

基于人脸图像的年龄函数计算



学 院 软件学院

专 业 软件工程

年 级 2005 级

姓 名 吴志

指导教师 张坤龙

2009 年 6 月 20 日

摘 要

人脸识别技术，是近几年发展较快的计算机识别技术。它可以从静态或动态的图像中提取出人脸，并加以识别。由于人脸识别技术在数据收集方面的便利性等原因，受到越来越多的重视。

随着年龄的变化，人的面部会发生巨大的变化。与其他人脸识别遇到的问题相比，老龄化所带来的面部特征的改变有着自己的特点。本文在基于主成分分析(PCA)提取出来的人脸特征值的基础上，使用最小二乘回归算法计算年龄函数。

本文分别探讨了线性最小二乘算法和非线性最小二乘算法计算全局年龄函数的方法。而后考察了在最小二乘法下，通过简单的特征降维和改进数据分布计算全局年龄函数的方法与效果。最后以年龄为依据，对输入数据进行分类，按年龄段对年龄函数进行计算。

关键词：人脸识别；主成分分析；回归分析；最小二乘法；年龄估计

ABSTRACT

Face recognition technology, as a computer recognition technology, has a fast development in recent years. It can extract face from static images or dynamic images, and recognition information from face images. As a result of face recognition technology in data collection and convenience, it gets more and more attention.

The process of ageing causes significant alterations in the facial appearance of individuals. Compared with other problems in face recognition, the changes from the process of ageing in facial features has its own characteristics. In this paper, based on principal component analysis (PCA) extracted eigenvalues Face, algorithm the function of age using Least-squares regression.

This paper discusses the least squares linear and nonlinear least squares algorithm for calculating the global function of age. Then inspected under the least squares method, the characteristics of a simple reduced-data distribution and improve the global age calculation methods and the effect of function. Finally, the base on age, Classification of input data, calculation function of age According to different age groups.

Key Words: Face Recognition; principal component analysis(PCA); regression analysis; least square method; age estimates

目 录

第一章	绪论	1
1.1	人脸识别与人脸年龄估计	1
1.2	研究背景及意义	4
1.3	国内外发展状况	5
1.4	研究目标与研究内容	5
第二章	人脸的特征提取	6
2.1	人脸检测与提取	6
2.2	使用 PCA 方法提取人脸图像特征	7
第三章	使用最小二乘法计算全局年龄函数	12
3.1	最小二乘法原理	12
3.2	使用线性最小二乘法计算全局年龄函数	14
3.3	最小二乘曲面拟合计算全局年龄函数	17
3.4	验证并分析最小二乘算法	20
3.5	简单的特征降维	22
第四章	依据年龄分布计算年龄函数	25
4.1	改进数据分布估算全局年龄函数	25
4.2	线性最小二乘法分段计算年龄函数	27
4.3	最小二乘曲面拟合分段计算年龄函数	29
第五章	工作总结与展望	31
5.1	工作总结	31
5.2	展望	31

参考文献.....	33
外文资料	
中文译文	
致谢	

第一章 绪论

1.1 人脸识别与人脸年龄估计

1.1.1 人脸识别

随着自动化人机交互系统的需求不断增加,作为人机交互重要组成部分的人脸识别领域也不断发展。人脸识别是通过静止或动态的人脸图像,利用已有的算法和数据库对人脸的特征进行识别,从而判断出图像中的一人或多人的一门技术。人脸识别涉及图像处理、模式识别、计算机视觉、生理学、心理学等诸多学科的知识。人脸识别技术与其他生物特征身份识别鉴别方法(比如指纹识别、虹膜识别、DNA 识别等)是现代的人机交互的核心技术的重要组成成分,而人脸识别被认为是生物特征识别领域与人工智能领域最困难的研究课题之一^[6]。

不同个体的人脸面部特征区别不大,有着相似的结构,人脸上的器官有着相近的外形。这些特点有利于对人脸进行定位,对于处理与识别人脸的姿态、面部表情、遮挡、图像方向等研究是一个很好的切入点,但对于用人脸区分人类个体,对人的身份进行识别是非常不利的。因此对于通过人脸进行人类的个体识别的关键就是找到个体之间与众不同却有稳定的特征,不过人脸没有直观的,可以直接用数学语言进行描述的外观特征。而且人脸的外形特征很不稳定,特别是随着年龄的增长,人的面部会产生非常大改变,这使得原本有意义的特征随着时间的变迁发生变换,使得区分能力下降^[17]。

人脸识别的研究始于上世纪六十年代,经历了非自动识别,人机交互与自动识别三个阶段^[17]。最早的非自动识别阶段,主要研究如何提取人脸识别所需的特征,而通过这些特征进行人类个体的识别完全依赖手工干预,不是一种自动识别的系统。在那之后的人机交互阶段,实现了一定的自动化,但仍然需要操作人员的某些先验知识。进入九十年代后,随着高速度、高性能计算机的发展,人脸识别有了较大的突破,提出了多种计算机全自动识别系统。除了基于 K-L 变换的特征脸方法与奇异值特征为代表的代数特征方法取得发展外,人工神经网络、隐马尔可夫模型、小波变换、支持向量机等也在人脸识别研究中得到了广泛的应用。

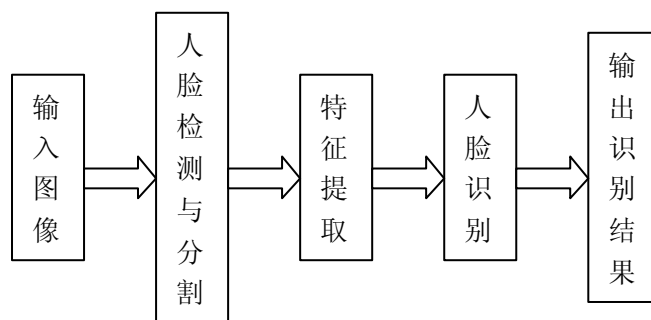


图 1-1 人脸识别流程示意图

人脸识别过程的重点有两个方面，分别是人脸检测与分割、特征提取与人脸识别。

1. 人脸检测。

人脸检测是指从一个简单或复杂的背景中分割出单一的人脸区域。在某些可以控制拍摄条件的场合，如在背景简单的证件照的场合，人脸定位相对比较简单。但在一般情况下，人脸在图像中的位置是未知的，比如在背景复杂的情况下拍摄的照片，人脸的检测与定位将受姿势、光照、表情以及噪声等因素的影响，定位有一定地困难。人脸检测的方法大致分为基于统计和基于知识两类，基本思想是用统计的方法或知识的方法对人脸进行建模，比较所有可能的待检测区域与人脸模型的匹配程度，从而得到可能存在人脸的区域。基于统计方法有区分人脸样本与非人脸样本的两类模式分类问题的样本学习方法，分别是特征脸方法和模板匹配方法。知识的方法是基于人们对一张典型人脸的构成知识，即人脸局部特征空间位置和几何关系来定位人脸。这种方法的主要缺点是很难满足所有的规则，需要通过实验确定规则的权重，而且如果规则太简单，可能得到太多的人脸。

2. 特征提取与人脸识别

根据人脸识别的方法的需要，会有与之对应的特征提取方法。

1) 基于人脸几何特征的人脸识别方法。

常用的人脸几何特征有人脸的局部形状特征、脸型特征以及五官在脸上分布的几何特征。提取特征需要用到人脸结构的先验知识。利用几何特征来识别人脸主要应用于正脸图像识别。识别所采用的几何特征是以人脸器官的形状和几何关系为基础的特征矢量，其分量通常包括人脸指定两点间的欧式距离、曲率、角度等。这种方法要求提取准确的人脸特征，对图像质量要求较高，使其应用受到较大限制。

2) 基于模板的人脸识别方法。

模板匹配法是一种经典的模式识别方法，首先设计人脸模板，然后计算样本与标准模板之间的匹配程度，通过设定的阈值来进行识别，一般分为固定模板和弹性模板。弹性模板根据待检测人脸特征的先验形状信息，定义一个参数描述的

形状模型，该模型的参数反应了对应特征形状可变部分，如位置、大小、角度等，它们最终通过模型与图像的特征动态地交互适应来进行修正。基于模板匹配的识别方法的效果要优于基于人脸特征点的识别方法。

3) 线性子空间法。

线性子空间法就是利用线性的空间变换，将原始的人脸图像数据压缩到一个低维的子空间中。常见的线性子空间法有：主成分分析法（PCA），线性判决分析，独立元分析。1991 年，Turk 和 Pnetinad 首次明确的提出了基于主元分析的人脸识别方法——特征脸方法。将主元分析用于人脸识别，可以得到人脸在低维空间中的线性表示，并且使人脸具有较好的可分性。图像在高维空间经过主元分析后，可以得到一组新的正交基，这些正交基张成了人脸的低维特征空间。如将低维特征空间的正交基按照图像阵列排列，则这些正交基呈现人脸形状，因此这些正交基称为特征脸，这样的人脸识别方法就叫做特征脸方法。

4) 其他方法。

除了上面介绍的几种，还有很多新的方法应用于人脸识别，如隐马尔可夫模型方法、基于神经网络的方法、小波变换方法等。

1.1.2 人脸年龄估计

在一个人从小孩长成青年人的过程中，人脸部骨骼不断生长，从而脸型不断变长变大，相貌也随之变化，而由青年人不断衰老变成老年人的过程中，皮肤的肤色、皱纹也会随之变化。因此年龄识别率会随着人的年龄变化而急剧下降。

Burt 和 Perrett 使用训练图像去为不同的年龄组制作平均人脸。例如，他们制作一张 20-24 岁年龄组平均人脸，然后为 25-29 岁年龄组制作等。由于图像是由平均法制作，他们表现出了该年龄组的典型面容，但是没有个人面部特征。为了改变人脸图像样本的年龄，他们将它与一张平均图像合并，这张图像是从一个年龄较高的年龄组中选取的。该系统是用于研究对年龄的感知。这种做法并不考虑到年龄变化导致的面部变形对于不同的人是不同的^[1]。

奥图尔等人利用三维面部信息，建立了一个三维人脸模型。他们使用一个夸张算法，以夸大或削减某些独特的面部特征；其结果是，依据夸张水平，观感上的年龄会有所增加或减少，这表明独特的面部特征在更加年长的人脸会更加突出。他们的发现由大量的对人的观察而予以证实^[1]。

有些计算机系统可用于根据年龄变化对人脸照片中的人脸进行变形。在此过程中，用户可以手动修改那些被定义为受年龄影响的特征。通过观察照片的近亲中的个人形象，用户修改其他的特征，使之包含遗传特征。这种系统已用于“老化”失踪儿童的照片的人脸图像。这种类型的系统的成效依赖于输入数据的用户的主观感受^[1]。

为了使模拟人脸受年龄影响而产生变化的过程自动化，通过模拟训练系统，

在任何新的图像上采用最适当年龄变化变形，而不是依赖于用户的技能。而这些工作的核心就是全局年龄函数。

基于人脸图像的全局年龄函数的计算，可以使我们得到一个函数，这个函数可以通过一张人脸图像提取出的特征值计算年龄。

通过人脸图像计算人的年龄，不可能出现完全精确的算法，因此只能先由计算机对已经掌握的数据进行学习，计算出年龄函数，而后通过年龄函数估算年龄。通过使用计算机进行自动估计，面对的困难就更加的多了。

现今主流的方法就是基于人脸模型和年龄函数的方法，其中年龄函数由多组已知年龄的人脸图像训练获得。基于这个方法，主要有以下四类^[1]：

1. 构建全局年龄函数：通过大量的训练图像，构建一个全局年龄函数。通过直接使用这个函数，计算出目标人脸图像的年龄。
2. 外观特定的年龄函数：这种年龄函数以面部特征划分输入。在这种方法中，我们认为对于面部特征类似的个体，其老化过程也是十分相似的。通过输入一张人脸图像，寻找与之有着相似特征的年龄函数，然后计算其年龄。
3. 加权年龄函数：与上一条提到的方法类似，但是通过输入图像，找不到与之有着相似特征的图像所计算出的年龄函数，这时使用所有年龄函数对该人脸图像的年龄进行估算，然后对其进行加权，从而得到估算结果。
4. 加权个人特殊年龄函数：这个方法对输入要求较高，不仅要求有人脸图像，还要有性别，健康，生活水平，经济状况，压力水平，工作条件，婚姻状况，居住地点和接触到的天气条件等要素。计算过程包含方法 2 和 3，同时对非图像信息的影响进行加权计算。

1.2 研究背景及意义

随着科学技术的发展和社会的进步，进行快速、有效、自动的人身辨别的实际需求日益迫切。随着计算机网络技术的广泛应用，人们生产生活中的信息越来越多的存储在计算机网络中，对这些信息的获取和支配也需要进行身份验证。因此，身份验证将在人们的生产、生活中扮演越来越重要的角色。现今常用的密码，证件等形式的验证手段具有可复制、可伪造、可丢失等不安全的特性。

基于生物特征的自动识别系统可有效的解决这类问题。生物特征作为人的内在属性，具有很强的自身稳定性和个体差异，因此是身份验证最理想的依据。基于指纹、虹膜、声音等生物特征的自动识别系统已经广泛的应用到很多行业中。而人脸特征，具有易于采集、不易仿冒、随身携带等特点，较之虹膜，指纹更适于人们的日常生活。

具有年龄变化的人脸识别系统，有很多特殊方面的应用。自动年龄核查，通

过个人照片自动统计年龄，可以用于对年龄精确度没有过大需求但数量巨大的人口年龄统计任务；法医学应用，对于长年失踪的人，可以通过全局年龄函数计算出特征值的偏移量，通过将偏移量加入已有的图像，可以估算出某一年龄时的脸部图像。

1.3 国内外发展状况

近年来，随着军事和民用的需要，一些国家对人脸识别的重视度不断提高。人脸识别的技术已经达到了比较高的应用水平，截止 2007 年，人脸识别的市场份额已经占到了生物识别市场的 12.9%，仅次于指纹识别，并且比重还在不断增加，打破了国际生物识别市场上“指纹”一统天下的局面。

在我国，863 计划、国家科技支撑计划、自然科学基金都拨出专款资助人脸识别的相关研究。在这种环境下，国内一些科研院所和院校在人脸识别技术方面取得了很大进展。如中科院自动化所，清华大学，中科院计算所自主开发的人脸识别技术已经达到了国际先进的水平。我国有 13 亿人口，这从根本上决定了我国人脸识别系统的巨量需求的同时，也决定了我国研究人脸识别系统的紧迫感。

1.4 研究目标与研究内容

从给定的训练样本集合中提取需要的人脸特征。设计合适的回归算法，用于计算出全局年龄函数。将提取的人脸特征进行处理，作为回归算法的输入，从而计算出全局年龄函数。最后运用待测年龄的人脸图像检验全局年龄函数的准确性。

本课题选用最小二乘算法来得出全局年龄函数，数据为 82 个人的 0 到 69 岁的人脸图像，总数据量为 906 条。首先使用 PCA 特征提取，从训练图像中提取人脸特征，然后使用最小二乘算法进行拟合，从而得出全局年龄函数。最后使用函数估算年龄，验证其准确度。

第二章 人脸的特征提取

由于一幅完整的人脸图像数据量过大，无法直接用来计算年龄函数，因此在计算年龄函数之前，首先要对图像进行处理，进行特征降维，提取出特征值。本章介绍人脸特征提取的两个组成部分，人脸检测与分割和使用 PCA 方法提取人脸特征。其中人脸检测与分割是使用 PCA 方法提取人脸特征的重要准备活动，可以有效地降低图像背景对特征提取的干扰。

2.1 人脸检测与提取

在利用 PCA 提取图像特征之前，需要对图像预处理，将人脸图像识别并提取出来。

在 OpenCV 中，库中自带了一个利用 harr 特征的人脸检测训练及检测函数：`cvHaarDetectObjects`。它利用训练好的检测器，在图片中间检测你想要的物体，如人脸。opencv 自带了很多检测器，在`%opencv%data/haarcascades`目录下。选择基于 24x24 离散 adaboost 正面人脸检测器 `haarcascade_frontalface_default`。检测后对图像进行分割，去除背景等干扰因素并统一大小为 40*50 像素。

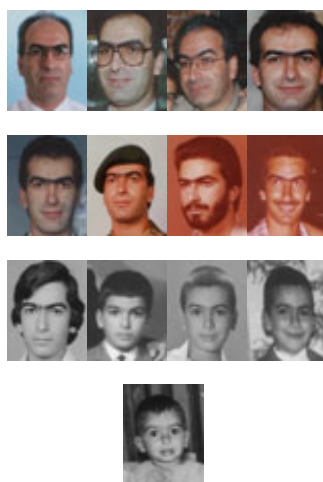


图 2-1 处理前图像



图 2-2 处理后图像

如上图所示，对人脸进行检测与分割处理后，尽量去除了背景图案，头发这些与人脸无关的元素，为下面的人脸图像特征提取做准备。为了便于调用，对已经创建好的完成人脸检测与分割处理的图像建立数据库索引。

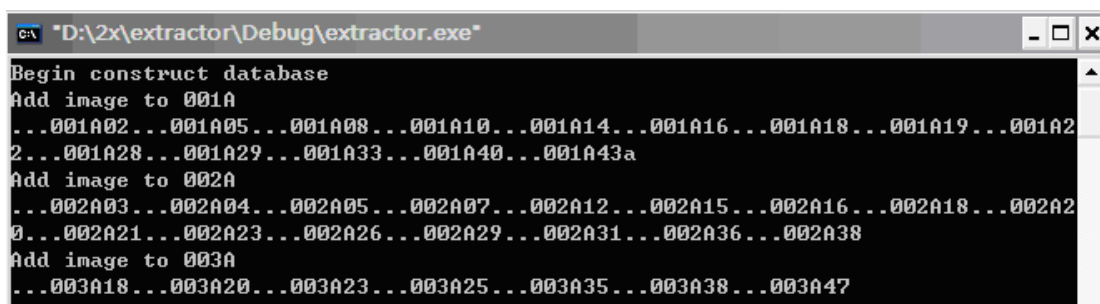


图 2-3 创建人脸图像数据库

原始的带背景的人脸图像分文件夹保存，每个文件夹保存一个人不同时期的人脸图像。通过人脸识别与分割之后，将图像以同样的文件结构保存到 FaceDB 文件夹中。生成的数据库文件 FaceDB.dll 内包含所有处理后的人脸图像的文件夹的名字，每个文件夹中保存一个人不同年龄时期的人脸处理后的图像，而与文件夹对应的.imobj 文件保留这些图像的路径。访问数据库时，先访问 FaceDB.dll 中的文件夹名，即用来标识人的个体的名称，然后通过访问各个与文件夹同名的.imobj 文件即可访问这些文件。

2.2 使用 PCA 方法提取人脸图像特征

主成分分析（principal components analysis (PCA)），是统计学中的一种对数据进行降维的技术。在实际工作中，经常会遇到由于原始数据量过大，使得数据

分析十分困难。在用统计分析方法分析之前，需要减少数据量。这里的数据量指的不是数据的数量，而且单条数据的变量数目。在很多情形，变量之间是有一定相关性的，即变量所反映出的信息有一定的重叠。主成分分析就是要通过原始数据生成一组新变量，这组新变量数量比原始数据少，尽可能多地反映原始数据的信息，同时这组新变量之间互不相关。

主成分分析是一个线性变换的过程，它通过变换，将原始数据转换到一个新的坐标系中，从而使原始数据在新的坐标系中的投影的最大方差在第一坐标上，第二大方差在第二坐标上，依次类推。其中第一坐标上的值称为第一主成分，第二坐标上的值称为第二主成分。主成分分析常用来对数据进行降维，同时保证了体现数据的特征。

主成分分析计算步骤如下：

1. 原始输入数据的一共含有 m 维随机向量 $x = (x_1, x_2, \dots, x_m)^T$ 的 n 个样品

$$x_i = (x_{i1}, x_{i2}, \dots, x_{im})^T, i = 1, 2, \dots, n \quad 2-1$$

$n > m$ ，构造样本矩阵，对样本矩阵的元素进行如下变换：

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, i = 1, 2, \dots, n; j = 1, 2, \dots, m \quad 2-2$$

其中

$$\bar{x}_j = \frac{\sum_{i=1}^n x_{ij}}{n}, s_j^2 = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n-1} \quad 2-3$$

得矩阵 Z 。

2. 对矩阵进行如下变换

$$R = \left[r_{ij} \right]_m \quad r_{ij} = \frac{Z^T Z}{n-1} \quad 2-4$$

其中

$$r_{ij} = \frac{\sum z_{kj} \cdot z_{ki}}{n-1}, i, j = 1, 2, \dots, m \quad 2-5$$

即可得到 Z 的相关系数矩阵。

3. 解相关矩阵 R 的特征方程

$$|R - \lambda I_m| = 0 \quad 2-6$$

得 m 个特征值，这样就确定了主成分。

按照下面的判断依据

$$\sum_{j=1}^k \lambda_j / \sum_{j=1}^p \lambda_j \geq a\% \quad 2-7$$

得到 k 值，使信息利用大于等于总体信息量的 $a\%$ ，对每个 λ ， $j=1, 2, \dots, k$ ，解方程组得单位特征向量。

4. 将标准化后的指标变量转换为主成分

$$U_{ij} = z_i^T b_j^o, j = 1, 2, \dots, m \quad 2-7$$

U_1 称为第一主成分， U_2 称为第二主成分，.....， U_m 称为第 m 主成分。

5. 对 k 个主成分进行综合评价

对 k 个主成分进行加权求和，即得最终评价值，权数为每个主成分的方差贡献率。

在图像识别领域，主成分分析是快速而又准确的对图像进行特征降维的方法之一。对图像进行主成分分析可以得到由有限个数构成的图像的特征值，这些特征值代表了图像的特征，可以使用这些值进行图像处理、人脸识别和年龄估计等方面的研究。

本题目所使用的 PCA 算法，目的是为了提取图像的特征值。

一组人脸图像的平均图像，即平均脸，可以表示这组人脸图像的大部分特征，如人脸的形状，面部器官位置等。PCA 计算出的特征值，排除了平均特征，主要体现的是人脸图像的差异。由人脸图像判断人的年龄，只需使用这些表示差异的特征值计算年龄函数。因此 PCA 提取的特征满足年龄函数计算的需要。

数据库创建完成后，使用 PCA 提取人脸图像的特征。而后将提取出来的特征值存入文件 ageFmaritx.dat 中，方便以后将数据导入最小二乘算法中计算年龄函数。

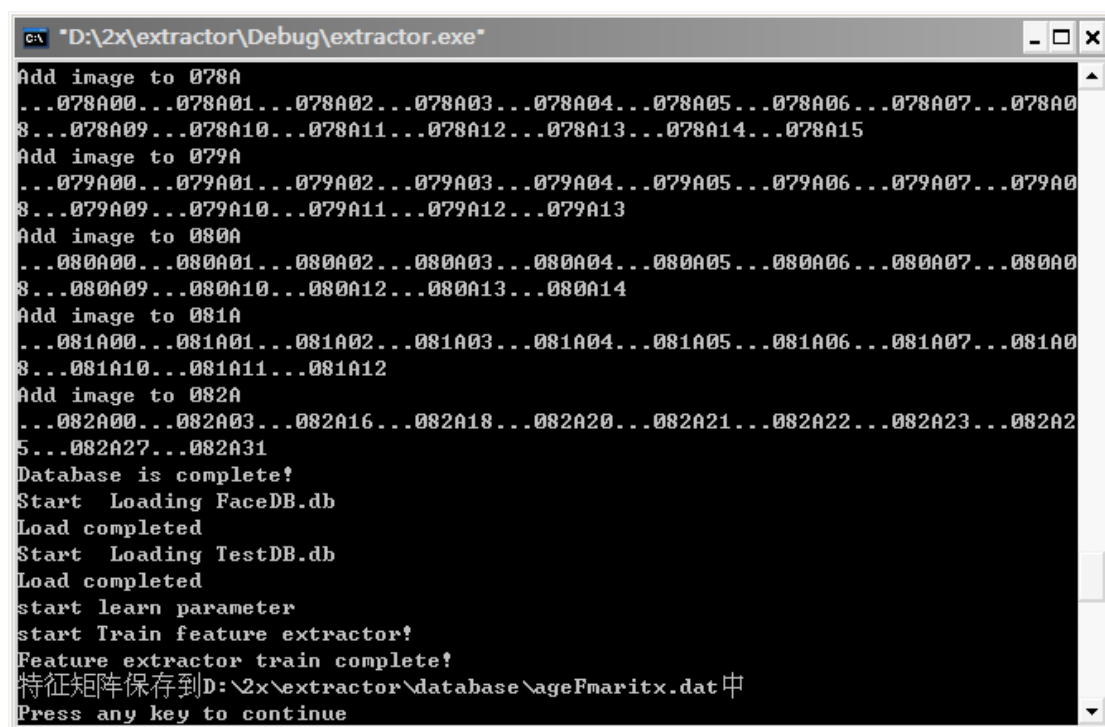


图 2-4 计算并保存特征矩阵

以第一个人 001A 的四张人脸图像为例，通过数据库将人脸图像，人脸年龄，特征值三者联系起来，如下表所示。

表 2-1 部分人脸图像、年龄、特征值表

人脸图像	年龄	特征值
	2	-638.454 570.85 -170.96 -257.505 195.892 -88.9589 340.436 -9.61584
		16.5065 -33.7262 -250.781 -42.0162 -167.722 -199.921 123.96 24.6372
		32.1848 170.67 -48.6846 -142.255
	8	-374.923 1403.82 -524.692 11.4001 -271.885 79.2988 175.984 -23.8743
		-39.6859 33.4455 34.5668 -10.1961 -19.3004 -50.2028 -72.8598 -125.209
		-26.8609 85.422 -32.8226 55.6571
	10	-687.615 724.501 -1024.13 -351.388 -56.8476 -98.4099 -6.1921 83.6636
		-372.402 67.0399 -432.21 -12.8008 38.9965 13.2004 -306.691 8.99788
		-66.3821 36.7255 -107.711 143.147
	16	-1079.74 751.569 -368.17 259.631 216.85 606.034 97.4241 275.898
		-216.259 110.969 27.7777 17.9757 172.479 159.323 -165.695 -23.3216
		12.3026 25.3715 166.009 362.904

每个人都有若干张不同年龄的人脸图像，每张图都对应着一个特征特征向量。这样就得到了总共 82 个人的总计 906 条，年龄范围为 0 岁到 69 岁的人脸图

像的特征值，每张人脸图像对应着由 20 个数组成的特征向量，这些数据就是下面试验的基础。后面要做的就是通过使用最小二乘法，求出年龄函数，即求出 20 个数组成的特征向量与年龄的关系的函数。

PCA 提取的特征值具有互不相关的性质，即特征值具有独立性，因此可以考虑将特征值直接作为年龄函数的自变量，而将年龄作为因变量，对年龄函数进行求解。通过参考文献，选取求解一阶多项式的线性最小二乘法和求解高阶多项式的非线性最小二乘法这两个方向估算年龄函数，并对比各个方法的效果。

第三章 使用最小二乘法计算全局年龄函数

在上一章中，已经介绍了如何通过 PCA 方法提取出了图像的特征值。本章主要介绍利用最小二乘法通过特征值计算全局年龄函数。首先用最基本的线性最小二乘法计算全局年龄函数，而后在此基础上讨论通过最小二乘曲面拟合的方法计算全局年龄函数。最后通过对最小二乘法的分析，提出一种基于简单降维的线性最小二乘法求解全局年龄函数，并对其结果进行分析。

3.1 最小二乘法原理

实际测量得到的数据或通过处理得到的数据，要对这些数据进行分析，确定变量之间相互依赖的关系，是研究工作中经常会遇到的问题。很多时候，无法确定这些变量之间关系的函数类型，因此很难直接求出变量的函数关系。大部分情况下，数据的误差也会影响对变量函数关系的求解。

具有相关关系的变量间虽然不具有确定的函数关系，但是通过大量的观测数据，可以发现它们之间存在一定的统计规律。在统计学中，研究这些统计规律即研究变量之间相关关系的方法称为回归分析。回归分析要对一组数据内变量间的关系建立数学函数，而后估计函数中的未知参数的值。估计参数的常用方法是最小二乘法。

回归分析中，按照变量的数量划分，可分为一元回归分析和多元回归分析。当需要进行分析的数据只涉及因变量和一个自变量时，称为一元回归分析；当需要进行分析的数据涉及因变量和两个或两个以上自变量时，叫做多元回归分析。

如果按照自变量和因变量的函数表达式是线性的还是非线性，可分为线性回归分析和非线性回归分析。线性回归分析是最基本的分析方法，遇到非线性回归问题可以借助数学手段将其化为线性回归问题进行处理。

最小二乘法（又称最小平方法）是一种数学优化技术。它通过最小化误差的平方和寻找数据的最佳函数匹配。利用最小二乘法可以简便地求得未知的数据，并使得这些求得的数据与实际数据之间误差的平方和为最小^[14]。

对于实验提供的大量数据，由于误差等方面的原因不能要求拟合函数在数据点处的偏差，严格为零，但为了使近似曲线尽量反映所给数据点的变化趋势，需对偏差有所要求。通常要求偏差的平方和最小，此即为最小二乘原理。

以一元线性最小二乘举例，解法如下：

假设要求解的是两个变量(x , y)之间的相互关系。通常输入数据有着下面的这种形式

$$(x_1, y_1), (x_2, y_2) \dots (x_m, y_m) \quad 3-1$$

可以设这个线性函数为

$$Y = a_0 + a_1 X \quad 3-2$$

其中： a_0 、 a_1 是任意实数。

为建立这直线方程就要确定 a_0 和 a_1 ，应用最小二乘法原理，将实测值 Y_i 与计算值的差值的平方和最小为优化判据，于是得出差值的平方和为：

$$\Phi = \sum_{i=1}^m (Y_i - Y)^2 \quad 3-3$$

把式 3-2 代入式 3-3 中得：

$$\Phi = \sum_{i=1}^m (Y_i - a_0 - a_1 X)^2 \quad 3-4$$

要使函数值最小，可用函数 Φ 对 a_0 、 a_1 求偏导数，并令其等于零。得：

$$\frac{\partial \Phi}{\partial a_0} = -2 \sum_{i=1}^n (Y_i - a_0 - a_1 x_i) = 0 \quad 3-5$$

$$\frac{\partial \Phi}{\partial a_1} = -2 x_i \sum_{i=1}^n (Y_i - a_0 - a_1 x_i) = 0 \quad 3-6$$

亦即：

$$ma_0 + (\sum_{i=1}^m X_i) a_1 = \sum_{i=1}^n Y_i \quad 3-7$$

$$(\sum_{i=1}^m X_i) a_0 + (\sum_{i=1}^m X_i^2) a_1 = \sum_{i=1}^m (X_i Y_i) \quad 3-8$$

得到的两个关于 a_0 、 a_1 为未知数的两个方程组，解这两个方程组得出：

$$a_0 = (\sum_{i=1}^m Y_i) / m - a_1 \sum_{i=1}^m X_i / m \quad 3-9$$

$$a_1 = [\sum_{i=1}^m X_i Y_i - \sum_{i=1}^m X_i \sum_{i=1}^m Y_i / m] / [\sum_{i=1}^m X_i^2 - (\sum_{i=1}^m X_i)^2 / m] \quad 3-10$$

这时把 a_0 、 a_1 代入式 3-1 中，此时的(式 3-1)就是要计算回归函数。

本题目通过对由 PCA 算法提取出的 20 个自变量(特征值)和 1 个因变量(年

龄）使用最小二乘法，进行多元线性回归分析和多元非线性回归分析，从而求出特征值与年龄间关系的年龄函数，而后使用年龄函数，以特征值作为输入，估算人的年龄。

3.2 使用线性最小二乘法计算全局年龄函数

3.2.1 算法设计

最小二乘法是一种数学优化技术，它通过最小化误差的平方和寻找数据的最佳函数匹配^[14]。

利用最小二乘法可以简便地求得未知的数据，并使得这些求得的数据与实际数据之间误差的平方和为最小。最小二乘法还可用于曲面拟合。

通过调用 PCA 程序提取的特征值为 20 个元素的特征矩阵。将这 20 个元素用 x_1 、 x_2 、 x_3 ， $\dots$ ， x_{20} 表示，年龄用 y 表示。假设它们的关系为线性方程，即可得到方程：

$$y = a_0 + a_1x_1 + a_2x_2 + \dots + a_{20}x_{20} \quad 3-11$$

线性最小二乘法既为通过已有的大量数据，求解式 3-11。

基于最小二乘法原理，实际值 y_j 与使用函数计算出的值的差的平方和：

$$\Phi = \sum_{j=1}^n (y_j - f(x_1, x_2, x_3, \dots, x_{20}))^2 \quad 3-12$$

最小为优化判断依据。式中的 j 表示的是一共拥有的数据的条数。

将式 3-12 最小为优化判断依据，既为将：

$$\Phi = \sum_{j=1}^n [y_j - (a_0 + a_1x_1 + a_2x_2 + \dots + a_{20}x_{20})]^2 \quad 3-13$$

最小为优化判断依据。可对式 3-13 中的 a 分别求偏导数，并另其等于零。即求方程：

$$-2x_{jk} \sum_{j=1}^n [y_j - (a_0 + a_1x_1 + a_2x_2 + \dots + a_{20}x_{20})] = 0 \quad 3-14$$

式中 $k=0, 1, 2, 3, \dots, 20$ 。

当 $k=0$ 时，求得：

$$a_0 = \bar{y} - \bar{a_1x_1} - \bar{a_2x_2} + \dots + \bar{a_{20}x_{20}} \quad 3-15$$

将式 3-15 代入式 3-14 得出：

$$\begin{aligned}
 & a_1(\sum_{j=1}^n x_{j1}x_{jk} - \bar{x}_1\sum_{j=1}^n X_{jk}) + a_2(\sum_{j=1}^n x_{j2}x_{jk} - \bar{x}_2\sum_{j=1}^n X_{jk}) + \dots \\
 & + a_{20}(\sum_{j=1}^n x_{j20}x_{jk} - \bar{x}_{20}\sum_{j=1}^n X_{jk}) = \sum_{j=1}^n X_{jk}y_j - \bar{y}\sum_{j=1}^n X_{jk}
 \end{aligned} \quad 3-16$$

当 $k=1, 2, \dots, 20$ 可以得到 20 个方程。相当于求解形如 $AX = b$ 的方程。其中

$$b = [\sum_{j=1}^n X_{j1}y_j - \bar{y}\sum_{j=1}^n X_{j1}, \sum_{j=1}^n X_{j2}y_j - \bar{y}\sum_{j=1}^n X_{j2}, \dots, \sum_{j=1}^n X_{j20}y_j - \bar{y}\sum_{j=1}^n X_{j20}] \quad 3-17$$

将 $AX = b$ 使用高斯消去法进行求解。求解过程实际上是将中学学过的消元法标准化，将一个方程乘或除以某个常数，然后将两个方程相加减，逐步减少方程中的未知数，最终使每个方程只含有一个未知数，从而得出所求的解。整个过程分为消元和回代两个部分。

高斯消去法解方程组基本思想是设法消去方程组的系数矩阵 A 的主对角线下的元素，而将 $AX = b$ 化为等价的上三角形方程组，然后再通过回代过程便可获得方程组的解。在本程序中，回代过程用通过将上三角矩阵计算成对角矩阵的方法。

通过以上方法即可求出 a 的值。

最后通过计算误差均值、方差、标准差衡量计算结果。

3.2.2 实验结果

使用生成的方程对训练集中的所有人脸图像的年龄进行估算，将其结果输出到文件中，输出包括 a 的值，906 条数据的实际值，估算值，差值，平均误差，方差，标准差。

求出的全局年龄函数的系数 a 的值如下图所示，偏移量 $a_0=19.78$ 。

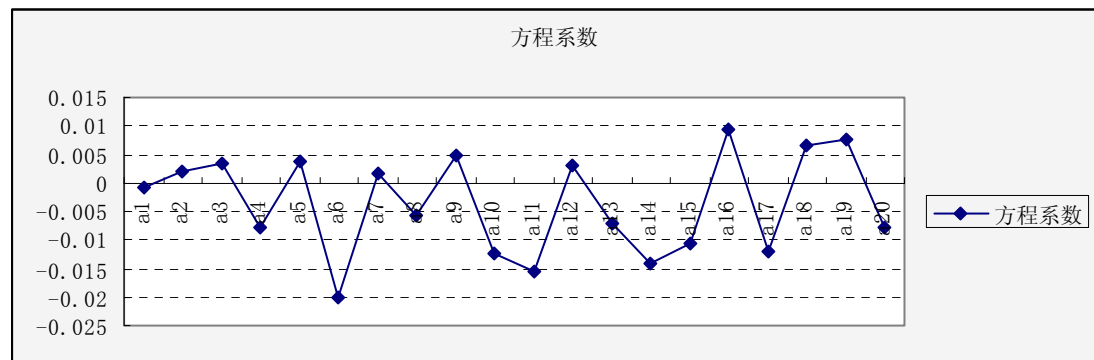


图 3-1 线性最小二乘计算全局年龄函数系数

使用本次计算出的全局年龄函数对年龄估算的统计结果如下：

表 3-1 线性全局年龄函数年龄估算统计结果

平均误差	方差	标准差
8	113	11.5326

3.2.3 结果分析

总体上看平均误差为 8，已经取得了比较好的效果，但是仔细观察年龄估算结果就会发现，本次计算的年龄函数有很严重的局限性。

表 3-2 部分估算结果表

人脸图像	实际年龄	估算年龄	差值
	2	16	14
	5	10	5
	8	10	2
	10	19	9
	16	17	1
	18	21	3
	19	18	-1
	22	27	5
	28	18	-10

经过观察后发现，对不同年龄段估算的准确度是不同的，修改程序，得出统计结果如下：

表 3-3 线性最小二乘分年龄段统计标准差

年龄段	数据数量	标准差
0~9	326	10.3441
10~19	316	5.56776
20~29	129	8.83176
30~39	71	15.8745
40~49	41	22.5389
50~59	15	31.7175
60~69	8	39.0384

通过观察以上结果发现，不同年龄段估算误差的差别很大，分析后得出可能有以下原因：

1. 算法问题：线性最小二乘法的局限性导致了估算误差值在不同年龄段的不平衡。
2. 用来计算年龄函数的人脸图像的原始数据的不平衡导致了在不同年龄段的估算值的不平衡。

因此，可以考虑通过改进最小二乘算法和改进读入数据两方面对年龄函数算法进行估算。本章剩下部分讨论改进最小二乘算法，而在下一章讨论改进读入数据，即对数据按年龄段分类后计算年龄函数。

3.3 最小二乘曲面拟合计算全局年龄函数

3.3.1 算法设计

设方程形如

$$y = a_0 + a_1x_1 + a_2x_2 + \dots + a_{20}x_{20} + a_{21}x_1^2 + a_{22}x_2^2 + \dots + a_{40}x_{20}^2 \quad 3-18$$

通过推导，式 3-18 的最小二乘曲面拟合可以转换为求：

$$y = a_0 + a_1z_1 + a_2z_2 + \dots + a_{20}z_{20} + a_{21}z_{21} + a_{22}z_{22} + \dots + a_{40}z_{40} \quad 3-19$$

的最小二乘的线性拟合。其中

$$z_1 = x_1, z_2 = x_2, \dots, z_{20} = x_{20}, z_{21} = x_1^2, z_{22} = x_2^2, \dots, a_{40} = x_{20}^2 \quad 3-20$$

求解线性最小二乘的方法如同 3.2 节中所示。可以将二次项的 x 视为特征向量，在应用最小二乘法求解前，将其计算出来，并存入矩阵相应位置。而后应用最小二乘法，求出形如 $AX = b$ 的矩阵，最后应用高斯消去法求解全局年龄函数的系数 a 。最小二乘二次曲面拟合计算可以转化为最小二乘线性拟合，对于本次

试验，可以将二次项视为新的特征值，这样就相当于对有 40 个特征值的数据求解全局年龄函数。

3.3.2 程序结果

求出的全局年龄函数的系数 a 的值如下所示，由于二次项系数与一次项系数数量级不同，因此不在同一张图统计，偏移量 $a_0=16.781$ 。在统计一次项系数时，将线性最小二乘与非线性最小二乘进行对比。

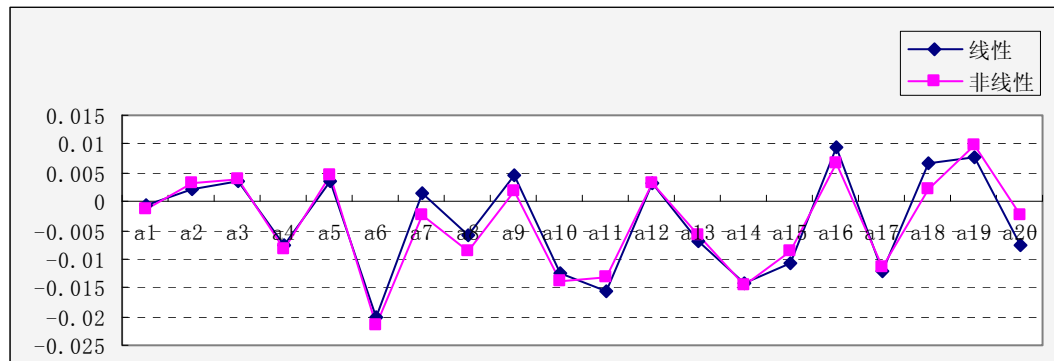


图 3-2 线性最小二乘与非线性最小二乘计算年龄函数一次项系数对比

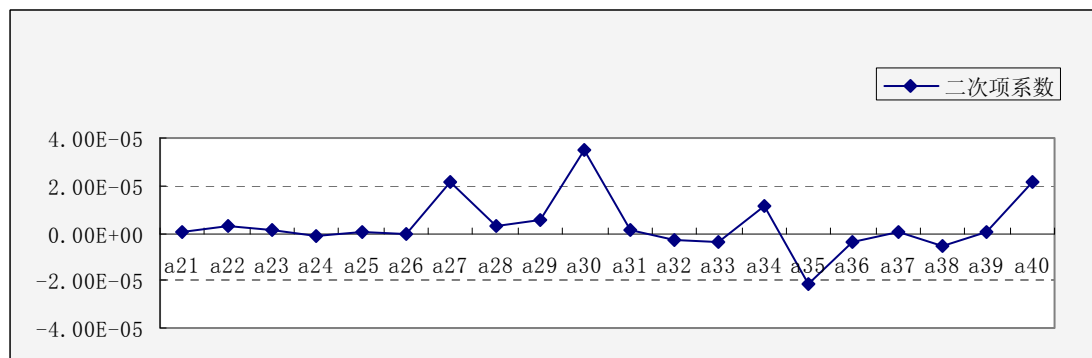


图 3-3 非线性最小二乘计算全局年龄函数二次项系数

使用本次计算出的全局年龄函数对年龄估算的统计结果如下，并将之与线性最小二乘全局年龄函数对年龄估算的统计结果进行对比：

表 3-4 线性与非线性最小二乘年龄估算标准差对比

年龄段	数据数量	线性最小二乘标准差	非线性最小二乘标准差
0~9	326	10.3441	9.89949
10~19	316	5.56776	5.65685
20~29	129	8.83176	8.60233
30~39	71	15.8745	15
40~49	41	22.5389	21.5407
50~59	15	31.7175	31.5595
60~69	8	39.0384	37.6431

对非线性最小二乘的总体的年龄估算进行统计，平均误差 8，方差 125，标准差 11.1803。

3.3.3 结果分析

经过观察发现，最小二乘的二次曲面拟合有一定的改进，较线性最小二乘改进了 6%。因此考虑次数更高的最小二乘曲面拟合。使用最小二乘三次曲面拟合，计算出的全局年龄函数对年龄估算的统计结果如下：

表 3-5 二次与三次曲面拟合最小二乘年龄估算标准差对比

年龄段	数据数量	二次曲面拟合最小二乘标准差	三次曲面拟合最小二乘标准差
0~9	326	9.89949	9.79796
10~19	316	5.65685	5.83095
20~29	129	8.60233	8.3666
30~39	71	15	14.4222
40~49	41	21.5407	20.8327
50~59	15	31.5595	31.257
60~69	8	37.6431	37.4299

对三次曲面拟合最小二乘的总体的年龄估算进行统计，平均误差 8，方差 121，标准差 11。

可以发现又略有改进，当使用更高次数的最小二乘拟合时，不过从趋势上可以看出虽然会有一定改进，但改进效果越来越有限。

最小二乘的二次曲面拟合相对于最小二乘线性拟合的改进率为 6%，分析原因后可能有两点：

- 1、在公式推导或算法程序含义不合理的地方。
- 2、公式推导和算法程序设计均正确，最小二乘算法本身某方面的局限性。

3.4 验证并分析最小二乘算法

本节的目的是验证最小二乘算法本身的正确性，并初步分析最小二乘算法的局限性。

1. 实例 1

下面的数据来源于文献的实际例题。

表 3-6 实例 1 数据

Y	21.6	15.5	10.4	31.0	13.0	32.4	19.0	10.4	19.0	11.8
X	208	152	113	227	137	238	178	104	191	130

如上表所示，Y 与 X 成线性关系，求线性方程组 $y=a_0+a_1x$ 。

例题给出的答案使用的是手工计算， $a_1=0.1612$ ， $a_0=-8.6394$

使用程序从 2x1.txt 中读取 X 和 Y 的值，对方程中 a 的值进行估算。

得出

A1=0.161234

A0=-8.64507

排除计算精度的影响，程序运行结果与文献中答案一致，说明程序运行良好。

2. 实例 2

本例中的数据依然是来自文献，目的是为了验证最小二乘曲面拟合可以转化为最小二乘线性拟合的可行性。

表 3-7 实例 2 数据

Y	10	5	4	2	1	1	2	3	4
X	1	3	4	5	6	7	8	9	10

Y 与 X 呈二次拟合，拟合方程为 $a_0 + a_1x + a_2x^2 = y$

文献中给出的答案使用的是手工计算， $a_0=13.4597$ ， $a_1=-3.6053$ ， $a_2=0.2676$ 。

使用程序从 2x1_.txt 中读取 X 和 Y 的值，应用 3.3.1 节的原理进行运算，对方程中 a 的值进行估算。

得出

A1=-3.60531

A2=0.267571

A0=13.4597

排除计算精度的影响，程序运行结果与文献中答案一致，说明程序运行良好。

3. 实例 3

本例编写程序按照方程

$$y=51+72x_1+39x_2+40x_3+5x_4-46x_5+37x_6-88x_7+49x_8+110x_9+14.8x_{10}$$

使用 1000 以内的随机的 X 值生成 100 组数据,然后使用最小二乘估算程序,估算 X 的系数。

表 3-8 实例 3 估算结果分析

系数 a_n	方程原始系数	方程估算系数
a_0	51	51.1975
a_1	72	72
a_2	39	38.9999
a_3	40	39.9999
a_4	5	4.99997
a_5	-46	-46.0001
a_6	37	37
a_7	-88	-88
a_8	49	49
a_9	110	110
a_{10}	14.8	14.7999

估算结果良好,证明算法设计与程序运行无误。

以上三个实例证明最小二乘算法设计和编写的正确性。由此可见,最小二乘法对于线性方程或 n 次曲面方程的估算效果是比较准确的。

估算结果一方面是由于全局年龄函数的本身拟合性能的限制,另一方面,用于估算全局年龄函数的数据是使用 PCA 从图像中提取的特征值,并不是所有数据都是与年龄有关。使用最小二乘法对全局年龄函数进行计算时,会受到与年龄无关的数据的影响。为了排除影响,可以考虑仅使用部分特征值对全局年龄函数进行估算的这种简单特征降维的方法。

3.5 简单的特征降维

本题目使用的 PCA 特征提取的特征是 20 个元素的特征向量，即含有 20 个元素的数组。尝试使用其中的 R 个元素进行全局年龄函数的计算。

3.5.1 算法设计

从 20 个元素提取 R 个元素是数学中的组合问题。通过分析，可以很容易地通过递归算法实现。

举例：A = {1, 2, 3, 4, 5}

从 A 中 5 个元素取 2 个得到 12, 13, 14, 15, 23, 24, 25, 34, 35, 45 一共 10 个组合。这里可以看到组合出现是有规律的，每个组合总是按 N 中元素顺序出现（组合不是排列，12 和 21 属于同一组合），其次，相邻组合后者在前者基础上从最末元素开始按 N 中元素顺序递增。

递归算法的程序十分简洁，本段程序的递归函数包含 5 个参数：数组 a 用于保存结果，n 为参加组合的数的个数，r 为要从 n 中取的数的个数，k1 为已经取到的数的个数，初始为 0，k2 为可填入数据的最小值，初始为 1。

递归函数由两部分组成：

- 1、如果已经抽出 R 个元素则输出保存在数组 a 中的结果。
- 2、向数组 a 的 k1+1 位填入所有可能的数，然后调用递归填入下一位的数。因为不存在类似于 11 的这种组合，也就是说按照算法的设计，下一位的数不可能小于上一位，因此递归调用时将 k2 加 1。

通过以上两个过程，可以得到从 N 个元素中抽取 R 个元素的组合方法。对于本程序来说，要从 20 个元素中，抽取 R 个元素，R 的值从 1 到 20。因此通过循环，计算出 R=1, 2, 3....., 20 时的所有组合情况。

使用递归的算法解决问题虽然极大地简化程序设计，但是消耗较多的内存和运行时间，效率不高，容易引起堆栈溢出。因此，不将年龄函数的计算直接地加入到递归函数中。由递归函数将组合的结果输出到 txt 文件中，然后在通过提取文件中记录的组合的结果来计算全局年龄函数。

将年龄计算和结果输出等过程函数化，方便主函数调用。不断组合各种调用方式时，以方差为衡量标准，将最小方差的组合方式作为结果输出。

3.5.2 试验结果

最后筛选出 18 位数字，分别为：

1 2 3 4 5 6 7 8 10 11 13 14 15 16 17 18 19 20

计算后求出的全局年龄函数的系数 a 的值如下所示，将 a 重新命名，因此 a 的角标范围从 0 到 18，偏移值 $a_0=19.687$ ：

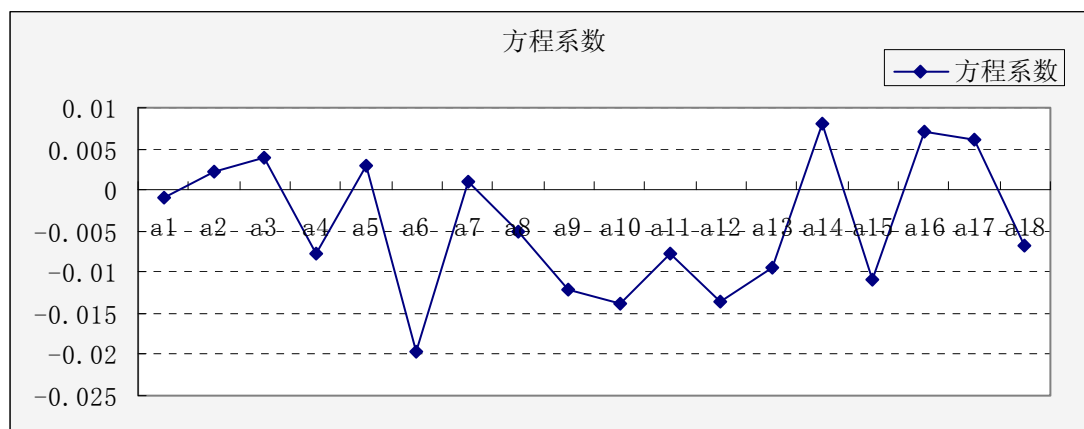


图 3-4 通过简单降维计算年龄函数系数

使用本次计算出的全局年龄函数对年龄估算的统计结果如下，与原始的线性最小二乘法进行对比：

表 3-9 原始最小二乘与简单降维最小二乘年龄估算标准差对比

年龄段	数据数量	线性最小二乘标准差	简单降维最小二乘标准差
0~9	326	10.3441	10.3441
10~19	316	5.56776	5.56776
20~29	129	8.83176	8.7178
30~39	71	15.8745	15.6844
40~49	41	22.5389	22.9347
50~59	15	31.7175	31.607
60~69	8	39.0384	39.281

对简单降维的最小二乘的总体的年龄估算进行统计，平均误差 8，方差 133，标准差 11.5326。

3.5.3 结果分析

简单的特征降维产生的结果与最原始的线性最小二乘算法直接估算年龄基本相同，不如最小二乘曲面拟合算法。从平均误差，方差，标准差这些衡量标准上可以看出，简单的特征降维几乎没有改进。

如果条件允许的话，可以考虑把这种简单的特征降维作为程序的一个附加组合。例如可以将特征降维与最小二乘曲面拟合算法结合。最小二乘曲面拟合算法对全局年龄函数的计算较之线性最小二乘算法有一定的改进，改进率为 6%，而将简单的特征降维与最小二乘曲面拟合算法结合有可能会得到进一步的改进。

但是权衡这种简单特征降维效果和所花费的代价来看，将特征降维作为程序

的一个附加组合是很不值得的。原因有以下两个方面：

- 1) 根据最小二乘法的原理，方程中的某一项的系数虽然不会是 0，但是可以是接近于 0 的一个非常小的小数。特征降维的作用就是分辨取小数与取 0 到底哪一个使得函数更加精确。在原始数据十分有限的情况下，这种做法可能具有一定效果，但是本题目中输入数据有 906 条，足以保证最小二乘法所要求的数据量，因此从最终效果来看，应用这种简单的特征降维是不值得的。
- 2) 在本次试验过程中，产生的保存这 20 个数所有组合的 TXT 文件为 30MB，程序运行时间为 87 分钟。如果对最小二乘曲面拟合算法应用这种简单的特征降维，就需要对 40 个数的所有组合方式进行尝试。组合算法复杂度的增长不是线性的，而是指数的，因此从花费的代价上来说，应用这种简单的特征降维也是不值得的。

特征降维是一个十分重大的课题，不仅在图像处理领域，在其他的模式识别领域是一个重要的分支。特征降维的目的就是为了使得模式识别的学习算法具有可行性，降低程序运行时间。通过特征降维可以极大地提高学习算法输入数据的质量，从而改进算法。

利用数学的组合这种算法效率比较低，而特征降维也不是本题目的重点，因此，不再考虑和使用通过简单的特征降维提高年龄函数估算精度。

第四章 依据年龄分布计算年龄函数

在上一章中，利用线性最小二乘法和非线性最小二乘法对全局年龄函数进行了计算。在不优化输入数据的情况下，通过改进最小二乘法算法，改进年龄函数的计算。在本章中，通过对改进数据分布和按年龄进行分段两个方面优化全局年龄函数的计算。最后，将按年龄分段计算年龄函数与最小二乘曲面拟合相结合，计算全局年龄函数。

4.1 改进数据分布估算全局年龄函数

4.1.1 算法设计

观察以上几次试验的数据可以发现，用全局年龄函数对年龄进行年龄估算，其结果的标准差随着年龄段的不同有着巨大的差异。平均上，30 岁以下（不含 30 岁）的年龄估算的标准差低于总体年龄估算的平均标准差；30 岁以上（含 30 岁）的年龄估算的标准差远高于总体年龄估算的平均标准差，以上的所有试验中，对 60 岁以上（含 60 岁的）的年龄估算最好的最小二乘曲面拟合的标准差为 37.6431，误差过大。对于全局年龄函数，不应该出现对不同年龄段的年龄估算有如此大的差别。通过分析发现，全部数据有 906 条，30 岁以下（不含 30 岁）的数据一共有 711 条，占全部数据的 85.10%，这应该是最小二乘法计算出的全局年龄函数估算高年龄时误差过大的原因。因此考虑在各个年龄段的输入数据条数一致的情况下使用最小二乘法对全局年龄函数进行计算。

通过观察发现，0 到 9 岁的数据最多，有 326 条，因此需要让其他年龄段的数据也达到 326 条。通过随机抽取本年龄段内的数据加入计算全局年龄函数所用的数据中，使得所有年龄段的数据都达到 326 条。最终，数据翻倍完成后一共有数据 2282 条。然后使用这 2282 条数据，通过线性最小二乘法计算全局年龄函数。

4.1.2 试验结果

计算后得出的全局年龄函数的系数 a 的值如下所示，偏移量 $a_0=33.2064$ ，在统计的同时，与 3.2 节的线性最小二乘法计算出的全局年龄函数的系数 a 的值进行对比：

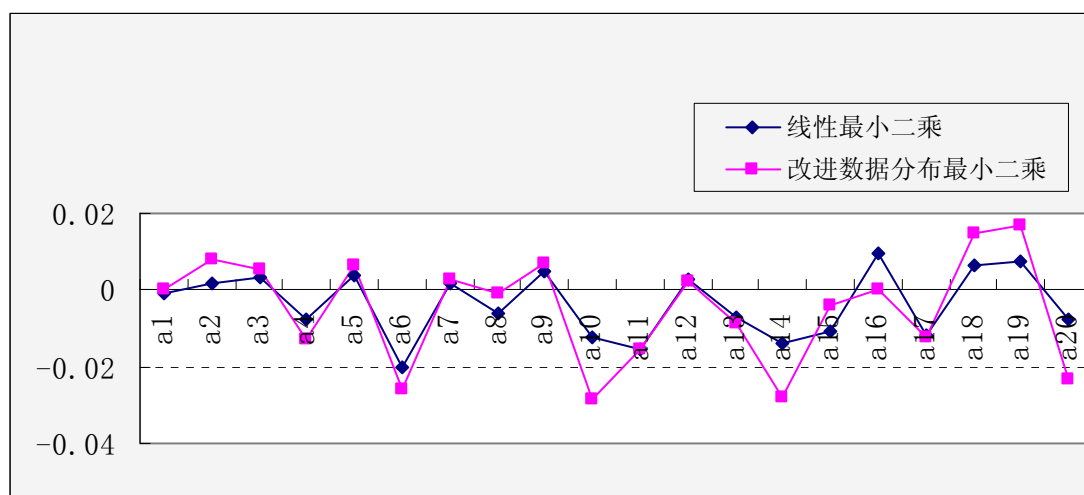


图 4-1 线性最小二乘与改进数据分布最小二乘计算年龄函数的系数

使用本次计算出的全局年龄函数对年龄估算的统计结果如下，与简单的线性最小二乘法计算结果进行对比：

表 4-1 原始最小二乘与改进数据分布的最小二乘年龄估算标准差对比

年龄段	数据数量	线性最小二乘标准差	改进数据分布最小二乘标准差
0~9	326	10.3441	22.3383
10~19	316	5.56776	15.5563
20~29	129	8.83176	11.7047
30~39	71	15.8745	9.05539
40~49	41	22.5389	22.9347
50~59	15	31.7175	13.2288
60~69	8	39.0384	16.5831

对改进数据分布的最小二乘的总体的年龄估算进行统计，平均误差 14，方差 305，标准差 17.4642。

4.1.3 结果分析

本次试验是从输入数据方向上着手，尽量减少因为输入数据在年龄分配上的不平均对计算全局年龄函数的影响。通过使用计算出的全局年龄函数对已有数据进行年龄估计得出的结果显示，全局年龄函数对不同年龄段年龄估算的差异已经有了很大的改善。但与此同时，总体的年龄估算的误差大幅上升。经分析，原因主要有以下两个方面：

1、本次试验的这种简单的数据翻倍对全局年龄函数计算造成的影响。重复利用数据并不能代表真正的数据，使用数据翻倍的方法虽然有效地降低了由于输入数

据不平衡造成的影响，但是却使得输入数据的质量下降，从而对全局年龄函数的估算造成不良影响。

2、计算全局年龄函数本身的复杂度。使用 PCA 可以提取出图像的特征值，但是不能认为某些特征值与年龄有着直接的关系，PCA 的目的是对图像进行压缩，降维从而提取主要特征。而且在不同年龄段，体现年龄的特征不一定是不变的。

直接计算全局年龄函数算法不仅复杂，精确度也很低。对于最小二乘法来说，很难直接估算出质量较高的全局年龄函数。因此，考虑对不同年龄段分别进行年龄函数的计算。

4.2 线性最小二乘法分段计算年龄函数

4.2.1 算法设计

应用线性最小二乘法估算全局年龄函数会产生很大的误差，因此考虑对不同年龄段分别进行年龄函数的计算。分段进行年龄函数的计算可以解决由于不同年龄段，数据量的不同对最小二乘法的影响。同时，可以预见，由于年龄函数估算的年龄范围的缩小，计算精度会有所提高。

将数据划分为 0 岁到 19 岁，20 到 39 岁，40 岁以上这 3 部分，应用线性最小二乘法估算年龄。

4.2.2 试验结果

1. 0 岁到 19 岁部分，有数据 642 条。

应用线性最小二乘法，计算后得出的年龄函数的系数 a 的值如下所示，偏移量 $a_0 = 9.20038$ ，同时与全局年龄函数的系数 a 的值进行对比：

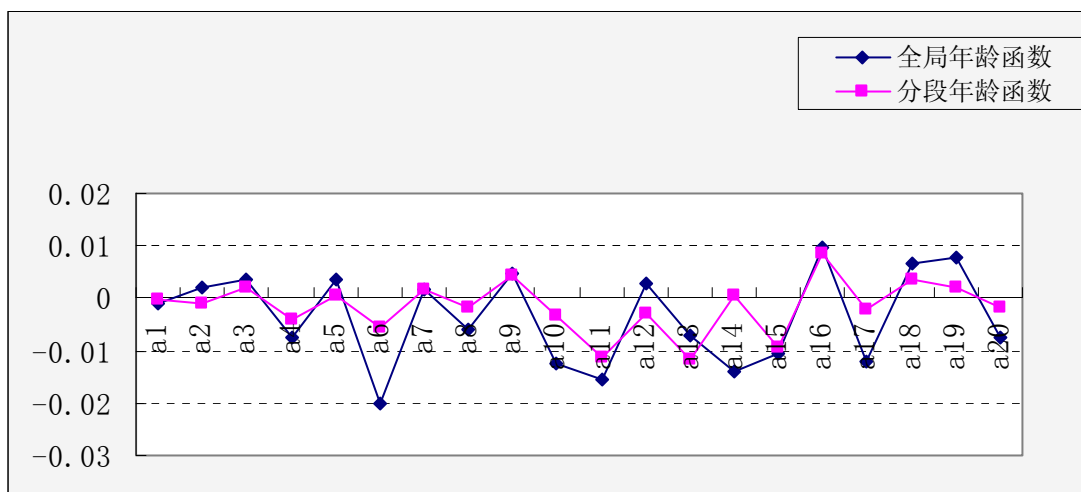


图 4-2 全局年龄函数与 0 岁到 19 岁年龄段年龄函数的系数

使用本次计算出的年龄函数对范围内的人脸图像进行年龄估算的统计结果为，平均误差 4，方差 27，标准差 5.19615。

2. 20 岁到 39 岁部分，有数据 200 条。

应用线性最小二乘法，计算后得出的年龄函数的系数 a 的值如下所示，偏移量 $a_0 = 27.1805$ ，同时与全局年龄函数的系数 a 的值进行对比：

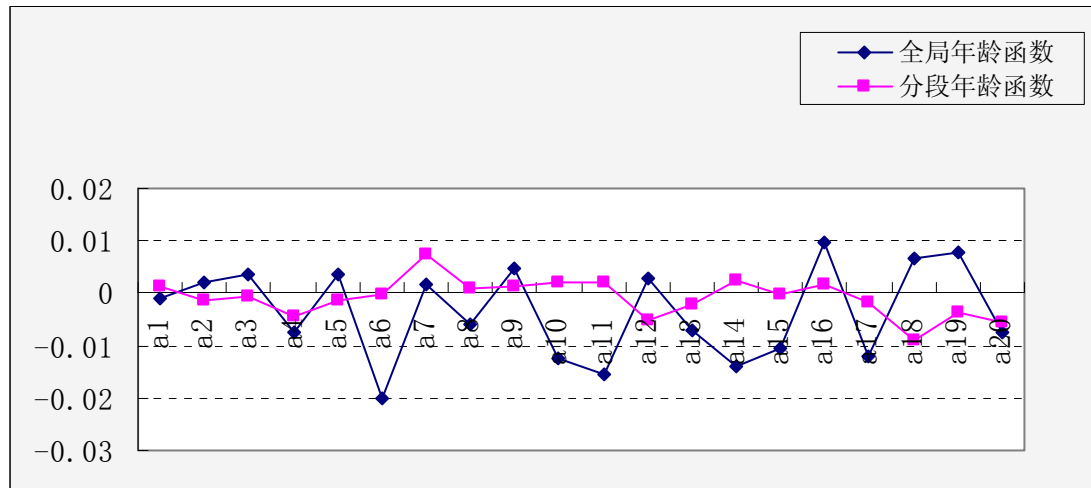


图 4-3 全局年龄函数与 20 岁到 39 岁年龄段年龄函数的系数

使用本次计算出的年龄函数对范围内的人脸图像进行年龄估算的统计结果为，平均误差 4，方差 24，标准差 4.89898。

3. 40 岁以上部分，有数据 64 条。

应用线性最小二乘法，计算后得出的年龄函数的系数 a 的值如下所示，偏移量 $a_0 = 48.2144$ ，同时与全局年龄函数的系数 a 的值进行对比：

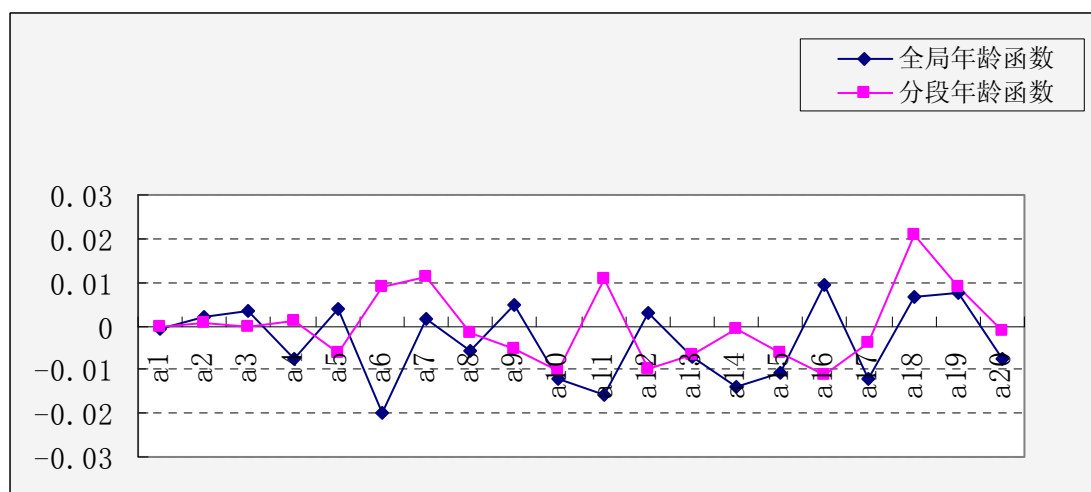


图 4-4 全局年龄函数与 40 岁到 69 岁年龄段年龄函数的系数

使用本次计算出的年龄函数对范围内的人脸图像进行年龄估算的统计结果为，平均误差 4，方差 30，标准差 5.47723。

4.2.3 结果分析

对年龄分段后进行年龄函数的计算，有效地消去了由于不同年龄段数据严重不平衡所造成的影响。同时，在有限的年龄范围内进行年龄函数的计算，使得方程更加有效地拟合数据，从而提高了年龄函数的精度。

不同年龄段，年龄函数的系数 a 的值变化很大，这说明不同年龄段，自变量 x 的权重是不同的，也就是说明在不同年龄段，年龄变化对人脸的主要影响是不同的。在不同的年龄段，受年龄变化影响最大的人脸特征是不同的。

现代的很多自动化识别系统都需要人的参与。一般需要人参与进行简单的判断，而后再交给识别系统进行识别。分段年龄函数在实际使用中也可以考虑由人来判断人脸图像的大致年龄范围，指定人脸图像所处的年龄段，再由程序进行判断。

总体上看，通过以年龄为依据对数据进行分段，对不同年龄段分别求年龄函数，其精确度已经可以让人接受。可以考虑使用最小二乘曲面拟合算法估算年龄，进一步提高精确度。

4.3 最小二乘曲面拟合分段计算年龄函数

4.3.1 算法设计

使用最小二乘曲面拟合算法分段计算年龄函数，将最小二乘曲面拟合算法与分段计算的算法相结合。已经证明，最小二乘曲面拟合算法较之最小二乘线性算法有一定的改进，因此通过本次试验，考察最小二乘曲面拟合在分段计算年龄函数时的效果。

在分段计算年龄函数的前提下，应用最小二乘曲面拟合，计算最高为最小二乘三次曲面拟合。

4.3.2 试验结果

1. 0 岁到 19 岁部分，最小二乘曲面拟合统计结果如下：

表 4-2 0 岁到 19 岁最小二乘曲面拟合结果

使用方法	平均误差	方差	标准差
一次线性	4	27	5.19615
二次曲面	3	22	4.69042
三次曲面	3	21	4.58258

2. 20 岁到 39 岁部分，最小二乘曲面拟合统计结果如下：

表 4-3 20 岁到 39 岁最小二乘曲面拟合结果

使用方法	平均误差	方差	标准差
一次线性	4	24	4.89898
二次曲面	3	21	4.58258
三次曲面	3	18	4.24264

3. 40 岁以上部分，最小二乘曲面拟合统计结果如下：

表 4-3 20 岁到 39 岁最小二乘曲面拟合结果

使用方法	平均误差	方差	标准差
一次线性	4	30	5.47723
二次曲面	3	15	3.87298
三次曲面	0.5	0.625	0.790569

4.3.3 结果分析

通过在分段年龄函数计算中应用最小二乘高次曲面拟合，可以改善年龄函数的计算。但注意 40 岁以上部分，这一部分表面上看是优化效果最好的，但是实际上是由于数据不足造成的。通过前面的统计可知，40 岁以上部分只有数据 64 条，数据的不充分，使得高次的曲面可以较容易地拟合输入数据。这同时也说明，盲目地使用高次的曲面会使年龄函数发生病态，对于外来数据，这个函数的准确度会降低。因此，本题目并非无限制地使用最小二乘高次曲面拟合，而是限制在三次曲面拟合之内。

第五章 工作总结与展望

5.1 工作总结

本文在广泛阅读关于年龄变化和回归算法的参考文献的基础上,通过比较分析,选取最小二乘法作为年龄函数的计算方法。通过合理的算法设计,进行了如下几部分工作:

- 1) 实现了线性最小二乘法:线性最小二乘法是非线性最小二乘法的基础,通常通过公式变换,将非线性最小二乘划为线性最小二乘进行计算求解。在使用线性最小二乘法计算出全局年龄函数之后,对训练集中的人脸图像的年龄进行估算,通过观察分析年龄估算结果,确定了从改进最小二乘法和改进输入数据两方面改进年龄函数的计算。
- 2) 在改进最小二乘法的过程中,首先尝试了最小二乘的曲面拟合算法,并对改进效果进行了分析。而后通过三组数据对其进行测试,排除了对最小二乘法算法的疑问。尝试使用简单的特征降维结合线性最小二乘法对年龄函数进行计算。对其结果进行分析后得出最小二乘法计算全局年龄函数是有一定局限性的结论。
- 3) 在改进输入数据方面,首先通过数据翻倍的方式,使得各个年龄段的数据取得平衡,而后通过最小二乘法计算全局年龄函数。随后以年龄段为依据,为 0 岁到 19 岁,20 岁到 39 岁,40 到 69 岁分别计算年龄函数。通过观察这三个年龄段的年龄函数的系数,得出了在不同的年龄段,受年龄变化影响最大的人脸特征是不同的这一结论。最后,将最小二乘曲面拟合与分段计算年龄函数结合,计算高次的年龄函数。

通过以上工作,充分讨论了最小二乘法对于年龄函数计算的成果与不足,并探讨了在使用最小二乘法的前提下,对其改进的几种方法。

5.2 展望

继续研究基于人脸的全局年龄函数的计算具有重要的理论意义和实际应用价值。本文仅在不改变特征提取方法的前提下,应用最小二乘法对全局年龄函数进行计算。今后,还可以从以下两个方面展开进一步的研究工作。

- 1) 从特征提取方法上改进。PCA 特征提取是线性子空间方法中的一种。线性子空间方法还有两个比较著名的特征提取方法:线性判决分析和独立元分析。在人脸识别这一领域,特征提取方法与识别方法是有一定联系的,对于不同的识别方法,适用的最佳特征提取方法是不同的,没有一种特征提取方法会被认为是最好的,因此可以从尝试不同的特征提取方法上改进年龄函数的计算。

2) 可以考虑在年龄函数的形式上改进。本文选取了一次多项式、二次多项式和三次多项式对全局年龄函数进行计算。通过改进年龄函数的形式,可以获得不同的类型的年龄函数,从而得到不同的拟合结果。

参考文献

- [1]Andreas Lanitis, Chris J. Taylor, Timothy F. Cootes. Modeling the process of ageing in face images[J]. in Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on. 1999.
- [2]Gandhi, M. . a method for automatic synthesis of aged human facial images[J]. Department of Electrical & Computer Engineering. McGill University, 2004.
- [3]Shaohua Kevin, Z. , et al. Image based regression using boosting method[J]. in Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. 2005.
- [4]鲍正祥. 浅谈人脸识别技术[J]. 科技创新导报, 2008, 31 (2): 57-59.
- [5]王俊艳, 苏光大, 林行刚. 人脸图像年龄估计[J]. 清华大学学报, 2007, 47 (4): 526-529.
- [6]孙晔. 具有年龄变化的人脸识别[D]. 江苏: 江苏大学, 2006.
- [7]吴洪森. 基于特征区域分析的人脸自动识别应用研究[J]. 中国人民公安大学学报, 2004 (39): 68.
- [8]余庆, 杜吉祥. 基于一种改进 NMF 算法的人脸年龄估计方法[J]. 中国图象图形学报, 2008, 13 (10): 1865-1868.
- [9]胡澜, 夏利民. 基于 Boosting RBF 神经网络的人脸年龄估计[J]. 计算机工程, 2006, 32 (19): 119-201.
- [10]陈君. 基于 HMM 具有年龄变化的人脸识别技术研究[D]. 江苏: 江苏大学, 2008.
- [11]闫娟, 程武山, 孙鑫. 人脸识别技术研究与发展概况[J]. 电视技术, 2006 (12): 81-84.
- [12]张宇, 周志华. 基于集成的年龄估计方法[J]. 自动化学报, 2007, 34 (8): 526-529.
- [13]聂翔. 最小二乘法在曲线拟合中的实现[J]. 陕西工学院学报, 2000, 16 (3): 79-82.
- [14]冯天祥. 多元线性回归最小二乘法及其经济分析[J]. 经济师, 2003 (11): 129.
- [15]田玉刚. 非线性最小二乘估计的遗传算法研究[D]. 武汉: 武汉大学, 2003.
- [16]赖小平. 正定二次规划的投影最小二乘算法[J]. 山东大学学报(理学版), 2004, 39 (4): 62-67.
- [17]祝媛. 基于多分类器融合的 3D 人脸年龄识别[D]. 北京: 中国科学技术大学, 2008.

致 谢

本课题的研究工作是在张坤龙老师的悉心指导下完成的。在整个毕业设计过程中，多次得到张老师的帮助，使我受益匪浅，明确了研究方向。张老师严谨的思想使得本文的逻辑更加严谨，结构更加合理。

感谢金波学长就 OpenCV 的使用和筛选实验用人脸图像库提供帮助，并对图像处理和 PCA 特征提取的相关问题的耐心指导。