****



**本 科 毕 业 设 计**

院 系 计算机科学与技术系

专 业 计算机科学与技术

题 目基于K匿名及差分隐私的定域功能性影像隐私保护

年 级 2013级 学 号 131220124

学生姓名 肖鹏

指导老师 华景煜 职 称 讲师

论文提交日期 2017.05.22

**南京大学本科生毕业论文（设计、作品）中文摘要**

题目： 基于K匿名及差分隐私的定域功能性影像隐私保护

计算机科学与技术 院系 计算机科学与技术 专业 2013 级本科生姓名： 肖鹏

指导教师（姓名、职称）： 华景煜、讲师

摘要：

近些年来，随着机器学习及数据挖掘研究的深入与进步，大数据产业变得十分火热，无论是学术机构或者是社会部门，都十分关注大数据分析带来的收益与相关产业的发展。就目前的数据交换现状来看，数据的交互过程中存在严重泄露的风险，使得拥有大量数据的部门如刑警部门在需求相关学术机构对其所持有的数据进行分析时会有对数据泄露风险的担忧。

本文的主要内容是针对影像数据（照片，视频资料等），在大数据交互分析中会产生泄漏风险的部分，人脸信息，寻求了一种较为安全的隐私保护构想。通过借助差分隐私模型，通过对人脸数据添加干扰噪音，从而达到即便攻击者已经掌握了除必要数据以外的辅助数据，该攻击者仍然无法从影像资料中推测出具体到个体的行为。同时为了更好的满足数据可用性，保护手段中同时保障了数据发布的K匿名性。为此我们将该算法的实现分成三部分，第一部分利用人脸识别算法，将影像资料中的人脸数据提取到标准尺寸，并按规则划分。预先提取需要处理的资料。第二部分是本保护算法的重点，依据不同规则划分出的人脸单位进行内部面部相似度比较，每个个体提取出前K个与之最相似的人脸。将每个个体与之最相似的K－1个个体进行加权线性叠加，建立新的对应个体。第三部分，将新生成的个体与原数据影响进行面部相似度比较，重现提取前K个分析是否达到保护要求，并将新生成的个体重新覆盖到原始的影像资料中。

为了验证算法的有效性，收集356个不同的个体数据划分后进行交叉测试。实验结果证明经过标准化提取以及规则划分后加权线性重叠处理后，新生个体与原始个体平均相似度只有。。％,在所有数据个体最相似排名中基本都大于K，满足K匿名的要求。同时，整个算法执行过程在单规则分类下，167个个体处理完成及输出比对分析结果需要15min，这些结果说明本文提出的算法是高效可行的。

关键词：影像隐私保护；差分隐私；K匿名；图像识别

**南京大学本科生毕业论文（设计、作品）英文摘要**

THESIS：The Design and Implementation of A CFO-Feature Estimation Algorithm for Wireless Network Devices

DEPARTMENT：Computer Science and Technology

SPECIALIZATION：Computer Science and Technology

UNDERGRADUATE：Peng Xiao

MENTOR：JingYu Hua

ABSTRACT：

KEY WORDS:

**目录**

[第1章 绪论 1](#_Toc476846183)

[1.1研究背景和意义 1](#_Toc476846184)

[1.2研究现状 2](#_Toc476846185)

[1.3本文研究内容 4](#_Toc476846186)

[1.4论文结构安排 4](#_Toc476846187)

[第2章 相关工作与背景介绍 6](#_Toc476846188)

[2.1灰度直方图 6](#_Toc476846189)

[2.1.1三通道RGB图像及MAT矩阵 6](#_Toc476846190)

[2.1.2灰度直方图转换算法 8](#_Toc476846191)

[2.2直方图对比算法 8](#_Toc476846192)

[2.2.1 Correlation 8](#_Toc476846193)

[2.2.2 Chi-Square 9](#_Toc476846194)

[2.2.3 Intersection 10](#_Toc476846195)

[2.2.4 Bhattacharyya距离 10](#_Toc476846196)

[2.3 图像可调加权线性合成算法 10](#_Toc476846197)

[2.4 隐私模型 11](#_Toc476846198)

[2.4.1 K匿名 12](#_Toc476846199)

[2.4.2 差分隐私 12](#_Toc476846200)

[2.5小结 13](#_Toc476846201)

[第3章 具体实现方案 15](#_Toc476846202)

[3.1方案概览 15](#_Toc476846203)

[3.2数据预处理 16](#_Toc476846204)

[3.2.1按规则划分数据集 16](#_Toc476846205)

[3.2.2人脸识别提取标准化 18](#_Toc476846206)

[3.3 相似度对比 21](#_Toc476846207)

[3.3.1 人脸图像灰度直方图化 21](#_Toc476846208)

[3.3.2 直方图交叉对比 24](#_Toc476846209)

[3.3.3 提取最相似的前K个主体 26](#_Toc476846210)

[3.4 图像加权线性叠加 26](#_Toc476846211)

[3.4.1 差分隐私噪音选择 26](#_Toc476846212)

[3.4.2 权值选择 26](#_Toc476846213)

[3.4.3 叠加生成新主体 26](#_Toc476846214)

[3.5 生成样本测试 26](#_Toc476846215)

[3.5.1 K匿名检测 26](#_Toc476846216)

[3.5.2 差分隐私测试 26](#_Toc476846217)

[3.6算法实现总体流程图 27](#_Toc476846218)

[第4章 实验 28](#_Toc476846219)

[4.1实验环境与数据集 28](#_Toc476846220)

[4.2算法精度 29](#_Toc476846221)

[4.3算法时间效率 31](#_Toc476846222)

[4.4实验结论 32](#_Toc476846223)

[第5章 总结与展望 33](#_Toc476846224)

[5.1本文工作总结 33](#_Toc476846225)

[5.2本文的不足与展望 33](#_Toc476846226)

[参考文献 V](#_Toc476846227)

[致谢 VIII](#_Toc476846228)

第1章 绪论

1.1研究背景和意义

近些年来，随着机器学习及数据挖掘研究的深入与进步，大数据产业变得十分火热，无论是学术机构或者是社会部门，都十分关注大数据分析带来的收益与相关产业的发展，数据持有者与研究机构拥有大量的合作机会。但同时，数据发布所面临的安全风险就值得格外关注。从大数据产业的角度来看，数据拥有庞大的内在价值，同时从数据收集的个体来讲，数据中又同时携带有大量的个体隐私，如搜索纪录，银行卡信息，生理信息。相较于其他而言，文字数据类信息要相对容易进行匿名化处理，但是对于带有重要生理信息如人脸等的视频，照片影像类资料在发布过程中，其需要既要兼顾个人信息匿名化与数据可用性，这就使得数据隐私保护变得较为困难。

举个例子，刑警发现一起小区内盗窃案件，确定不是外来盗贼，刑警部门想要通过摄像头拍下来的视频资料分析小区内住户的日常行为有无异常。假设现在有多家研究机构可以帮助警方来进行分析，但是警方就会十分担忧这些影像资料发布后会带来的住户人脸数据泄露风险。

为了有效解决影像资料在发布时的隐私泄漏问题，人们如今采取了许多措施，比如直接把人像位置扣去，增加对应的编号等。还有给人脸面部打马赛克进行模糊化处理的。然而影像资料和普通的字符串数据资料不同，在分析起端需要进行图像识别处理工作，以上的处理方式会严重的造成数据可用性的下降，给分析工作造成困，另外大部分的处理发布也没有考虑满足差分隐私模型。在种种方案都不尽如人意的情况下，我们考虑结合K匿名及差分隐私的构想，同时兼顾数据可用性的前提下，选择通过在确定工作范围下，依据具体规则（如某特定地点）划分，提取人脸特征，比对后加权线性合成，替换原始人脸图像特征。

为了完成提取，对比，合成新的人脸图像同时兼顾完成差分隐私及K匿名的要求，需要寻求合适的人脸图像特征比对度量单位，还有合成中权值的选择。故接下来会介绍下目前图像特征相似度量以及一些图像隐私工作的研究现状。

1.2研究现状

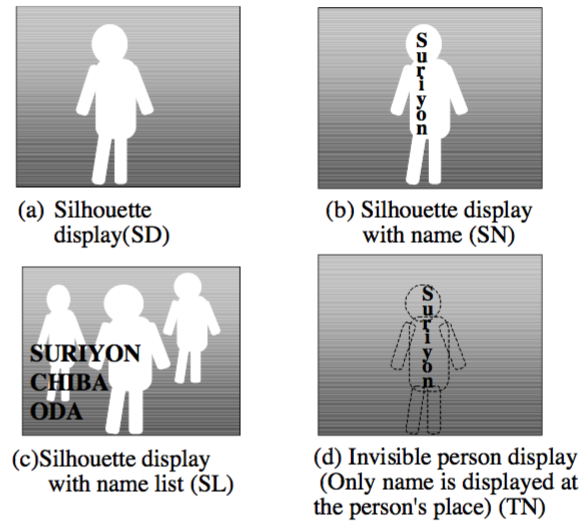
图像相似度比较是人脸识别中一个重要组成部分，也同时广泛应用于大数据分析，机器学习，信息检索，统计学等多个领域。图像相似度对比的算法有很多，应用场景与要求也差异颇大，最常用的是数字图片灰度直方图化后，通过四种直方图对比标准来判定两张图片的相似程度。

对比的要求首先要将图像转化为灰度直方图，灰度直方图是灰度级的函数，描述图像中该灰度级的像素个数（或该灰度级像素出现的频率）：其横坐标是灰度级，纵坐标表示图像中该灰度级出现的个数（频率）。这里要求图像是单通道的灰度图像，对应建立一维直方图，对于彩色图像的例如RGB图像，可以对应不同通道建立多维直方图。之后可以应用Correlation，Chi-Square，Intersection，Bhattacharyya距离来度量相似程度。对于上面的四种对比标准，属于统计学领域范畴，考虑到本次实验的数据标准参考标准照片库，为保证试验的有效性，将对实验数据结果通关以上四种标准混合检验。

图像发布，传输中出现的隐私泄漏问题在近几年社交平台发展，大数据分析技术提升中越发引人关注，也有很多学者提出了自己的保护方式及观点。在一份方案中（Privacy-CNH: A Framework to Detect Photo Privacy with Convolutional Neural Network Using Hierarchical Features），探究了在facebook上发布私人照片的隐私泄漏问题，其探究的面较广，不仅限于人脸信息，还有信用卡，身份证等。作者选择通关图像识别技术将敏感部分直接制空后再发布。但是作者并没有探讨这些影像资料在学术研究及大数据挖掘中的价值，其保护方式会对影像数据的可用行严重破坏。如下图

**图1-1 （1）中图片处理样例**

还有一篇论文中考虑到了视频资料的研究价值（Privacy protection by concealing persons in circumstantial video image），给出了逐步递进的四种处理图像的方式，Silhouette display，Silhouette display with name，Silhouette display with name list，Invisible person display (Only name is displayed at the person's place)，其核心思想是依据不同的保护等级对视频资料进行保护，可以说很大程度上保胡了图像的分析可用性，但是没有细致的探讨对差分隐私的考虑，可以说攻击者如果掌握基本的附加信息如身份ID等还是可以很容易的确定具体主体的行为数据。同时，他保护的方式是对整个人体形态进行处理，故对微小的行为信息有一定的干扰。如下图：



**图1-1 （2）中图片处理方案**

另外还有一篇论文中（Preserving Privacy by De-identifying Facial Images）详细探讨了图片隐私保护的分级依据对图片数据可用性，考虑了差分隐私模型定义了完整一套规则去度量不同的图片模糊手段对于图像数据破坏程度。文章中比对了几种基本的图片保护处理方式，加马赛克，直接抹黑，图像聚合，线性叠加等不同方式下度量的高低，依据其试验结果，线性叠加保护的效果远好于普遍认为聚合处理较优的结果。但是文中仅仅考虑了单通道的灰度图像没有考虑三通道情况下的叠加效果，另外也没有探讨在具体环境及应用场景。

1.3本文研究内容

本文为了获得统一标准化的处理后的图像，首先对所有图片进行人脸识别并且将提取的人脸标准化提取，按照模拟的具体规则进行归类，然后基于opencv直方图算法将图像转化为灰度直方图，在同规则类下的人脸直方图互相按照直方图比对规则选择前K个最相似的主体。

然后开始进行加权线性偶合，权数的选择依据四种不同的直方图比较规则进行优化，合成后的图片重新规则话后依据同上规则划分后保存。

接着，为了检验合成后图片的K匿名性和差分隐私性，将合成后的图像重新按照之前的比较方式进行比较，重新选择前K个最相似的合成后主体，其汇总结果与第一轮进行比对看加入噪音后的新主体的K匿名性是否达到要求。为了检验十分满足差分隐私，我们选择在图像库中随机删除一部分图像，然后依据按照上述规则生成新的图像重新对比验证，调整权制参数。

最后，将训练好的参数投入系统重新生成最后的加入噪音后的图像并发布， 并统计最终比对结果。

1.4论文结构安排

第一章为绪论，主要介绍了目前影像资料在发布中可能出现隐私泄漏的风险于目前的研究现状。同时简单的对本文研究的关于在考虑差分隐私以及K匿名安全的保护系统进行了描述，并介绍了其核心的图像对比算法标准以及目前前人关于影像隐私保护的研究现状。

第二章简单地介绍了与本文相关的工作背景和概念定义。首先介绍三通道图像的概念及灰度直方图转换算法。然后介绍图像相似度对比的算法基础，并分别对K匿名性，差分隐私概念进行了介绍，为后面算法据其进行参数选择做好准备。

第三章将本文将图像加密算法分成了3部分进行详细描述，图像标准化提取按规则划分、内部相似度排序加权叠加和轮换检测方法调整参数。并分析了相关部分的运行效率，最后给出了算法结构图。

第四章对实验结果进行分析，统计了在各权值下各个数据集的测试结果，总结了在权值变化下得到的验证结果，在目前数据集上该算法的K匿名成功率92.7%以上，处理效率都较高。

第五章总结了本文的主要研究内容，并对其进行总体评价。盘点了本算法的优势与限制，并指出了就不足之处接下来的工作方向，提出了对于算法的加速可能方案。

第2章 相关工作与背景介绍

2.1灰度直方图

直方图英文名称是(HISTOGRAM)，相机上显示的直方图和PHOTOSHOP使用的直方图都是灰度直方图，从图形上说，它是一个二维图，用坐标表示。横坐标表示图象中各个像素点的灰度级。(0到255个级别,一般人眼能够分辨的只有32个级别，人眼对光的强度变化非常敏感，而对颜色的变化就比较弱，目前，流行的视频压缩软件都是应用这一原理，比如RM)它是多种空间域处理技术的基础。直方图操作能够有效用于图像增强；提供有用的图像统计资料，其在软件中易于计算，适用于商用硬件设备。纵坐标为各个灰度级上图象各个像素点出现的次数或概率.各个软件细分程度不同。

1.表征了图像的一维信息。只反映图像中像素不同灰度值出现的次数（或频数）而未反映像素所在位置。

2.与图像之间的关系是多对一的映射关系。一幅图像唯一确定出与之对应的直方图，但不同图像可能有相同的直方图。

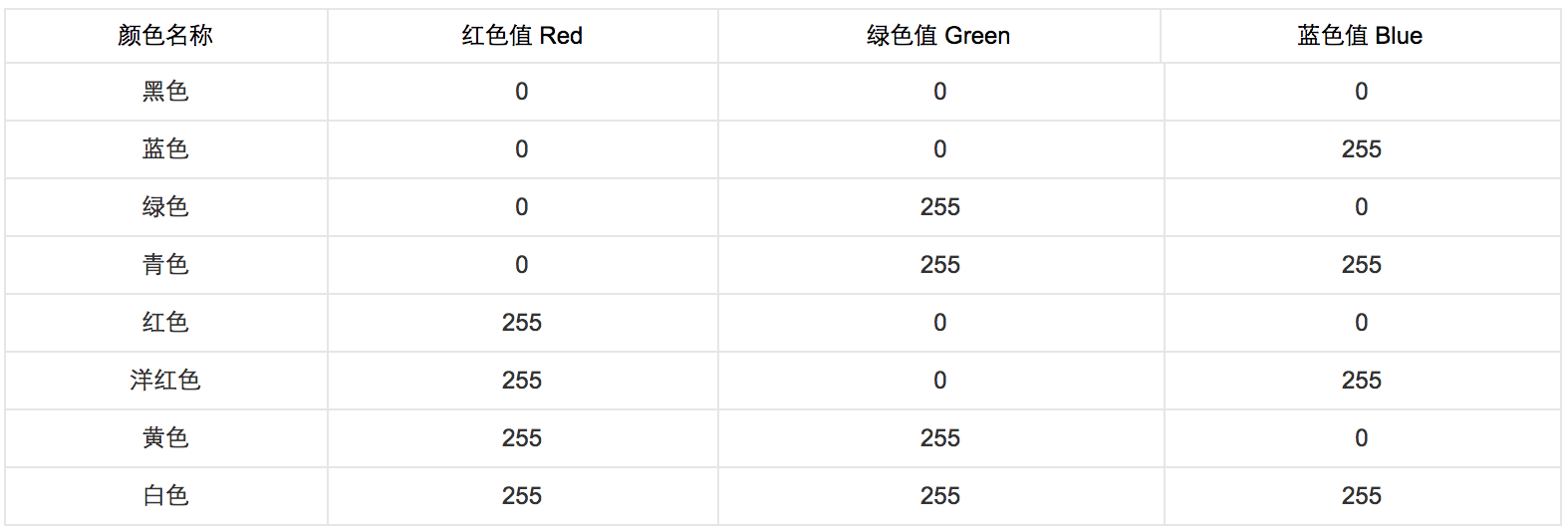
3.子图直方图之和为整图的直方图。

（以上来源于资料）

2.1.1三通道RGB图像及MAT矩阵

RGB色彩模式是工业界的一种颜色标准，是通过对红(R)、绿(G)、蓝(B)三个颜色通道的变化以及它们相互之间的叠加来得到各式各样的颜色的，RGB即是代表红、绿、蓝三个通道的颜色，这个标准几乎包括了人类视力所能感知的所有颜色，是目前运用最广的颜色系统之一。

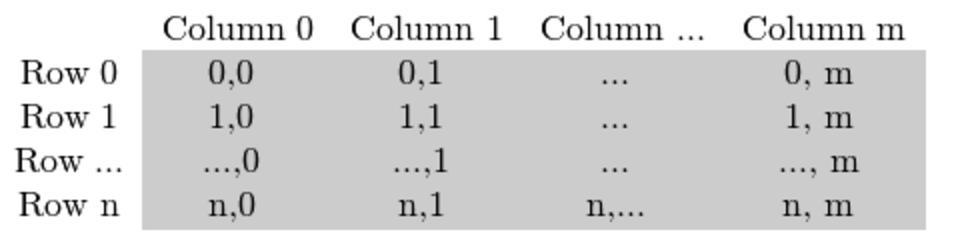
RGB 颜色称为加成色，因为您通过将 R、G 和 B 添加在一起（即所有光线反射回眼睛）可产生白色。加成色用于照明光、电视和计算机显示器。例如，显示器通过红色、绿色和蓝色荧光粉发射光线产生颜色。绝大多数可视光谱都可表示为红、绿、蓝 (RGB) 三色光在不同比例和强度上的混合。这些颜色若发生重叠，则产生青、洋红和黄。



**图2-1 基本颜色三通道数值**

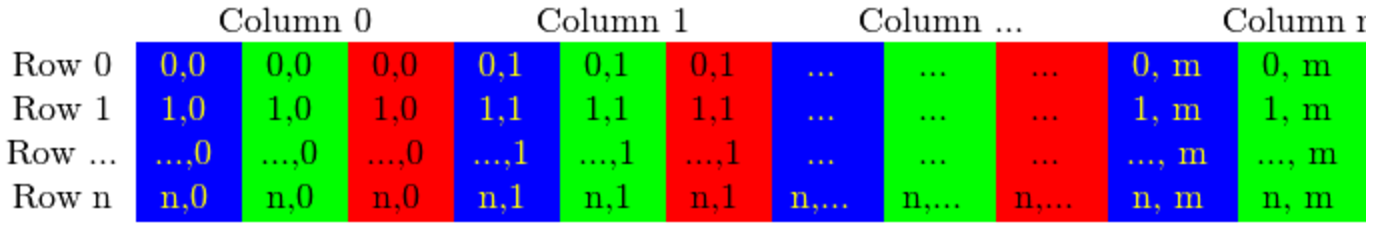
**在**对图片处理时，例如访问每个像素点，修改色彩等通常也是将图片每个像素点分离出三通道数值再进行操作，Opencv中的Mat矩阵给我们提供了方便快捷的操作方式。

如果图像是一幅灰度图像，他就像这样，从左到右，从上到下，依次是矩阵的每一行每一列，这时候矩阵Mat(i,j)的值自然就是当前点的灰度值了。



**图2-2 灰度图像Mat矩阵构成**

而对于一幅彩色图像，由于它的像素通道并不是一个，所以每一列又分为了通道。拿常见的RGB图像来说，就像这样：



**图2-3 RGB图像Mat矩阵构成**

从上图可以看出内存中是如何存放多通道图像，但是实际上在RGB模型中，Mat每一列依次是Blue，Green，Red顺序与名称相反。在本文中，在我们讨论图片相似度对比时往往会优先把RGB图像灰度化，而在讨论加权线性合成时往往是三通道图像。

2.1.2灰度直方图转换算法

灰度直方图是灰度级的函数，描述图像中该灰度级的像素个数（或该灰度级像素出现的频率）：其横坐标是灰度级，纵坐标表示图像中该灰度级出现的个数（频率）。其算法为：

其中，代表某像素个数，代表总像素数，高维直方图可以理解为图像在每个维度上灰度级分布的直方图。常见的是二维直方图。如红-蓝直方图的两个分量分别表示红光图像的灰度值和蓝光图像灰度值的函数。其图像坐标（Dr,Db）处对应在红光图像中具有灰度级Dr同时在蓝光图像中具有灰度级Db的像素个数。这是基于多通道数字图像，二维中对应每个像素统计个变量。本文中在对比图像相似度时是将RGB图像转化为灰度图像，求得其灰度直方图再进行对比的。

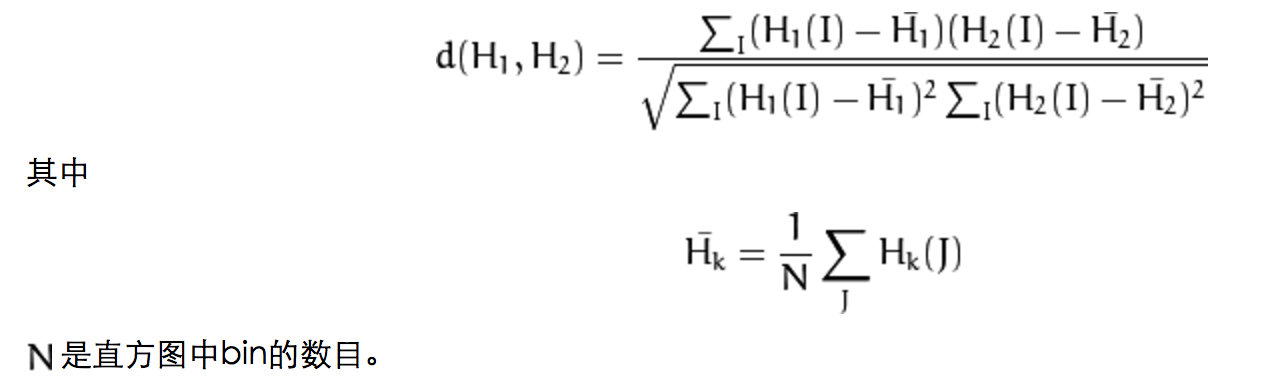
2.2直方图对比算法

为了度量多幅图片之间的相互相似程度，我们在将图片灰度直方图化后，需要选择相应的标准进行度量，以下将对统计学中四种比较标准进行介绍。

2.2.1 Correlation

Correlation相关系数，或称线性相关系数、皮氏积矩相关系数（Pearson product-moment correlation coefficient，PPCC）等，是衡量两个随机变量之间线性相关程度的指标。它由卡尔·皮尔森（Karl Pearson）在1880年代提出，现已广泛地应用于科学的各个领域。依据相关现象之间的不同特征，其统计指标的名称有所不同。如将反映两变量间线性相关关系的统计指标称为相关系数（相关系数的平方称为判定系数）；将反映两变量间曲线相关关系的统计指标称为非线性相关系数、非线性判定系数；将反映多元线性相关关系的统计指标称为复相关系数、复判定系数等。

相关系数是衡量观测数据之间相关程度的一个指标，相关关系是一种确定性关系，相关系数是研究变量之间线性相关程度的量，一般情况下，相关系数越大表明相关程度越高，其算法公式如下图：



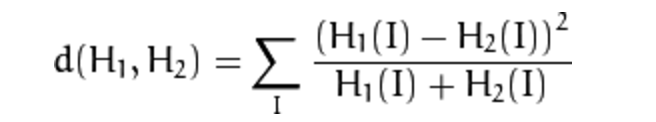
**图2-4 相关系数计算方法**

以上来源Opencv官方手册

2.2.2 Chi-Square

Chi-square test方检验是用途非常广的一种假设检验方法，它在分类资料统计推断中的应用，包括：两个率或两个构成比比较的卡方检验；多个率或多个构成比比较的卡方检验以及分类资料的相关分析等。

卡方检验就是统计样本的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，卡方值越大，越不符合；卡方值越小，偏差越小，越趋于符合，若两个值完全相等时，卡方值就为0，表明理论值完全符合，其算法公式如下图：

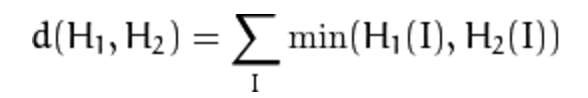


**图2-4 卡方检验计算方法**

以上来源Opencv官方手册

2.2.3 Intersection

Intersection直方图交叉距离是一种较为特殊的集合差异检验，通过取两个类中相同属性值中较小的值累加，其累加之值为度量两个集合直方图差异的数据，值越大，两个集合越相似，其计算方式如下图：



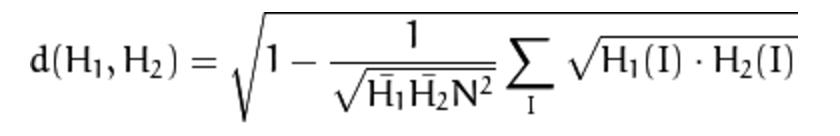
**图2-6 交叉距离计算方法**

2.2.4 Bhattacharyya距离

巴氏距离，它在统计学中用于测量两种离散概率分布的可分离性。在直方图相似度计算时，巴氏距离获得的效果最好，但计算是最为复杂的。设为的巴氏距离计算结果，其值完全匹配为1，完全不匹配则为0。

Bhattacharyya系数（Bhattacharyya Coefficient, 巴氏系数）是对两个统计样本的重叠量的近似计算。巴氏系数可用来对两组样本的相关性进行测量。

计算巴氏系数涉及到对该两个样本的重叠部分进行基本形式的积分。两个样本值的积分被分成指定数目的部分。而每一个样本的每一个部分的成员数被用于下式中：



**图2-6 巴氏距离计算方法**

2.3 图像可调加权线性合成算法

两个图片的融合，首先两个图片大小必须一致，如果不一致怎么办，要使用ROI区域，找到大图片中要和小图片进行融合的尺寸。两个图片的像素值，按照给定的权值进行叠加，就可以将两个图片融合起来。

其算法表达式如下：

dst= α\*src1 + β\*src2 + γ

在三通道RGB图像下，其算法模型如下：

B = psrc1[j\*channels] \*alpa + psrc2[j\*channels] \*beta + gama;

G = psrc1[j\*channels + 1] \*alpa + psrc2[j\*channels + 1] \*beta + gama;

R = psrc1[j\*channels + 2] \*alpa + psrc2[j\*channels + 2] \*beta + gama;

if (B>255)

B = 255;

if (G>255)

G = 255;

if (R>255)

R = 255;

pdst[j\*channels] = B;

pdst[j\*channels + 1] = G;

pdst[j\*channels + 2] = R;

（以上代码转换成执行图）

2.4 隐私模型

随着信息技术的发展,个人隐私的概念发生了很大的变化,人们对隐私信息的保护也越来越重视。另一方面,信息的大量增长,使人们陷入了数据丰富而知识匮乏的困境,数据挖掘技术能够对海量数据进行处理,发现数据中潜在的关联关系。然而,数据挖掘技术的应用增加了隐私泄露的风险,因此数据需要在发布前进行匿名化处理,以提供对个人隐私信息的保护。（以上来源《隐私保护模型研究》－重庆大学－2009－石香灵）为了度量某种保护手段所起到的作用，近些年研究出现了很多种隐私保护模型，例如K匿名，差分隐私，l-Diversity等，本文中设计的影像保护算法模型兼顾数据的可用性及匿名性，以下就对涉及到的K匿名及差分隐私模型进行简单叙述。

2.4.1 K匿名

数据表的k-匿名化( k-anonymization) 是数据发布时保护私有信息的一种重要方法。 k-匿名技术是1998 年由Samarati和Sweeney [1]提出的 ,它要求发布的数据中存在一定数量(至少为k) 的在准标识符上不可区分的记录,使攻击者不能判别出隐私信息所属的具体个体,从而保护了个人隐私, k-匿名通过参数k指定用户可承受的最大信息泄露风险。k-匿名化在一定程度上保护了个人的隐私,但同时会降低数据的可用性。因此, k-匿名化的研究工作主要集中在保护私有信息的同时提高数据的可用性。

假设一条微数据记录对应于一个个体，从广义集合论的角度出发，K-匿名就是要求在一个集合中只能以不大于1/k的概率确定任何一个元素，即要求任何一个元素，在集合中至少存在k-1个和其相同副本元素。个体特征通过相应的属性值来区分。将上述K-匿名概念引申到关系型数据库中，称一个关系表是K匿名化的，当且仅当对于任何一条微数据，在所有准标识属性的投影中，该表中存在至少k-1条记录与之完全相同。（以上来源于网络）

k-匿名的基本思想是通过匿名化（如泛化和隐匿）原始数据集中的某些属性值形成满足一定匿名要求的匿名数据集并可用于数据发布。而在信息发布时实施数据隐私保护要考虑两个方面问题:（1）确保数据发布过程中隐私不泄露；（2）发布的匿名共享数据具有实用性。（《基于K-匿名的隐私保护方法研究》《哈尔滨工程大学》 2012年 任向民）

本文中描述的影像资料隐私保护算法正是借助k匿名的概念，保证分组发布的用户数据即使攻击者掌握了用户的身份ID信息，也只有小于等于1/K的概率对应到相应的主体信息。

2.4.2 差分隐私

为解决当前信息越来越发达的社会所带来的用户隐私泄露问题，差分隐私模型是一种被广泛认可的严格的隐私保护模型。它通过对数据添加干扰噪声的方式保护所发布数据中潜在的用户隐私信息，从而达到即便攻击者已经掌握了除某一条信息以外的其他信息，该攻击者仍然无法推测出这条信息。因此，这是一种从数据源头彻底切除隐私信息泄露可能性的方法。然而，该模型的最基础来源是基于抽象的数学理论，使得该领域的入门需要较高的门槛。

数据集D 发布时，通过删除标识符属性（例如姓名、ID号等）能够在一定程度上保护个人隐私，但这远远不够，要特别注意到，数据集中还有其他属性，例如：生日、性别、居住地、是否抽烟、是否饮酒等，用这些信息来猜测个人身份，是不是类似数据挖掘中的分类？即将一个没有类别标签的条目识别归类，训练集看成敌手A 从其它地方获得的具有标识符属性的与 D 有属性交集的数据集T，类别标签当然就是某个人。在这方面，美国曾经有几个经典案例，被起诉方通常都是赔了几百万。

在隐私保护方面，k-anonymity 及其扩展模型影响深远且被广泛研究，但随着研究的深入，该系列模型也面临着许多新型攻击的挑战，于是2006年就出现了DP 。DP 在攻击者A 在拥有最大背景知识条件下，仍能抵御各种攻击。

通常来说，差分隐私机制的实现是通过对数据集添加拉普拉斯噪声来实现的，其核心构想就是对于两个数据集 ，两者假设只有一定数量的数据差异，此时通过一个满足差分隐私（ϵ-DP）查询算法A，去获取具体的条目，其应满足：



本文中通过在图片合成过程中添加噪音，并设计相应的书去查询函数来保障数据发布的差分隐私。

2.5小结

图像相似度是一个很大的学术问题，关于此问题的学术讨论也曾出不穷，尤其是针对各种复杂的光线环境，角度变换等情况。本文讨论的测试样本均建立与ID照片等，故为了算法的效率选择了直方图比对，同时为了算法的严谨性将四种比较标准交叉检验。在本节对图像，直方图算法，对比标准进行了简单的阐述，为后面在交叉测试中参数的选择进行铺垫。

数据的隐私发布也是一个十分火热的话题，K匿名及差分隐私是目前最适用的数据发布隐私模型，但同时两者的应用效果也略有差异，在不同问题的优缺性也存在差异，本文在设计保护模型中同时兼顾这两种保护模型，故在本节对以上所需概念进行了阐述，为后面算法的扩展及噪音的添加说明打下基础。