(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115274093 A
(43)申请公布日 2022.11.01

(21)申请号 202210882919.2

(22)申请日 2022.07.26

(71)申请人 华东师范大学
地址 200241 上海市闵行区东川路500号

(72)发明人 李庆利 李逸殊 林凡力 胡雨婷

(74)专利代理机构 北京盛询知识产权代理有限公司 11901
专利代理师 相黎超

(51)Int.Cl.
G16H 50/20(2018.01)
G16H 30/20(2018.01)
G16H 30/40(2018.01)
G06V 10/762(2022.01)
G06V 10/764(2022.01)
G06N 20/00(2019.01)

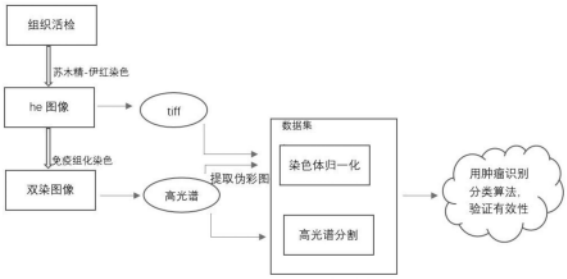
权利要求书2页 说明书6页 附图5页

(54)发明名称

生成包含自动标注文件的基准病理数据集的方法及系统

(57)摘要

本发明公开了生成包含自动标注文件的基准病理数据集的方法及系统，包括以下步骤：获取病理图像，其中所述病理图像包括：目标图像和高光谱图像；对所述高光谱图像提取伪彩图；对所述高光谱图像中的像素进行识别标注，得到病理数据集的标签部分；对所述目标图像和所述伪彩图进行聚类，基于聚类结果对所述目标图像和所述伪彩图进行染色归一化处理，得到病理数据集的图像部分；基于所述标签部分和所述图像部分，得到病理数据集。通过以上技术方案，本发明能够生成了包含自动标注文件的基准病理数据集，缓解了计算机辅助诊断方面研究数据集紧缺的情况。



CN 115274093 A

1.一种生成包含自动标注文件的基准病理数据集的方法,其特征在于,包括以下步骤:

获取病理图像,其中所述病理图像包括:目标图像和高光谱图像;对所述高光谱图像提取伪彩图;

对所述高光谱图像中的像素进行识别标注,得到病理数据集的标签部分;

对所述目标图像和所述伪彩图进行聚类,基于聚类结果对所述目标图像和所述伪彩图进行染色归一化处理,得到病理数据集的图像部分;基于所述标签部分和所述图像部分,得到病理数据集。

2.根据权利要求1所述的生成包含自动标注文件的基准病理数据集的方法,其特征在于,

得到病理数据集的标签部分的过程包括:

通过病理识别模型对所述高光谱图像中的像素进行识别标注,得到病理数据集的标签部分,其中所述病理识别模型为LightGBM模型。

3.根据权利要求2所述的生成包含自动标注文件的基准病理数据集的方法,其特征在于,

对所述高光谱图像中的像素进行识别标注之前还包括:

提取所述高光谱图像的特征像素,基于所述特征像素的通道值及标签构建病理识别模型,通过决策树算法对所述病理识别模型进行训练,直到输出的误差减小到期望值,得到训练好的病理识别模型,通过训练好的病理识别模型对所述高光谱图像中的像素进行识别标注。

4.根据权利要求1所述的生成包含自动标注文件的基准病理数据集的方法,其特征在于,

对所述目标图像和所述伪彩图进行聚类的过程包括:

基于所述目标图像和所述伪彩图中的像素值,对所述目标图像和所述伪彩图分别进行聚类,得到聚类结果。

5.根据权利要求4所述的生成包含自动标注文件的基准病理数据集的方法,其特征在于,

得到病理数据集的图像部分的过程包括:

基于所述聚类结果,分别计算第一亮度值和第二亮度值,将所述第一亮度值和所述第二亮度值进行比较,若所述第一亮度值小于所述第二亮度值,则将所述伪彩图中像素点的红绿蓝三通道的值替换成所述目标图像中像素点的红绿蓝三通道的值;否则不替换;基于比较结果得到病理数据集的图像部分;

其中所述第一亮度值为所述伪彩图中像素点的亮度值,所述第二亮度值为所述目标图像中像素点亮度值。

6.一种生成包含自动标注文件的基准病理数据集的系统,其特征在于,包括:病理图像获取模块、数据集标签获取模块及数据集图像获取模块;

所述病理图像获取模块,用于获取病理图像,其中所述病理图像包括:目标图像和高光谱图像;对所述高光谱图像提取伪彩图;

所述数据集标签获取模块,用于对所述高光谱图像中的像素进行识别标注,得到病理数据集的标签部分;

所述数据集图像获取模块，用于对所述目标图像和所述伪彩图进行聚类，基于聚类结果对所述目标图像和所述伪彩图进行染色归一化处理，得到病理数据集的图像部分。

7.根据权利要求6所述的生成包含自动标注文件的基准病理数据集的系统，其特征在于，

所述数据集标签获取模块包括模型构建单元；

所述模型构建单元，用于提取所述高光谱图像的代表性像素，基于所述代表性像素的通道值及标签构建病理识别模型，通过决策树算法对所述病理识别模型进行训练，直到输出的误差减小到期望值，得到训练好的病理识别模型，其中所述病理识别模型为LightGBM模型。

8.根据权利要求7所述的生成包含自动标注文件的基准病理数据集的系统，其特征在于，

所述数据集标签获取模块还包括标签获取单元；

所述标签获取单元，用于通过所述训练好的病理识别模型对所述高光谱图像中的像素进行识别标注，得到病理数据集的标签部分。

9.根据权利要求6所述的生成包含自动标注文件的基准病理数据集的系统，其特征在于，

所述数据集图像获取模块包括图像处理单元；

所述图像处理单元，基于所述目标图像和所述伪彩图中的像素值，对所述目标图像和所述伪彩图分别进行聚类，得到聚类结果。

10.根据权利要求9所述的生成包含自动标注文件的基准病理数据集的系统，其特征在于，

所述数据集图像获取模块还包括亮度值比较单元；

所述亮度值比较单元，基于所述聚类结果，分别计算第一亮度值和第二亮度值，将所述第一亮度值和所述第二亮度值进行比较，若所述第一亮度值小于所述第二亮度值，则将所述伪彩图中像素点的红绿蓝三通道的值替换成所述目标图像中像素点的红绿蓝三通道的值；否则不替换；基于比较结果得到病理数据集的图像部分；

其中所述第一亮度值为所述伪彩图中像素点的亮度值，所述第二亮度值为所述目标图像中像素点亮度值。

生成包含自动标注文件的基准病理数据集的方法及系统

技术领域

[0001] 本发明属于医学肿瘤识别领域，特别是涉及生成包含自动标注文件的基准病理数据集的方法及系统。

背景技术

[0002] 在人工智能(AI)的大趋势下，实现人工智能与数字病理学的结合可能是该领域的趋势。对于病理学家来说，许多癌症的检测和分析越来越依赖于数字病理学，越来越多的深度学习模型被提出，用来评估和预测肿瘤。然而，这些智能算法都需要大量的带有标注的高质量数据集。

[0003] 目前数据集都需要专业人员手工标注，成本高，准确率低，可获取量很少，识别分析结果的准确率和算法鲁棒性不足以满足实际应用中计算机辅助病理诊断的要求。虽然已有学者意识到了带有标注信息数据集地重要性，一些数据集也被制作出来。但是，所有这些已经得到的数据集仍然需要专业人员手工进行标注，即使是专业病理医生，也容易遗漏小区域的肿瘤，收集带有大量注释信息的数据集来训练深度学习模型变得不切实际。

发明内容

[0004] 本发明的目的是提供一种生成包含自动标注文件的基准病理数据集的方法，以解决上述现有技术存在的问题。

[0005] 为实现上述目的，本发明提供了一种生成包含自动标注文件的基准病理数据集的方法，包括以下步骤：

[0006] 获取病理图像，其中所述病理图像包括：目标图像和高光谱图像；对所述高光谱图像提取伪彩图；

[0007] 对所述高光谱图像中的像素进行识别标注，得到病理数据集的标签部分；

[0008] 对所述目标图像和所述伪彩图进行聚类，基于所述聚类结果对所述目标图像和所述伪彩图进行染色归一化处理，得到病理数据集的图像部分；基于所述标签部分和所述图像部分，得到病理数据集。

[0009] 优先地，得到病理数据集的标签部分的过程包括：

[0010] 通过病理识别模型对所述高光谱图像中的像素进行识别标注，得到病理数据集的标签部分。

[0011] 优先地，对所述高光谱图像中的像素进行识别标注之前还包括：

[0012] 提取所述高光谱图像的特征像素，基于所述特征像素的通道值及标签构建病理识别模型，通过决策树算法对所述病理识别模型进行训练，直到输出的误差减小到期望值，得到训练好的病理识别模型，通过训练好的病理识别模型对所述高光谱图像中的像素进行识别标注。

[0013] 优先地，对所述目标图像和所述伪彩图进行聚类的过程包括：

[0014] 基于所述目标图像和所述伪彩图中的像素值，对所述目标图像和所述伪彩图分别

进行聚类，得到聚类结果。

[0015] 优先地，得到病理数据集的图像部分的过程包括：

[0016] 基于所述聚类结果，分别计算第一亮度值和第二亮度值，将所述第一亮度值和所述第二亮度值进行比较，若所述第一亮度值小于所述第二亮度值，则将所述伪彩图中像素点的红绿蓝三通道的值替换成所述目标图像中像素点的红绿蓝三通道的值；否则不替换；基于比较结果得到病理数据集的图像部分；

[0017] 其中所述第一亮度值为所述伪彩图中像素点的亮度值，所述第二亮度值为所述目标图像中像素点亮度值。

[0018] 另一方面，为了实现上述技术目的，本发明提供了一种生成包含自动标注文件的基准病理数据集的系统，包括：病理图像获取模块、数据集标签获取模块及数据集图像获取模块；

[0019] 所述病理图像获取模块，用于获取病理图像，其中所述病理图像包括：目标图像和高光谱图像；对所述高光谱图像提取伪彩图；

[0020] 所述数据集标签获取模块，用于对所述高光谱图像中的像素进行识别标注，得到病理数据集的标签部分；

[0021] 所述数据集图像获取模块，用于对所述目标图像和所述伪彩图进行聚类，基于所述聚类结果对所述目标图像和所述伪彩图进行染色归一化处理，得到病理数据集的图像部分。

[0022] 优选地，所述数据集标签获取模块包括模型构建单元；

[0023] 所述模型构建单元，用于提取所述高光谱图像的特征像素，基于所述特征像素的通道值及标签构建病理识别模型，通过决策树算法对所述病理识别模型进行训练，直到输出的误差减小到期望值，得到训练好的病理识别模型。

[0024] 优选地，所述数据集标签获取模块还包括标签获取单元；

[0025] 所述标签获取单元，用于通过所述训练好的病理识别模型对所述高光谱图像中的像素进行识别标注，得到病理数据集的标签部分。

[0026] 优选地，所述数据集图像获取模块包括图像处理单元；

[0027] 所述图像处理单元，基于所述目标图像和所述伪彩图中的像素值，对所述目标图像和所述伪彩图分别进行聚类，得到聚类结果。

[0028] 优选地，所述数据集图像获取模块还包括亮度值比较单元；

[0029] 所述亮度值比较单元，基于所述聚类结果，分别计算第一亮度值和第二亮度值，将所述第一亮度值和所述第二亮度值进行比较，若所述第一亮度值小于所述第二亮度值，则将所述伪彩图中像素点的红绿蓝三通道的值替换成所述目标图像中像素点的红绿蓝三通道的值；否则不替换；基于比较结果得到病理数据集的图像部分；其中所述第一亮度值为所述伪彩图中像素点的亮度值，所述第二亮度值为所述目标图像中像素点亮度值。

[0030] 本发明的技术效果为：本发明获取目标图像、原图像及高光谱图像；利用高光谱图像实现了病理影像的自动标注，得到数据集的标签部分；通过对目标图像和原图像进行聚类计算，进一步实现染色体归一化，得到数据集的图片部分，不需要大量数据提前训练网络，本发明充分利用高光谱图像的光谱信息，生成了包含自动标注文件的基准病理数据集，缓解了计算机辅助诊断方面研究数据集紧缺的情况。

附图说明

[0031] 构成本申请的一部分的附图用来提供对本申请的进一步理解，本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。在附图中：

[0032] 图1为本发明实施例中的方法流程图；

[0033] 图2为本发明实施例中的数据集中标签变化过程示意图；

[0034] 图3为本发明实施例中的染色标准化流程图；

[0035] 图4为本发明实施例中的数据集中图像变化过程示意图；

[0036] 图5为本发明实施例中的系统示意图。

具体实施方式

[0037] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

[0038] 需要说明的是，在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0039] 实施例一

[0040] 如图1所示，本实施例中提供一种生成包含自动标注文件的基准病理数据集的方法，包括以下步骤：

[0041] 获取病理图像，其中病理图像包括：目标图像和高光谱图像；对高光谱图像提取伪彩图；

[0042] 对高光谱图像中的像素进行识别标注，得到病理数据集的标签部分；

[0043] 对目标图像和伪彩图进行聚类，基于聚类结果对目标图像和伪彩图进行染色归一化处理，得到病理数据集的图像部分；基于标签部分和图像部分，得到病理数据集。

[0044] 在一些实施例中，得到病理数据集的标签部分的过程包括：

[0045] 通过病理识别模型对高光谱图像中的像素进行识别标注，得到病理数据集的标签部分。

[0046] 在一些实施例中，对高光谱图像中的像素进行识别标注之前还包括：

[0047] 提取高光谱图像的代表性像素，基于代表性像素的通道值及标签构建病理识别模型，通过决策树算法对病理识别模型进行训练，直到输出的误差减小到期望值，得到训练好的病理识别模型，通过训练好的病理识别模型对高光谱图像中的像素进行识别标注。

[0048] 在一些实施例中，对目标图像和伪彩图进行聚类的过程包括：

[0049] 基于目标图像和伪彩图中的像素值，对目标图像和伪彩图分别进行聚类，得到聚类结果。

[0050] 在一些实施例中，得到病理数据集的图像部分的过程包括：

[0051] 基于聚类结果，分别计算第一亮度值和第二亮度值，将第一亮度值和第二亮度值进行比较，若第一亮度值小于第二亮度值，则将伪彩图中像素点的R、G、B的值替换成目标图像中像素点的R、G、B的值；否则不替换；基于比较结果得到病理数据集的图像部分；

[0052] 其中第一亮度值为伪彩图中像素点的亮度值，第二亮度值为目标图像中像素点亮度值。

[0053] 生成病理数据集的具体实施步骤包括：

[0054] (1)组织活检并制备病理切片，得到单染，双染和高光谱图像；

[0055] 该步骤具体为：

[0056] ①、组织活检，制备苏木精-伊红染色(h&e)的病理切片以及h&e与免疫组化染色(CAM5.2)的双染病理切片；

[0057] ②、使用全玻片彩色扫描仪得到单染图其放大20倍后的全玻片彩色图像；

[0058] ③、用显微高光谱成像平台得到双染图的高光谱图像；

[0059] ④、从高光谱图像中抽三个近似R、G、B波段得到伪彩图；

[0060] (2)利用XGBoost的改进集成算法LightGBM训练分类器，通过高光谱图像多个通道的光谱信息，区分出两个感兴趣区域：病变区域非病变区域。得到数据集的标签部分；

[0061] 该步骤具体为：

[0062] ①、利用ENVI软件挑选部分具有代表性的像素，将这些像素所有通道的值及其标签(癌症区域为0，正常区域为1)生成excel表；

[0063] ②、将生成的excel表输入LightGBM进行训练，LightGBM采用了基于直方图(Histogram)的决策树算法，基本思想是：把连续的浮点特征值离散化成K个整数；遍历数据，根据离散化后的值作为索引在直方图中累积统计量；然后根据直方图的离散值，遍历寻找最优的分割点。相比基于预排序(pre-sorting)的XGBoost，直方图算法有占用内存小和时间复杂度低的优点。

[0064] 在Histogram算法之上，LightGBM进行进一步的优化。首先它抛弃了大多数GBDT工具使用的按层生长(level-wise)的决策树生长策略，而使用了带有深度限制的按叶子生长(leaf-wise)算法，降低了更多误差，提升了精度。但可能会长出比较深的决策树，产生过拟合，所以叶子节点数是其中最重要的参数；

[0065] ③、调整参数，叶子节点数的值越大准确率越高，但是太大会出现过拟合，将其从30调为60后，准确率从0.97上升为0.98；

[0066] ④、用训练好的模型对整幅图中每个像素进行预测，>0.5作黑色正常区域，<0.5作白色癌症区域；

[0067] ⑤、对得到的图像进行中值滤波(窗口大小为36)优化结果。数据集中标签过程示意图，如图2所示。

[0068] (3)提出了一种结合K-means聚类和Wasserstein距离的染色标准化算法，得到数据集的图像部分。

[0069] 该步骤具体为：

[0070] ①、对原图像(双染图像)和目标图像(h&e单染图像)分别进行kmeans聚类，各聚20类；

[0071] ②、根据原图像和目标图像灰度图中每个像素的值所属的区间(0到255分成255个区间)，得到上述40类的分布；

[0072] ③、由上述的分布利用wasserstein距离为原图中的每一类找到目标图中最相近的一类，公式如下：

[0073] $$\mathbb{W}(P1,\ P2)=\inf_{\gamma\sim\Pi(P1,P2)} E_{(x_1y)\sim\gamma}[\|x-y\|]$$

[0074] 其中，γ为每一个可能的联合分布，

[0075] x和y为从γ中采样得的样本，

[0076] ||x-y||为这对样本的距离，

[0077] Wasserstein距离就是在所有可能的联合分布中样本对距离的期望值能够取到的下界。

[0078] ④、比较原图中每个像素点的亮度(0.299*R+0.587*G+0.114*B)和该像素点所属类别对应目标图中类别中心的亮度，如果小于，则将该点处R、G、B的值换为对应目标图中类别中心处R、G、B的值；否则将该点处R、G、B的值不变。具体染色标准化流程图，如图3所示；数据集中图像过程示意图，如图4所示。

[0079] 本实施例有益效果：

[0080] 本实施例提出的生成包含自动标注文件的基准病理数据集的方法，通过制备多标记的病理切片，利用高光谱图像多出的一维光谱信息实现了病理影像的自动标注，得到数据集的标签部分；使用kmeans等无监督的方法实现染色体归一化，得到数据集的图片部分，不需要大量数据提前训练网络。本发明充分利用高光谱图像的光谱信息，生成了包含自动标注文件的基准病理数据集，缓解了计算机辅助诊断方面研究数据集紧缺的情况。

[0081] 实施例二

[0082] 如图5所示，本发明提供了一种生成包含自动标注文件的基准病理数据集的系统，包括：病理图像获取模块、数据集标签获取模块及数据集图像获取模块；

[0083] 病理图像获取模块，用于获取病理图像，其中病理图像包括：目标图像和高光谱图像；对高光谱图像提取伪彩图；

[0084] 数据集标签获取模块，用于对高光谱图像中的像素进行识别标注，得到病理数据集的标签部分；

[0085] 数据集图像获取模块，用于对目标图像和伪彩图进行聚类，基于聚类结果对目标图像和伪彩图进行染色归一化处理，得到病理数据集的图像部分。

[0086] 在一些实施例中，数据集标签获取模块包括模型构建单元；

[0087] 模型构建单元，用于提取高光谱图像的代表性像素，基于代表性像素的通道值及标签构建病理识别模型，通过决策树算法对病理识别模型进行训练，直到输出的误差减小到期望值，得到训练好的病理识别模型。

[0088] 在一些实施例中，数据集标签获取模块还包括标签获取单元；

[0089] 标签获取单元，用于通过训练好的病理识别模型对高光谱图像中的像素进行识别标注，得到病理数据集的标签部分。

[0090] 在一些实施例中，数据集图像获取模块包括图像处理单元；

[0091] 图像处理单元，基于目标图像和伪彩图中的像素值，对目标图像和伪彩图分别进行聚类，得到聚类结果。

[0092] 在一些实施例中，数据集图像获取模块还包括亮度值比较单元；

[0093] 亮度值比较单元，基于聚类结果，分别计算第一亮度值和第二亮度值，将第一亮度值和第二亮度值进行比较，若第一亮度值小于第二亮度值，则将伪彩图中像素点的R、G、B的值替换成目标图像中像素点的R、G、B的值；否则不替换；基于比较结果得到病理数据集的图像部分；其中第一亮度值为伪彩图中像素点的亮度值，第二亮度值为目标图像中像素点亮

度值。

[0094] 以上所述，仅为本申请较佳的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到的变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应该以权利要求的保护范围为准。

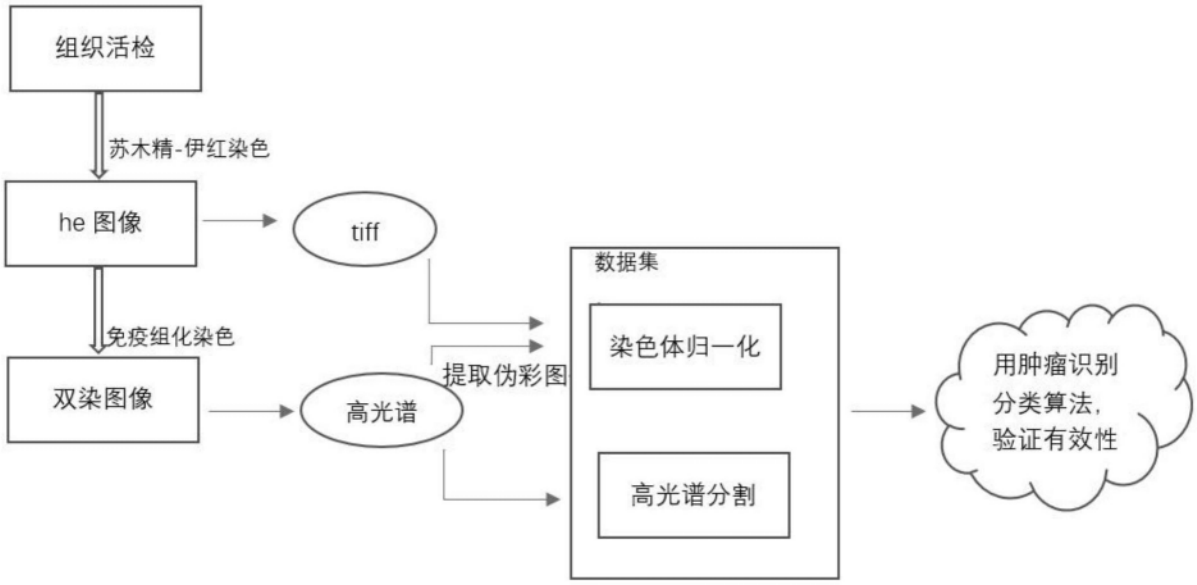

组织活检
苏木精-伊红染色
he 图像
tiff
免疫组化染色
双染图像
高光谱
提取伪彩图
数据集
染色体归一化
高光谱分割
用肿瘤识别
分类算法，
验证有效性

图1

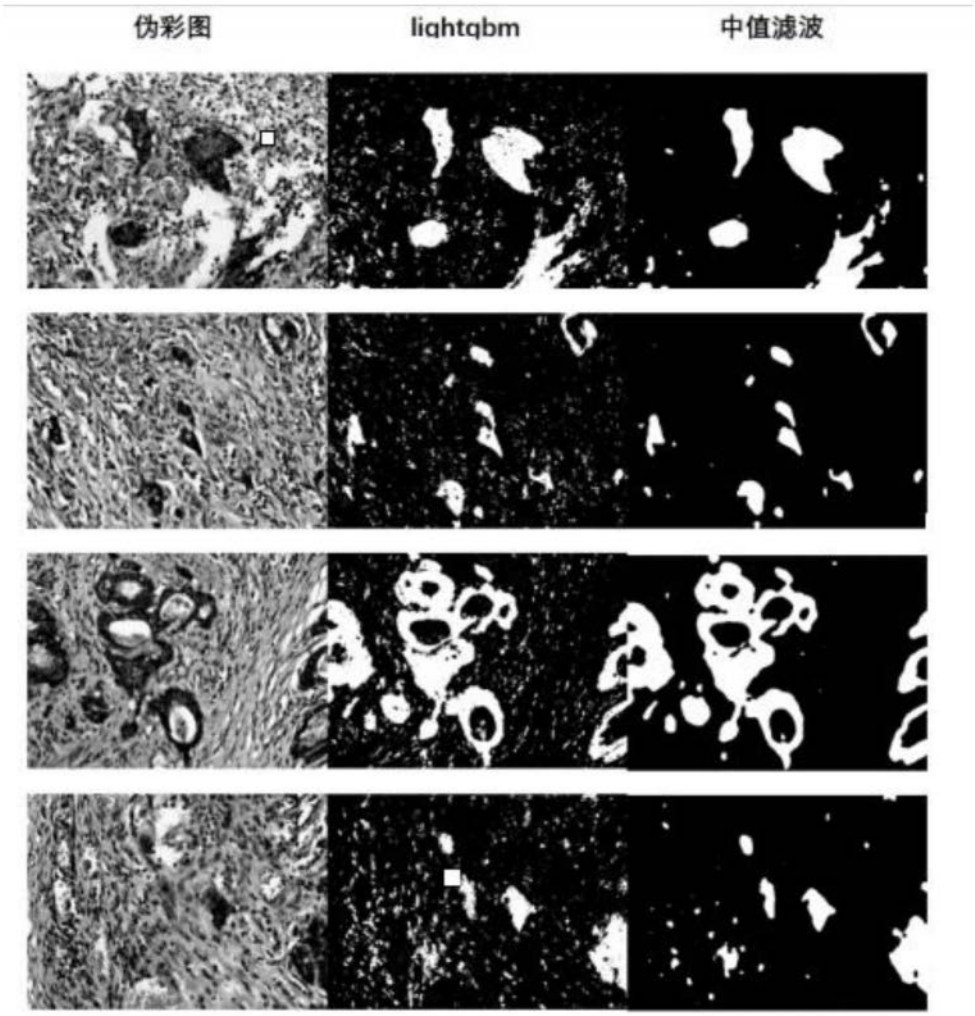


图2

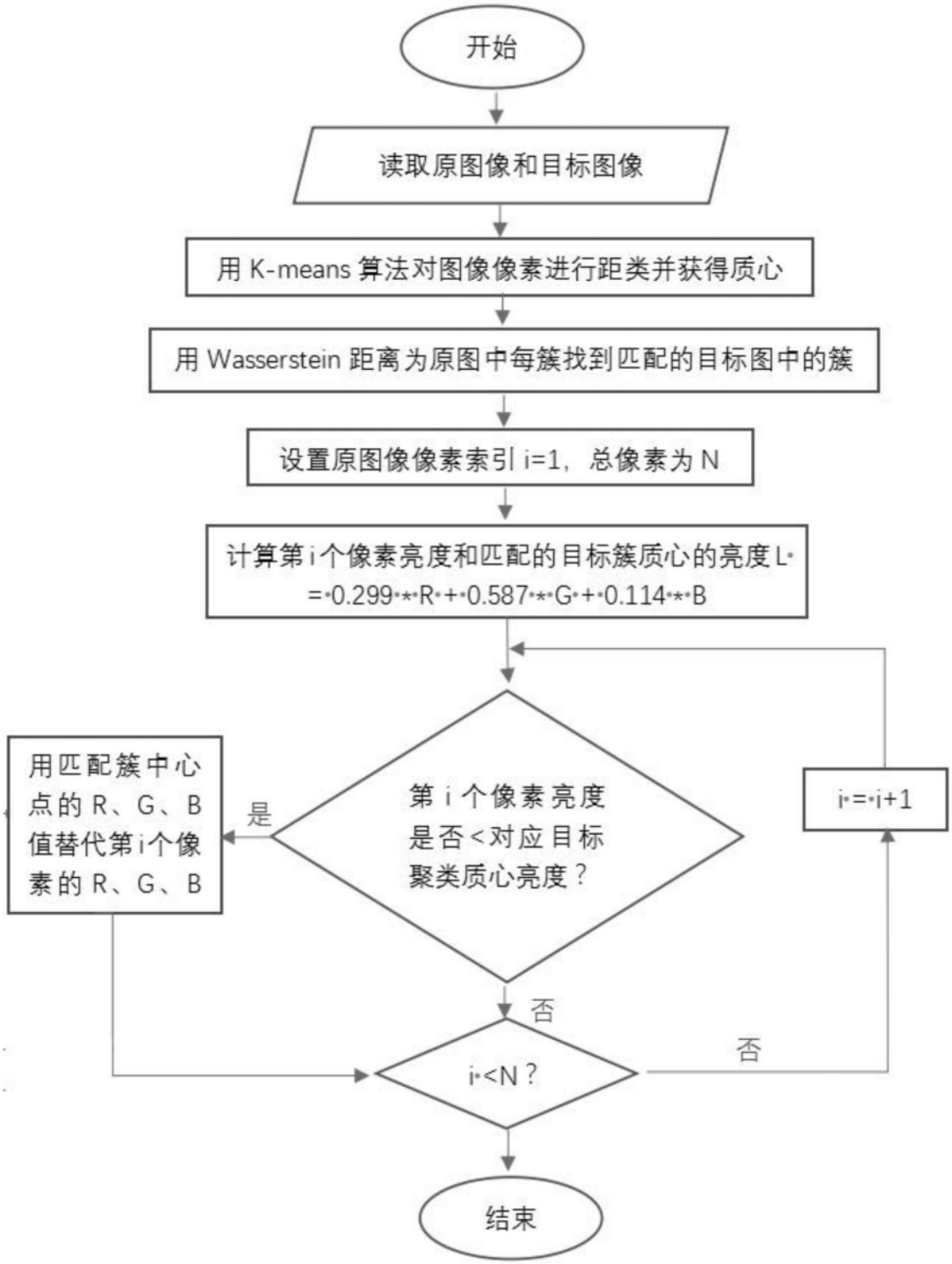

开始
读取原图像和目标图像
用 K-means 算法对图像像素进行距类并获得质心
用 Wasserstein 距离为原图中每簇找到匹配的目标图中的簇
设置原图像像素索引 i=1，总像素为 N
计算第 i 个像素亮度和匹配的目标簇质心的亮度 L
=0.299*R++0.587*G++0.114*B
第 i 个像素亮度是否＜对应目标聚类质心亮度？
是
用匹配簇中心点的 R、G、B 值替代第 i 个像素的 R、G、B
否
i=i+1
i<N？
否
结束

图3

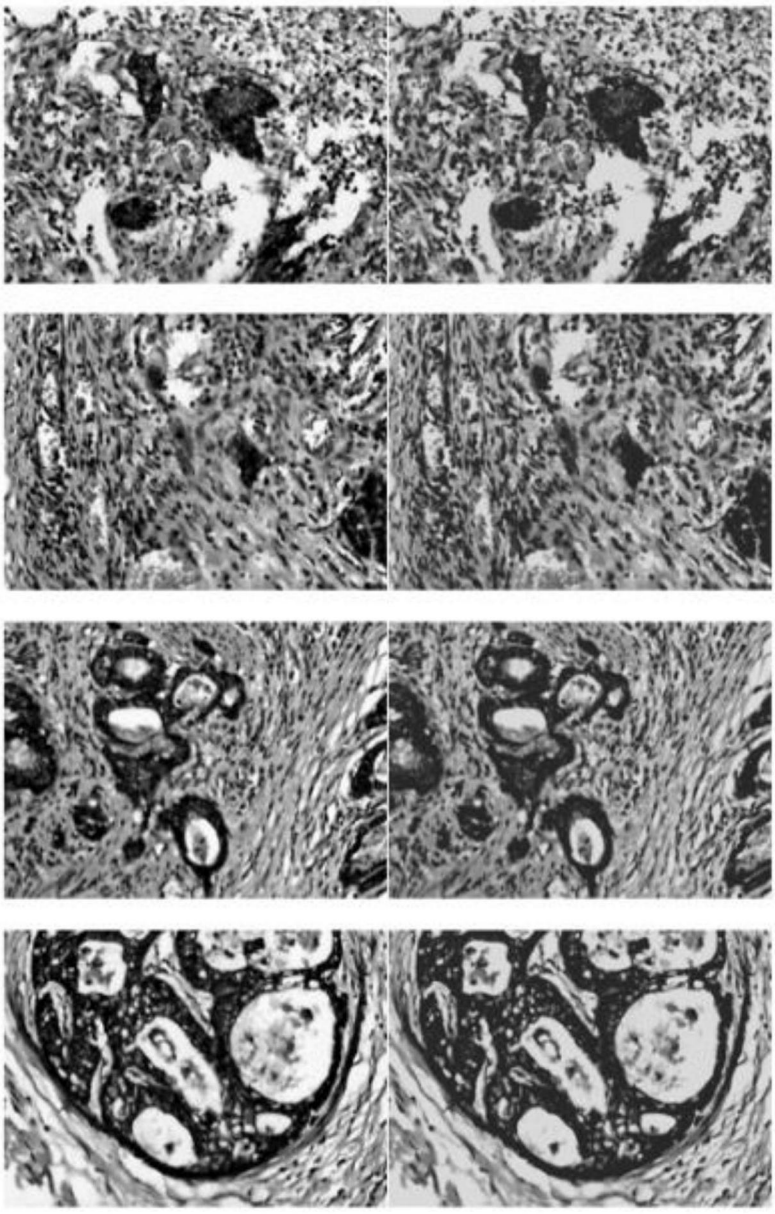


图4

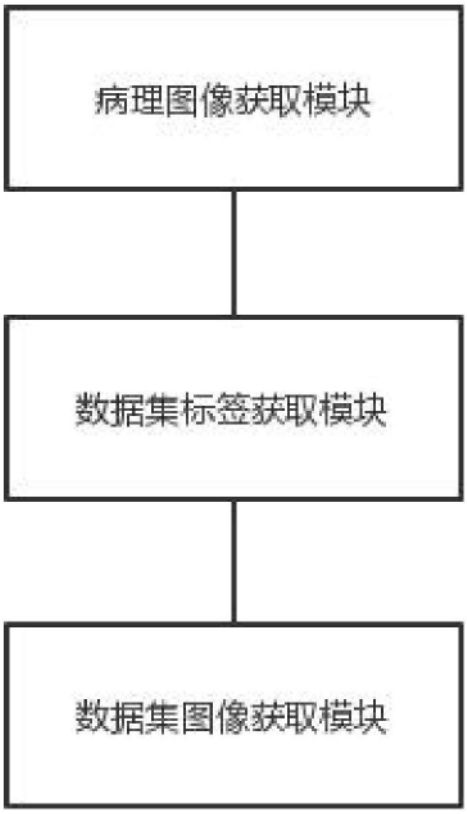


图5