

# 实验报告四——人脸识别

21122897 王子棠 21122895 蒋鑫怡 21122896 徐蓓蓓

**摘要：**本实验使用 ORL 人脸数据集，通过数据预处理、PCA 降维、KNN 训练实现了人脸识别，并在此过程中探究各参数对准确率的影响，不断提高准确率。首先，利用 PCA 对训练集图像进行特征提取和降维，为确定最佳主成分数，本实验通过两种方法分别计算不同主成分数量下的预测准确率和累积可解释方差贡献率。接着，使用 KNN 算法对降维后的特征进行训练，利用 KNN 来评估上一阶段中得到的 4 个不同最佳主成分数量下的预测准确率。通过上述部分，得到猜想：在降维维度为 24，k 值为 1 时，准确率最高。然后，为了验证猜想，我们枚举了所有降维维度（1-320）与 k 值（1-10）的排列组合，计算其准确率，并使用三维散点图展示主成分数量、k 值和准确率的关系。最后得到结论，猜想正确，最佳配置模型的测试集准确率为 0.9625。

**关键词：**人脸识别，PCA，KNN

## 1. 项目背景和意义

人脸识别是一种重要的生物特征识别技术，具有广泛的应用，包括安全监控、身份验证、访问控制等领域。随着科技的进步，人脸识别技术在社会生活中发挥着越来越重要的作用。

主成分分析（PCA）作为一种常用的降维技术，可以有效地降低数据维度，减少计算量和存储空间，并且可以保留大部分的信息，有助于加快后续分类算法的速度。K 近邻（KNN）作为一种基本的分类算法，简单易实现，对于小规模数据集效果良好，并且对数据分布没有假设，适用于各种类型的数据。

本实验使用 PCA+KNN 算法实现人脸识别，同时设计实验寻找最佳主成分数量、k 值，并总结各参数对准确率的影响，不断提高准确率。通过本实验，我们加深了对于 PCA 和 KNN 算法的原理，优势和局限性的理解，以及设计实验以提高准确率的动手能力和对于实验结论的分析总结能力。

## 2. 项目实施过程及结果分析

### 2.1 数据预处理

数据预处理阶段是为了准备数据以便后续的特征提取和模型训练。我们的预处理阶段主要包括文件重命名、字典存储像素和标签信息、大小归一化等操作。

为解决原始文件名包含多余字符、排列顺序混乱、未明确分组等问题，我们采取文件重命名等操作，使图片从 x\_1.bmp~x-10.bmp 按顺序排序。随后将读取的像素信息转化为 numpy 数组，将像素信息、标签信息存储在字典里，方便后续的访问和使用。接着，调用 `normalize(matrix)` 函数实现了对矩阵的归一化处理，使得每个行向量都变成了单位向量，加速算法收敛，消除量纲影响。

### 2.2 划分数据集

本次实验用到的数据集为 ORL 人脸数据集。共包含 40 个不同人的 400 张图像，是在 1992 年 4 月至 1994 年 4 月期间由英国剑桥的 Olivetti 研究实验室创建。此数据集下包含 40 个目录，每个目录下有 10 张图像，每个目录表示一个不同的人。所有的图像是以 PGM 格

式存储，灰度图，图像大小宽度为 92，高度为 112。

划分数据集时，我们按照 80%训练集，20%测试集的比例收取了训练集 320 张，测试集 80 张。为实现训练集与测试集中均有 40 个类别的人脸照片，我们没有直接从 400 张照片中随机抽取，而是从每个类别中选取 80%的照片放入训练集。即每个类别在训练集中的样本数量为 8，每个类别在测试集中的样本数量为 2。每个类别中的样本按照图片在输入数据中的顺序抽取。

## 2.3 利用 PCA 对训练集图像进行特征提取和降维

PCA 是一种常用的降维技术，它通过线性变换将高维数据转换为低维数据，同时保留数据的主要特征。在这个阶段，我们利用 PCA 对训练集图像进行特征提取和降维。为确定主成分数，本实验通过两种方法分别计算不同主成分数量下的预测准确率和累积可解释方差贡献率来得到最佳主成分数。

### 2.3.1 计算预测准确率求最佳主成分数

第一种方法的思路是用训练集和测试集之间的余弦相似度作为准确度，基于相似度进行样本分类，记录不同主成分数下的准确率。这种做法的优势在于不会引入其他超参数，避免调参。具体做法为：首先导入了 `cosine_similarity` 函数用于计算余弦相似度。然后，设定了主成分数量为 320，并创建了一个空字典 `component_acc_dict` 用于存储不同主成分数量下的预测准确率。接着，通过一个循环遍历不同的主成分数量。在每次迭代中，使用 PCA 模型对训练数据进行拟合，并将训练数据转换为主成分空间。然后对训练集和测试集之间的主成分数据进行归一化处理。接下来，利用 `cosine_similarity` 函数计算训练集和测试集之间的余弦相似度，找到与每个测试样本最相似的训练样本的索引。然后根据这些索引获取训练样本的标签，作为对测试样本的预测结果。最后，计算预测准确率，并将其存储在 `component_acc_dict` 字典中，以主成分数量为键，准确率为值。

由以上方法绘制而成的折线图如下：

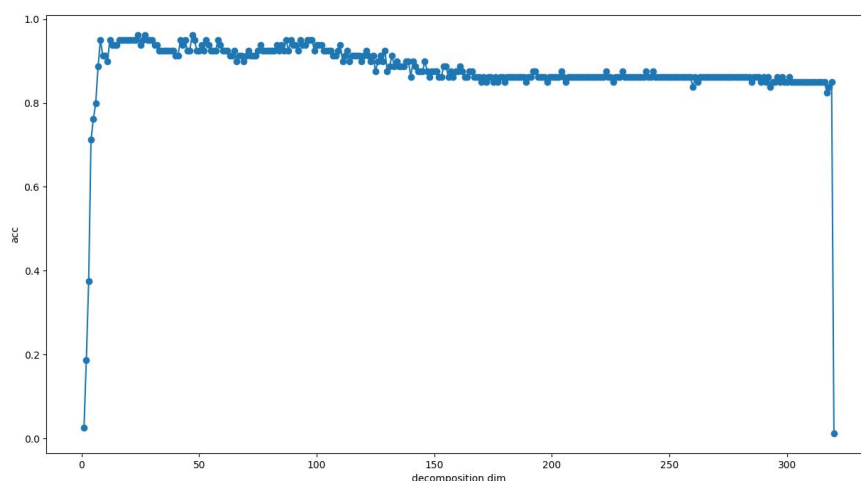


图 2 不同主成分数下的预测准确率折线图

如图 2 所示，我们发现最大准确率时的维度值即主成分数为 24，即当主成分数量为 24 时，模型在测试数据集上达到了最高的预测准确率。

2.3.2 通过累积可解释方差贡献率求最佳主成分数

这种方法的思路是利用 PCA 模型拟合训练数据，然后分析主成分的累计解释方差比率。首先，创建一个 PCA 对象并使用 fit 方法对训练数据进行拟合，得到主成分。接着，获取主成分的解释方差比率，并计算累计解释方差比率。然后，绘制累计解释方差比率与主成分个数的关系图，横坐标表示主成分个数，纵坐标表示累计解释方差比率。最后，设定一个阈值，通过自动找到第一个超过该阈值的主成分数，确定最佳的主成分数量，并将其打印出来。

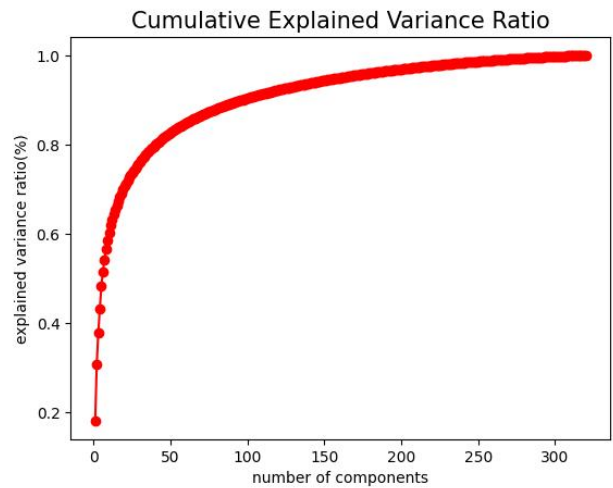


图 3 累计解释方差比率与主成分个数关系图

根据关系图，我们得到如下数据：

阈值	主成分数
0.95	161
0.9	97
0.85	61

表 1 阈值与主成分数关系表

2.4 使用 KNN 对降维后的特征进行训练

KNN (K-最近邻) 是一种简单有效的分类算法，它基于特征空间中样本之间的距离来进行分类。在这个阶段，我们使用 KNN 算法对降维后的特征进行训练。通过 2.3 中两个方法，我们得到了不同的最佳主成分数，因此我们不妨利用 KNN 来评估这 4 个不同主成分数量下的预测准确率。

首先，设置好 4 个不同的主成分数量（24、61、97、161），然后通过循环遍历这些主成分数量。在每次迭代中，对训练数据进行 PCA 降维处理，并使用 K 最近邻分类器进行训练。在训练过程中，通过尝试不同的近邻数量来评估模型在测试数据上的预测准确率。最后，绘制了近邻数量与准确率之间的关系图，并通过图例展示了不同主成分数量对应的曲线。使用降维后的训练数据和对应的标签来训练 KNN 模型。

由折线图我们得到，随着 K 值的增加，分类器对图片的类别判断会带来更大误差。同时我们可以初步看到，主成分数，即降维维度为 24 时，总体准确率相对较优。因此我们猜想，在降维维度为 24，k 值为 1 时，准确率最高。

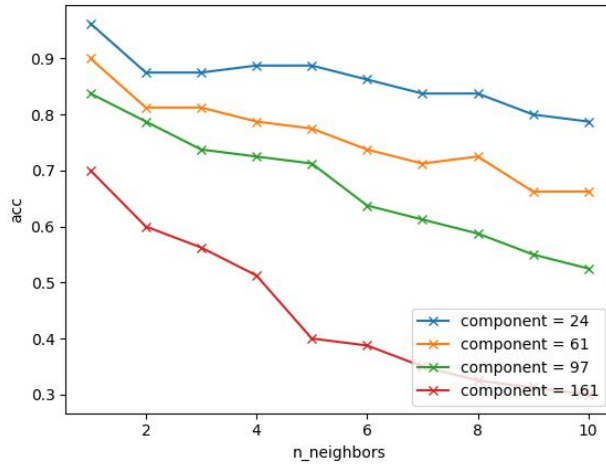


图 4 4 种主成分数量下 k 值从 1-10 的分类器准确率

## 2.5 枚举法验证猜想

在 2.4 中，我们猜想，在降维维度为 24，k 值为 1 时，准确率最高。为验证猜想，我们将所有降维维度（1-320）与 k 值（1-10）的排列组合枚举出来，计算其准确率，并使用三维散点图展示主成分数量、k 值和准确率的关系。

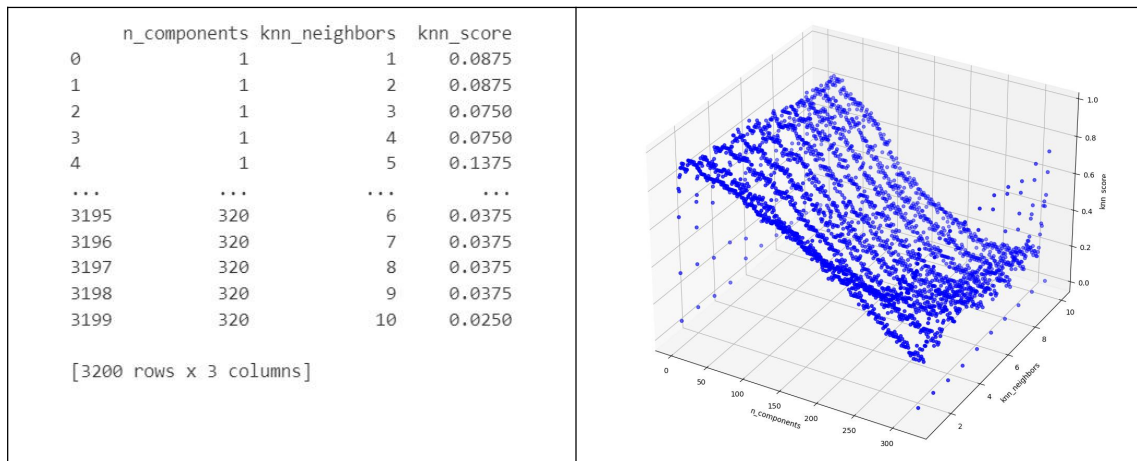


图 5 主成分数量、k 值和准确率的三维散点图

	n_components	knn_neighbors	knn_score
180	19	1	0.9625
190	20	1	0.9625
210	22	1	0.9625
220	23	1	0.9625
230	24	1	0.9625
240	25	1	0.9625
260	27	1	0.9625

表 2 准确率最高时的主成分数以及 k 值数据表

经验证，猜想正确。在主成分数为 24，k 值为 1 时，模型性能最佳。

### 2.6 在测试集上评估模型的准确率

我们使用训练好的模型在测试集上进行预测，并评估模型的准确率，得到 `knn_score` 为 0.9625，模型在测试集的表现良好。部分结果如图所示：

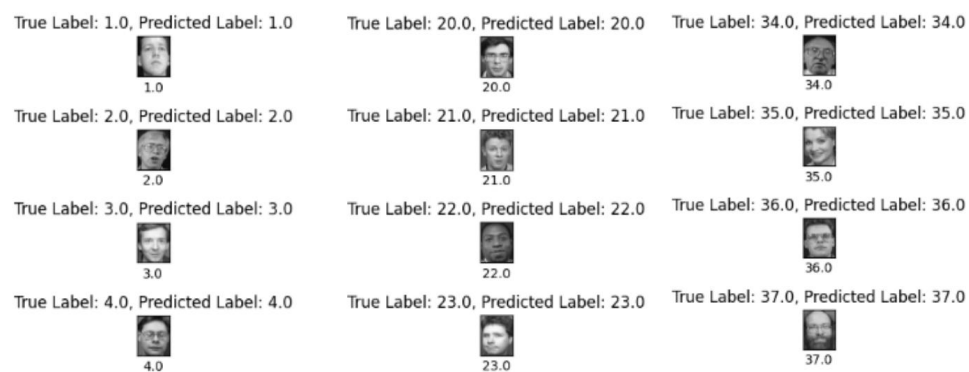


图 6 测试集分类结果

### 参考文献

- [1] Maćkiewicz A, Ratajczak W. Principal components analysis (PCA)[J]. Computers & Geosciences, 1993, 19(3): 303-342.
- [2] Guo G, Wang H, Bell D, et al. KNN model-based approach in classification[C]//On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings. Springer Berlin Heidelberg, 2003: 986-996.
- [3] Tolba A S, El-Baz A H, El-Harby A A. Face recognition: A literature review[J]. International Journal of Signal Processing, 2006, 2(2): 88-103.
- [4] 边子宸. 基于数字图像处理技术的人脸识别算法研究 [J]. 电子元器件与信息技术, 2023, 7(11): 87-89+94.
- [5] Guo X. A KNN classifier for face recognition[C]//2021 International Conference on Communications, Information System and Computer Engineering (CISCE). IEEE, 2021: 292-297.
- [6] 郭磊, 周文杰, 高少伟. 基于 PCA 算法的人脸识别技术研究[J]. 现代信息科技, 2021, 5(05): 108-112+117.
- [7] Chen J, Jenkins W K. Facial recognition with PCA and machine learning methods[C]//2017 IEEE 60th international Midwest symposium on circuits and systems (MWSCAS). IEEE, 2017: 973-976.