在扩展黎曼度量学习在视频人脸识别中的应用研究方面开展了一些开发性和探索性的工作。本文取得的主要创新性成果有：

(1) 基于视频序列的线性子空间建模，提出了一种投影度量学习方法来解决视频-视频人脸识别问题。为了能在以线性子空间为基本元素的格拉斯曼流形上进行判别学习，该方法提出了一个将原始的格拉斯曼流形变换到一个更具判别能力的新格拉斯曼流形上的度量学习框架来学习线性子空间之间的更具判别性的投影度量。由于学习到的类马氏矩阵可以分解成变换矩阵，新方法除了是一种度量学习方法以外同时也是一种施加于格拉斯曼流形的降维技术。大量的实验不仅表明了该方法在几个有代表性的视频人脸数据库上的性能均达到了与当前最优的方法同等可比的水平，也验证了它作为降维方法与其它方法结合之后可以进一步提高原方法的性能。

(2) 基于视频序列的双阶统计量建模，提出了一种跨欧氏-黎曼度量学习框架来同时解决三种不同的视频人脸识别问题，即视频-图像、图像-视频和视频-视频人脸识别。

该方法首先采用均值和协方差双阶统计量来对视频数据进行建模，然后将这三种视频人脸识别问题统一形式化成异质数据的匹配/融合问题，最后提出一个新型的基于多视图学习的异质度量学习统一框架来匹配/融合异质数据的。由于这一新的度量学习框架可以包容不同的二阶统计量，如协方差、线性子空间和仿射子空间，因此它也成为一个通用的异质度量学习框架。大量的视频-图像/图像-视频和视频-视频人脸识别实验表明新提出的方法在四个极具挑战性的数据库上均达到了明显优于现有方法的性能。

(3) 基于视频序列的高斯分布函数建模，提出了一种在对称正定矩阵流形上的对数欧氏度量学习方法来解决视频-视频人脸识别问题。该方法基于信息几何理论将高斯模型转化为对称正定矩阵，然后通过学习一个切映射使得在新切空间上的对称正定矩阵的对数欧氏度量更具判别性。大量的实验验证了新方法在视频人脸识别和视频人脸确认两个任务上能够取得与同类型的基于对称正定矩阵建模的分类方法相当甚至更优的性能，从而表明其可以有效地解决视频人脸识别问题。

(4) 基于视频序列的多种统计建模，提出了一种混合欧氏-黎曼度量学习方法来有效融合样本均值、样本协方差和高斯模型这三种统计模型，从而能更鲁棒地解决视频-视频人脸识别问题。为了减少高斯模型所在的空间与其它两种统计模型的空间之间的异质性，该方法同样首先借鉴了经典的几何理论将高斯分布的空间嵌入到一个对称正定矩阵流形上，然后设计了一个混合度量学习框架同时学习多个马氏矩阵来对这些异质的多统计模型进行有效融合。大量的实验结果表明了该方法在这四个数据库上的两个不同视频人脸识别任务上均达到了当前最好的性能，同时也验证了融合这三种统计模型可以有效提高单独采用单种统计模型的视频人脸识别性能。

从方法前端对视频人脸数据的统计建模角度来说，本文主要开发了四种不同的统计模型来建模视频序列中的人脸变化模式：利用线性子空间模型刻画了视频中的人脸图像集合数据的二阶信息（变化模式），采用双阶统计量来同时刻画视频中的人脸数据的一阶和二阶信息，开发高斯概率分布函数模型来描述视频数据的完整概率模型，采用均值、协方差和高斯概率模型三种统计模型的融合建模方法。从方法后端对统计模型（表示为黎曼数据）进行黎曼度量学习角度来说，本文提出的投影度量学习方法和对数欧氏学习方法均在流形-流形映射学习的框架下对黎曼数据学习判别函数，而跨欧氏-黎曼度量学习方法和混合欧氏-黎曼度量学习方法均采用的是流形-核空间映射学习框架。从实验中可以发现前者具有更强的泛化能力，而后者的拟合能力更好。

7.2 下一步研究方向

本文面向基于视频的人脸识别问题对基于统计建模的黎曼度量学习方法展开了深入和细致的研究并取得了一定的研究成果。然而，基于视频的人脸识别问题和基于统计建模的黎曼度量学习方法均是机器学习、计算机视觉领域和模式识别领域内新兴的研究课题，不管是在问题研究层面还是在理论研究层面都需要再进一步地开发与拓展。

本文目前的工作主要针对基于统计建模的黎曼度量学习方法在视频人脸识别中的应用研究进行了相应的探索。下一步的研究工作可以围绕以下几个方面来开展：

(1) 针对基于线性子空间建模的投影度量学习方法，该方法在格拉斯曼流形的投影度量的框架下，研究了进行降维和度量学习时开发流形上对应的黎曼几何的重要性。在后续研究工作中可以考虑将本工作的投影度量学习框架应用于其它格拉斯曼流形上的度量如比奈-柯西度量。此外，为了更好地求解格拉斯曼流形上的度量学习问题，开发一种更加有效的可以在理论上能证明收敛的求解算法是值得深入研究的方向。

(2) 针对基于双阶统计量建模的跨欧氏-黎曼度量学习方法，由于该方法只能通过实验的定量分析来证明其对应的优化算法可以收敛到一个稳定的解，因此在后续研究工作中可以研究开发一个可理论证明收敛的求解算法。另外，该方法是基于双阶统计量模型来对视频序列进行建模，探索更高阶的统计量或者将双阶统计量融合为高斯分布函数模型是未来值得深入研究的方向。

(3) 针对基于高斯分布函数建模的对数欧氏度量学习方法，其中一个可能的改进方案是将该方法所采用的Bregman投影算法扩展成经典的黎曼共轭梯度求解算法，从而可以更加有效地求解此度量学习问题对应的类马氏矩阵。此外，将其它类型的度量如仿射不变度量和基于Stein散度的度量包容在这一度量学习的框架内也是值得尝试的一个改进方向。

(4) 针对多统计模型的混合欧氏-黎曼度量学习方法，由于高斯模型中的观察协方差是样本协方差的一个最大似然估计（即样本协方差的松驰版本）。因此，直接将高斯模型、样本均值和样本协方差进行融合的提升幅度相对比较有限。在后续研究工作中可以再继续探索一种更高阶的统计量来真正弥补样本均值和样本协方差不能刻画其它数据信息的不足。

(5) 本文针对视频人脸识别问题开发了一些统计模型，这些模型主要采用一阶统计量和二阶统计量来刻画视频序列中的人脸多视空间信息，而忽略了视频中的人脸动态时序信息。未来可以考虑采用自动回归与移动平均（Auto-Regressive and Moving Average, ARMA）模型来刻画视频的时序信息。由于ARMA模型通常也可以表示成格拉斯曼流形数据，因此可以采用本文提出的投影度量学习方法来为这一模型的分类问题提供判别学习策略。

(6) 本文基于不同的统计建模策略提出了四种不同的在特定黎曼流形上的黎曼度量学习方法，虽然不同的方法有其特性，但是部分方法之间也有一定的共性，如第四章和第六章均采用了流形-核空间的黎曼度量学习策略而第三章和第五章均采用了流形-流形的黎曼度量学习策略，因此后续工作可以从这些方法的共性出发提出一个统一的黎曼度量学习框架。

(7) 本文提出的几种黎曼度量学习方法主要是基于不同的统计模型来解决视频人脸识别问题。在计算机视觉与模式识别领域，基于统计模型的黎曼流形数据表示已经被

广泛地应用于各个不同领域，比如行人检测、行为识别、纹理分类以及人脸识别等任务上。因此，考虑如何将本文所提的黎曼度量学习方法应用于其它实际问题上是值得未来探索的一个研究方向。

参 考 文 献

[1] Faces96 database. *available at http://cswww.essex.ac.uk/mv/allfaces /faces96.html*.

[2] University of cambridge face database. *available at http://mi.eng.cam.ac.uk/oa214/academic/*.

[3] 王瑞平. 流形学习方法及其在人脸识别中的应用研究. 博士学位论文, 北京：中国科学院研究生院, 2010.

[4] 邱慧宁. 基于流形学习和稀疏表示的人脸识别方法研究. 博士学位论文, 中山大学, 2011.

[5] 崔振. 稀疏与流形表示及其在人脸识别中的应用研究. 博士学位论文, 北京：中国科学院大学, 2014.

[6] 陈维桓. 微分流形初步. 高等教育出版社, 1998.

[7] P.A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

[8] G. Aggarwal, A. Chowdhury, and R. Chellappa. A system identification approach for video-based face recognition. In *Int. Joint Conf. Neural Networks*, 2004.

[9] M. Al-Azzeh, A. Eleyan, and H. Demirel. PCA-based face recognition from video using super-resolution. In *Computer and Information Sciences*, 2008.

[10] S. Amari and H. Nagaoka. *Methods of information geometry*. Oxford University Press, 2000.

[11] L. An, B. Bhanu, and S. Yang. Boosting face recognition in real-world surveillance videos. In *Advanced Video and Signal-Based Surveillance*, pages 270–275, 2012.

[12] O. Arandjelovic and R. Cipolla. A manifold approach to face recognition from low quality video across illumination and pose using implicit super-resolution. In *IEEE Proc. Int. Conf. Comput. Vision*, 2007.

[13] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 581–588, 2005.

[14] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, 2005.

[15] V. Arsigny, P. Fillard, X. Pennec, and N Ayache. Fast and simple computations on tensors with log-euclidean metrics. *Res. Rep. RR-5584,INRIA*, 2005.

[16] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Analysis and Applications*, 29(1):328–347, 2007.

[17] J. H. Barr, K.W. Boyer, P. Flynn, and S. Biswas. Face recognition from video: A review. *Int. J. Pattern Recog. Arti. Intell.*, 26(5), 2012.

[18] J. R. Barr, K. W. Bowyer, and P. J. Flynn. Detecting questionable observers using face track clustering. In *Proc. IEEE Workshop Appl. Comput. Vision*, pages 182–189, 2011.

[19] M. Barry and E. Granger. Face recognition in video using a what-and-where fusion neural network. In *Int. Joint Conf. Neural Networks*, 2007.

[20] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural computation*, 12(10):2385–2404, 2000.

[21] L. Benedikt, D. Cosker, P. Rosin, and D. Marshall. 3D facial gestures in biometrics: From feasibility study to application. In *Int. Conf. Biometrics: Theory, Applications and Systems*, 2008.

[22] C. Berg, J. P. R. Christensen, and P. Ressel. Harmonic analysis on semigroups. *The Annals of Mathematics*, 117(2):267–283, 1983.

[23] S. Berrani and C. Garcia. Enhancing face recognition from video sequences using robust statistics. In *IEEE Conf. Advanced Video and Signal Based Surveillance*, 2005.

[24] J.R. Beveridge, P.J. Phillips, D. Bolme, B. A. Draper, G.H. Givens, Y.M. Lui, M.N. Teli, H. Zhang, W.T. Scruggs, K.W. Bowyer, P.J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *Proc. Biometrics: Theory, Applications and Systems*, pages 1–8, 2013.

[25] L.M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[26] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, 2010.

[27] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff. SVM based speaker verification using a gmm supervector kernel and nap variability compensation. In *Proc. Int. Conf. Acoust. Speech, Signal Process.*, pages 97–100, 2006.

[28] J. Carreira, R. Caseiro, J. Caseiro, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *Proc. Euro. Conf. Comput. Vision*, 2012.

[29] R. Caseiro, P. Martins, J. Henriques, Silva L., and J. Batista. Rolling riemannian manifolds to solve the multi-class classification problem. In *IEEE Proc. Comput. Vision Pattern Recog. Conf*, pages 41–48, 2013.

[30] Y. Censor and S. Zenios. *Parallel optimization: Theory, algorithms, and applications*. Oxford University Press, 1997.

[31] H.E. Cetingul and R. Vidal. Intrinsic mean shift for clustering on stiefel and grassmann manifolds. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, 2009.

[32] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 2567–2573, 2010.

[33] H. Cevikalp, B. Triggs, and R. Polikar. Nearest hyperdisk methods for high-dimensional classification. In *Proc. Int. Conf. Mach. Learn.*, 2008.

[34] L. Chen, H. Liao, and J. Lin. Person identification using facial motion. In *Int. Conf. Image Processing*, 2001.

[35] S. Chen, C. Sanderson, M.T. Harandi, and B.C. Lovell. Improved image set classification via joint sparse approximated nearest subspaces. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 452–459, 2013.

[36] J. Chien and C. Wu. Discriminant waveletfaces and nearest feature classifiers for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(12):1644–1649, 2002.

[37] J. Chien and C. Wu. Remote identification of faces: problems, prospects, and progress. *Pattern Recognition Letters*, 33(15):1849–1859, 2012.

[38] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, 2013.

[39] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen. Image sets alignment for video-based face recognition. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 2626–2633, 2012.

[40] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proc. Int. Conf. Mach. Learn.*, 2007.

[41] A. Edelman, T.A. Arias, and S.T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[42] S. Eickeler, F. Wallho, U. Lurgel, and G. Rigoll. Content based indexing of images and video using face detection and recognition methods. In *Int. Conf. Acoustics, Speech and Signal Processing*, 2001.

[43] W. Fan and D. Yeung. Face recognition with image sets using hierarchically extracted exemplars from appearance manifolds. In *Int. Conf. Automatic Face and Gesture Recognition*, 2006.

[44] K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. In *Robotics Research*, pages 192–201. Springer, 2005.

[45] K. Gallivan, A. Srivastava, X. Liu, and P. Van Dooren. Efficient algorithms for inferences on grassmann manifolds. In *Statistical Signal Processing Workshop*, pages 315–318, 2003.

[46] R. Goh, L. Liu, X. Liu, and T. Chen. The CMU face in action (FIA) database. In *Proc. Int. Conf. Autom. Face Gesture Recog.*, 2005.

[47] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *Proc. Neu. Infor. Proces. Sys.*, 2004.

[48] G.H. Golub, L. Van, and F. Charles. *Matrix computations*, volume 3. JHU Press, 2012.

[49] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Research*, 13:723–773, 2012.

[50] M. Grgic, K. Delac, and S. Grgic. SCface–surveillance cameras face database. *Multimedia Tools Appl. J*, 51(3):863–879, 2011.

[51] R. Gross and J. Shi. The CMU motion of body (MoBo) database. *Tech. Rep. CMU-RI-TR-01-18*, 2001.

[52] B. Gunturk, A. Batur, Y. Altunbasak, M. Hayes, and R. Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE Trans. Image Process*, pages 597–606, 2003.

[53] A. Hadid and M. Pietikainen. Selecting models from videos for appearance-based face recognition. In *Int. Conf. Pattern Recognition*, 2004.

[54] J. Hamm. Subspace-based learning with grassmann kernels. *Thesis*, University of Pennsylvania, 2008.

[55] J. Hamm and D. D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proc. Int. Conf. Mach. Learn.*, pages 376–383, 2008.

[56] J. Hamm and D.D. Lee. Extended grassmann kernels for subspace-based learning. In *Proc. Neu. Infor. Proces. Sys.*, pages 601–608, 2008.

[57] M. T. Harandi, C. Sanderson, R. Hartley, and B. C Lovell. Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach. In *Proc. Euro. Conf. Comput. Vision*, 2012.

[58] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 2705–2712, 2011.

[59] M.T. Harandi, M. Salzmann, and R. Hartley. From manifold to manifold: Geometry-aware dimensionality reduction for spd matrices. In *Proc. Euro. Conf. Comput. Vision*. 2014.

[60] M.T. Harandi, M. Salzmann, S. Jayasumana, R. Hartley, and H. Li. Expanding the family of grassmannian kernels: An embedding perspective. In *Proc. Euro. Conf. Comput. Vision*, 2014.

[61] M.T. Harandi, M. Salzmann, and F. Porikli. Bregman divergences for infinite dimensional covariance matrices. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 1003–1010, 2014.

[62] M.T. Harandi, C. Sanderson, C. Shen, and B.C. Lovell. Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution. In *IEEE Proc. Int. Conf. Comput. Vision*, 2013.

[63] M.T. Harandi, C. Sanderson, A. Wiliem, and B.C. Lovell. Kernel analysis over riemannian manifolds for visual recognition of actions, pedestrians and textures. In *Applications of Computer Vision Workshop*, pages 433–439, 2012.

[64] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[65] U. Helmke, K. Hüper, and J. Trumpf. Newton's method on grassmann manifolds. *arXiv preprint arXiv:0709.2205*, 2007.

[66] P. H. Hennings-Yeomans, S. Baker, and B.V. Kumar. Recognition of low-resolution faces using multiple still images and multiple cameras. In *Biometrics: Theory, Applications and Systems*, pages 1–6, 2008.

[67] Y. Hu, A.S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 121–128, 2011.

[68] S. Jain and V.M. Govindu. Efficient higher-order clustering on the grassmann manifold. In *IEEE Proc. Int. Conf. Comput. Vision*, pages 3511–3518, 2013.

[69] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M.T. Harandi. Combining Multiple Manifold-valued Descriptors for Improved Object Recognition. In *Int. Conf. Digital Image Comput.: Techni. and Appli.*, 2013.

[70] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M.T. Harandi. Kernel methods on the riemannian manifold of symmetric positive definite matrices. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, 2013.

[71] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M.T. Harandi. Optimizing over radial kernels on compact manifolds. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 3802–3809, 2014.

[72] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M.T. Harandi. Kernel Methods on Riemannian Manifolds with Gaussian RBF Kernels. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015.

[73] K. Jia and S. Gong. Multi-modal tensor face for simultaneous super-resolution and recognition. In *IEEE Proc. Int. Conf. Comput. Vision*, volume 2, pages 1683–1690, 2005.

[74] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv: 1408.5093*, 2014.

[75] R. R. Jillela and A. Ross. Adaptive frame selection for improved face recognition in low-resolution videos. In *Int. Joint Conf. Neural Networks*, 2009.

[76] S. Jung, I.L. Dryden, and J.S. Marron. Analysis of principal nested spheres. *Biometrika*, 99(3):551–568, 2012.

[77] M. Kan, S. Shan, D. Xu, and X. Chen. Side-information based linear discriminant analysis for face recognition. In *British Machine Vision Conference*, 2011.

[78] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 1–8, 2008.

[79] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 1–8, 2008.

[80] T.K. Kim, O. Arandjelović, and R. Cipolla. Boosted manifold principal angles for image set-based recognition. *Pattern Recognition*, 40(9):2475–2484, 2007.

[81] T.K. Kim, J. Kittler, and J. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):1005–1018, 2007.

[82] T.K. Kim, J. Kittler, and R. Cipolla. Incremental learning of locally orthogonal subspaces for set-based object recognition. In *British Machine Vision Conference*, pages 559–568, 2006.

[83] B. Kulis, M.A. Sustik, and I. S. Dhillon. Low-rank kernel learning with bregman matrix divergences. *J. Mach. Learn. Research*, 10:341–376, 2009.

[84] H. Le. On geodesics in euclidean shape spaces. *J. Lond. Math. Soc.*, pages 360–372, 1991.

[85] K. Lee, J. Ho, M. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 313–320, 2003.

[86] K. Lee, J. Ho, M. Yang, and D Kriegman. Visual tracking and recognition using probabilistic appearance manifolds. *Comput. Vision Image Underst.*, 99(3):303–331, 2005.

[87] P. Li, Q. Wang, W. Zuo, and L. Zhang. Log-euclidean kernels for sparse representation and dictionary learning. In *IEEE Proc. Int. Conf. Comput. Vision*, 2013.

[88] X. Li, K. Fukui, and N. Zheng. Boosting constrained mutual subspace method for robust image-set based object recognition. In *International Joint Conference on Artificial Intelligence*, pages 1132–1137, 2009.

[89] X. Li, K. Fukui, and N. Zheng. Image-set based face recognition using boosted global and local principal angles. In *Asian Conference on Computer Vision*, pages 323–332. Springer, 2010.

[90] L. Liu, L. Zhang, H. Liu, and S. Yan. Toward large-population face identification in unconstrained videos. *IEEE Trans. on Circuits and Systems for Video Technology*, 24(11):1874–1884, 2014.

[91] X. Liu and T. Chen. Face mosaicing for pose robust video-based recognition. In *Asian Conference on Computer Vision*, 2007.

[92] X. Liu and T. Cheng. Video-based face recognition using adaptive hidden markov models. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 340–345, 2003.

[93] M. Lovrić, M. Min-Oo, and E. A. Ruh. Multivariate normal distributions parameterized as a riemannian symmetric space. *Journal of Multivariate Analysis*, 74(1):36–48, 2000.

[94] J. Lu, G. Wang, and P. Moulin. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *IEEE Proc. Int. Conf. Comput. Vision*, pages 329–336, 2013.

[95] J. Masci, M. Bronstein, A. Bronstein, and J. Schmidhuber. Multimodal similarity-preserving hashing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(4):824–830, 2013.

[96] F. Matta and J.-L. Dugelay. Video face recognition: A physiological and behavioural multimodal approach. In *Int. Conf. Image Processing*, 2007.

[97] B. McFee and G. Lanckriet. Learning multi-modal similarity. *J. Mach. Learn. Research*, 12:491–523, 2011.

[98] A. Mian. Unsupervised learning from local features for video-based face recognition. In *Unsupervised learning from local features for video-based face recognition*, 2008.

[99] H.Q. Minh, M.S. Biagio, and V. Murino. Log-hilbert-schmidt metric between positive definite operators on hilbert spaces. In *Proc. Neu. Infor. Proces. Sys.*, 2014.

[100] S. Mitra, M. Savvides, and B. V. K. V. Kumar. Human face identification from video based on frequency domain asymmetry representation using hidden markov models. *Multimedia Content Representation, Classification and Security, Lecture Notes in Computer Science*, 2006.

[101] M. Nishiyama, O. Yamaguchi, and K. Fukui. Face recognition with the multiple constrained mutual subspace method. In *Audio-and Video-Based Biometric Person Authentication*, pages 71–80. Springer, 2005.

[102] A.J. O'Toole, P.J. Phillips, S. Weimer, D.A. Roark, J. Ayyad, R. Barwick, and J. Dunlop. Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach. *Vision Research*, 51(1):74–83, 2011.

[103] U. Park, H. Chen, and A. Jain. 3d model-assisted face recognition in video. In *Canadian Conf. Computer and Robot Vision*, 2005.

[104] U. Park and A. Jain. 3d model-assisted face recognition in video. In *Int. Conf. Biometrics*, 2007.

[105] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *Int. J. Comput. Vision*, 66(1):41–66, 2006.

[106] X. Pennec, P.Fillard, and N. Ayache. A riemannian framework for tensor computing. *Int. J. Comput. Vision*, 66(1):41–66, 2006.

[107] D. Pham and S. Venkatesh. Robust learning of discriminative projection for multicategory classification on the stiefel manifold. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, 2008.

[108] J. Phillips. Video Challenge Problem Multiple Biometric Grand Challenge: Preliminary Results of Version 2. *Nat. Inst. Standards Technol.*, 2009.

[109] P. Phillips, J. Beveridge, B. Draper, G. Givens, A. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer. An introduction to the good, the bad, the ugly face recognition challenge problem. In *Proc. Int. Conf. Autom. Face Gesture Recog.*, 2011.

[110] N. Quadrianto and C. H Lampert. Learning multi-view neighborhood preserving projections. In *Proc. Int. Conf. Mach. Learn.*, 2011.

[111] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *J. Mach. Learn. Research*, 9(11), 2008.

[112] C. Sanderson. Biometric person recognition: Face, speech and fusion. *available at http://www.itee.uq.edu.au/ onrad/vidtimit/*, 2008.

[113] I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[114] G. Shakhnarovich, J.W. Fisher, and T. Darrell. Face recognition from long-term observations. In *Proc. Euro. Conf. Comput. Vision*, pages 851–865, 2002.

[115] G. Shakhnarovich, J.W. Fisher, and T. Darrell. Face recognition from long-term observations. In *Proc. Euro. Conf. Comput. Vision*, 2002.

[116] A. Sharma and D.W. Jacobs. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, 2011.

[117] A. Sharma, A. Kumar, H. Daume, and D.W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, 2012.

[118] J. Sherman and W. J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, pages 124–127, 1950.

[119] R. Sivalingam, V. Morellas, D. Boley, and N. Papanikolopoulos. Metric learning for semi-supervised clustering of region covariance descriptors. In *ICDSC*, 2009.

[120] S. Sra. A new metric on the manifold of kernel matrices with application to matrix geometric means. In *Proc. Neu. Infor. Proces. Sys.*, 2012.

[121] A. Srivastava and E. Klassen. Bayesian and geometric subspace tracking. *Advances in Applied Probability*, pages 43–56, 2004.

[122] J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen. Video-based face recognition on real-world data. In *IEEE Int. Conf. Computer Vision*, 2007.

[123] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *J. Mach. Learn. Research*, pages 1027–1061, 2007.

[124] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *J. Mach. Learn. Research*, 8:1027–1061, 2007.

[125] D. Thomas, K. Bowyer, and P. Flynn. Multi-factor approach to improving recognition performance in surveillance-quality video. In *IEEE Conf. Biometrics: Theory, Applications and Systems*, 2008.

[126] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, 61(3):611–622, 1999.

[127] M. Tistarelli, M. Bicego, and E. Grosso. Dynamic face recognition: From human to machine vision. *Image Vis. Comput*, 2009.

[128] D. Tosato, M. Farenzena, M. Cristani, M. Spera, and V. Murino. Multi-class classification on riemannian manifolds for video surveillance. In *Proc. Euro. Conf. Comput. Vision*, 2010.

[129] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa. Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(11):2273–2286, 2011.

[130] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. Euro. Conf. Comput. Vision*, pages 589–600. Springer, 2006.

[131] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 1–8, 2007.

[132] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(10):1713–1727, 2008.

[133] K. Varshney and A. Willsky. Learning dimensionality-reduced classifiers for information fusion. In *ICIF*, 2009.

[134] R. Vemulapalli and D. W. Jacobs. Riemannian metric learning for symmetric positive definite matrices. In *arXiv*, 2015.

[135] R. Vemulapalli, J. Pillai, and R. Chellappa. Kernel learning for extrinsic classification of manifold features. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 1782–1789, 2013.

[136] P. Vincent and Y. Bengio. K-local hyperplane and convex distance nearest neighbor algorithms. In *Proc. Neu. Infor. Proces. Sys.*, pages 985–992, 2001.

[137] H. Wang, Y. Wang, and Y. Cao. Video-based face recognition: A survey. *World Academy of Science, Engineering and Technology*, 60:293–302, 2009.

[138] R. Wang and X. Chen. Manifold discriminant analysis. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 429–436, 2009.

[139] R. Wang, H. Guo, L. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 2496–2503, 2012.

[140] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-Manifold distance with application to face recognition based on image set. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 2940–2947, 2008.

[141] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Research*, 10:207–244, 2009.

[142] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 529–534, 2011.

[143] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *J. Mach. Learn. Research*, 4:913–931, 2003.

[144] Y. Wong. Differential geometry of grassmann manifolds. *Proceedings of the National Academy of Sciences of the United States of America*, 57(3):589, 1967.

[145] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B.C. Lovell. Patch-based probablistic image quality assessment for face selection and improved video-based face recognition. In *Proc. IEEE Workshop Comput. Vision Pattern Recog. Conf.*, pages 81–88, 2011.

[146] L. Wu, H. Xiong, L. Du, B. Liu, G. Xu, Y. Ge, Y. Fu, Y. Zhou, and J. Li. Heterogeneous metric learning with content-based regularization for software artifact retrieval. *IEEE International Conference on Data Mining*, 2014.

[147] P. Xie and E. P. Xing. Multi-modal distance metric learning. In *International Joint Conference on Artificial Intelligence*, 2013.

[148] Y. Xu, A. Roy-Chowdhury, and K. Patel. Integrating illumination, motion and shape models for robust face recognition in video. *EURASIP J. Adv. Signal Process*, 2008.

[149] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *Proc. Int. Conf. Autom. Face Gesture Recog.*, pages 318–323, 1998.

[150] Y. Yan and Y. Zhang. State-of-the-art on video-based face recognition. *Chinese Journal of Computers*, 32(5):878–886, 2009.

[151] J. Yang, D. Zhang, A.F. Frangi, and J. Yang. Two-dimensional pca: a new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(1):131–137, 2004.

[152] M. Yang, P. Zhu, L.V. Gool, and L. Zhang. Face recognition based on regularized nearest points between image sets. In *Proc. Int. Conf. Autom. Face Gesture Recog.*, 2013.

[153] N. Ye and T. Sim. Towards general motion-based face recognition. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, pages 2598–2605, 2010.

[154] X. Zhai, Y. Peng, and J. Xiao. Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In *Association for the Advancement of Artificial Intelligence*, 2013.

[155] H. Zhang, J. Yang, Y. Zhang, N.M. Nasrabadi, and T.S. Huang. Close the loop: Joint blind image restoration and recognition with sparse representation prior. In *IEEE Proc. Int. Conf. Comput. Vision*, pages 770–777, 2011.

[156] Y. Zhang and A. M. Martínez. A weighted probabilistic approach to face recognition from multiple images and video sequences. *Image Vis. Comput*, 24(6):626–638, 2006.

[157] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv*, 35(4):399–458, 2003.

[158] S.K. Zhou and R. Chellappa. Beyond one still image: Face recognition from multiple still images or a video sequence. *Face Processing: Advanced Modeling and Methods*, pages 547–567, 2005.

[159] S.K. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Trans. Image Process.*, 13(11):1491–1506, 2004.

[160] S.K. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *Comput. Vision Image Underst.*, 91(1):214–245, 2003.

[161] X. Zhou and B. Bhanu. Human recognition based on face profiles in video. In *IEEE Proc. Comput. Vision Pattern Recog. Conf.*, 2005.

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

[163] P. Zhu, W. Zuo, L. Zhang, S. Shiu, and D. Zhang. Image set based collaborative representation for face recognition,” ieee trans. on information forensics and security. *IEEE Trans. on Information Forensics and Security*, 9(7):1120–1132, 2014.

[164] W.W. Zou and P.C. Yuen. Very low resolution face recognition problem. *IEEE Trans. Image Process*, 21(1):327–340, 2012.

致　　谢

时光荏苒白驹过隙，五年的读博生涯即将结束，如今的我不仅体会到了做学问中“衣带渐宽终不悔，为伊消得人憔悴”的滋味，而且也收获到了心态、学识、眼界等各方面的成长。值此论文付梓之际，谨向所有鼓励过、支持过、帮助过我的老师、同学以及亲友表示衷心的感谢!

首先，我要诚挚感谢我的导师山世光教授。几年来，山老师无论在博士论文的选题还是在日常的工作进展讨论上都给予了我悉心的指导和无私的帮助，从而引领我进入了计算机视觉和模式识别研究领域，踏入了奇妙无穷的科学研究殿堂。在科研道路上，我的每一步成长都饱含了山老师无尽的教诲和全力的支持。特别是在论文写作方面，山老师都会耐心地指导我如何科学地撰写论文，经常不惜牺牲休息时间为我修改文章。另外，山老师广博的科学知识，敏锐的学术眼光，严谨求实的科学精神，永远是我学习的榜样。

真诚感谢陈熙霖教授为我们提供了高标准的科研条件和良好的生活环境。衷心感谢陈老师一直以来对我的殷切鼓励和悉心培养。当我在工作上遇到困难时总是能得到陈老师的支持与帮助。针对我在科学研究方法和表达能力上的一些不足和欠缺，他总是循循善诱并给我提出一些建设性的意见。正是在陈老师的悉心教导下，我才在科学研究的道路上得以快速成长并成熟起来。除此之外，陈老师对科研坚定、热情的精神不断影响和激励我奋发向上、不断进取，他宽广的视野、敏锐的见解、平易近人的教导让我受益终生。

衷心感谢王瑞平博士带领我进入本文的主要研究课题。在这些研究工作中，瑞平师兄至始至终参与了我的每一篇学术论文创新思路的讨论与修改，耐心指导我对每一次论文评审意见的斟酌与反馈。因此，我取得的每个学术成果都无不凝结着瑞平师兄的汗水与辛劳。瑞平师兄既是对我严格要求的良师也是与我志同道合的益友，特别感谢他一直以来对我的宽容与信任，为我创造了宽松的环境下让我不急不躁的稳步开展工作。他对科学研究的热爱、严谨的作风，求实的态度，勤奋的精神以及真挚的人格魅力带给我极为深远的影响。

还有更多的老师需要感谢。感谢实验室的黄庆明老师、蒋树强老师、常虹老师、苗军老师、柴秀娟老师给予的无私帮助和宝贵的意见，与各位老师的交流使我受益匪浅。他们无私奉献、兢兢业业的工作态度和丰硕的科研成果令我钦佩和敬仰。感谢办公室王晓彪老师、感谢蔡光辉老师、胡兰萍老师，他们的辛勤劳动为实验室提供了有序的研究环境，是我们能够安心科研的坚强后盾。感谢研究生部周世佳老师、李丹老

师、张平老师、李琳老师、冯刚老师和宋守礼老师在我入学、答辩、就业等方面给予我热情的帮助。

同时也感谢师兄师姐们。感谢谭庆师兄、肖欣延师兄、马丙鹏师兄、秦磊师兄、洪晓鹏师兄、王琪师兄、谢术富师兄、韩琥师兄、李安南师兄、郑伟师兄、马志国师兄、翟德明师姐、王丹师姐在研究过程中给予我的指导和帮助，他们的关心和建议，让我受益颇多。特别感谢阚美娜师姐和崔振师兄帮助我顺利度过科研入门阶段，他们俩对科研的热爱与追求、敏捷的思维方式、勤奋的工作态度都给我留下了深刻的印象；也特别感谢赵小伟师兄在我们一起合作的国际计算机视觉会议论文撰写过程中给予我实验设计和论文写作方面的指导，他那严谨缜密的写作风格对我后来的学术论文写作带来了深刻的影响。

感谢我的同年级博士生同学吉娜烨、王茜、李亮、刘洛麒、狄晓菲、陈振宇、陈荔城、李阳、尚田丰和颜成钢等，与他们的交流讨论使我受益匪浅，与他们一起度过了许多快乐而难忘的日子，与他们的友谊让我的生活丰富多彩。感谢我的师弟师妹李绍欣、褚令洋、李岩、王雯、刘梦怡、刘昕、张杰、尹芳、林宇顺、李显求、朱文涛、方振鹏、刘文献、邬书哲、李健超、谢广志、姜华杰、尹肖贻、王汉杰、宋松林、淮静、袁善欣、李瑞平、严灿祥、项翔、李浩、唐毅力、王占孔、严兴、缑丹、吴俊婷、李哲、黄煜、陈静、池晨、王崇秀、梁孔明等，他们为我的博士生涯提供了一个快乐的舞台和永久而美好的回忆。

感谢陪我一路走来的各位好友们詹志敏、王芳、宋国石、柯吓伟、洪金坤、王晓勇、宋志贤、林志钦、林智威、洪达君、赵志强、陈雪平、林渊灿、韩智雪、蒋绪团、林成竹、林武、黄志刚、龚元浩、谢源等，他们在精神上一直默默支持鼓励我，无论在我得意还是失意的时候都给予我最诚挚的帮助和开导，祝我们之间的友谊长存!

最后，我要特别感谢我的家人，在外求学漂泊时亲情永远是我最温柔的港湾和最重要的支撑。感谢我的父亲母亲一如既往地给予我精神上和物质上的鼎力支持和无私奉献! 感谢我的哥哥、嫂嫂和妹妹对我的理解和支持，帮助我分担家庭的责任! 感谢我的岳父岳母对我的包容和理解以及在生活学习上提供强有力的支持和帮助! 特别感谢我的妻子在这么多年来以自己深挚宽厚无私的爱充当我的精神支柱并默默支持和照顾我的学业和生活，使我在求学的道路上不畏艰辛坚持至今!

作 者 简 历

基本情况

姓名：黄智武　　**性别**：男　　**出生日期**：1985年3月　　**籍贯**：福建省莆田市

教育经历

2010年9月— 至今，中国科学院计算技术研究所，计算机应用技术专业，博士

2007年9月— 2010年7月，厦门大学，计算机软件与理论专业，硕士

2003年9月— 2007年7月，华侨大学，计算机科学与技术专业，学士

【攻读 博士 学位期间发表的论文】

[1] **Zhiwu Huang**, Shiguang Shan, Ruiping Wang, Haihong Zhang, Shihong Lao, Alifu Kuerban, Xilin Chen. A Benchmark and Comparative Study of Video-based Face Recognition on COX Face Database. IEEE Transactions on Image Processing (T IP), 2015. (Accepted)

[2] **Zhiwu Huang**, Ruiping Wang, Shiguang Shan, Xilin Chen. Face Recognition on Large-scale Video in the Wild with Hybrid Euclidean-and-Riemannian Metric Learning. Pattern Recognition (PR), 2015. (Accepted)

[3] **Zhiwu Huang**, Ruiping Wang, Shiguang Shan, Xianqiu Li, Xilin Chen. Log-Euclidean Metric Learning on Symmetric Positive Definite Manifold with Application to Image Set Classification. ACM International Conference on Machine Learning (ICML), 2015.

[4] **Zhiwu Huang**, Ruiping Wang, Shiguang Shan, Xilin Chen. Projection Metric Learning on Grassmann Manifold with Application to Video based Face Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[5] **Zhiwu Huang**, Ruiping Wang, Shiguang Shan, Xilin Chen. Learning Euclidean-to-Riemannian Metric for Point-to-Set Classification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. **Oral**.

[6] **Zhiwu Huang**, Ruiping Wang, Shiguang Shan, Xilin Chen. Hybrid Euclidean-and-Riemannian Metric Learning for Image Set Classification. Asian Conference on Computer Vision (ACCV), 2014.

[7] **Zhiwu Huang**, Xiaowei Zhao, Shiguang Shan, Ruiping Wang, Xilin Chen. Coupling Alignments with Recognition for Still-to-Video Face Recognition. IEEE International Conference on Computer Vision (ICCV), 2013.

[8] **Zhiwu Huang**, Shiguang Shan, Haihong Zhang, Shihong Lao, Alifu Kuerban, Xilin Chen. Benchmarking Still-to-Video Face Recognition via Partial and Local Linear Discriminant Analysis on COX-S2V Dataset. Asian Conference on Computer Vision (ACCV), 2012.

[9] **Zhiwu Huang**, Shiguang Shan, Haihong Zhang, Shihong Lao, Xilin Chen. Cross-view Graph Embedding. Asian Conference on Computer Vision (ACCV), 2012.

[10] Yan Li, Ruiping Wang, **Zhiwu Huang**, Shiguang Shan, Xilin Chen. Face Video Retrieval with Image Query via Hashing across Euclidean Space and Riemannian Manifold. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[11] Wen Wang, Ruiping Wang, **Zhiwu Huang**, Shiguang Shan, Xilin Chen. Discriminant Analysis on Riemannian Manifold of Gaussian Distributions for Face Recognition with Image Sets. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[12] J. Ross Beveridge, Hao Zhang, Bruce A. Draper, Patrick J. Flynn, Zhenhua Feng, Patrick Huber, Josef Kittler, **Zhiwu Huang****, Shaoxin Li**, Yan Li, Meina Kan, Ruiping Wang, Shiguang Shan, Xilin Chen, Haoxiang Li, Gang Hua, Vitomir Struc, Janez Krizaj, Changxing Ding, Dacheng Tao, P. Jonathon Phillips. Report on the FG 2015 Video Person Recognition Evaluation. IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2015. (**: equal contribution）

[13] Mengyi Liu, Ruiping Wang, Shaoxin Li, Shiguang Shan, **Zhiwu Huang**, Xilin Chen. Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild. ACM International Conference on Multimodal Interaction (ICMI), 2014.

[14] Mengyi Liu, Ruiping Wang, **Zhiwu Huang**, Shiguang Shan, Xilin Chen. Partial Least Squares Regression on Grassmannian Manifold for Emotion Recognition. ACM International Conference on Multimodal Interaction (ICMI), 2013.

【攻读 博士 学位期间参加的科研项目】

[1] 2014.1-至今，参与国家自然科学基金重大项目“图像视频的群体数据协同结构化表达与处理”

[2] 2012.11-至今，参与国家杰出青年基金项目“图像与视频处理”

[3] 2011.1-2012.10，参与国家自然科学基金项目“跨姿态人脸识别研究”

[4] 2010.9-2011.12，参与国家973计划课题“面向视频编码的视觉计算模型与方法研究”

【攻读 博士 学位期间的获奖情况】

[1] 2015年度，IEEE FG VPRE 2015 The First Position（共同第一参与人）

[2] 2014年度，国家奖学金

[3] 2014年度，中国科学院大学三好学生

[4] 2014年度，IEEE CVPR 2014 Doctoral Consortium Student Travel Grant

[5] 2014年度，ACM ICMI EmotiW 2014 Winner Award（主要参与人之一）

[6] 2013年度，中国科学院计算技术研究所所长奖学金特别奖

[7] 2013年度，IEEE ICCV 2013 Student Travel Grant

[8] 2013年度，ACM ICMI EmotiW 2013 Second Runner-Up Award（主要参与人之一）