黑背景下收获前棉花图像色特征生成及其品级聚类分析

王 玲[1]，姬长英[1*]，陈兵林[2]
（1.南京农业大学工学院，南京 210031；2.南京农业大学农业部作物生长调控重点开放实验室，南京 210095）

摘要：为了客观地评价田间收获前棉花品级，依据我国子棉收购文字标准，运用机器视觉技术，对棉花的尺寸和色泽特征进行了研究，其色泽特征包括6个常用颜色空间下棉花和带壳棉花的黄色区域、黄色深度、白度、色差等8个色泽特征。结果表明，RGB、NTSC、亨特、HSI颜色空间下的特征参数区分度较好；棉花尺寸与色泽特征的相关性表明棉花白度为无效特征。在各颜色空间下，基于有效图像特征对7个品级的棉花样本进行了K均值聚类分析。分析结果说明，棉花品级的聚类结果独立于颜色空间；铃壳色泽对品级的贡献显著；由于其聚类品级与图像特征的相关性普遍较高、较均衡，且算法运行时间较短，HSI颜色空间可能是棉花品级聚类的最优颜色空间。与人工分级相比，机器视觉的区分度更好，因此机器视觉技术可用于提高收获前棉花分级精度。

关键词：图像特征；颜色空间；棉花品级；聚类；有效性
中图分类号：S562 **文献标识码**：A
文章编号：1002-7807(2007)02-0119-05

Color Features Selecting and Grades Clustering for Preharvest Cotton Images with Dark Background

WANG Ling[1], JI Chang-ying[1*], CHEN Bing-lin[2]
(1. *College of Engineering, Nanjing Agricultural University, Nanjing* 210031, *China*; 2. *Key Laboratory of Crop Regulation, Ministry of Agriculure, Nanjing Agricultural University, Nanjing* 210095, *China*)

Abstract: In order to assess the grades of preharvest cottons objectively, boll size and fiber color features, including yellow region, yellow degree, white degree and brightness contrast of cottons with/without bracteoles in six typical color spaces, were investigated based on Chinese government grading standards and machine vision technologies. The results show that better discriminations of feature parameters can be obtained in RGB, NTSC, Hunter and HSI color spaces, white degree is unvalued according to the correlation analysis between boll size and fiber color features. K-means clustering were performed based on valid image features to grade cotton samples into seven categories in each color. The results indicate that clustering results are independent of color spaces, bracteoles color contributes to their grades obviously and HSI color space maybe the best color spaces for grade clustering because of higher and more uniform correlations between grades and features as well as shorter runtime can be obtained in this space. Machine vision can be used to improve sorting accuracy of preharvest cottons since its more exact discriminability comparing with that of human eyes.

Key Words: image features; color space; cotton grade; clustering; validity

---

收稿日期：2006-06-08 **作者简介**：王玲(1966-)，女，在读博士生；*通讯作者，chyji@sohu.com
基金项目：国家863项目(2006AA10Z259)、2005年江苏省农机基金资助(GXZ05013)

我国子棉收购文字标准的品级条件可概括为棉瓣的大小、棉花的色泽、杂质含量以及是否僵瓣、雨锈、轻霜等[1]。长期以来，我国棉花收购以感官检验为主，检验结果易受人为因素影响，必须加快收获前棉花品级检验仪器化的步伐，基于机器视觉提取棉花图像特征参数、建立收获前棉花品级分级模型是其中的必要环节[2]。本研究直接在彩色测量空间选择反映棉花品级条件的图像特征，并依此进行收获前棉花品级聚类。

1 材料和方法

1.1 样本采集

2005年8—11月，在南京农业大学棉花试验田自然光照条件下，用数码相机分批次拍摄黑丝绒布背景下苏棉12号带铃壳的棉花图像402张，拍摄时人为地将其分为7个品级。

1.2 特征选择与处理方法

根据我国子棉收购文字标准，考虑到棉瓣的尺寸可间接反映僵瓣情况，而色泽特征则兼顾杂质含量、雨锈、轻霜等，本试验仅选择尺寸、色泽特征进行如下数据处理：(1) 初步筛选颜色空间，用相关分析剔除无效特征；(2) 基于图像特征对棉花品级聚类，根据聚类品级与图像特征的相关性选取颜色空间及其聚类品级，并与人工分级进行比较。

2 结果与分析

2.1 图像特征生成

2.1.1 *尺寸特征*。利用数学形态学原理将棉花与铃壳分割[3]，统计棉花与带壳棉花面积之比、棉花与铃壳面积之比，这两个特征平均值分别为0.73，3.37，变异系数分别为0.14，0.66，后者变异系数较大，可作为尺寸特征。

2.1.2 *色泽特征*。通常用于区别颜色的特性是彩色(色调、饱和度)和亮度，大量实践表明：棉花的基本色调很接近孟塞尔(Munsell)色轮的10YR[4]，处于采摘期的铃壳色调也基本如此，因此，用饱和度和亮度就可以生成收获前棉花色泽特征：就带壳棉花而言，黄色深度、区域特征是基于饱和度分量分别统计其平均值、黄色区域占总面积的比值，反映棉花光泽的白度、色差特征则基于亮度分量分别统计其平均值、标准差；类似地，可统计棉花的这4个色泽特征。下面分别在线性颜色空间RGB、NTSC、YCbCr和非线性颜色空间HSI、L* a* b*、亨特下进行实验。

RGB颜色空间基于R、G、B三维笛卡儿坐标系统，用于彩色监视器和视频摄像机。Ohta等人通过统计大楼、海滨等多种不同类型的彩色图像的方差[5]，归纳出亮度分量(R+G+B)/3和二个正交的彩色分量(R-G)/2、(2B-R-G)/4，类似地，还可以归纳出另外二组正交的彩色分量(G-B)/2、(2R-G-B)/4和(R-B)/2、(2G-R-B)/4。试验证明(R-G)/2、(2R-G-B)/4、(R-B)/2都可作为饱和度分量，本试验选择变异系数相对较大的(R-G)/2(表1)，全部样本运行时间为288 min。

NTSC彩色制式在美国用于电视系统，由亮度分量Y和二个正交的彩色分量I、Q组成[6]。YCbCr彩色空间广泛应用于数字视频，由亮度分量Y和二个正交的彩色分量Cb、Cr[6]。试验证明这两个颜色空间的饱和度分量I、Cr区分度较好(表1)，运行时间分别为174、236 min。由于YCbCr颜色空间下色泽特征的变异系数近似于RGB颜色空间，不必进一步重复讨论。

表1 棉花色泽特征变异系数

Table 1 Variations coefficients of cotton color features

色泽特征		颜色空间					
		亨特	数字视频	彩色电视	调色板	均匀	彩色监视器
		Hunter	YCbCr	NTSC	HSI	L* a* b*	RGB
黄度	带壳棉花黄色区域	0.83	0.77	0.51	0.45	—	0.79
	带壳棉花黄色深度	0.75	0.66	0.63	0.37	0.02	0.70
	棉花黄色区域	1.50	1.22	0.71	0.76	—	1.28
	棉花黄色深度	1.24	1.00	0.97	0.33	0.02	1.10
亮度	带壳棉花白度	0.13	0.07	0.07	0.07	0.07	0.07
	带壳棉花色差	0.16	0.12	0.12	0.12	0.12	0.12
	子棉白度	0.12	0.05	0.05	0.05	0.05	0.05
	子棉色差	0.18	0.13	0.13	0.13	0.13	0.13

HSI颜色空间是人们用来从调色板或颜色轮中挑选颜色所用的彩色系统之一，它将亮度分量

I 与二个正交的彩色分量 H、S 分开[7]，试验证明其饱和度分量 S 区分效果良好（表 1），运行时间为 259 min。

$L^*a^*b^*$ 是国际照明委员会最通用的均匀颜色空间，强调的是人类能够观察到的颜色区别，也是一个优秀的亮度 L^* 与彩色 a^*、b^* 分离器[7-8]。试验证明其饱和度分量 a^*、b^* 在黄色深度上的变异系数近似（表 1），但黄色区域区分度差（表 1 "—"），运行时间为 1821 min，不必进一步讨论。

亨特（Hunter）颜色空间是美国棉花测色仪 HVI900 使用的颜色系统[4]，由亮度分量 Rd、彩色分量＋a（红色）、-a（绿色）、＋b（黄色）、-b（蓝色）组成，收获前棉花色调介于＋a、＋b 之间。试验证明饱和度分量＋a 区分效果良好（表 1），其变异系数远远大于＋b，运行时间为 1176 min。

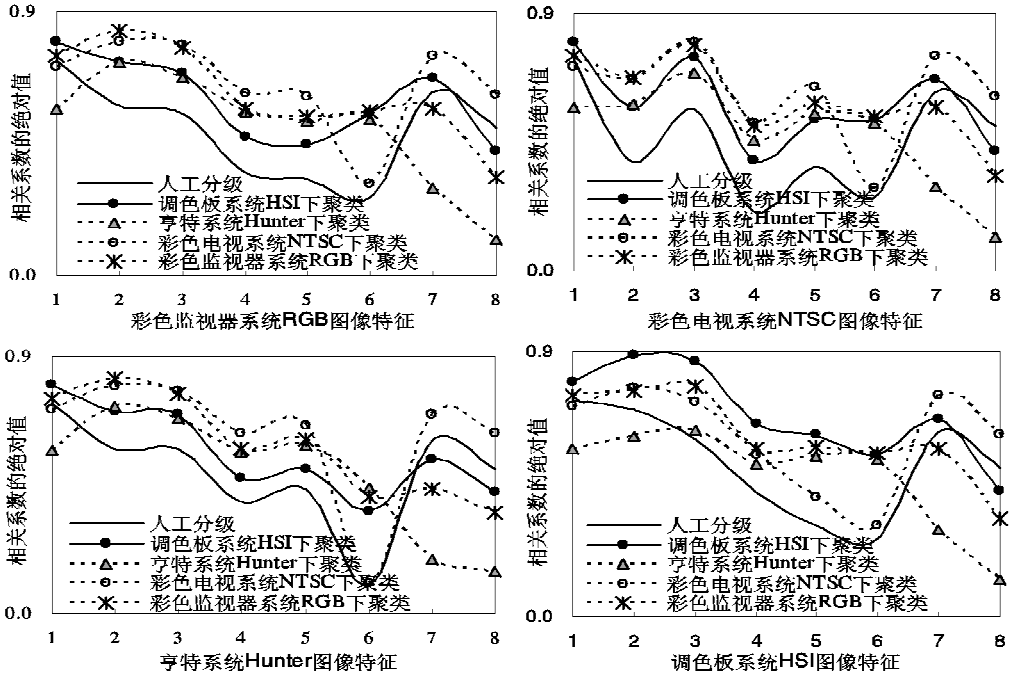
2.1.3 *棉花尺寸与色泽特征相关分析*。不同颜色空间下提取的棉花色泽不尽相同，有必要以不受制于颜色空间的棉花尺寸特征为参照物来考察色泽特征的有效性，理论上，棉花尺寸应与白度特征正相关，与其余色泽特征负相关，试验结果显示：(1)棉花白度与棉花尺寸特征呈虚假负相关，其灵敏性差受光照条件的影响，进一步的讨论将舍去棉花白度特征；(2)其余色泽特征与棉花尺寸特征相关高度显著（$n=400$，$r_{0.01}=0.128$），相关方向与理论上保持一致。

2.2　聚类分析

为了考察不同颜色空间下黄色区域、深度特征是否单调一致，可利用 Wilcoxon 符号秩和检验两总体分布的一致性原假设[9]，试验证明：各黄色特征显著性概率全部为 0，样本落在拒绝域内，说明同一个黄色特征在不同颜色空间下分布差异显著。由表 1 可知，RGB、NTSC、HSI 颜色空间亮度特征一致，亨特颜色空间亮度特征与其它颜色空间差异显著，因此，有必要在各颜色空间下进行棉花品级聚类分析。

一个聚类任务所选特征应尽可能多地包含任务关心的信息，并使信息冗余最小化[10]，鉴于此，本试验选择上述有效的图像特征，基于迭代下降的 K-means 聚类算法进行棉花品级聚类，采用平方欧氏距离为相似度度量，它较多地强调了差异较大的对象[11]。实验中，样本引入顺序为人工对棉花品级的排序，初始中心点由计算机选择，采用在线修改类中心点的方式将标准化后的样本聚为 7 类。

2.2.1 *棉花品级与图像特征相关分析*。通常颜色空间之间有许多可能的坐标转换，使用一套坐标代替另一套坐标很难有明显的改进[12]，重要的问题并非在哪个坐标系下度量颜色，而在于如何计算其差异[13]，这一点可通过人工品级（Manual）、各颜色空间下聚类品级（RGB_cluster 等）与各颜色空间下图像特征之间的相关性来考察（图 1）。



1 棉花尺寸，2～5 带壳棉花和棉花的黄色区域、深度，6～7 带壳棉花白度、色差，8 棉花色差

图 1　棉花品级与图像特征相关性

Fig. 1　Spearman's correlation coefficients between cottons grades and image features

很自然地，聚类时应保证所有选中的特征具有相同的邻近性，并且没有占支配地位的特征[10]，由图 1 可知：各颜色空间聚类品级与图像特征之间的相关性分布呈现一致性，总体上，聚类品级相关性高于人工品级，带壳棉花相关性高于不带壳棉花；NTSC、亨特颜色空间下的聚类品级与 4 个颜色空间下的白度或色差特征相关性较弱；RGB、HSI 颜色空间下的聚类品级与 4 个颜色空间下的图像特征之间的相关性普遍较高、较均衡，相对而言，HSI 颜色空间较优，其运行时间更短，进一步研究应采用 HSI 颜色空间及其聚类品级。

2.2.2 *聚类有效性检验*。聚类算法的共同特点是给数据集强加一个聚类结构，数据集可能完全是随机的，是否具有聚类趋势必须经过验证，由于 K 均值在线聚类方式对样本参与聚类的顺序敏感，可基于相对准则[10]将当前聚类、不同顺序样本参与聚类的结果进行比较。

试验中，为了验证较优的 HSI 颜色空间下以人工品级为序进行的聚类 $C=\{i:d_{ij}<d_{ik}$，对所有的 $j\in C$，$k\notin C\}$ 是否有效，可定义原假设 H_0：样本集的聚类结构 C(大小为 7)具有随机性，*Gordon* 研究了一种基于 U 统计量的聚类有效性检验方法[14]，对有序点对 $(i,j)\in W$ 和 $(k,l)\in B$ 组成的子集 $W(i\in C, j\in C)$ 和 $B(k\in C, l\notin C)$ 有：

$$U_{ijkl}=\begin{cases}0 & d_{ij}<d_{kl}\\ 1/2 & d_{ij}=d_{kl}\\ 1 & d_{ij}>d_{kl}\end{cases}$$

$$U=\sum_{(i,j)\in W}\sum_{(k,l)\in B}U_{ijkl}$$

具体检验算法如下：

(1) 对聚类 C 计算统计量 $U^*=43944915$。

(2) 用随机数生成算法[15]产生 1～402 之间的随机数，并以此为序产生一个 402×8 阶的随机顺序样本矩阵，使用与产生聚类 C 相同的算法将其聚为 7 类，并计算 U。

(3) 重复步骤(2)，直到 U 有 m-1＝20 个值(图 2，U 由小到大排序)。

(4) 由于 U^* 小于 U 的第 j＝2 个最小值 U_j，则以 p＝j/m＝0.095 的显著性水平拒绝随机原假设。

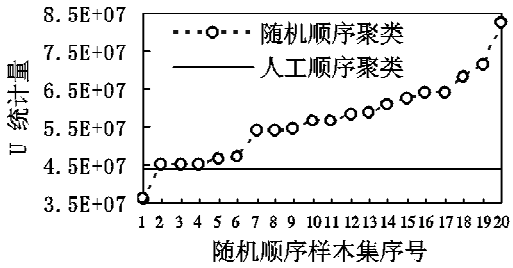


图 2 随机样本的 U 统计量

Fig. 2 Stat. U of randomicity cases

2.2.3 *聚类品级与人工品级的差异*。Wilcoxon 符号秩和检验[9]结果表明，HSI 颜色空间下的聚类品级与人工品级差异显著，与人工品级相比：聚类后的平均品级由 4.59 提高到 4.19，各级样本含量由 9，29，91，67，67，69，70 变为 9，46，100，104，46，41，56(图 3)，其中仅 100 个样本品级没变，占 25%；108 个样本品级下降，占 27%，主要集中于 5～7 级，其中 37 个跨级；194 个样本品级上升，占 48%，主要集中于 2～4 级，其中 88 个跨

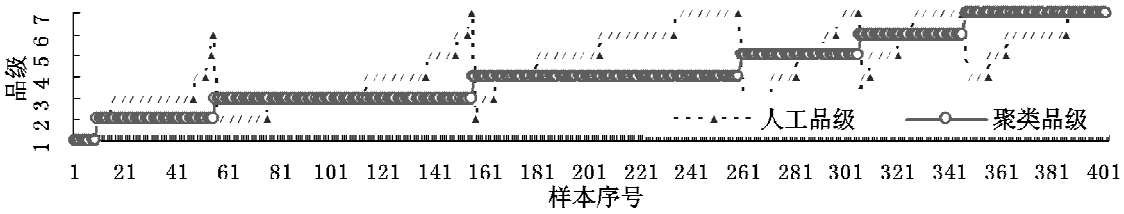


图 3 聚类品级与人工品级的分布差异

Fig. 3 The differences between clustering grades and manual grades

级。由图 3 还可知，优质棉的人眼分级精度较高，人眼易低估一般棉花的品级而高估劣质棉的品级，说明人眼识别范围比机器视觉窄，且易受心里因素的影响。

3 结论

根据我国棉花收购文字标准，基于机器视觉在各颜色测量空间下生成特征参数，利用相关分析、聚类分析并结合算法运行时间所选取的 HSI 颜色空间可能是棉花品级聚类的最优颜色空间。在 HSI 颜色空间试验中，按人工品级排序后进行的聚类借助人眼的先验知识，聚类结构在 α＝0.10显著性水平下拒绝随机原假设，聚类结果是真实有效的。聚类的意义在于：对现有数据集进

行聚类分析,用形成该聚类的样本特征表示该聚类;接下来,如果给出一个未知模式,可以决定它最可能属于哪类,并用相应聚类的特征表示[11];因此,聚类品级是进一步研究收获前棉花分级模型的基础,对我国子棉收购检验仪器化具有重要的现实意义。

棉花品级的聚类结果独立于颜色空间。总体上,棉花聚类品级与图像特征的相关性普遍高于人工品级,与人工分级相比,机器视觉比人眼的辨识范围更宽;带壳棉花聚类品级与图像特征的相关性总是高于不带壳棉花,铃壳色泽对品级贡献显著。

光照条件的变幻莫测导致棉花白度区分度很差,进一步的研究应考虑这一重要特征。

参考文献:

[1] 李　宁,刘东波,臧英明.中国棉花分级标准与国外棉花分级标准差异的研究[J].大连轻工业学院学报,2001,20(4):309-312.

[2] 王　玲,姬长英.农业机器人采摘棉花的前景展望与技术分析[J].棉花学报,2006,18(2):124-128.

[3] 王　玲,姬长英,陈兵林.基于形态学的黑背景下收获前棉花图像自动分割技术研究[J].棉花学报,2006,18(5):299-303.

[4] 熊宗伟.棉花色特征[J].中国棉花,1995,22(1):39-40.

[5] OHTA Y I, Kanade T, Sakai T. Color information for region segmentation[J]. Comp. Graphics & Image Processing, 1980, 13: 222-241.

[6] GONZALEZ R C, Woods R E, Eddins S L. 数字图像处理. Matlab 版[M]. 北京:电子工业出版社,2005:144-178.

[7] GONZALEZ R C, Woods R E. 数字图像处理(第二版)[M].北京:电子工业出版社,2003: 224-254.

[8] MACADAM D L. Visual Sensitivites to Color Differences in Daylight[J]. Opt Soc Am, 1942, 32: 247-274.

[9] 茆诗松,周纪芗.概率论与数理统计[M].北京:中国统计出版社,1996:300-398.

[10] THEODORIDIS S, Koutroumbas K.模式识别(第二版)[M].北京:电子工业出版社,2004,8:257-283.

[11] HASTIE T, Tibshirani R, Friedman J. 统计学习基础:数据挖掘、推理与预测[M]. 北京:电子工业出版社, 2004: 305-356.

[12] FAIRCHILD M D. Color appearance models[M]. Addison Wesley, Reading, Massachussets, 1998.

[13] FORSYTH D A, Ponce J. 计算机视觉:一种现代方法[M].北京:电子工业出版社, 2004: 83-109.

[14] GORDON A D. Identifying genuine clusters in a classification[J]. Computational Statistics and Data Analysis, 1994,18:561-581.

[15] 王　玲.一种生成随机数的新算法:Josephus 算法[J].微型机与应用,2000,19(12):10-11.