# 基于机器学习的色彩快速检测系统

梁早清 陈宏彪 林文娜 吴尉溶

# 摘要

随着互联网在当今世界的应用越来越普及，各类人机交互的智能方向层出不穷，机器学习越来越成为当代互联网的主流。机器学习，是以一种方式来赋予机器以学习的能力，通过不断地输入数据，产生处理大数据的算法，然后使用模型进行预测的一种人工智能。基于机器学习的色彩检测也是近几年互联网和工业结合关注的热点之一，值得研究。

在色彩检验中，通过调用手机摄像头，通过获取触定点的RGB值，通过相关公式规则及其算法计算直接得到该点相关特性，比如灰度值，pH值等，本次项目是通过机器学习算法如k-nearest-neighbours算法 k-means算法来得出检测结果。

本项目采用Java高级语言搭建了快速检测系统，用Python高级语言实现机器学习算法，实现了色彩检验技术。

## 关键词：机器学习 色彩检测 多元线性回归算法 k-means算法

# 第一章 绪论

## 1.1选题背景和研究意义

机器学习是一门研究如何让计算机程序来进行类似人类学习的多领域交叉学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。在我们的现实生活中，有一些问题，比如如何让计算机识别我们手写的数字，无人驾驶技术，还有我们经常使用的购物平台根据我们的浏览记录来对我们的喜好进行猜测并推荐物品等等，这些只用编程技术是无法实现的。计算人不像人类一样有思维，它不可以根据实际情况进行判断，而人类在一次经历之后会总结经验，经验不断累积，在下一次遇到同类问题的时候，会根据自己的经验对实际情况进行判断。人类在积累丰富的经验之后使自己的行为产生变化，这一过程可以称为学习。机器学习便是用大量的已有的数据的对计算机进行训练，训练出一种算法模型，最终达到某种技能。

简言之，机器学习就像一个小孩子的学习过程，从刚开始的一无所知到最后逐渐掌握技巧，人类学习的技巧就相当于机器自身创造了一种普适算法，只要你给机器一堆大数据，它能从大数据中挖掘出自己的逻辑，从而模拟或实现人类的行为。

随着颜色工业的发展，对颜色产品的要求越来越高，颜色检测系统在工业中发挥着日益重要的作用，在食品行业和医疗行业等领域也有着不可替代的地位。传统的颜色识别依靠人眼识别，误差极大，速率极慢，已经不能满足现如今高速发展的工业需求了。不仅如此，在目测法的方法下，数据的记录及其整理及其麻烦，不能形成一个有序的大数据库。因此，快速颜色检测技术应运而生。快速颜色检测是计算机视觉的一个热点，具有重要的应用价值，普遍使用于工业、食品安全检测等方面。同时使用机器视觉用机器代替人眼来做测量和判断，再把人工智能学习运用到规则自动学习，使用算法来解析数据、从中学习，然后对真实世界中的检测事件做出决策和预测，具有重要的理论研究意义。同时，把采集到的数据及测试出的数据规范整理，形成完整的数字信息体系。

## 1.2 研究现状

色彩检测目前是国内外工业上的一种非常重要的检查目标，随着科技的不断发展，使用高精度高速率的色彩检测代替传统的人眼对颜色检测的估计是一个重要的进步。借助机器的客观检测，不仅能统一工业检测标准，也能使产品的颜色信息能以数字化的形式呈现在数据库中。

我国工业基础相对发达国家薄弱。发达国家在工业发展的过程中就意识到了颜色检测的重要性。早在上世纪50-80年代，美国AppliedColorSystems、英国的Instru-mentalColorSystems和瑞士Datacolor是世界上三个最主要的色控仪器公司(这三家公司于1991年被瑞士一财团收购合并成现在的DatacolorInternational)。另外一些重要的公司如美国Macbeth、SheLyn、Hen-negan、BYK)Garder等,德国Zeiss、Optronic等以及英国等国家著名光学公司均研制实用的测色与配色系统,并且相当数量出口。国内色彩检测仪器基本被国外产品所垄断，但由于价格高昂，并且国内对相关色彩检测的理论和实践相当落后，至今仍有一些生产单位色彩检测仍要依靠工作人员的经验和目视判断，不利于我国的工业发展。因此，研究高精度、高速率、低价格的颜色检测方式成为国内的必然方式。

## 1.3 研究的难点和要点

（1）自定义相机校准环境

检测之前，需要对环境进行校准，比如不同光照条件下结果也要相同，不能因为亮度不同结果就相差很大。所以需要一种校准环境的算法，来避免环境影响。

本项目拟通过自定义相机的预览，对焦，取景过程，在预览画面中实现对焦，获取该点RGB，然后继续预览。

（2）图片定点数据的获取

要实现相机画面的实时预览，对焦点的显示，需要自绘控件View，Android自带的控件是不能够绘制触摸点的对焦点的。通过截取那一刻的数据流再分析坐标来实现定点数据的获取。

（3）用户自定义检测规则的识别

采用人工输入，当用户使用检测的时候先把要使用的规则输入系统，系统会保存规则，用户之后可以自行选择不同规则使用，当检测出RGB值后系统会根据所选规则来计算结果。

用户输入的公式是一条String字符串，需要对这条字符串进行解析，结合检测的RGB值计算结果。一般检测的公式是含有比较多的加减乘除的，有一些带有平方开根号等的就更复杂了，正确解析这条字符串是一个需要解决的关键问题。

（4）机器的学习

这种通过机器学习来实现自动学习规则，自动制定规则，减少手工输入规则，增加规则准确性。要解决的关键问题在于要实现基于机器学习的规则学习。机器学习是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。

# 第二章 研究方法

## 2.1 颜色检测系统的概述

许多人工检测的方式中，是存在缺陷的，比如pH试纸检测，是通过比对显色卡颜色，人眼对比会导致一定的误差，用滴定法测试过程又太过复杂。一些检测方法中，检测过程往往复杂耗时，检测出一些物质属性后往往还需要大量的多位小数点的计算，例如检测灰度值，吊白块浓度等。检测还容易受到人为因素、环境因素的影响而使得检测结果出现偏差或者效率低、速度慢的问题。此项目中所研究的快速检测系统基于解决这些问题而开发，实现高效并且准确的智能快速检测。

快速检测系统通过调用手机摄像头，通过获取触定点的RGB值，通过相关公式规则计算直接得到该点相关特性，比如灰度值，pH值等，提高了检测的准确性，避免了一些误差，提高了检测速度。

使用机器视觉用机器代替人眼来做测量和判断，再把人工智能学习运用到规则自动学习，使用算法来解析数据、从中学习，然后对真实世界中的检测事件做出决策和预测。



## 2.2 总体思路

（1）自定义相机，surfaceView和ImageView：相机和surfaceView实现相机实时预览， ImageView实现焦点小圆圈效果，通过机器视觉识别待测样品的RGB。

（2）新建规则：新建规则有两种方法，一种是用户输入规则，另一种是机器学习规则，而在颜色检测时大多时候用户无法知道检测的规则，便采用机器学习的方法让计算机学习规则。

（3）机器学习规则：用大量的已有的实验数据不断训练计算机，系统会自动进行机器学习，自动识别，得出一种算法模型，再让计算机进行识别，用户判断识别结果是否准确，如果不准确，再进行进一步学习。



## 2.3 实验数据的获取

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| photoshop采集的rgb | | | | 平均rgb | 平均rgb（平均） | PH |
| 122 | 148 | 85 | 122,148,85 | 123,148,,87 | 124,148,88 | 8 |
| 121 | 149 | 90 | 121,149,90 |
| 125 | 146 | 87 | 125,146,87 |
| 124 | 148 | 86 | 124,148,86 |
| 125 | 149 | 91 | 125,149,91 |
|  | | | | | | |
| 58 | 126 | 150 | 58,126,150 | 58,126,148 | 59,126,147 | 9.00 |
| 59 | 127 | 146 | 59,127,146 |
| 60 | 125 | 145 | 60,125,145 |
| 57 | 125 | 146 | 57,125,146 |
| 58 | 126 | 147 | 58,126,147 |
| 60 | 128 | 149 | 60,128,149 |
|  | | | | | | |
| 34 | 111 | 155 | 34,111,155 | 35,110,154 | 36,109,154 | 9.6 |
| 38 | 113 | 155 | 38,113,155 |
| 37 | 108 | 154 | 37,108,154 |
| 37 | 105 | 152 | 37,105,152 |
| 35 | 108 | 149 | 35,108,149 |
| 34 | 111 | 157 | 34,111,157 |
|  | | | | | | |
| 204 | 165 | 34 | 204,165,34 | 204,165,34 | 204,165,34 | 6 |
| 203 | 165 | 32 | 203,165,32 |
| 205 | 167 | 34 | 205,167,34 |
| 205 | 166 | 35 | 205,166,35 |
| 203 | 164 | 35 | 203,164,35 |
| 203 | 165 | 32 | 203,165,32 |
|  | | | | | | |
| 165 | 83 | 61 | 165,83,61 | 166,84,62 | 167,85,62 | 7.1 |
| 160 | 82 | 58 | 160,82,58 |
| 166 | 81 | 60 | 166,81,60 |
| 178 | 88 | 62 | 178,88,62 |
| 170 | 90 | 67 | 170,90,67 |
| 164 | 88 | 65 | 164,88,65 |
|  | | | | | | |
| 54 | 21 | 50 | 54,21,50 | 54,21,50 | 54,21,51 | 10 |
| 56 | 21 | 51 | 56,21,51 |
| 55 | 22 | 51 | 55,22,51 |
| 54 | 21 | 52 | 54,21,52 |
| 50 | 21 | 49 | 50,21,49 |
| 54 | 21 | 50 | 54,21,50 |
|  | | | | | | |
| 187 | 49 | 10 | 187,49,10 | 188,49,10 | 188,50,10 | 4 |
| 188 | 50 | 11 | 188,50,11 |
| 186 | 53 | 10 | 186,53,10 |
| 189 | 47 | 9 | 189,47,9 |
| 187 | 50 | 8 | 187,50,8 |
| 191 | 49 | 13 | 191,49,13 |

## 2.4 环境影响的消除

检测之前，需要对环境进行校准，比如不同光照条件下结果也要相同，不能因为亮度不同结果就相差很大。所以需要一种校准环境的算法，来避免环境影响。

## 2.5 K-means算法的自学习(离散规则学习采用的方法)

**2.5.1 K-Means算法概述**

K-平均算法（K-Means clustering）是一种聚类算法（主要用于将相邻的样本点都自动归到一个聚类中心），是一种典型的无监督学习算法，基于原型的目标函数聚类方法的代表，它是数据点到原型的某种距离作为优化的目标函数，利用函数求极值的方法得到迭代运算的调整规则。K-Means算法以欧式距离作为相似度测度，它是求对应某一初始聚类中心向量V最优分类，使得评价指标J最小。算法采用误差平方和准则函数作为聚类准则函数。



K-Means算法是很典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。

工作原理：基于K-Means算法思想，首先确定k个聚类，并随机选取k个样本点为初始质心（聚类中心），计算每个数据点与样本点的欧氏距离，根据欧氏距离把各个数据点归到各自的质心，每次归类质心后，都要重复计算新的质心，直至质心不再改变或者低于规定阈值，最终确定的质心即为所求的聚类中心。基于大规模的数据集，K-Means算法收敛的速度慢。

**2.5.2 K-Means算法的流程**

输入：聚类个数k{*m1(1) ,…,mk(1)*}，以及包含n个数据对象的数据库{*x1,…,xn*}

输出：满足方差最小标准的k个聚类{*m1(t) ,…,mk(t)*}

聚类准则函数：一般都采用均方差作为聚类准则函数。K个聚类具有以下特点：各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开。

处理流程：

1. 从n个数据对象任意选择k个对象作为初始质心（聚类中心）；

{*m1(1) ,…,mk(1)*}⊆{*x1,…,xn*}

1. 根据每个聚类对象的质心，计算每个对象与这些质心的欧式距离，并根据最小距离重新对相应对象进行划分；其中mit是t次迭代之后的结果。



1. 重新计算每个（有变化）质心；



1. 循环（2）（3）直至每个聚类不再发生变化为止。

示例图如下：



**2.5.3 K-Means算法的优缺点**

1、优点：

1. 算法快速，简单；
2. 对大数据集有较高的效率并且是可伸缩性的；
3. 时间复杂度近于线性，而且适合挖掘大规模数据集。K-Means聚类算法的时间复杂度是O（nkt）（n代表数据集中对象的数量，t代表着算法迭代的次数，k代表着簇的数目）。

2、缺点：

1. K值的选定是非常难以估计的；
2. 初始聚类中心的选择对聚类结果又较大的影响；
3. K-Means算法在计算过程中不断调整聚类中心，当有庞大的数据量时，算法的时间开销是非常大的。

**2.5.4 K-Means算法的实现**

（1）数据准备：基于2.3的实验数据

（2）检测类型：k=2（阴性和阳性）//或者更多

（3）对数据集进行测试

摆放原始数据集的图

（4）代码展示

## 2.6 多元线性回归算法的自学习(连续规则学习采用的方法)

在回归分析中，如果有两个或两个以上的自变量，就称为多元回归。事实上，一种现象常常是与多个因素相联系的，由多个自变量的最优组合共同来预测或估计因变量，比只用一个自变量进行预测或估计更有效，更符合实际。

社会经济现象的变化往往受到多个因素的影响，因此，一般要进行线性回归分析，我们把包括两个或两个以上自变量的回归成为多元线性回归。多元线性回归与一元线性回归类似，可以用最小二乘法估计模型参数，也需对模型及模型参数进行统计检验。选择合适的自变量是正确进行多元回归预测的前提之一，多元回归模型自变量的选择可以利用变量之间的相关矩阵来解决。

1. 多元线性回归建立的模型如下 ：

 

其中，*β*是自变量的系数。把常数项看成是一个虚变量的系数，在参数估计过程中该虚变量的样本观测值始终为1。模型中的自变量的数目为 。可表示为如下矩阵：



2.参数估计

使用普通最小二乘法进行参数估计如下：

随机抽取因变量和自变量的组样本观测值：



得到模型的参数估计值后，可得到



根据最小二乘法原理，下列方程组的解就是参数估计值：



其中





得到待估参数估计值的正规方程组如下：



解以上个方程便可得到个待估参数的估计值。

寻找一组参数估计值使得残差平方和



最小。即求解方程组



得： ,

故 

随机误差项的方差的无偏估计量为：



又 

故 

3.拟合优度检验

一个回归模型的拟合程度体现在总体平方和与回归平方和的接近程度，即中越小越好。计算公式如下：



为使拟合指标能反应模型中自变量个数的影响，调整为。计算公式为：



式中为残差平方和的自由度，为总变差平方和的自由度。于是有：

本文针对PH的测量问题，以多元回归为理论基础，结合已知数据，让计算机分析出PH值变化的一般规律。首先，确定影响PH值的因素。其次，将影响因素作为自变量，PH值作为因变量让计算机进行多元回归分析，得到最优的回归方程。

### 2.6.1 影响因素确定

本文确定了三个因素会对PH造成影响。分别为待测样品的R（x1）、G（x2）以及B（x3）。

R：待测样品的红颜色通道；G：待测样品的绿颜色通道；B：待测样品的蓝颜色通道。

获得各影响因素的数据值如下：

|  |  |  |  |
| --- | --- | --- | --- |
| R/x1 | G/x2 | B/x3 | ph值/y |
| 170 | 5 | 47 | 0.5 |
| 160 | 18 | 68 | 1 |
| 188 | 1 | 22 | 1.5 |
| 169 | 2 | 32 | 2 |
| 235 | 25 | 27 | 2.5 |
| 212 | 20 | 33 | 3 |
| 230 | 49 | 28 | 3.5 |
| 221 | 57 | 30 | 4 |
| 232 | 110 | 33 | 4.5 |
| 223 | 112 | 33 | 5 |
| 211 | 112 | 27 | 5.5 |
| 193 | 120 | 52 | 6 |
| 144 | 104 | 42 | 6.5 |
| 133 | 99 | 51 | 7 |
| 47 | 58 | 28 | 7.5 |
| 46 | 53 | 35 | 8 |
| 28 | 35 | 47 | 8.5 |
| 29 | 35 | 35 | 9 |
| 18 | 19 | 47 | 9.5 |
| 39 | 39 | 67 | 10 |
| 26 | 22 | 57 | 10.5 |
| 24 | 24 | 60 | 11 |
| 20 | 12 | 51 | 11.5 |
| 10 | 2 | 41 | 12 |
| 33 | 10 | 54 | 12.5 |
| 41 | 18 | 64 | 13 |
| 25 | 12 | 42 | 13.5 |
| 26 | 13 | 43 | 14 |

### 2.6.2 模型建立

建立的PH值模型如下：

，

……

### 2.6.3 PH值模型的检验

采用拟合优度检验回归方程的拟合程度。

……

### 2.6.4 PH值模型的测试

用已有的大量实验数据对模型进行测试，求出误差。

……

# 第三章 APP应用的整体结构

# 第四章 APP结果实验

实验一：离散规则的规则学习与使用，（找王锦）；

实验二：PH值的检测的规则学习与使用；

第一步：规则学习

第二部：检测结果统计

# 第五章 总结

# 第六章 参考引用