K-means 聚类数的确定及在细胞图像颜色校正中的应用①

罗丽丽[1], 蔡坚勇[1,2,3,4], 蔡荣太[1,3], 林李金[1], 蔡 娟[1]

[1](福建师范大学 光电与信息工程学院, 福州 350007)
[2](福建师范大学 医学光电科学与技术教育部重点实验室, 福州 350007)
[3](福建师范大学 福建省光子技术重点实验室, 福州 350007)
[4](福建师范大学 智能光电系统工程研究中心, 福州 350007)

摘 要: 针对大量瑞氏染色细胞图像, 通过 YCbCr 颜色空间进行 K-means 聚类, 观察各分量聚类中心差值变化规律, 提出了一种新的确定 K-means 聚类数的颜色校正算法. 该算法首先是将瑞氏染色细胞图像中不同目标分别准确地聚集在相应类当中, 再与标准图像中的每类进行配比, 并利用直方图规定化进行直方图调整, 得到颜色校正结果. 经大量实验证明, 尤其在细胞图像中目标颜色特征较接近的情况下, 该算法通过确定合适的聚类数可大大提高颜色校正结果的准确率.

关键词: K-means; 中心差值; 聚类数; 颜色校正; 准确率

Determination of Number of Clusters in K-means and Application in Color Correction of Cell Image

LUO Li-Li[1], CAI Jian-Yong[1,2,3,4], CAI Rong-Tai[1,3], LIN Li-Jin[1], CAI Juan[1]

[1](College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou 350007, China)
[2](Key Laboratory of Optoelectronic Science and Technology for Medicine Ministry of Education, Fujian Normal University, Fuzhou 350007, China)
[3](Fujian Provincial Key Laboratory for Photonics Technology, Fujian Normal University, Fuzhou 350007, China)
[4](Intelligent Optoelectronic Systems Research Centre, Fujian Normal University, Fuzhou 350007, China)

Abstract: To observe the changing rule of clustering center's value in K-means by the YCbCr color space, against a lot of cell image by Wright Stain, this paper proposed a new method to determine the number of clusters in K-means to get a good result of color correction. Different target of cell image could be firstly gathered in the corresponding class accurately, then by matching between classes, and using histogram specification for adjusting histogram, the results of color correction could at last be achieved. It has been proved by many experiments that this algorithm can greatly improve the accuracy of color correction results, especially the target with closer color features in cell.

Key words: k-means; center's value; number of clusters; color correction; accuracy

颜色信息是彩色图像的一个重要特征之一, 它可以作为图像分割、图像识别等的重要依据[1]. 光源、被拍摄的物体与图像采集设备是图像形成的三要素[2]. 物体所呈现出的颜色与光源特性、光照条件等有关; 随着光照、图像采集设备、成像设备等的不同, 物体被感知的颜色也不尽相同.

瑞氏染色法是医学临床上最常用的染色方法. 血涂片常常会因为在瑞氏染色期间, 染色剂调配的不一致, 或者染色并不总是在同一时间进行, 以及光照、成像设备的不同等原因, 导致最后采集到的瑞氏血细胞图像颜色不同; 又因为医学图像是当代医学研究的重要手段和临床诊断的主要依据, 医学图像的准确分析对判断有无疾病、疾病种类及严重程度是至关重要的. 然而同类细胞的颜色特征不同, 必然会造成后续图像分割等处理工作的困扰. 为了消除该困扰, 保持瑞氏染色后细胞图像中颜色的一致性, 提

① 基金项目:福建省高校产学合作科技重大项目(2011H6010);国家自然科学基金(61179011)
收稿时间:2013-07-10;收到修改稿时间:2013-08-19

出一种将偏色图像校正为标准图像的颜色校正算法是很有必要的，该标准图像为便于图像后续分割处理的参考图像.

颜色校正是图像预处理分析中十分重要的步骤之一，也是国内外研究的热点课题，已经在医学图像、遥感图像、壁画图像等众多图像处理领域中得到了广泛应用[3]. 目前，出现的颜色校正方法有很多，常见的有: (1)基于映射的颜色校正方法，它需要通过确定源颜色空间到目标颜色空间的映射关系，来实现颜色空间的转换[4]，其方法复杂，不易实现; (2)基于 BP 神经网络的颜色校正算法[5]，其主要思想是根据标准图像像素值来确定其光源情况，以输入层和输出层作为学习层机制，多次模拟来实现颜色校正，该算法运行速度比较慢，且不能保证学习结果达到均方误差的全局最小点，没有知识积累性等; (3)基于白平衡处理的颜色校正算法，其主要思想[6,7]是假设图像中有“白色区域”存在，以“白”色为参照标准，将图像的 R、G、B 三个通道的“白色区域”点校正成标准光源下的参考白点，以此来实现目标的颜色校正，但该方法存在一定的局限性，当图像中不存在白色部分或高光部分时，白平衡算法失效.

上述颜色校正算法都具有各自的适用范围. 为满足实际工作需要，结合细胞图像自身的特征，本文提出了一种基于 K-means 聚类分量中心差值来确定聚类数，实现类与类之间的精确配比，再通过调整相应类的直方图，从而实现颜色校正结果.

1 K-means聚类的血细胞颜色校正算法

1.1 K-means 聚类

K-means 聚类算法[8]是在图像分割中被广泛使用的一种聚类算法，它实际上是一种基于样本间相似度量的统计方法. 该算法的主要思想是先选定聚类数和初始聚类中心，然后根据最小距离原则(即样本间的相似度)，遍历每个样本点，将各数据点分配到相应的类别中，通过多次迭代更换聚类中心，使表示聚类性能的准则函数达到最优.

距离测度的选择对 K-means 聚类效果的影响很大. 针对不同类型的图像，采用的距离测度也不相同. 常见的距离测度有欧氏距离、曼哈顿距离以及马氏距离等等. 本文根据瑞氏染色后的细胞图像本身的特性，采用的距离计算方法为欧氏距离.

设 X, Y 为两个向量样本，$X=(x_1,x_2...x_n)^T$，$Y=(y_1,y_2...y_n)^T$，则 X, Y 的欧式距离定义[9]如下:

$$D_e(x,y)=\|X-Y\|=\sqrt{\sum_{i=1}^{d}|x_i-y_i|^2} \quad (1)$$

其中 D 为 n 维空间中 X 与 Y 之间的距离. 由上述公式可知: D 越小时，表明 X 与 Y 就越相似.

1.2 颜色空间的选择

瑞氏染色后的细胞图像通常包括三个明显的部分: 白细胞、红细胞和背景，根据此特点，在初次 K-means 聚类时，聚类初值应选择 3. 在此类图像当中，背景的颜色较单一，且亮度最大，而白细胞中细胞核染色最深，故亮度最小. 基于亮度特征，可将背景准确聚类. 但红细胞与白细胞的细胞质亮度较接近，若单一从亮度角度考虑聚类，必然会将细胞质与红细胞聚为一类. 为了将红细胞和白细胞进行区分，可以从它们之间的颜色差异进行考虑.

在 YCbCr 彩色空间中, Y 表示亮度分量, Cb 表示蓝色分量，Cr 表示红色分量[10]，三者之间是互相独立的，对染色血细胞图像的颜色分布区域可以起到良好的限制作用. 该模型主要特点是不仅提取了亮度分量，而且只需要考虑图像的两个基本色调(这两个色调值的大小又反应了图像中颜色色温的分布情况). 对于后续校正过程，该空间避免了多分量颜色的干扰，故 K-means 聚类应在 YCbCr 颜色空间下进行，即可将细胞图像较好地聚成白细胞，红细胞和背景三类. RGB 颜色空间到 YCbCr 颜色空间的转换公式如下:

$$\begin{bmatrix}Y\\C_b\\C_r\end{bmatrix}=\begin{bmatrix}16\\128\\128\end{bmatrix}+\frac{1}{256}\begin{bmatrix}65.738 & 129.057 & 25.064\\-37.945 & -74.494 & 112.439\\112.439 & -94.154 & -18.285\end{bmatrix}\bullet\begin{bmatrix}R\\G\\B\end{bmatrix} \quad (2)$$

1.3 类与类的配比

由于 K-means 每次聚类时，类是无序的，故在类与类直方图调整前，必须将标准图像聚成的三类与偏色图像三类一一对应. 而在一幅细胞图像中，背景区域比较单一，染色均匀，所以相对于白细胞和红细胞区域而言，背景部分的颜色变化是最小的. 其中颜色变化值可通过先求出每类中 Y、Cb、Cr 三分量的均值，然后计算出每类每分量中每个点的对应值与相应类中每分量的平均值之差的平均值，进而求出类与类之间的颜色变化值. 公式如下式(3)所示.

$$J[i] = (\sum_{j=1}^{N} | Y_{[i][j]} - Y\_Mean_i | /N + \sum_{j=1}^{N} | Cb_{[i][j]} - Cb\_Mean_i | /N + \sum_{j=1}^{N} | Cr_{[i][j]} - Cr\_Mean_i | /N)/3 \quad (3)$$

上式中 $i \in (1,2,3)$ 为类型标注值; N 为对应类的像素总数量; $Y_{[i][j]}$、$Cb_{[i][j]}$、$Cr_{[i][j]}$ 分别是第 i 类的亮度、蓝色色度和红色色度值; Y_Mean_i、Cb_Mean_i、Cr_Mean_i 分别是第 i 类的平均亮度、平均蓝色色度以及平均红色色度值. 利用上式可获到 3 类各自的 J 值, 其中 J 最小的值所对应的类为背景区域.

根据标准图像与偏色图像中同类细胞的特征值相近, 提取红细胞、白细胞的二值化图像的几何特征, 利用几何特征中的面积均值、周长均值和似圆度均值计算出各类之间的欧拉加权距离[11], 并由欧拉加权距离判断出相似程度. 欧拉加权距离越小则相似程度越大, 视为同一类, 即可识别出标准图像和偏色图像中的白细胞及红细胞的类.

1.4 算法流程图

针对本文所研究的瑞氏染色细胞图像的分布特点, 可基于YCbCr颜色空间, 利用K-means聚三类得到3×3的聚类中心矩阵. 由 Y 分量的三类聚类中心值, 可判断出亮度最大值所在的类为背景类, 再去除背景类在 Cb 分量中的聚类中心值, 接着通过计算剩下两类的 Cb 聚类中心差值△来确定聚类数, 从而更精确地将白细胞和红细胞分别聚在相应类当中. 算法流程图如图 1 所示.

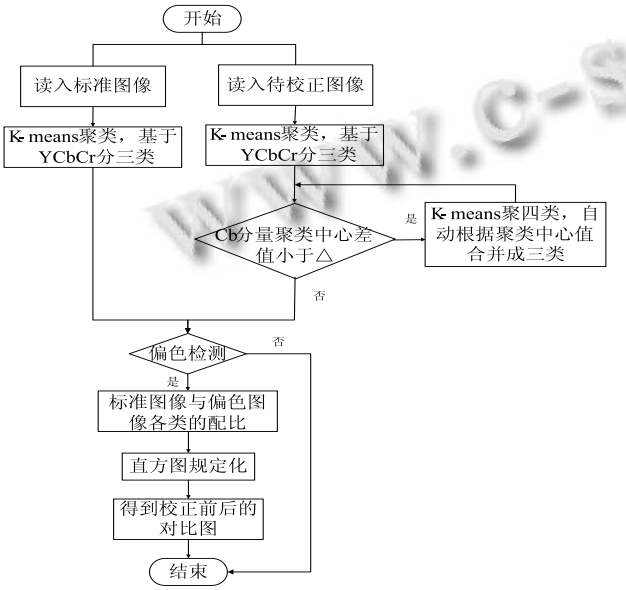


图 1　基于K-means聚类的血细胞颜色校正算法流程图

2　K-means聚类数的确定

2.1 K-means 聚类算法

K-means 算法是聚类技术中一种基于划分的方法, 是一种无监督的学习算法. 其优点是简单易行, 具有高效性. 利用 K-means 算法聚类时, 对于类内的相似度大、类间差异大的聚类结构, 聚类效果比较好. 然而对于不含此特征的图像, 利用 K-means 算法聚类时, 对初始值的依赖性很强, 初值选取的不同直接影响着聚类结果的好坏. 算法[12]具体步骤如下:

1) 指定聚类数目 i, 按照某种方法选择初始聚类中心 $\{A_j\}_{j=1}^{i}$;

2) 计算每个数据到各初始聚类中心的欧氏距离 D, 按距离就近原则把每一个数据归入一个类别中, 然后计算此时各类别的中心值, 以此作为新的聚类中心 $\{A_j^*\}_{j=1}^{i}$;

3) 用新的聚类中心 $\{A_j^*\}_{j=1}^{i}$ 重新聚类, 聚类完成后继续计算各类别的中心值, 聚类中心不断迭代, 直到聚类中心不再变化或小于某个事先规定的值为止.

2.2 聚类数确定依据

K-means 聚类方法最大的难点就是聚类的类数选择. 根据细胞图像的目标种类, 一般情况下联合 YCbCr 颜色空间, 将待校正细胞图像基于 K-means 聚成三类,得到 3×3 矩阵的聚类中心值:

$$A\_ctrs = \begin{matrix} Y & Cb & Cr \end{matrix} \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}$$

由亮度分量 Y 的聚类中心值, 可将背景很好的聚成一类. 但由于红细胞和白细胞的亮度有部分重叠, 故不能简单地从 Y 分量进行聚类. 大量实验观察可知, 当Cb分量中红白细胞的聚类中心差值很小时, 直接聚三类容易将部分红细胞归入白细胞类当中, 聚类效果不佳. 这时应加大聚类类数, 提高聚类中心的精度.

令 $Cb = (A_{12}; A_{22}; A_{32})$;

$Cb_{\max} = \mathrm{Max}(Cb)$; $Cb_{\min} = \mathrm{Min}(Cb)$;

$Cb_{\mathrm{mid}} = Cb_{[i][1]}$, $Cb_{[i][1]} \neq Cb_{\max}$,

$Cb_{[i][1]} \neq Cb_{\min}$, $i \in (1,2,3)$;

当 $Cb_{\max}$ 与 Cb_{mid} 的差值△大于 10 时, 聚类数仍为 3; 差值△小于等于 10 时, 聚类数应为 4. 其中 10 为多次实验得出的经验阈值. 通过选择聚类数 4 进行迭代

判断, 足以达到所需分类效果, 而且计算量相对较小, 故无需再增大聚类数. 当偏色图像需选择聚类数 4 时, 将 YCbCr 中任意分量的聚类中心值从小到大排列, 自动合并中间两类, 组合成新的三类聚类中心值, 再次判断其去除背景类的 Cb 聚类中心差值△与经验值 10 的关系, 直到差值△大于 10 为止.

2.3 聚类中心值的组合及对应序号的修正

由于标准图像直接聚三类可得到较好的聚类效果, 故其得到的聚类中心值一定是 3×3 的矩阵. 如果待校正的细胞图像需聚 4 类时, 得到 4×3 矩阵的聚类中心值需自动组合成 3×3 矩阵的聚类中心值.

令 $Y=(A_{11};A_{21};A_{31};A_{41})$ 为聚四类的 Y 分量聚类中心值. 找出 Y 分量聚类中心值中最大值与最小值所对应的行数 m, p. 最大值所在的行 m 为背景类, 最小值所在的行 p 为白细胞类. 公式(4)可计算出另外两类中心值的平均值, 将其作为重组后的一类, 即为红细胞所对应的类.

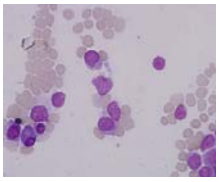
$$\begin{cases} Y_{[i_{\min}][1]}=(\sum_{i\text{-}1}^{\mathrm{N}}Y_{[i][1]})/2 \\ Y_{[i_{\min}][2]}=(\sum_{i\text{-}1}^{\mathrm{N}}Y_{[i][2]})/2\ ,\ (i\neq m且i\neq p,\mathrm{N}=4) \\ Y_{[i_{\min}][3]}=(\sum_{i\text{-}1}^{\mathrm{N}}Y_{[i][3]})/2 \end{cases} \tag{4}$$

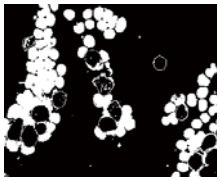
式中 $i\in(1,2,3,4)$ 为聚类矩阵的行数. 为了方便后续类与类的配比, 公式(5)可将合并后的位置序号进行规定重组, 其余序号顺序不变. 设 j 为聚类中心矩阵的列数:

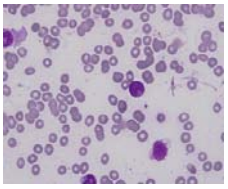
$$A_{[3][j]}=\begin{cases} A_{[m][j]}\ ,m=4 \\ A_{[p][j]}\ ,p=4 \end{cases},\ j\in(1,2,3) \tag{5}$$

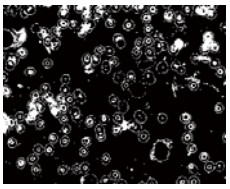
3 实验结果与讨论

本文仿真实验是在 Windows7 环境下, 用 MATLAB 编程实现的. 瑞氏染色后的细胞被存储为 241×320 像素的 JPEG 真彩色图像. 图 2 为标准图像以及待校正图像基于 K-means 聚为三类的分类效果对比图.

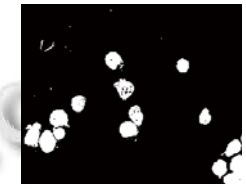

(a) 标准图像

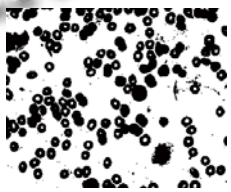

(b) 红细胞类


(c) 背景类


(d) 白细胞类


(a) 待校正图像


(b) 红细胞类


(c) 背景类

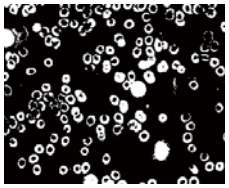
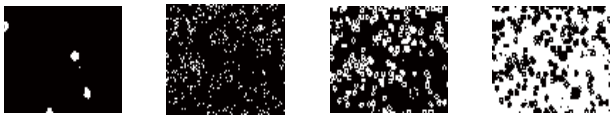

(d) 白细胞类

图 2 标准图像以及待校正图像基于 K-means 聚为三类的分类效果对比图

观察可知, 标准图像聚三类的分类效果良好, 但待校正图像聚三类时, 其中部分红细胞聚到了白细胞的类别当中, 效果明显不佳.

观察 K-means 聚类中心值时, 发现 Cb 分量聚类中心值中最大值与中间值的差值为 4.79. 可知此细胞图像在蓝分量中白细胞与红细胞的聚类中心值非常接近, 很可能导致错聚类. 根据本文提出的方法, 由于其差值△小于等于 10, 故此细胞图像需聚四类, 然后组合成三类, 得到重组后的 Cb 分量聚类中心差值为 25.19, 此时差值△大于 10, 即可得到较好的分类效果. 如图 3 所示.


(a) 待校正图像聚四类效果图

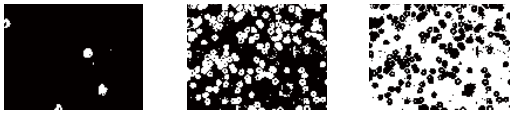
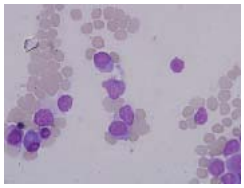
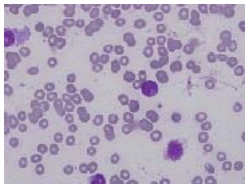

(b) 待校正图像聚四类组合三类后的效果图

图 3 待校正图像聚四类效果图及组合三类后的效果图

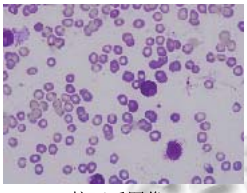
下面图 4 是将待校正细胞图像直接聚三类的校正前后对比图，以及根据本文的方法，先将其聚成四类，后合并成三类，得到的校正前后对比图.



标准图像

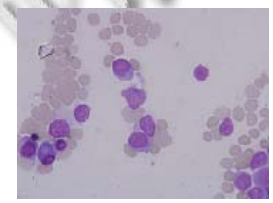


待校正图像

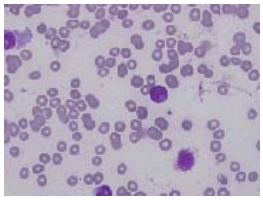


校正后图像

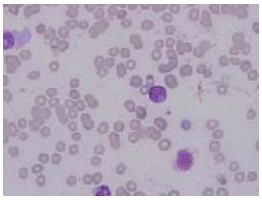
(a)　待校正图像直接聚三类校正效果图



标准图像



待校正图像



校正后图像

(b)　本文算法校正效果图

图 4　待校正图像聚三类校正效果图与本文算法校正效果图的对比

由表 1 校正前后颜色平均偏差值的对比表格可知，聚四类后重组的三类效果图比直接聚三类效果图准确很多. 正因为聚类的准确率高，才获得了较好的颜色校正效果.

表 1　聚三类及本文算法校正前后颜色偏差值的对比

校正前颜色平均偏差值	聚三类校正后颜色平均偏差值	本文方法校正后颜色平均偏差值
7.1970	5.4578	0.7820

4　结语

K-means 聚类算法最大的难点是需要用户根据先验知识提供聚类数. 本文针对细胞图像中白细胞与红细胞颜色较近难聚类的问题，提出了一种基于 K-means 聚类分量中心差值来确定聚类数的血细胞颜色校正算法. 通过观察去除背景类的 Cb 分量聚类中心差值的大小来确定合适的聚类数，其差值小，表明红白细胞的颜色特征较接近，这时应加大聚类数，提高聚类精度. 聚类数并非越大越好，能达到实验较好效果即可. 实验对比结果表明，本文方法简单、易行，且大大提高了颜色校正结果的准确率.

参考文献

1 胡威捷,汤顺青,朱正芳.现代颜色技术原理及应用.北京:北京理工大学出版社,2007.5–32.

2 徐晓昭,蔡轶珩,刘长江,等.基于图像分析的偏色检测及颜色校正方法.测控技术,2008,27(5):10–12.

3 Bertalmio M, Caselles V, Provenzi E, et al. Perceptual color correction through variational techniques. IEEE Trans. on Image Processing, 2007, 16(4): 1058–1072.

4 徐晓昭,沈兰荪,刘长江.颜色校正方法及其在图像处理中的应用.计算机应用研究,2008,25(8):2250–2254.

5 赵忠旭,沈兰荪,卫保国,等.基于人工神经网络的彩色校正方法研究.中国图象图形学报,2000,5(9):785–789.

6 印蔚蔚.数字图像自适应白平衡算法的研究与改进[硕士学位论文].镇江:江苏大学,2011.

7 Xiong W, Funt B, Shi L, et al. Automatic white balancing via gray surface identification. Color and Imaging Conference. The Society for Imaging Science and Technology, 2007, 2007(1): 143–146.

8 王易循,赵勋杰.基于K均值聚类分割彩色图像算法的改进.计算机应用与软件,2010,27(8):127–130.

9 汪嘉,姜明富,李友国.一种基于改进的K-Means算法的聚类分析方法.农业网络信息,2009,10:41.

10 Koschan A, Abidi M.章毓晋译.彩色数字图像处理.北京:清华大学出版社,2012.1–70.

11 Yin J, Cooperstock JR. Color correction methods with applications for digital projection environments. Journal of the Winter School of Computer Graphics, 2004, (12): 499– 506.

12 周世兵,徐振源,唐旭清.K-means 算法最佳聚类数确定方法.计算机应用,2010,(8):1995–1998.