第 1 章 绪论

随着 20 世纪 40 年代计算机出现，50 年代人工智能兴起，模式识别在 20 世纪 60 年代初迅速发展并成为一门新学科。本章介绍模式识别的基本概念与方法。

1.1 模式的基本概念

人们在日常生活中不时地对环境中的事物进行识别。比如，辨认出房子、道路、树木；辨识出汽车、电动车、自行车；看出熟人的面孔；听出电话中的熟人的声音；区分出马达声、喇叭声；闻出泄露的煤气味、变质食品的异味等。拿汽车的辨识作为一个例子，人通过反复观察各种形状与用途的汽车，包括轿车、客车、货车、救护车、消防车、警车、越野车、跑车、房车等等，可以学习得到汽车的性质和特点，具备准确辨识汽车的能力。

一般地，把自然界或社会生活中的相同或相似的事物称为模式。在对个别的具体事物实例进行观察的基础上，人可以获得对此类事物整体性质和特点的认识，从而具备正确辨认此类事物的能力，即具备模式识别学习能力。人类的模式识别能力是视觉、听觉、嗅觉等感知能力的基础，使得人类能够看似比较轻松地完成各种感知认知任务。

本书讨论的模式识别就是用计算机实现人类的模式识别能力，也称为计算机模式识别或机器识别。目前，我们对人脑的模式识别过程尚不完全清楚，让计算机做人类较容易做到的模式识别还是非常困难的，计算机模式识别能力在多数方面还远不如人类。人脑的模式识别过程研究将有利于计算机模式识别模型与算法的研究，反之亦然。

在模式识别学科中，“模式”可以理解为一种相同或相似事物，即模式类；也可以理解为对具体的个别事物进行观测所得到的观测数据，即样本。本质上，模式是指对一种相同或相似事物进行大量观测所得到的数据所具有的性质和特点，即模式分布。以手写体数字图像识别问题为例，10 个数字对应着 10 种模式类，图 1-1-1 显示了手写体数字 0 的 200 幅图像样本。从图 1-1-1 可以看到，图像样本的变化复杂，模式分布研究具有挑战性。

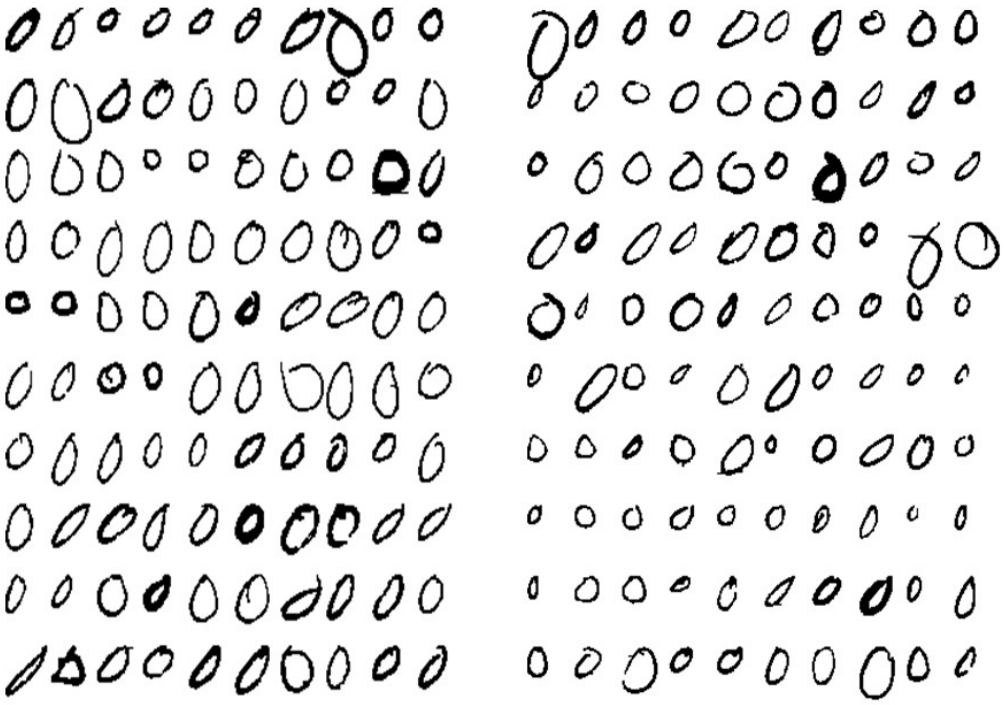

图 1-1-1　手写体数字图像样本示例

一个统计模式识别系统可以分成两个阶段：一是训练阶段，输入的样本包含了样本数据及其相应的类别

真实标签，从样本数据中深入挖掘模式分布，通过最小化训练样本的标签预测与真实标签的预测误差，学习得到在某种意义下最优的模式表示与模式分类器，见图 1-1-2 中的实线部分与虚线部分；二是识别阶段，输入的待识别样本只包含了样本数据，按训练阶段所学习得到的最优模式表示与模式分类器对样本做标签预测，即将其归入一个模式类，见图 1-1-2 中的实线部分。

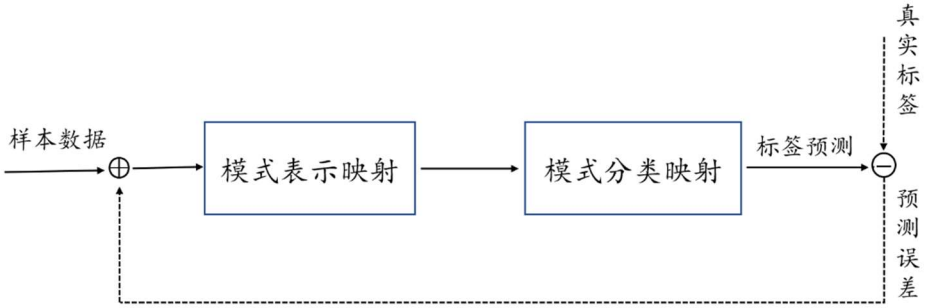

图 1-1-2　一个统计模式识别系统示意图

在图 1-2 中，输入的样本数据，都是指通过传感器对模式所获取的观测空间中的数据点，并进行了预处理；模式表示是从观测空间到表示空间的一个映射，模式分类是从表示空间到标签预测空间的一个映射。模式识别基础研究主要集中在对于训练阶段的研究，可以看成为需要求解最优模式表示映射与最优模式分类映射的联合优化问题。BP 神经网络是该联合优化问题的一个经典示例，常见的模式识别研究工作包括最优模式表示映射研究与最优模式分类映射研究。

1.2 模式表示学习

人脑在每天感知外界事物的过程中，首先面临着一个模式表示的问题，即利用大约 3 万听觉神经纤维和 100 万视觉神经纤维等，从高维的感官输入信息（如视频图像、音频信号等）中抽取便于管理的很小数量的感知关联特征来完成对事物的认识。模式表示问题可以定义为：在高维的观察样本空间中寻找其隐藏的有意义的低维数据结构，并藉此分析和探索事物的内在规律。因此，模式表示的过程本质上可以看成在一定准则下的优化问题，不仅从原始模式信息中得出了最有利于模式分类的特征，而且这些特征与原始样本信息相比，极大地降低了模式样本的维数。近年来，大规模的高维数据在科学、工程和社会生活中激增，这必然为众多学科领域（如图像视频和多媒体数据的处理、网络数据分析和搜索、生物信息学和测量学、智能机器人等）的发展带来机遇和挑战。在这些领域里，数据的维数成千上万甚至高达数亿，样本的个数也达到了几乎同样的数量级，一方面含有丰富的信息可供挖掘利用，另一方面增加了挖掘这些数据的困难与成本，模式表示问题尤显重要，亟待解决。

模式表示包含了模式特征抽取与特征选择。假设在模式观测样本空间中有一个样本数据集，其中每一个样本数据都是一个高维向量。因为这些高维向量的各维度之间可能存在一些信息冗余，所以模式可能分布在一些低维子空间上。为了挖掘模式分布，需要寻找比较好的模式表示空间，即需要寻找从高维观测空间到低维表示空间的一个映射，在实现降维的同时，使新的模式表示具有一定的最优性与可解释性。

最简单的映射是线性映射、投影映射。按向子空间投影的模式表示映射是否线性可以把最优模式表示方法简单分成线性子空间方法、非线性子空间方法。在样本集中，可能对于每个样本数据，都没有类别标记信息；也可能对于每个样本数据，都有类别标记信息；或者，对于部分样本数据，都有类别标记信息，但对于其它样本数据，都没有类别标记信息。因此，按样本集中类别标记信息的多少，可以把最优模式表示方法分成非监督方法、监督方法、半监督方法。

1.2.1 线性子空间分析

主成分分析（Principal Component Analysis, PCA）是一种广泛应用的非监督子空间学习方法[1]，旨在寻找具有最大方差的模式表示投影方向，即在投影后所得到的新特征的方差最大。方差大小刻画了数据的散布情况，即描述了数据偏离均值的散布程度情况。如果方差小，就说明样本在该维度上偏离均值的变化

小，将不利于描述数据在该维度上的可区分性。在求出方差最大的第一个投影方向后，按正交条件可以定义方差最大的第二个投影方向。以此类推，可以定义方差最大的一组投影方向，其中后续的投影方向与已有的投影方向正交，并且方差最大。这组方差最大的投影方向集就被称为主成分向量集。容易证明，它是样本数据的协方差矩阵对应于大本征值的本征向量集，并且也可以看成为在正交约束下的一个最小重构误差问题的解集。

线性鉴别分析（Linear Discriminant Analysis，LDA）是经典的监督子空间学习方法[2-4]，旨在寻找具有最大鉴别力的模式表示投影方向。线性鉴别分析的基本思想最早是由Fisher提出[2]，其目的是选择使得Fisher准则函数达到极值的向量作为最佳投影方向，从而使得样本在该方向上投影后，达到最大的类间离散度和最小的类内离散度，即得到最大的鉴别力。在求出鉴别力最大的第一个投影方向后，按共轭正交条件可以定义鉴别力最大的第二个投影方向[5]。以此类推，可以定义鉴别力最大的一组投影方向，其中后续的投影方向与已有的投影方向共轭正交，并且鉴别力最大。这组鉴别力最大的投影方向集就被称为具有统计不相关性的最佳鉴别向量集。容易证明，它就是样本数据的类内协方差矩阵之逆矩阵与类间协方差矩阵的乘积矩阵对应于大本征值的本征向量集，即它与Wilks[3]和Duda[4]在Fisher思想的基础上分别提出的使得类间散布矩阵行列式与类内散布矩阵行列式之比最大的线性鉴别向量集是等价的。在模式表示理论中，要求表示之间尽可能是不相关的，其出发点是有利于最大限度地消除维度之间的相关性。如果参考主成分分析的做法，在要求后续的投影方向与已有的投影方向正交的条件下求解鉴别力最大的投影方向所得到的最优鉴别向量集在性能上逊色于经典的线性鉴别分析最优鉴别向量集。

Fisher 线性鉴别分析有很多拓展[6]。在模式空间维数高而训练样本数少的情况下，很难精确地估计协方差矩阵，类内协方差矩阵的逆矩阵还可能不存在，子空间分析方法遇到了困难，小样本子空间分析方法研究引起了人们的高度关注[7]。由于常规的子空间学习技术都是针对向量数据而言的，对于计算机视觉中的图像数据矩阵，简单的处理方法是将图像的各个象素数据叠加形成一个向量数据，其维数相当高，小样本问题普遍出现。另外，将图像矩阵转化向量，只能部分保持图像像素的邻近关系，难以保持图像的结构信息。因此，催生了面向图像数据的子空间学习技术研究，直接利用图像矩阵数据构造图像协方差矩阵，不但保持了图像的结构信息，还提高了算法的效率，其代表性工作包括二维主成分分析[8]、二维线性鉴别分析等等。

1.2.2 基于流形、稀疏与低秩假设的模式表示

在数据表示建模时，必须诉诸于数据固有的结构。认知科学的研究提供了假设与启迪，引领着模式表示理论与方法的发展。有三种假设占主导地位：（i）流形假设，即数据存在于内嵌的低维流形上，该假设导致了流形学习理论的产生和蓬勃发展；（ii）稀疏假设，即数据在超完备基底上的表示是稀疏的，该假设带来了稀疏学习、稀疏表示和压缩感知理论的兴起；（iii）低秩假设，即数据矩阵在代数空间本质上具有比较低的秩,在国际著名学者斯坦福大学 Candes 教授推动下，一种崭新的理论，即低秩矩阵恢复理论，应运而生并势如破竹地发展。

理论与实验都证明，复杂模式的特征之间往往存在着高阶的相关性，因此数据集呈现明显的非线性性，并且往往是由一组维度远远低于样本维度的隐含变量决定的。在数学上，具有上面性质的模型是流形。以流形为模型，利用流形的基本假设和性质来研究高维空间中的数据分布，达到简约数据，降低维度，探寻复杂模式的内部规律的学习方法被称为流形学习[9-10]。2000 年，Seung 提出感知以流形方式存在[9]，并通过实验证明了人脑中的确存在着稳态的流形。这为统计模式识别与人类感知架起了一座桥梁，使得流形学习具有了更加坚实的理论基础。流形学习的核心，是如何合理有效地对数据进行流形建模，即如何找到一个好的流形模型，能够较好地逼近数据，使得数据的内在结构性质能够在流形上良好地保持下来，以便研究者通过对流形模型的研究，获得对数据集内在结构的深刻认识。

科学工作者们揭示了在低层和中层的人类视觉系统中，视觉通道中的许多神经元对大量的具体的刺激，比如目标的颜色、纹理、朝向和尺度等，具有选择性[11-12]。若将这些神经元视为视觉阶段的超完备集中的信号基元的话，神经元对于输入图像的激活机制具有高度的稀疏性。图像模式的稀疏性不仅体现在模式内部，也体现在模式之间。模式内部的稀疏性刻画为特征抽取提供了依据，模式之间的稀疏性则为分类器的设计提供了可能。

矩阵低秩性刻画了数据的内在低维结构，揭示了真实数据的变化往往由少数的重要因子线性决定。低秩

假设源远流长，从最早的主成分分析到线性鉴别分析，几乎每一种子空间分析方法都在她的辐射下诞生。传统的子空间方法对含有小高斯噪音的数据比较有效，但对于含野点、含大的稀疏噪音的数据比较敏感。对于大规模的高维数据，数据矩阵中常常含有野点、孤立点，有部分元素受噪声污染、甚至缺失，低秩矩阵恢复理论与方法成为有效的数据处理手段[13-14]。主成分分析、矩阵补全和稳健主成分分析[14]可分别看作稠密小噪声、缺失数据和稀疏大噪声假设下的低秩矩阵学习模型。针对日益涌现的多模态数据，作为低秩矩阵学习模型的多线性拓展，低秩张量学习模型、低秩张量关联主成分分析、张量恢复的快速求解以及面向轻量级神经网络设计的张量分解方法等研究正在蓬勃发展。

1.3 模式分类

设有 c 个模式类别：$\omega_1, \omega_2, \cdots, \omega_c$，$X$为一个待识别样本，模式分类问题就是将 X 映射到决策空间上,即将其分类到一个最相近或相似的模式类别。

1.3.1 贝叶斯分类器

设各模式类别的先验概率与条件分布密度函数分别为$P(\omega_i)$与$P(X|\omega_i)$ $(i = 1, \cdots, c)$。对于待识别样本X，计算出样本X属于各模式类的后验概率$P(\omega_i|X)$，最小错误率 Bayes 判决规则根据概率最大原则进行分类判决如下：

$$P(\omega_j|X) = \max_{1 \le i \le c} P(\omega_i|X) \rightarrow X \in \omega_j \tag{1.3.1}$$

在$c = 2$的情形下，假设条件分布服从正态分布，当两个条件协方差矩阵不相等时，最小错误率 Bayes 判决规则所对应的判决函数是一个二次函数，如图 1-3-1 中的实线所示；当两个条件协方差矩阵相等时，最小错误率 Bayes 判决规则所对应的判决函数是一个一次函数，其函数曲线为两个分布中心连线（如图 1-3-1 中的水平虚线）的垂直平分线（如图 1-3-1 中的竖直虚线），即各类以其中心为代表，按待识别样本到各类代表的距离最小原则进行分类。

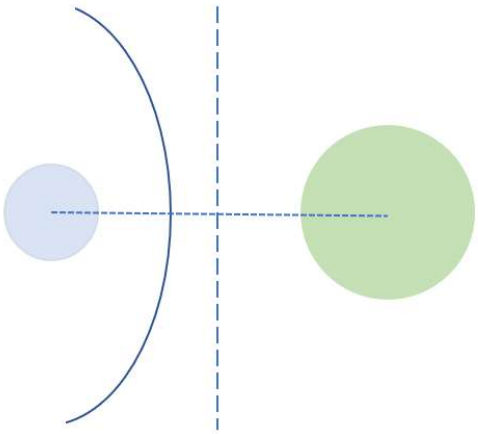

图 1-3-1　两类问题的 Bayes 判决函数曲线示意图

1.3.2 最小距离分类器

在实际应用中，类条件概率密度是未知的，需要大量样本进行估计。为此，我们可以利用样本集直接设计分类器。

将ω_i类模式的训练样本记为X_{ij} $(j = 1, \cdots, M_i)$，其样本均值记为$m_i = \frac{1}{M_i}\sum_{j=1}^{M_i} x_{ij}$。最小距离分类器对于各模式只选取一个代表$m_i$，对于待识别样本$X$，计算出样本$X$与各模式代表$m_i$的欧氏距离$\|X - m_i\|$，根据距离最小原则进行分类判决如下：

$$\|X - m_i\| = \min_{1 \le i \le c} \|X - m_i\| \rightarrow X \in \omega_i \tag{1.3.2}$$

在$c = 2$的情形下，最小距离分类器的判决函数曲线就是两个样本均值连线的垂直平分线。

欧氏距离刻画了模式样本之间的相似程度，样本相关系数也是能够刻画模式样本之间相似程度的一种度量方法。采用样本相关系数代替欧氏距离，可以得到最大相关分类器。最大相关分类器对于各模式只选取一个代表m_i，对于待识别样本X，计算出样本X与各模式代表m_i的样本相关系数$\rho(X, m_i)$，根据相关系数最大原则进行分类判决如下：

$$\rho(X, m_i) = \max_{1 \le i \le c} \rho(X, m_i) \rightarrow X \in \omega_i \tag{1.3.3}$$

1.3.3 最近邻分类器

最近邻距离分类器也是按距离最小原则分类，但各模式不是只选一个代表，而是把其所有训练样本都作为代表。对于待识别样本X，计算出样本X与各模式代表X_{ij}的欧氏距离$\|X - X_{ij}\|$，进行分类判决如下：

$$\|X - X_{ij}\| = \min_{\substack{1 \le i \le c \\ 1 \le j \le M_i}} \|X - X_{ij}\| \rightarrow X \in \omega_i \tag{1.3.4}$$

最近邻相关分类器也是按相关系数最大原则分类，但各模式不是只选一个代表，而是把其所有训练样本都作为代表。对于待识别样本X，计算出样本X与各模式代表X_{ij}的样本相关系数$\rho(X, X_{ij})$，进行分类判决如下：

$$\rho(X, X_{ij}) = \max_{\substack{1 \le i \le c \\ 1 \le j \le M_i}} \rho(X, X_{ij}) \rightarrow X \in \omega_i \tag{1.3.5}$$

1.3.4 BP 神经网络

在 1986 年 Rumelhart[16-17]提出了一种按误差反向传播（Back-propagation， BP）算法训练的三层前馈网络，简称之为 BP 神经网络，其拓扑结构包括了输入层、隐含层和输出层，见图 1-3-2。

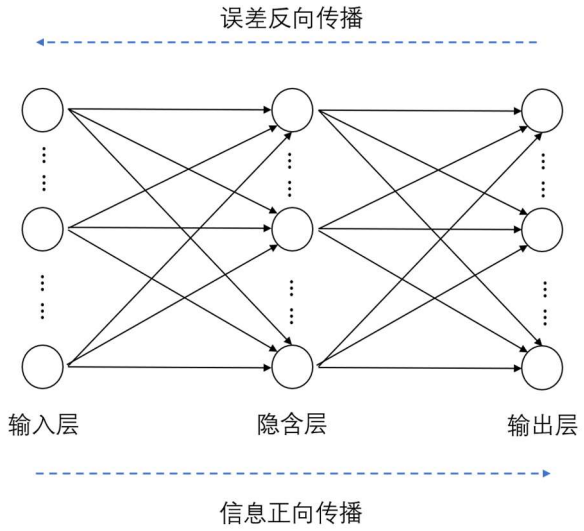

图 1-3-2　三层 BP 神经网络结构示意图

在图 1-3-2 中，信息正向传播通过全连接线性组合的激活函数分别计算隐含层的输出、输出层的输出，而误差反向传播可以利用输出层神经元的误差调节隐含层-输出层的全连接参数与阈值，并利用隐含层神经元的误差调节输入层-隐含层的全连接参数与阈值。BP 神经网络可以逼近任意连续函数，具有很强的非线性映射能力。近几年出现的深度神经网络可以看成是 BP 神经网络的发展。

1.3.5 支持向量机

支持向量机是上世纪末九十年代在统计学习理论的基础上发展起来的模式识别方法，在解决小样本高维模式识别问题中表现出特色与优势[18-19]。

给定一个二分类问题，假设在模式表示空间用一个线性分类映射$\Psi(X) = w^T X + b$可以将训练样本集中的两个类无误地分开，即假设训练样本集是线性可分的。不妨将分类映射$\Psi(X)$规范化，使得对于所有的训练样

本都满足$|\Psi(X)| \geq 1$。此时，可以说明分类间隔为$2/\|w\|$。经典的支持向量机是如下优化问题的解：

$$\min_{w,b} \frac{1}{2}\|w\|^2 \ \text{s.t.} \ y_i(w^T X_i + b) - 1 \geq 0 \ (\mathrm{i} = 1, \cdots, M) \tag{1.3.6}$$

其中，$y_i \epsilon \{1, -1\}$是训练样本X_i的类别标签$(\mathrm{i} = 1, \cdots, M)$。超平面$|\Psi(X)| = 1$上的训练样本被称为支持向量。$\|\mathrm{w}\|^2$越小，分类间隔越大，从而支持向量机的泛化能力越强。针对非线性可分的情况，对于每一个训练样本可以引入松弛变量$\xi_i \geq 0$以确定软边界，优化问题为

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M}\xi_i \ \text{s.t.} \ y_i(w^T X_i + b) - 1 + \xi_i \geq 0 \ (\mathrm{i} = 1, \cdots, M) \tag{1.3.7}$$

其中C为权值。对于多分类问题，支持向量机有很多解决方案。

1.3.6 分类器组合

不同性质的特征往往反映模式的不同方面，在一种特征空间很难区分的两种模式可能在另一种特征空间上很容易分开；而对应于同一特征的不同分类器又从不同的角度将该特征映射到决策空间上，因此利用不同性质特征和不同分类器的组合就可能全面反映出一个模式，从而得到一个较好的分类结果。

多传感器数据融合可以有效地解决单传感器的不足，更精确地观察和解释环境，有着重要的理论研究价值和实际应用价值。从处理对象的层次上分，多源信息融合研究可分为低层（象素级）、中层（特征级）和高层（决策级）三个层次。从数据融合的角度看，可以认为多分类器的组合属于决策级一层的信息融合。

分类器组合方法很多[15]，主要有择多判决法（投票表决法、计分法），线性加权法，贝叶斯估计法，证据推理法，模糊推理法，以及将分类结果作为一种新的输入特征的神经网络组合方法。

1.4 应用算例

模式识别在人工智能、计算机视觉、大数据分析、智能机器人等领域具有广泛的应用。本节仅介绍在经典的手写体数字图像识别与人脸图像识别应用中的算例。

1.4.1 手写体数字图像识别

手写体数字识别是计算机模式识别领域的一个经典课题。国际上广泛使用的 Concordia University CENPARMI 手写体数字数据库有 4000 个训练样本与 2000 个测试样本。图 1-1-1 显示了数字 0 的 400 个训练样本中的前 100 个与 200 个测试样本中的前 100 个。

在手写体数字识别的研究中，对于二值化图像的表示方法有多种。通过对手写体数字图像作预处理，提取出如下四组图像表示[20-21]：Gabor 变换特征 X^G、Legendre 矩特征 X^L、Pseudo-Zernike 矩特征 X^P 与 Zernike 矩特征 X^Z。对于这四种图像表示，分别进行线性鉴别分析提取出 9 维的鉴别特征表示，分别记为 X^G+LDA、X^L+LDA、 X^P+LDA、X^Z+LDA。分别按最小距离、最近邻距离、最大相关、最近邻相关四种分类器对测试样本进行分类识别，识别错误率结果见表 1-4-1。

从表 1-4-1 可以看出，在 X^G、X^L、 X^P、X^Z、X^G+LDA、X^L+LDA、 X^P+LDA、X^Z+LDA 八种图像表示中，基于 Legendre 矩特征的图像表示具有优势；进行线性鉴别分析后，在分类阶段，如果每个类别只取一个代表，性能极大提升；如果将所有训练样本都取为类别代表，性能可能会稍有下降。

表 1-4-1 手写体数字识别实验的错误率

图像表示	维数	分类器			
		最小距离	最近邻距离	最大相关	最近邻相关
X^G	256	0.269	0.179	0.297	0.144
X^G+LDA	9	0.183	0.175	0.24	0.212
X^L	121	0.479	0.088	0.262	0.067

X^L+LDA	9	0.106	0.108	0.147	0.141
X^P	36	0.429	0.237	0.485	0.241
X^P+LDA	9	0.296	0.299	0.362	0.387
X^Z	30	0.449	0.245	0.517	0.254
X^Z+LDA	9	0.299	0.308	0.383	0.384

1.4.2 人脸图像识别

人脸识别是信息技术领域的前沿课题[22]。国际上广泛使用ORL人脸图像数据库，由40人的400幅92×112图像所组成，其中有些图像是拍摄于不同时期；人脸脸部表情与脸部细节有变化，例如：笑或不笑，眼睛睁着或闭着，带或不带眼镜；人脸姿态有变化，深度旋转与平面旋转可达到20°；人脸的尺度也有最多10%的变化。

(a) 92×112

(b) 46×56

(c) 23×28

(d) 12×14

(e) 6×7

(f) 3×4

图 1-4-1　多分辨率人脸图像

将图像分辨率92×112逐渐降低到46×56、23×28、12×14、6×7、3×4，可以得到不同分辨率图像，

见图 1-4-1。

现将每人 10 幅图像样本中的前 5 幅图像用作训练样本，另 5 幅图像用着测试样本，则训练样本总数为 200，测试样本总数也为 200。对于多种分辨率图像92×112、46×56、23×28、12×14、6×7、3×4的像素数据表示，分别按最小距离、最近邻距离、最大相关、最近邻相关四种分类器对测试样本进行分类识别，识别错误率结果见表 1-4-2 中的前 6 行。

对于分辨率6×7、3×4 的图像，像素数目分别为 42 与 12，而训练样本数目为 200，所以可以进行线性鉴别分析，分别抽取 39 维与 12 维的线性鉴别特征，并按最小距离、最近邻距离、最大相关、最近邻相关四种分类器对测试样本进行分类识别，识别错误率结果见表 1-4-2 中的后 2 行。另外说明一下，对于分辨率13×14的图像，像素数目为 168，而类内协方差矩阵的秩为训练样本数目-类别数目-1=201-41-1=159<168，其逆矩阵不存在，不能进行线性鉴别分析。

表 1-4-2 在多种分辨率下的人脸识别实验错误率

图像表示	维数	分类器			
		最小距离	最近邻距离	最大相关	最近邻相关
92×112	10304	0.16	0.09	0.205	0.12
46×56	2576	0.155	0.09	0.19	0.11
23×28	644	0.16	0.09	0.205	0.12
12×14	168	0.145	0.08	0.195	0.085
6×7	42	0.175	0.1	0.21	0.095
3×4	12	0.29	0.14	0.365	0.195
3×4+LDA	12	0.17	0.125	0.235	0.145
6×7+LDA	39	0.055	0.03	0.07	0.025

从表 1-4-2 可以看出，高分辨率图像像素之间的信息冗余严重，适当降低图像分辨率可以保持比较接近的识别性能；把图像分辨率降得过低，会影响识别性能；线性鉴别分析能够极大地提升识别性能，采用最近邻距离分类器或最近邻相关分类器，基于6×7人脸图像的 LDA 表示可以获得不少于 97%的正确率。

参考文献

[1] K. Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press, New York, 1990.

[2] R. A. Fisher. The use of multiple measurements in taxonomic problems. Annals Eugenics, 7:179-188, 1936.

[3] S. S. Wilks. Mathematical Statistics. Wiley, New York, 1962, pp. 577-578.

[4] R.O. Duda, P.E. Hart. Pattern Classification and Scene Analysis. Wiley, New York, 1973.

[5] Z. Jin, J. Y. Yang, Z. M. Tang, Z. S. Hu. A theorem on the uncorrelated optimal discriminant vectors. Pattern Recognition, 34(10):2041-2047, 2001.

[6] Ian D Longstaff. On extensions to Fisher's linear discriminant function. IEEE Trans. Pattern Anal. Machine Intell., 1987, 9(2): 321-324.

[7] Yang J, Frangi A F, Yang J. -y, et al. KPCA plus LDA: a complete kernel fisher discriminant framework for feature extraction and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(2): 230-244.

[8] Yang J, Zhang D, Frangi A F, et al. Two dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Pattern Analysis and Machine Intelligence, 2004, 26(1): 131-137.

[9] Seung H S, Lee D D. The manifold ways of perception. Science, 2000, 290: 2268-2269.

[10] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290: 2323-2326.

[11] Vinje W E, Gallant J L. Sparse coding and decorrelation in primary visual cortex during natural vision. Science, 2000, 287(5456):

1273-1276.

[12] Olshausen B A, Field D J. Sparse coding of sensory inputs. Current Opinion in Neurobiology, 2004, 14(4): 481-487.

[13] Candès E J, Tao T. The power of convex relaxation: Near-optimal matrix completion. IEEE Transactions on Information Theory, 2010, 56(5): 2053-2080.

[14] Candès E J, Li X, Ma Y, et al. Robust principal component analysis? Journal of the ACM, 2011, 58(3): Article No.11.

[15] Kittler J et al. On combining classifiers. IEEE Trans. Pattern Anal. Machine Intell., 1998, 20(3):226-239.

[16] David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams. Learning representations by back-propagating errors. Nature, 323, 533–536, 1986.

[17] David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams. Learning internal representations by error propagation. In D. Rumelhart & J. McClelland, (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 1, pp. 318–364). Cambridge, MA: The MIT Press. Foundations, 1986.

[18] V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. Automation and Remote Control, 24, 1963, pp. 774-780.

[19] V. N. Vapnik, S. Golowich, A. Smola. Support vector method for function approximation, regression estimation, and signal processing. in M. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems, 9, Cambridge, MA, MIT Press, 1997, 281-287.

[20] Zhong Jin, Jingyu Yang, Zhongshan Hu, Zhen Lou. Face recognition based on the uncorrelated discriminant transformation. Pattern Recognition, 2001, 34(7):1405-1416.

[21] Liao S X, Pawlak M. On image analysis by moments. IEEE Trans. Pattern Anal. Machine Intell., 1996, 18(3):254-266

[22] Rama Chellappa, Charles L. Wilson, and Saad Sirohey. Human and machine recognition of faces: A survey. Proceedings of the IEEE, 1995,83(5):705-740