

基于聚类分析的光谱重建样本分析

易文娟¹, 孙刘杰¹, 陈志文², 张雷洪¹, 王文举¹

(1.上海理工大学, 上海 200093; 2.上海城建职业学院, 上海 201415)

摘要: **目的** 解决目前光谱重建中因数量大而出出现的冗余和繁重等问题, 证明聚类算法可以很好地应用在光谱选择样本分析中, 并可以实现较高的重构色度精度和物理精度。 **方法** 采用主成分分析法进行仿真实验, 首先探究主成分个数, 再确定聚类个数, 然后比较聚类方法和3种常用的样本选择方法, 最后分析比较光源种类对重构结果的影响。 **结果** 通过实验确定主成分个数为6且聚类个数为20时, 在A光源下使用KFCM算法的重构效果最好, 此时平均色差为 $0.35 \Delta E_{00}$, 平均RMSE为0.0078, 平均GFC为99.94%。 **结论** 聚类方法可以应用于光谱成像过程中训练样本选择过程, 且有助于提高光谱重构的运算速度和精度。

关键词: 光谱反射率; 主成分分析法; 训练样本选择; 聚类分析

中图分类号: TP801.3 文献标识码: A 文章编号: 1001-3563(2019)17-0249-07

DOI: 10.19554/j.cnki.1001-3563.2019.17.036

Spectral Reconstruction Sample Analysis Based on Clustering Analysis

YI Wen-juan¹, SUN Liu-jie¹, CHEN Zhi-wen², ZHANG Lei-hong¹, WANG Wen-ju¹

(1.Shanghai University of Science and Technology, Shanghai 200093, China;

2.Shanghai Urban Construction Vocational College, Shanghai 201415, China)

ABSTRACT: The work aims to solve the problem of redundancy and heaviness caused by the large number of factors in current spectral reconstruction, so as to prove that the clustering algorithm can be well applied in spectral selection sample analysis, and achieve higher reconstruction chromaticity accuracy and physical accuracy. PCA method was used to carry out simulation experiments. Firstly, the number of principal components was explored and the clustering number was determined. Then, the clustering method and three common sample selection methods were compared. Finally, the effects of types of light source on the reconstruction results were analyzed and compared. When the number of principal components determined by experiment was 6 and the clustering number was 20, the reconstruction effect using KFCM algorithm under light source A was the best. At this time, the mean color difference was $0.35\Delta E_{00}$, the mean RMSE was 0.0078, and the mean GFC was 99.94%. Clustering algorithm can be well applied to the selection of training samples in spectral imaging process, and help improve the computing speed and accuracy of spectral reconstruction.

KEY WORDS: spectral reflectance; principal component analysis (PCA); training sample selection; clustering analysis

多光谱系统利用目标物体每个点在不同波段的响应曲线, 获得的颜色信息不受环境光渲染、显示设备的色彩特性干扰等外部条件的影响^[1]。光谱重构方

法已有多种研究, 其中主成分分析法 (Principal Component Analysis, PCA)这种通过获取数据基向量的方法, 是热门研究方向^[2-3]。建立多通道图像信号

收稿日期: 2019-01-21

基金项目: 上海市科学技术委员会科研计划 (18060502500)

作者简介: 易文娟 (1995—), 女, 上海理工大学硕士生, 主攻多光谱复制。

通信作者: 孙刘杰 (1965—), 男, 上海理工大学教授, 主要研究方向为光信息处理技术、印刷机测量与控制技术、数字印刷防伪技术等。

与光谱反射率之间转换矩阵, 离不开训练样本^[4], 由于光谱数据的高维特性, 训练样本的方法对光谱重建的精度和速度至关重要。

将聚类算法应用到样本选择中, 可以解决高维的光谱数据占用存储空间、计算复杂度高、运算时间长和运算效率低的问题。Mohammadi^[5]等通过计算光谱数据间的相关夹角距离, 对颜色在光谱空间进行聚类。李金城^[6]提出了一种基于色域分析与聚类分析结合的基色筛选算法, 筛选出4色域外的特征点, 结合主成分分析法可实现高保真色域的重现。Liu Zhen^[7]利用BP神经网络和分层聚类分析推导最佳样本个数, 但方法复杂且计算量庞大, 耗时较长。龙艳群^[8]将核模糊C聚类运用到样本优化中, 从光谱空间选点, 再在色度空间进行映射分类, 但是最终完成重构的样本个数较多。文中通过探究主成分个数、聚类个数、光源种类、聚类方法对光谱重构的影响, 并从色度精度和光谱精度探究聚类算法在样本选择中的适用性, 对今后光谱重建样本选择的研究提供参考。

1 主成分分析法重建

光谱重建的实质是建立成像系统的多通道响应值 M 与光谱反射率 r 之间的模型关系, 用 A^T 表示系数矩阵, 得到光谱图像正向转换模型:

$$M = A^T r \tag{1}$$

在对高维光谱进行降维处理时, 通常会用到一些多元统计方法, 如主成分分析(Principal Component Analysis, PCA)或独立成分分析(Independent Component Analysis, ICA)以及其他改进的算法。通过主成分分析法, 高维光谱数据可被表述成一系列特征向量的线性叠加, 然后通过前几个主成分和对应特征向量的线性叠加逼近原始的光谱数据, 从而进行重构。

将 $n \times m$ 维 (n 表示光谱维数, m 表示样本数) 的光谱反射率样本集 R 表示为 $R = (R_1, R_2 \cdots R_m)$, 设备响应值通道数为 k , f 为特征向量, 有:

$$R = \sum_{i=1}^n f_i d_i \tag{2}$$

$$P = A^T \sum_{i=1}^k f_i d_i \tag{3}$$

主成分对应系数为:

$$\begin{bmatrix} d_1 \\ \vdots \\ d_k \end{bmatrix} = (A^T [f_1 \cdots f_k])^{-1} P \tag{4}$$

累积贡献率 ρ_k 为:

$$\rho_k = \sum_i^k \sigma_i / \sum_i^N \sigma_i \tag{5}$$

式中: ρ_k 为前 k 个主成分的累计贡献率; σ_i 为第 i 个主成分的贡献量; N 为总主成分个数。

2 样本选择方法

为了提高光谱成像技术的工作效率, 克服光谱数据的高维度和冗余性, 实现颜色信息的精确快速再现, 已有相关学者进行研究, 其中基于最小条件数的 Hardeberg 方法^[9], 根据光谱空间距离进行聚类的 Mohammadi 方法^[5], 以及基于最小最大色度的 Cheung & Westland 法^[10], 是比较有代表性的算法。聚类算法正是分析和提取数据的内在结构的优秀算法, 有必要对其在样本选择中的适用性和实用性进行分析。

2.1 模糊 C-均值聚类

模糊 C-均值聚类算法(Fuzzy C-Means, FCM)是一种经典的动态聚类算法, 以实现样本与聚类中心误差平方和最小化为目的^[11], 每个样本点的类别归属都是其和所有聚类中心的隶属度不断优化的结果。

对训练样本集 $r = (r_1, r_2 \cdots r_m)$, 聚类个数表示为 c , v_o ($o=1, 2 \cdots c$) 代表第 o 个聚类的中心, u_{op} ($\forall o=1, 2 \cdots c, p=1, 2 \cdots m$) 是第 p 个样本对第 o 类的隶属度函数, $i > 1$ 为常数, FCM 的目标函数为:

$$J_i(U, v) = \sum_{o=1}^c \sum_{p=1}^m u_{op}^i \|x_p - v_o\|^2 \tag{6}$$

$\forall o=1, 2 \cdots c, p=1, 2 \cdots m$ 有约束:

$$u_{op} = \frac{(1/\|x_p - v_o\|^2)^{1/i-1}}{\sum_{j=1}^c (1/\|x_p - v_j\|^2)^{1/i-1}}, \quad v_o = \frac{\sum_{p=1}^m u_{op}^i x_p}{\sum_{p=1}^m u_{op}^i} \tag{7}$$

2.2 核模糊 C 均值聚类

核模糊 C 均值聚类方法(KFCM)的核心是使用非线性映射 Φ , 将输入空间中的数据点映射到高维特征空间中^[11], 在高维特征空间其目标函数为:

$$J_i(U, v) = \sum_{o=1}^c \sum_{p=1}^m u_{op}^i \|\Phi(x_p) - \Phi(v_o)\|^2 \tag{8}$$

约束:

$$u_{op} = \frac{(1/(K(x_p, x_p) + K(x_o, x_o) - 2K(x_p, x_o)))^{1/(i-1)}}{\sum_{j=1}^c (1/(K(x_p, x_p) + K(x_j, x_j) - 2K(x_p, x_j)))^{1/(i-1)}} \tag{9}$$

$$V_o = \frac{\sum_{p=1}^m u_{op} \tilde{K}(x_p, v_o) x_p}{\sum_{p=1}^m u_{op}^i \tilde{K}(x_p, v_o)} \tag{10}$$

核函数采用高斯核函数, 对高斯函数有:

$$\tilde{K}(x_p, v_o) = K(x_p, v_o) \tag{11}$$

其利用聚类中心的对偶表示, 通过非线性映射增

大了数据点线性可分的概率,实现更为准确的聚类和更快的算法收敛速度。

3 实验与分析

实验条件为 D65 光源 2° 视场,首先选择 Munsell 色卡^[12]作为训练样本,以 10 nm 间隔采集 400~700 nm 波长的样本光谱反射率,用 X-Rite 的 Eye-one pro 分光光度计测量实际的样本光谱反射率,使用 Matlab2018 版本进行仿真实验。实验步骤如下所述。

1) 确定主成分个数。使用 Matlab 中 pca 和 svds 函数,在无聚类条件下设定主元个数 i ($i=3, 4 \dots 9$) 根据累计贡献率确定主成分个数。

2) 探究聚类个数对重构结果的影响。从 10~80 以 5 为间隔设定聚类个数,最大迭代次数 $N=100$, 高斯核参数 $\sigma = 250$, 目标函数值改变量的最小阈值 $\epsilon < 1e-6$, 比较 FCM 和 KFCM 两聚类方法的重构结果,确定聚类个数。

3) 探究聚类方法对重构结果的影响,比较 2 种聚类算法和 3 种常用样本选择方法的适用性。

4) 探究光源对重构结果的作用。重构 ColorChecker RenditionChart24 色卡(以下简称 RC24),将光源条件设定为 A, D50, D65, 比较光源对聚类算法的重构结果的影响。

重构结果的评价采用均方根误差 RMSE、适应度系数 GFC^[14], 色差使用 CIE2000 标准色差公式 ΔE_{00} ^[15-16]。

均方根误差 (RMSE) 计算公式:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\bar{r}_i - r_i)^2} \quad (12)$$

适应度系数(GFC)表示重建光谱反射率 \bar{r}_i 与原始光谱反射率 r_i 之间的余弦角:

$$GFC = \frac{r_i^T \bar{r}_i}{\sqrt{r_i^T r_i} \sqrt{\bar{r}_i^T \bar{r}_i}} \quad (13)$$

式中: \bar{r}_i 为每一点像素的重构光谱反射率; r_i 为原始光谱反射率; n 代表样本维度。

3.1 主成分个数

每个主成分包含的信息量是递减的,首先获得不同主成分个数的累积贡献率。

采用无聚类方法重构时累计贡献率见图 1, 为了对样本信息进行高效提取,取光谱反射率累计贡献率超过 0.998 为标准,确定主成分个数为 6。

3.2 聚类个数

获得合适的聚类数目是聚类分析中一个关键的问题,文中实验通过均方根误差的均值来选择聚类个

数,得到的结果见图 2。

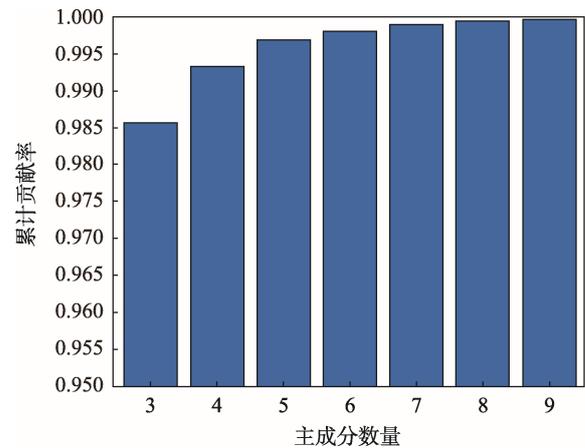


图 1 主成分个数为 3~9 时的累积贡献率
Fig.1 Cumulative contribution rate of 3~9 principal components

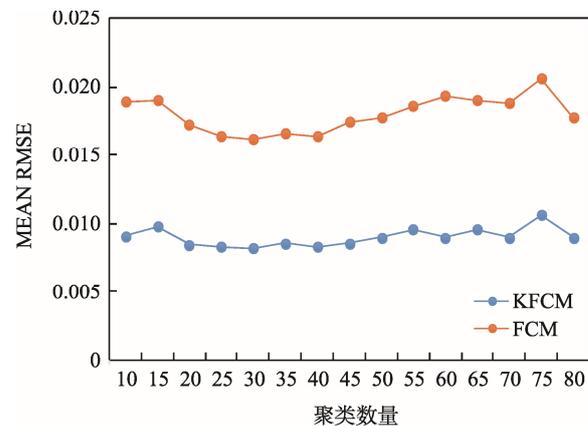


图 2 聚类个数为 10~80 的平均均方根误差
Fig 2 Mean RMSE of 10~80 clustering samples

KFCM 实现的平均均方根误差优于 FCM 算法。聚类样本个数从 10 增至 20 时,2 种方法得到的平均均方根误差明显减小,说明一定范围内增加聚类个数有益于提高重构精度,但当聚类个数超过 45 后重构精度随着聚类个数的增加而降低,当聚类个数在 20~40 时均方根误差变化趋于稳定。下一步由图 2 开始重点探讨聚类个数在 10~40 时的重构精度。

由表 2 可以看出,聚类个数选择为 20 即可以对样本进行良好重构,原因可能是后续选择的部分样本带有较大噪声,因此 2 种算法的均方根误差有所波动,呈现上升趋势,但是由于是重构自身,因此平均色差虽有波动,但变化非常,小人眼无法识别。通过 RMSE 和 GFC 判断此时采用 KFCM 算法得到物理精度优于 FCM 算法,此 KFCM 算法得到平均色差为 0.5489 ΔE_{00} , 均方根误差 0.0085, GFC 为 99.90%, 采用 FCM 算法得到的平均色差为 0.7824 ΔE_{00} , 均方根误差为 0.0087, GFC 为 99.83%。

表 1 聚类个数对重构的影响
Tab.1 Influence of clustering number on reconstruction

聚类数量	评价标准	KFCM	FCM
10	Mean ΔE_{00}	0.7089	0.7768
	Mean RMSE	0.0091	0.0098
	Mean GFC	99.88%	99.85%
15	Mean ΔE_{00}	0.6401	0.9170
	Mean RMSE	0.0098	0.0092
	Mean GFC	99.86%	99.83%
20	Mean ΔE_{00}	0.5489	0.7824
	Mean RMSE	0.0085	0.0087
	Mean GFC	99.90%	99.83%
25	Mean ΔE_{00}	0.5503	0.7395
	Mean RMSE	0.0083	0.0081
	Mean GFC	99.90%	99.84%
30	Mean ΔE_{00}	0.5709	0.7210
	Mean RMSE	0.0082	0.008
	Mean GFC	99.91%	99.84%
35	Mean ΔE_{00}	0.5422	0.6293
	Mean RMSE	0.0086	0.0079
	Mean GFC	99.90%	99.83%
40	Mean ΔE_{00}	0.5520	0.7338
	Mean RMSE	0.0083	0.0081
	Mean GFC	99.91%	99.83%

表 2 不同聚类方法的重构结果
Tab.2 Reconstruction results of different clustering methods

方法	ΔE_{00}			RMSE			GFC/%		
	mean	max	min	mean	max	min	mean	max	min
Hardeberg	0.846	3.487	0.011	0.01	0.026	0.002	99.89	100.00	96.38
Cheung&Westland	0.609	3.421	0.028	0.012	0.063	0.002	99.85	100.00	91.15
Mohammadi	0.914	4.106	0.024	0.012	0.037	0.002	99.84	100.00	96.47
KFCM	0.549	4.146	0.006	0.008	0.038	0.001	99.90	100.00	92.44
FCM	0.782	8.89	0.012	0.009	0.071	0.001	99.83	100.00	84.99

3.3 聚类方法

使用实验方法和目前常用的 Hardeberg 方法、Mohammadi 方法以及 Cheung&Westland 方法^[4,16]分别对 Munsell 色卡进行重构, 其结果见表 2。

由于此时是重构训练样本自身, 因此数据较小, 表明了主成分个数为 6 时主成分分析法对训练样本的再现效果良好。KFCM, FCM, Mohammadi 等 3

种聚类方法结果都较良好, 但 FCM 在个别色块的色度还原效果非常不理想, 由于光谱反射率高维特性 KFCM 通过映射实现了更优的色度精度还原。比较 5 种样本选择方法, 实验的 2 种方法获得的均方根误差均值最优, 对个别色块可能存在过拟合的情况。部分重构色块绘制的曲线见图 3, 实验的重构曲线和实际光谱贴合良好, 并可在 D65 光源、2°视角的条件下在 Photoshop 中模拟该得到该色块图像。

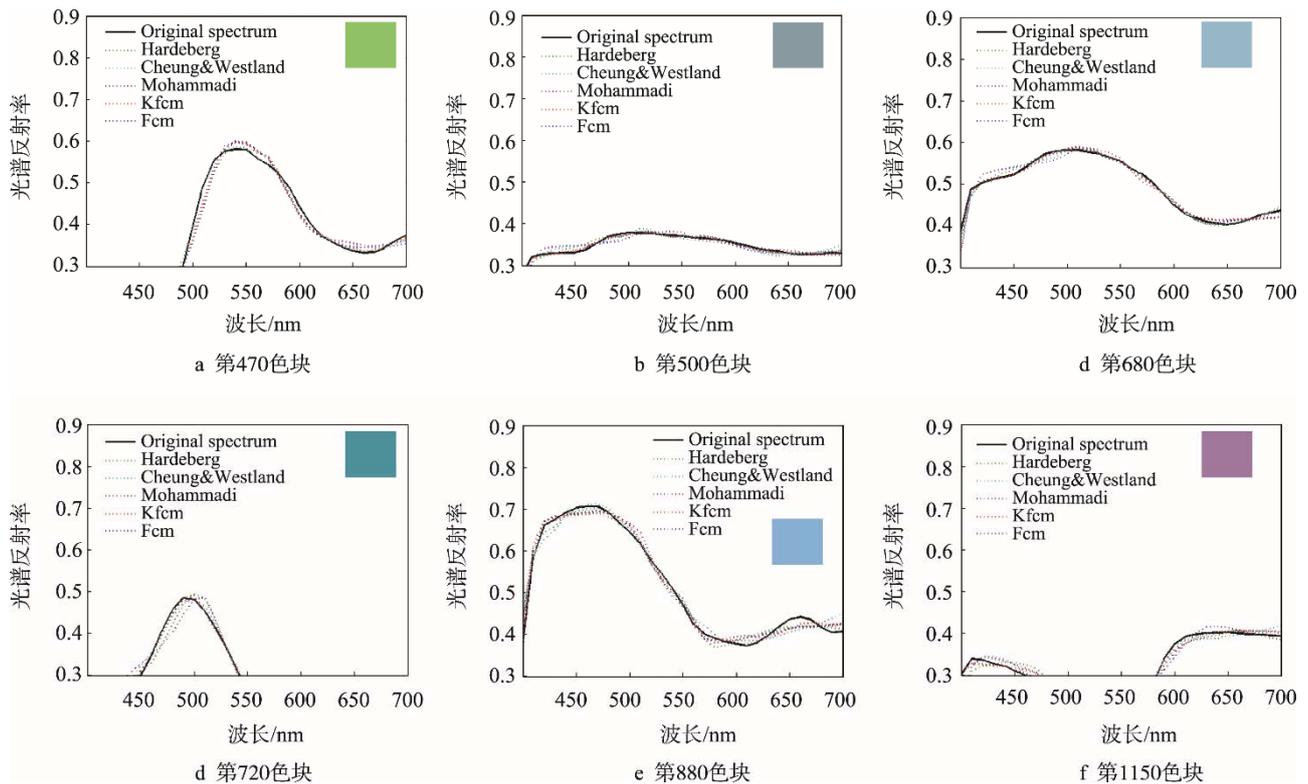


图 3 部分色块重构曲线
Fig.3 Reconstruction curve of some color blocks

3.4 光源种类

颜色是场景光照、观察者、物体的光谱反射率三者共同作用的结果，因此下一步探究场景光源的选择。检验样本选用常用的 RC24 色卡，2 种色卡在 D65 光源下的 Lab 分布见图 4，可以看出 2 种色卡的总体分布趋势相似，且 RC24 色卡部分色块不与 Munsell 色块重合。

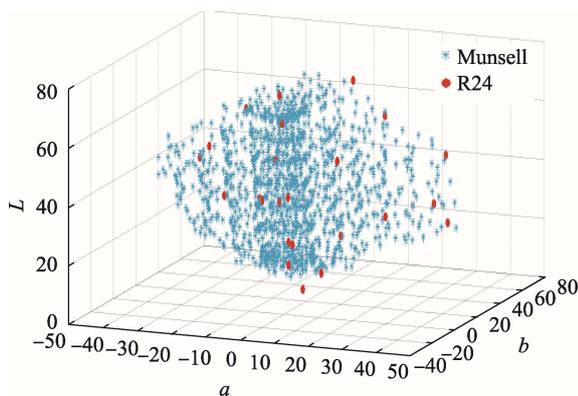


图 4 D65 光源下 2 种色的色度分布
Fig.4 Chromaticity distribution of two colors under D65 light source

在 A, D50, D65 光源条件下比较 3 种聚类算法的重构的色度精度，结果见图 5。

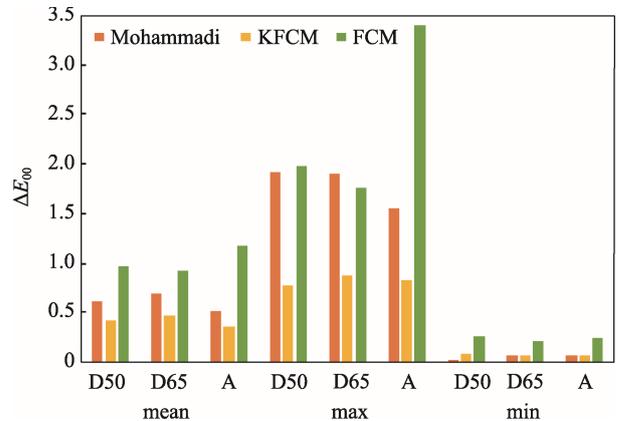


图 5 不同光源条件下的色差
Fig.5 Color difference under different light sources

由于 Munsell 色卡和 RC24 色卡整体分布相似，实验对 RC24 色卡色彩再现效果良好。3 种光源下，KFCM 通过映射使得保存了较多光谱数据，因此平均色差表现最优。虽然 FCM 算法明显减少了样本数量，但色差精度不太理想，尤其在 A 光源下最大色差为 $3.41 \Delta E_{00}$ 已经可以肉眼识别。在 A 光源下使用 KFCM 算法可得到最小平均色差为 $0.35 \Delta E_{00}$ ，最大平均色差为 $0.82 \Delta E_{00}$ ，证明了 KFCM 算法在颜色领域上的实用性。不同光源下 KFCM 和 FCM 算法实现的重构的色度分布见图 6，色度拟合良好，聚类方法既保证了色彩再现，又减少了样本数量。

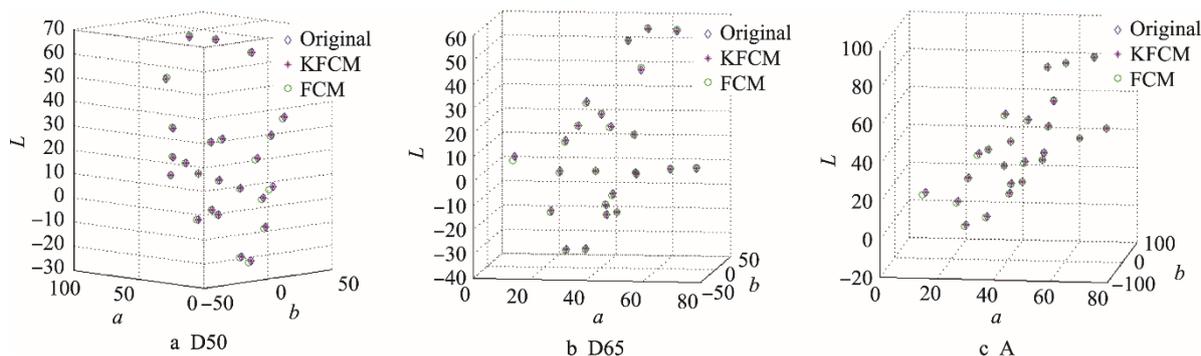


图6 不同光源下重构色度分布

Fig.6 Reconstruction chromaticity distribution under different light sources

4 结语

探讨了FCM和KFCM等2种聚类方法在样本选择中运用,并与3种常用的训练样本方法进行了比较。通过实验确定了主成分个数为6,聚类个数为20时,在A光源下KFCM算法可以对孟塞尔色卡选择最优子集。此研究证明聚类算法可以有效提取高维样本的特征,并降低运算量,在光谱重构、光谱数据压缩处理、色域映射、色彩转换等方向有应用前景。今后可以对聚类算法中的距离函数进行优化,有助于高保真色彩复制和艺术品复制、数字博物馆、可视化通讯等领域的拓展研究。

参考文献:

- [1] 赵杨. 多光谱光场成像系统与算法研究[D]. 南京: 南京大学, 2018.
ZHAO Yang. Study of Multispectral Light Field Imaging System and Algorithm[D]. Nanjing: Nanjing University, 2018.
- [2] 刘传杰, 李玉梅, 陈浩杰, 等. 不同算法模型对光谱重构精度的影响[J]. 包装工程, 2018, 39(1): 168—173.
LIU Chuan-jie, LI Yu-mei, CHEN Hao-jie, et al. Influence of Different Algorithm Models on the Accuracy of Spectral Reconstruction[J]. Packaging Engineering, 2018, 39(1): 168—173.
- [3] 徐鹏. 基于多光谱成像技术的光谱图像重构与显示研究[D]. 杭州: 浙江大学, 2017.
XU Peng. Study on Spectral Image Reconstruction and Display Based on Multispectral Imaging Technology[D]. Hangzhou: Zhejiang University, 2017.
- [4] 万晓霞, 梁金星, 刘强. 宽带光谱成像系统最优训练样本选择方法研究[J]. 北京理工大学学报, 2016, 36(6): 641—646.
WAN Xiao-xia, LIANG Jin-xing, LIU Qiang. Optimal Training Sample Selection for Broadband Spectral Imaging System[J]. Transaction of Beijing Institute of Technology, 2016, 36(6): 641—646.
- [5] MOHAMMADI M, NEZAMABADI M. A Prototype Calibration Target for Spectral Imaging[C]. In 10th Congress of the International Colour Association, 2005: 387—390.
- [6] 李金城, 刘真, 陈广学, 等. 一种基于色域分析与聚类分析的基色筛选[J]. 光学学报, 2012, 32(6): 321—325.
LI Jin-cheng, LIU Zhen, CHEN Guang-xue, et al. Colorant Selection Based on Gamut Analysis and Cluster Analysis[J]. Acta Optics Sinica, 2012, 32(6): 321—325.
- [7] LIU Zhen, LIU Qiang, GAO Gui-ai, et al. Optimized Spectral Reconstruction Based on Adaptive Training Set Selection[J]. Optics Express, 2017, 25(11): 12435—12445.
- [8] 龙艳群, 王慧琴, 王可, 等. 基于样本优化选取的光谱重建方法研究[J]. 影像科学与光化学, 2017, 35(1): 88—96.
LONG Yan-qun, WANG Hui-qin, WANG Ke, et al. Study on Spectral Reconstruction Method Based on Optimized Selected Samples[J]. Imaging Science and Photochemistry, 2017, 35(1): 88—96.
- [9] HARDEBEG J Y, SCHMITT F, BRETTEL H. Multi-spectral Color Image Capture Using A Liquid Crystal Tunable Filter[J]. Optical Engineering, 2002, 41(10): 2532.
- [10] CHEUNG V, WESTLAND S. Methods for Optimal Color Selection[J]. Journal of Imaging Science and Technology, 2006, 50(5): 481.
- [11] 刘双喜. 基于色彩聚类的玉米种子纯度识别算法研究[D]. 济南: 山东农业大学, 2018.
LIU Shuang-xi. Research on Identification Algorithm of Maize Seeds Purity Based on Color Clustering[D]. Jinan: Shandong Agricultural University, 2018.
- [12] COHEN J. Dependency of the Spectral Reflectance

- Curves of the Munsell Color Chips[J]. *Psychonomic Science*, 1964, 1(12): 369—370.
- [13] FANG J Y, XU H S, XU P, et al. Spectral Estimation of Fluorescent Lamps Using RGB Digital Camera and Standard Color Chart[J]. *Optik-International Journal for Light and Electron Optics*, 2017, 130: 50—60.
- [14] AMIRI M M, FAIRCHILD M D. A Strategy toward Spectral and Colorimetric Color Reproduction Using Ordinary Digital Cameras[J]. *Color Research & Application*, 2018, 43(5): 675—684.
- [15] DIERL M, ECKHARD T, FREI B, et al. Novel Accuracy Test for Multispectral Imaging Systems Based on ΔE Measurements[J]. *Journal of the European Optical Society-Rapid Publications*, 2018, 14: 1.
- [16] XU H S, YAGUCHI H, SHIOIRI S. Testing CIELAB-based Color-difference Formulae Using Large Color Differences[J]. *Optical Review*, 2001, 8(6): 487—494.