基于二分 K 均值聚类算法的数字档案优化

陈鹏[1,2] 程思[3] 鲍婷婷[1,2] 翟伶俐[1,2] 王宏斌[1,4]

(1 中国气象局交通气象重点开放实验室,南京 210008; 2 江苏省气象信息中心,南京 210009;
3 福建省泉州市气象局,泉州 362000; 4 江苏省气象科学研究所,南京 210009)

摘要 精细化预报服务和气象能源开发等需要时间序列长、空间和时间分辨率更高的气象资料,对逐小时资料的需求尤为突出。现存历史气象资料进行数字化扫描之后存在污点、褪色、模糊、字迹洇透等问题,不符合档案归档和服务的要求、同时也造成对图像进行数值提取的难度大大增加,提取结果的准确性也难以保证。本文提出一种基于 K 均值的图像优化算法,能够快速识别和区分图像背景和数据记录曲线,过滤图像中的噪点,统一数据记录曲线的颜色和粗细。经过优化之后的图像对比度和清晰度明显增加,体积明显缩小,实际应用中发现,经过优化之后的图像节约了存储资源和成本,同时清晰度有明显地提高,结果表明基于 K 均值的优化方法明显提高了气象数字化档案的质量和应用效果。

关键词 K 均值;气象档案;数字化;图像优化

中图分类号: P416.2 **DOI**: 10.19517/j.1671-6345.20180524 **文献标识码**: A

引言

在建设发展过程中,气象部门逐步积累下来大量的原始气象观测档案,特别是大量的自记纸档案。自记纸档案是气象台站进行温度、湿度、降水、风等要素观测时记录数据的原始纸质资料。各省都成立了专门的气象档案馆负责气象档案的管理和保护工作,仅以江苏省气象档案馆为例,目前馆藏建国至今的各类气象档案资料,包括各类观测报表 4393 册,各类观测簿(气簿)68051 本,各类自记纸 702.77 万张,各类气象天气图 1683 册,地面资料 1648 册等资料。由于年代长、保存条件差,许多纸张已经出现不同程度的变质、字迹变淡等现象,亟需对其开展拯救和保护工作。

现代化设备的快速发展与更新,信息技术的提高与逐渐完善,为以纸质为载体的大批量气象资料进行数字化处理提供了可行的基础[1-7]。2008 年以来,中国气象局开展了拯救历史气候资料的工作,先后组织了 7 期数字化扫描和录入工作,依托计算机技术,存储技术,扫描技术将过去的纸质气象档案资料转化为数字化气象档案资料。档案数字化之后不但可以节约保存的成本,还能极大地延长档案的保存时间和重复利用率,便于业务和科研人员随时调阅,为我省气象业务、科研和服务提供更多的信息化数据产品,进一步提升我省气象防灾减灾能力。

目前扫描后的档案存在图像歪斜、污点、褪色、模糊、字迹洇透等问题,既不符合档案归档和服务的要求、也不利于业务应用,特别是对图像进行数值提取的难度大大增加,数值提取准确性也难以保证。为了解决以上问题,本文提出一种基于二分 K 均值聚类的图像优化算法,通过聚类分析减轻甚至去除图像中的无效和错误部分,在此基础上进一步对图像修正和优化,最终获得清晰易读的档案图像,经过优化后的图像不但清晰度增加,同时图像体积最多能缩小一半以上。

1 档案数字化

档案数字化通过电子扫描技术将纸质档案资源转化为数字化的档案信息,通过图片加工存储、数据共享服务等手段,建立科学、高效、易用的数字档案

江苏省气象重点科研项目(基金编号:KZ201701)资助

作者简介:陈鹏,男,1986 年生,高工,主研领域气象信息化与数据应用,Email:409856986@88.com

收稿日期:2018 年 8 月 24 日;定稿日期:2019 年 7 月 9 日

库,便于检索和阅读,充分发挥档案的历史价值和实用价值,同时能避免经常翻阅对纸质档案的损害,利于档案的长期保存。

档案纸质资料数字化加工是以《DB32/T 1894—2011 档案数字化转换操作规程》为工作依据,按照规定的操作流程将各类纸质档案通过扫描仪输入到电脑中以电子格式存储的过程。档案数字化加工步骤主要包括领卷、拆卷、整理、扫描、图像处理、著录补录、导入数据库、目录挂接、光盘刻录等流程。整个处理流程充分利用扫描设备和图像处理技术,提高数字图像的质量,最终形成 TIF、PNG 等格式的数字档案。

对待扫描的档案案卷进行整理、检查,确认实体卷是否存在问题,并核对档案实体与电子目录的对应准确性,对电子目录中明显存在的问题依据相关标准规范进行校对。实体卷检查时如果发现页码标注错误等问题,则在整理时进行纠正,保证电子档案和实体档案完全一致。

档案扫描时,要根据纸质档案的状况包括纸张是否褪色、发黄、变薄等因素,设置最佳的扫描明暗度和对比度,保证原始扫描图像效果与原件高度吻合。对于纸张保存较好的档案,运用高速扫描仪进行档案资料的批量扫描。对于档案纸张质地脆弱,不适合反复拆分、装订的档案,则通过高速平板扫描仪进行不拆卷扫描,然后通过手工方式对电子图片进行处理。扫描要求保证电子档案图片的完整性、连续性及逻辑性;保证电子档案具体内容的完整,无漏码及缺页现象,内容连贯,不能前后颠倒(表 1)。

表 1　档案扫描要求

采集方式	档案质量	扫描色彩模式选择			分辨率
		黑白	灰度	彩色	
平板扫描仪、无边距扫描仪	纸张状况较差,过薄、过软或超厚的档案	档案页面为黑白两色,字迹清晰、不带插图	档案页面为黑白两色,字迹清晰度差或带有插图的档案,页面为多色文字	档案页面中有红头、印章或插有黑白照片、彩色照片、彩色插图	≥300 dpi (OCR 汉字识别)
高速扫描仪	纸张状况好的档案				≥300 dpi (OCR 汉字识别)

2　图像优化方法

档案图像优化的目的是为了解决原始档案图像中存在的污点、噪点、褪色等问题,通过一系列的颜色空间变换、聚类分析、降噪、锐化等手段修正或减少图像中无效或者错误的像素点,凸显图像主体如:坐标轴、数据曲线、手工标记等。优化后的图像应更加清晰,能在保证图像质量和分辨率的同时尽可能缩小图像存储空间。图像经过优化之后清晰易读,便于业务应用,提高数据处理应用及服务的效率。另外在优化过程当中发现有严重错误的文件,例如发现有图像不清晰、过度倾斜、重叠、漏扫等情况,可以通知扫描人员进行重扫或补扫。

档案图像内容包括标注栏、坐标轴、数据曲线、手工标记,其中标注栏及坐标轴为同色油墨印刷,数据曲线为探测仪器自动记录产生的蓝色墨迹曲线,手工标注为后期人工处理档案时添加的铅笔标注,一张高质量的档案图像理论上只应该存在以上几种内容对应的颜色,实际应用中发现并非如此。由于档案长时间存放引起的发黄、褪色、损坏,以及实际扫描中产生的图像噪点,使得数字化之后的档案图像中实际存在的颜色种类要远远多于理论值,影响图像的可辨识性。

针对以上问题,通过二分 K 均值聚类算法对扫描图像的所有像素进行聚类分析,从而将图像颜色分为背景色、印刷内容、墨迹曲线、手工标注四种颜色,将图像中的有效内容准确识别出来,再进一步对图像进行锐化、降噪等,达到凸显关键有用的信息,降低甚至去除噪点、污点等干扰信息,在提高图像表达能力的同时降低图像的大小。图像优化主要针对温度、湿度、降水、风等自记纸扫描档案,具体方法包括颜色空间转换、颜色聚类分析、图像强化 3 个步骤。

2.1　颜色空间转换

颜色空间(Colour Space)也称为色彩空间,是通过颜色空间上点的不同位置表示不同颜色的方

法，目前常用的颜色空间有 RGB，CMY，HSV，HSI 等，除此之外在图像处理的过程当中还经常会用到(Gray)灰度图。

优化前的原始档案图像是基于 RGB 的格式保存，RGB 空间中可以通过用 R、G、B 3 色不同分量的相加混合任意颜色。

传统图像分割方法是通过计算图像中每个像素点颜色与背景色在颜色空间中的欧式距离来实现图像分割，但直接使用该方法的效果并不理想，一方面是由于档案图像背景色噪点较多，直接通过像素与背景色的欧式距离进行判断和图像分割的误差较大，另一方面由于 RGB 空间的定义方式，有几类颜色与背景色的欧式距离非常接近，无法选取合适的距离阈值进行判断和分类。

为了有效地解决这个问题，需要将图片的 RGB 值映射到 HSV 空间，HSV 空间通过色相(Hue)、饱和度(Saturation)、明度(Value)3 个值表示不同的颜色。与 RGB 立方体的空间形状不同，HSV 的形状为圆椎体，色度(H)是指围绕圆柱体的中心轴旋转的角度(红色为 0°)；圆锥体的中心轴为亮度(V)，是指颜色的整体亮度，其变化范围从底部的黑色、中间的灰色渐变到顶部的白色；饱和度(S)对色彩的纯度，从中心往边缘递增，值越大色彩越纯，中心为 0，外圆周上鲜艳的颜色饱和度都为 1。将 RGB 值映射到 HSV 空间便于选择合适的阈值实现图像分割的任务(图 1)。

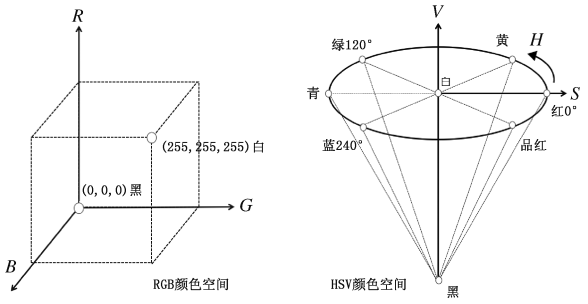

图 1 RGB 与 HSV 颜色空间示意图

2.2 聚类分析算法

聚类分析(cluster analysis，亦称为群集分析)是对于统计数据分析的一门技术，目前已在许多领域广泛应用[6-15]，包括机器学习，数据挖掘，模式识别，图像分析以及生物信息研究等。聚类的目的是把特征相似的对象通过分类的方法是把相似的对象通过静态分类的方法分成不同的组别或者更多的子集(subset)，这样让在同一个子集中的成员对象都有相似的一些属性，常见的包括在坐标系中更加短的空间距离等。

K 均值聚类过程为首先定义聚类的个数 k，随机产生 k 个聚类中心，按照点到每个聚类中心的距离，将点分配到离聚类中心距离最近的那个簇，完成所有点的分配后，重新计算每个簇的聚类中心；重复以上步骤直到满足收敛要求。但是传统的 K 均值聚类方法常常会收敛到局部最小值，而非全局最小值，实际聚类效果并不好。

为了解决局部收敛的问题，克服 K 均值算法收敛于局部最小值的问题，有人提出了另一个称为二分 K 均值的算法，该算法首先将所有点作为一个簇，然后将该簇一分为二，之后选择其中一个簇继续进行划分，选择的依据是划分之后的总误差最小，重复以上步骤，直至达到要求的聚类个数。与传统 K 均值方法相比，该算法的最大优势在于简洁和快速，并且能找到全局最优解(图 2)。

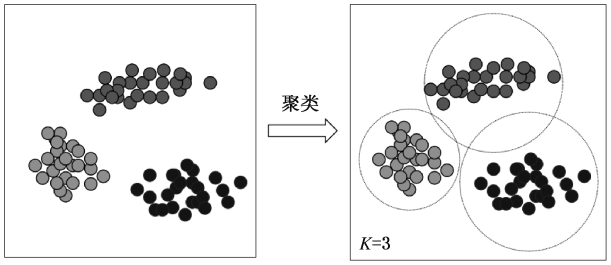

图 2 颜色聚类示意

通过二分 K 均值算法对图片所有像素的颜色进行聚类分析，得到几个颜色相对集中的簇分别对应自记纸表格、数据曲线、手工标注等需要提取的信息。基于聚类分析的结果将上述内容从背景色中分离出来，进一步地通过聚类算法获得每个颜色簇的中心点的颜色，利用每个簇中心点颜色代替整个簇的颜色，并在图像中做对应替换。这样根据图片内容整个图像只需要用 5～6 种颜色即可清晰地展现自记纸档案的内容，本文中初始 K 值为 3。

2.3 图像分析及图像强化

气象档案扫描图片为 3900×100 像素大小，分辨率为 300 DPI，除表格和标题等印刷内容外，数据记录曲线为纯蓝色墨水，手工标记为黑色铅笔，背景色应该是米白色。气象档案由于保存时间较长，自

记纸底色已经发黄，将档案扫描图片局部放大之后发现空白处并非纯色，而是包含黑、灰、红、蓝、黄、绿等多种颜色的组合，这些多余的颜色实际上是在扫描过程中产生的噪点。数据记录曲线同样存在噪点的问题，图片放大后发现整个数据曲线的颜色深浅不一、粗细不均，同时还存在少量非蓝色的像素点。从档案应用价值考虑，这些都不是因为实际观测产生的信息，而是扫描过程中产生的图像噪声，从图像分析应用的角度看，这些无效信息均应予以剔除。另一方面，通过聚类分析算法得到图像不同内容的代表色，并对图像内容进行替换，实际上是对图像中有效信息的强化，使得档案图像变得更加清晰易读（图 3）。

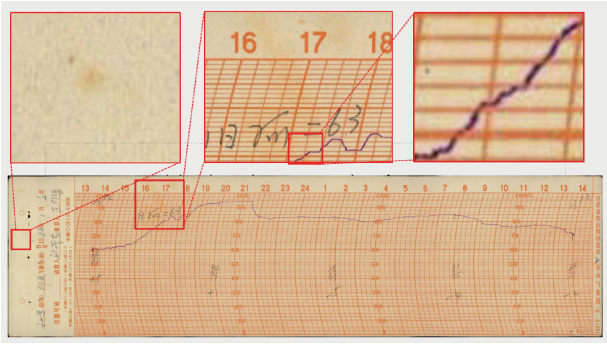

图 3　自记纸局部放大图像

2.4　应用对比

以湿度自记纸为例说明该算法的具体使用效果。首先通过对档案图像应用二分 K 均值聚类分析，通过聚类首先将背景色分离出来，另外颜色分布较为集中的 3 个簇分别是自记纸印刷内容、数据记录曲线、手工标注的颜色。在此基础上利用聚类算法获得颜色簇中心颜色，并用簇中心颜色替代整个簇的颜色，背景色用纯色替换，将原始 24 位彩色图像转成 8 位彩色图像（图 4）。

经过优化处理后的图像相比较原来的图像更加美观清晰，图像背景无噪点，图像内容更加凸显。通过局部放大发现，优化前后的区别主要体现在 3 个方面：①空白的部分变为纯色，之前图像空白处斑驳的“雪花点”已经消失不见；②图像中原先存在的发黄印记（图中蓝圈部位）已经消失，且不影响图中该部位的其他内容；③最重要的是原先深浅不一的数据记录曲线已变为深浅一致的深蓝色，数据曲线变得更加突出，易于辨识。

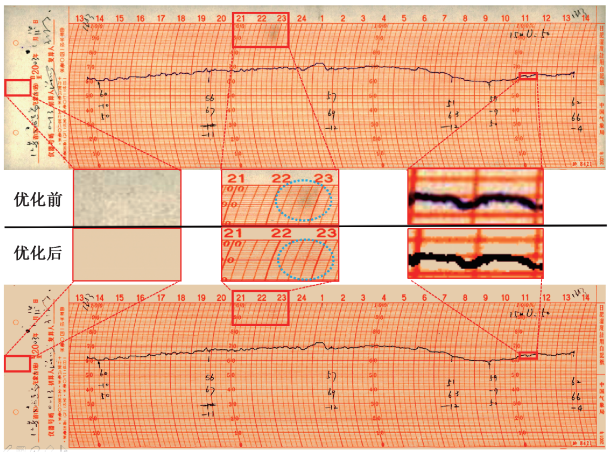

图 4　扫描图像优化前后对比

由于图片从 24 位彩色图像转成 8 位彩色图像，优化之后的图片存储空间也相应缩小，湿度图像未优化前每张大小约为 430 KB，经过优化之后的图像为 210 KB，在提高图片质量的同时还大大节约了存储空间。

经过优化之后的数据记录曲线从原来颜色深浅不一，线条粗细不一变成粗细基本均匀、颜色固定且与背景颜色及表格颜色有明显的区别，可以直接通过 RGB 值自动识别和提取数据曲线。

3　结论与讨论

目前国内对气象档案图像数字化的优化处理研究文献较少，主要研究方向都集中在如何进行图像数值提取[1,3,4,16-19]，虽然部分研究过程中有降噪的处理流程，但其目的是通过降噪、滤波等方法提取记录曲线，对自记纸档案整体图像质量的提升帮助不大，特别是图像由于年代久远出现的褪色，发黄，污损等档案保存中常常出现的问题，并未得到有效的解决。

本文提出了基于二分 K 均值聚类分析的图像优化算法，优化目的是对档案图像进行整体去噪的同时进行图像增强，优化之后的图像质量，无论是从人眼观察还是程序应用，其清晰度、可读性、易用性都得到较大的提升。应用结果表明，该算法能够较好地处理自记纸图像中存在的污损、发黄、模糊等问题，通过去除噪点，颜色聚类替换等方法，增加图像整体清晰度，提高图片识别率和准确率的同时大大减少图片占用空间，节约了存储和管理成本。聚类算法属于无监督算法，针对每张图片会产生不同的聚类结果，因此每张图片都基于不同的阈值进行图

像优化处理，适用性更强，优化效果也更明显。

数字档案优化为进一步解决档案资料规范化、准确性、可用性的问题打下了良好的基础。目前气象数字档案自动化提取软件正在研发当中。通过将扫描图像中的各类气象要素记录曲线转换成对应的数值，统一存储到数据库中，便于后期与各类科研和业务系统对接，提供历史数据在线服务，充分发挥历史资料的使用价值，有效提升气象档案管理和服务的水平。

参考文献

[1] 马宁，曹宁，马蕾. 数字图像处理技术在温度自记纸数字化识别中的应用[J]. 信息系统工程，2014(7)：74-75.

[2] 周玉文，姚双龙，翁窈瑶，等. 城市暴雨强度公式数据采样新方法[J]. 中国给水排水，2012，28(6)：9-12.

[3] 彭江华. 降水自记纸彩色图形数字化的技术处理[J]. 气象，2011，37(2)：249-253.

[4] 何志军，封秀燕，吴京生. 降水自记纸图形数字化处理资料分析[J]. 大气科学研究与应用，2008(1)：98-104.

[5] 吴名杰. 降水自记纸数据化处理常见问题及解决方法[J]. 气象研究与应用，2007，28(增刊2)：175-176.

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

[7] 贲志伟，赵勋杰. 基于改进的K均值聚类算法提取彩色图像有意义区域[J]. 计算机应用与软件，2010，27(9)：11-13.

[8] 韩最蛟. 基于数据密集性的自适应K均值初始化方法[J]. 计算机应用与软件，2014(2)：182-187.

[9] 刘广聪，黄婷婷，陈海南. 改进的二分K均值聚类算法[J]. 计算机应用与软件，2015(2)：261-263.

[10] 隋心怡，王瑞刚，张鸿翔. 一种改进的K-均值聚类算法[J]. 计算机与数字工程，2018，46(4)：682-685.

[11] 潘巍，周晓英，吴立锋，等. 基于半监督K-Means的属性加权聚类算法[J]. 计算机应用与软件，2017，34(3)：189-193.

[12] 郭占元，林涛. 面向大规模数据快速聚类K-means算法的研究[J]. 计算机应用与软件，2017，34(5)：43-47.

[13] 王振辉，夏鸿斌. 模糊加权多视角可能性聚类算法[J]. 计算机应用与软件，2017，34(4)：294-298.

[14] 崔红艳，曹建芳. 基于改进的分布式K-Means特征聚类的海量场景图像检索[J]. 计算机应用与软件，2016，33(6)：195-199.

[15] 朱真，杜轶诚，秦绪佳，等. 结构光条纹图像分割方法[J]. 计算机应用与软件，2016，33(8)：206-210.

[16] 王伯民，吕勇平，张强. 降水自记纸彩色扫描数字化处理系统[J]. 应用气象学报，2004，15(6)：737-744.

[17] 朱尽文，王志峻，汪青春. “降水自记纸数字化处理系统”简介及数字化处理时应注意的问题[J]. 青海气象，2006(2)：87-88.

[18] 孙力威，王艳，方晓，等. 降水自记纸数字化处理系统常见问题及解决方法[J]. 辽宁气象，2004(4)：40-40.

[19] 刘莎. 降水自记纸迹线提取有关异常处理的案例分析[J]. 气象与环境科学，2018，41(2)：126-130.

Digital Archive Optimization Based on *K*-Means Algorithm

Chen Peng[1,2] Cheng Si[3] Bao Tingting[1,2] Zhai Lingli[1,2] Wang Hongbin[1,4]

(1 Key Laboratory of Transportation Meteorology of China Meteorological Administration, Nanjing 210008; 2 Jiangsu Meteorological Information Center, Nanjing 210009; 3 Quanzhou Meteorological Service, Fujian, Quanzhou 362000; 4 Jiangsu Institute of Meteorological Sciences, Nanjing 21009)

Abstract: Meteorological forecasting services and meteorological energy development require data with longer time series, higher spatial and temporal resolution, especially for hourly data. Meteorological data scanned from recording papers have problems such as stains, fading, blurring, and smearing, which cannot meet the requirements of archiving and servicing, and also makes the numerical extraction of images greatly difficult, and the accuracy of extraction results is not guaranteed. This paper proposes an image optimization algorithm based on *K* means, which can quickly identify and distinguish the image background and data recording curves, filter noise in images, and unify the color and thickness of data recording curves. After optimization, the contrast and sharpness of the images are obviously increased, and the volume is obviously reduced. In practice, it is found that the optimized images save storage resources and cost, and the recognition rate is obviously improved. The result shows that the optimization method based on *K* means improves the quality and application effect of meteorological digital files.

Keywords: *K*-mean; weather archives; archives scanning; image optimization