作业六

关文聪 2016060601008

1 使用PCA对Yale人脸数据集进行降维，并分别观察前20、前100个特征向量所对应的图像。请随机选取3张照片来对比效果。数据集<http://vision.ucsd.edu/content/yale-face-database>

下载数据集并解压，先设置图片路径，根据路径读取图片的人脸数据，使用矩阵存储读取的数据。设定要保留的特征数量k为20、100，调用PCA函数进行主成分分析（PCA）。将得到的结果再输出为图像，与原图进行比较。

Python代码：

import numpy as np

import scipy.misc as misc

import matplotlib.pyplot as plt

from sklearn.decomposition import PCA

import os

# 数据的读取与初始化预处理

path = 'C:\\Users\\Eternity-Myth\\Desktop\\yalefaces'

for dirpath, subdir, file\_set in os.walk(path):

all\_img = [path + '\\' + f for f in file\_set] # 保存所有文件的路径

m, n = len(all\_img), len(misc.imread(all\_img[0]).ravel()) # 行和列的数据

data = np.zeros((m, n)) # 初始化数据为（m,n）形状的矩阵

for i, f in enumerate(all\_img):

img = misc.imread(f).ravel() # 将每个2D图像展平为1D阵列

data[i] = img

# 对数据进行主成分分析（PCA）处理

data\_centered = data - data.mean(axis=0) # 对所有数据进行中心化

data\_centered -= data\_centered.mean(axis=1).reshape(m, -1) # 对所有参数进行中心化

gap = data - data\_centered # 保存数据与中心化处理后的数据之间的关系

k = [20, 100] # 保留的特征数k，设定k为20与100

pca1, pca2 = PCA(n\_components=k[0]), PCA(n\_components=k[1])

r\_set, im\_set = [], [] # 保存每个pca的方差比，输出去中心1D数组

for pca in [pca1, pca2]:

lower\_data = pca.fit\_transform(data\_centered) # 形状是(166,k)

comp = pca.components\_ # 形状是(k,77760), 这是一个稀疏的二维数组

r\_set.append(np.sum(pca.explained\_variance\_ratio\_))

im\_set.append(np.dot(lower\_data, comp) + gap)

# 输出处理过后的数据图像

for j in range(1, 166):

# 原图

fig, [ax0, ax1, ax2] = plt.subplots(1, 3, figsize=(10, 2.2))

ax0.imshow(data[j].reshape((243, 320)), cmap=plt.cm.gray)

ax0.set\_title('primal')

ax0.axis('off')

# PCA降维后的图像

for i, ax in enumerate([ax1, ax2]):

ax.imshow(im\_set[i][j].reshape((243, 320)), cmap=plt.cm.gray)

ax.set\_title('k=%s, Variance-Ratio: %.3f' % (k[i], r\_set[i]))

ax.axis('off')

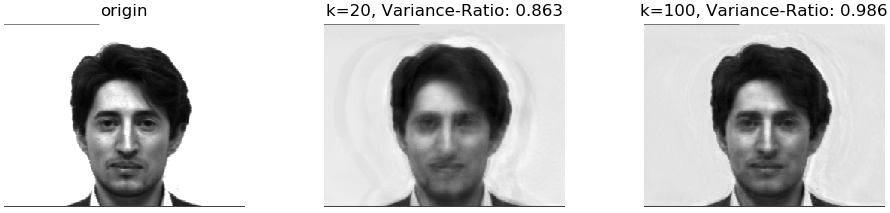
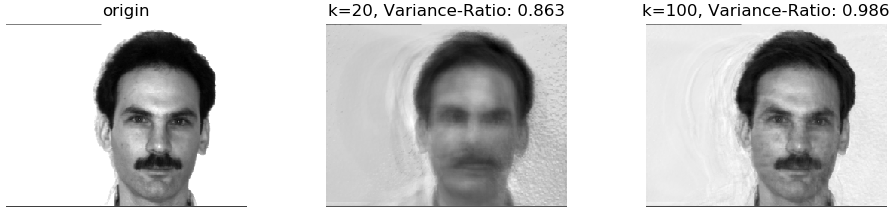
plt.subplots\_adjust(left=0.02, bottom=0.05, right=0.98, wspace=0)

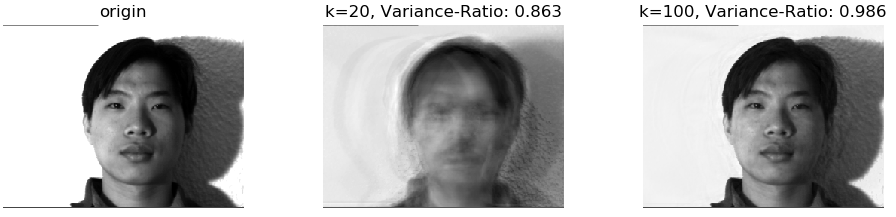
plt.savefig(r'C:\\Users\\Eternity-Myth\\Desktop\\output\\' + str(j) + '.png')





如图所示，在设置的output文件夹下已经生成了对应的输出图像。随机选取3张图片对比效果如下：





由对比结果可见，选取PCA的特征数越多，图像越清晰、明显，而在特征数不足时，会出现大量阴影轮廓，可以用pca.explained\_variance\_ratio\_来查看当前选择的最大k个特征向量的方差占比，方差占比越大则此特征表征的信息越多。可以发现当k从20增加到100时，选择的k个特征的累计方差占比已经接近于1了，而相应地，图像的特征已经与原始图像非常接近了。这意味着我们可以用大约100维的向量来描述一张原本维数达数万维的图像，可见PCA在这样的灰度人脸图像下的降维是非常有效的。