K-means算法学习报告

## 一、基本概念

1、图像分割：图像分割是指将一幅图像分解成若干互不相交区域的集合，可以看作是像素的聚类。

2、k-means算法：基于距离相似性的聚类算法，通过比较样本之间的相似性，按照距离将样本聚成不同的簇。两个点的距离越近，其相似度就越大。彩色图像中的每一个像素是三维空间中的一个点，三维对应红、绿、蓝三原色的强度，基于K-means聚类算法的图像分割，以图像的像素为数据点进行聚类，然后将每个像素点以其对应的聚类中心替代，重构该图像。

## 二、k-means算法的具体过程为：

* **确定聚类中心的个数k**，即该图像当中主要色调的个数为K。论文中给出了k值的选取准则——将图像转化到HSV色彩空间，通过计算色调直方图来判定彩色图像中主色调的个数，进而自动地确定K值的大小。（存疑）
* **随机初始化K个聚类中心**。由于K-means算法对初始聚类中心的选取较为敏感，若选择的聚类中心不好，得到的聚类效果会非常差，因此K-means算法有许多改进方法，其中一种算法如下：

1. 在数据集中随机选择一个样本点作为第一个初始化的聚类中心
2. 选择其余的聚类中心
3. 计算样本中每一个样本点与初始化的聚类中心之间的距离，并选择其中最短的距离
4. 以概率选择距离最大的样本作为新的聚类中心
5. 重复第（1）（2）步，直到k个聚类中心被确定

* **确定k个聚类中心**

1. 计算每个样本与每个聚类中心之间的相似度，将样本划分到最相似的类别中
2. 对每个簇，以所有样本的均值作为该簇新的聚类中心
3. 重复第（a）（b）步，直到聚类中心不再改变或达到设定的迭代次数

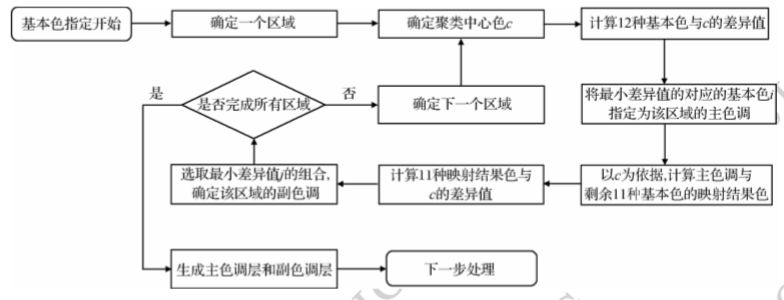
* **将每个像素点以其对应的聚类中心替代**，重构该图像，即可得到图像分割的结果。

## 三、k-means算法的缺点

从K-means算法框架可以看出，该算法的每一次迭代都要遍历所有样本，计算样本到所有聚类中心的距离。因而当样本规模非常大时，算法的时间开销是非常大的。

## 四、联系

通过k-means分割实现图像的预处理，对于图像中的每一区域，通过用户交互或者自动计算指定区域的基本色，即为该区域的主色调和副色调。具体流程如图所示：



在确定两种基本色之后，即可利用双色调映射的方法分别算出两个基本色层的颜色密度。