学号 20155467

密级

东北大学本科毕业论文

基于最小生成树的
宫颈癌病理图像聚类研究

学 院 名 称 ：中荷生物医学与信息工程学院

专 业 名 称 ：生物医学工程

学 生 姓 名 ：尚麟静

指 导 教 师 ：李晨 副教授

二〇一九年六月

郑 重 声 明

本人呈交的学位论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本学位论文的知识产权归属于培养单位。

本人签名： 日期：

摘　　要

组织病理学在临床上占有非常重要的地位，虽说目前也有一些非入侵性的测试和成像方法可以检测癌症，但组织病理学终究是不可避免的检验手段，并被医生在临床中看作是诊断肿瘤的“金标准”。然而，组织病理学家利用高精度的电子显微镜查看组织切片，根据所学的医学知识来分析组织结构，并进行诊断。这必定需要花费很长时间才能培育一名优秀的有经验的组织病理学医生。并且，这种肉眼观测病理图像的方式的客观性是不稳定的，医生难免受经验、工作量和情绪的影响。为此，越来越多的计算机辅助诊断（CAD）技术被应用到组织病理学图像分析领域。

本文提出了一种基于图论的方法，利用图形的特征来识别图像中不同组织结构的信息来帮助医生快速发现病灶区域，提高诊断的准确率。宫颈癌是女性发病率十分高的癌症之一，对于它的研究十分具有代表性。为了将病理图像中隐藏的重要拓扑信息应用于解决宫颈癌组织病理学聚类问题，本文首先基于图论方法应用颜色特征和 k 均值进行第一阶段的粗糙聚类。其次用骨架化的方法将生成的节点近似看做宫颈细胞核的分布。随后，基于所生成的节点构造最小生成树并提取几何特征。然后，再次应用 k 均值算法聚类。最后，提供鼠标手动标注的功能给医生。在实验中，本文所采用的宫颈癌组织病理学数据集有十张完整扫描的图像，取得可观的结果，并在癌症预防方面具有巨大潜力。

关键词：宫颈癌；病理图像；图论；骨架化；聚类

ABSTRACT

Histopathology takes an overwhelming role in clinic. Although there are some non-invasive testing and imaging methods to detect cancer at present, histopathology is still an unavoidable means of testing, which is regarded as the ‘golden standard’ of diagnosing tumors by clinical doctors. Histopathologists, however, use microscopes to observe tissue slices, analyze their structures and diagnose them based on their medical knowledge. It must take a long period to develop an excellent and experienced histopathologist. Moreover, the objectivity of this way of observing pathological images with naked eyes is unstable, which doctors are inevitably affected by experience, workload and emotion. To this end, an increasing number of computer-aided diagnosis (CAD) techniques have been applied to the field of histopathological image analysis.

In this paper, a method based on graphs theory is informed, which can identify information of different organization structures in images by using the characteristics of graphs to help doctors quickly find the lesion area and improve the accuracy of diagnosis. Cervical cancer is one of the most common cancers in woman, hence its research is very representative. In order to apply the important topological information hidden in pathological images to solve the clustering problem of cervical cancer tissues, the first stage of rough clustering is managed used color features and *k*-means based on graph theory in this paper. Secondly, applying the skeletonization method, the generated nodes are approximately regarded as the distribution of cervical nuclei. Then, the minimum spanning tree is constructed based on the generated nodes and the geometric features are extracted. Subsequently, *k*-means clustering algorithm is applied again. Finally, the function of manual correction using mouse is provided to the doctor. In the experiment, the dataset of histopathology of cervical cancer used in this paper has ten whole scanned images, which have achieved considerable results, and has great potential in cancer prevention.

Key words: Cervical cancer; Histopathology image; Graph theory; Skeletonization; Clustering

目　　录

摘　　要 I

ABSTRACT II

目　　录 III

1　绪论

1.1　研究背景及研究动机 1

1.2　国内外研究现状 1

1.2.1　医学层面相关研究现状 1

1.2.2　技术层面相关研究现状 2

1.3　研究意义与贡献 3

1.4　研究内容与主要工作 3

1.5　论文结构介绍 5

2　图论及机器学习基础知识

2.1　宫颈癌组织病理学图像 7

2.2　特征提取 8

2.2.1　颜色特征 8

2.2.1　形状特征 10

2.3　机器学习理论 11

2.3.1　有监督学习 11

2.3.2　无监督学习 12

2.3.3　k-means 算法 12

2.4　图论 14

2.5　鼠标取点 16

2.6　本章小结 16

3　基于最小生成树的病理图像聚类方法

3.1　预处理 17

3.1.1 大津阈值分割 17
3.1.2 k-means 图像分割 18
3.2 最小生成树构建 19
3.2.1 颜色直方图特征提取 19
3.2.2 连通区域筛选 20
3.2.3 骨架化 20
3.2.4 节点生成 21
3.2.5 最小生成树构建 22
3.3 k-means 聚类 23
3.3.1 特征值计算 23
3.3.2 k-means 聚类 24
3.4 手动标注 25
3.5 本章小结 25
4 实验结果与分析
4.1 实验设置 27
4.1.1 数据来源简介 27
4.1.2 数据设置 27
4.1.3 实验环境 28
4.2 实验结果及分析 28
4.2.1 基于最小生成树的病理图像聚类研究结果与分析 28
4.2.2 错误结果分析 36
4.3 本章小结 38
5 总结与展望 39
5.1 总结 39
5.2 展望 40
参考文献 41
致谢 45
个人简历 47

1 绪论

1.1 研究背景及研究动机

宫颈癌是一种全球性频发的癌症，发病率排名仅次于乳腺癌位列第二，死亡率排名第三，占女性子宫恶性肿瘤的一半以上，是困扰着许多女性健康的难题[1]。然而，近年来宫颈癌的发病率不断上升，发展中国家更是深受其害[2]。并且患者的患病的年龄越来越小[3]。作息不规律、妊娠次数频繁都是导致宫颈癌变的原因[4]。宫颈癌发病率高，是子宫恶性肿瘤中非常具有代表性的一种，因此研究宫颈癌具有其深刻意义，尤其是其生存率高，治愈效果好的特点，越来越多的研究学者开始投入研究。

传统的病理图像诊断方法就是凭借医生用肉眼观察病理图像，之后根据自己的判断来进行诊断。这使得诊断过程持续时间长，花费精力多，且结果存在许多不确定因素，诊断结果在不同医生之间常常差别很大，假阴性和假阳性时有发生[5]。随着现代科学技术的蓬勃壮大，计算机技术迅猛飞升，因此利用计算机来进行医疗方面的图像分析与辅助诊断成为可能。为此，计算机辅助诊断技术掀起了新一轮的研究热潮，被引入来研究宫颈癌组织病理学，帮助医生提高看病效率和准确性。而关于计算机辅助诊断方法应用于宫颈癌方面的研究是在近几年才出现的，仍然是一个新兴领域。当宫颈癌细胞发生癌变时，细胞的外形会发生改变，细胞核会随着病变程度的增加有所增大。随着细胞形态的变化，在组织中的分布和拓扑结构也会随之改变[6]。这一特性有助于区分正常组织和癌变组织，因此本研究通过对其拓扑结构进行研究，利用图论的方法找到病灶区域帮助医生诊断。

1.2 国内外研究现状

1.2.1 医学层面相关研究现状

随着科学技术的发展，应用于病理图像的成像设备得到进一步完善，高精度的数码显微镜的发明使得病理图像诊断有了新的出路。医生不再需要人工细胞测量，机器可以自动完成扫描病理切片等功能[7]。临床医疗也取得进一步发展，毛细式液基细胞

薄层染色、叶酸受体介导染色、巴氏涂片染色等不同种类的染色方法均已成功运用于宫颈癌的相关检测中[8-10]。而近年来，研究人员发现人类基因中的 p16 基因的第 1 个外显子与宫颈癌有着密切关系。P16 的基因表达产物随着宫颈癌癌变的程度而异常增高[11]。而 p16 对苏木精-伊红染色法的染色效用起到一定辅助效果[12]。

1.2.2 技术层面相关研究现状

随着信息化技术的提高，计算机系统应用已经渗透到各个行业之中，医院信息化的发展使得效率偏低的肉眼镜下观察逐渐被计算机智能识别所取代。

首先，要想利用计算机识别不同程度癌变的宫颈细胞，就必须要利用到不同细胞的特征。目前用于癌变细胞的特征提取方法可分为基于颜色、基于纹理和基于形状的特征提取方法[13]。颜色集通过转换图像颜色空间来实现特征提取[14]。颜色直方图则是通过对选择的整张图像进行统计来提取特征[15]。这两种都是较多应用于病理图像的基于颜色的特征提取方法。模型法和信号处理法则用于纹理特征提取，是病理图像常用的特征提取方法[16]。形状特征提取方法则有两类，一是利用所选区域的边缘特征提取，二是利用整个所选区域来提取。

在方法选择上面，人工神经网络（Artificial Neural Network）、*k*-均值（*k*-means）聚类、向量机（Support Vector Machine）都是较为主流的研究方向。Royan Dawud Aldian 等人将人工神经网络和学习矢量化相结合实现了对癌变细胞的分类检测[17]。而 Aabha S.Phatak 则是将 ANN 和 SVM 相融合来识别宫颈癌病理图像[18]。在国内，董建军、马瑾等也对人工神经网络在宫颈癌检测和识别方面做出相应研究[19,20]。鲁武警则是提供了一种支持向量机和 snake 分割的宫颈癌细胞识别的方法[21]。李文杰分别构造支持向量机（SVM）、*k*-近邻（*k*-Nearest Neighbor）、人工神经网络（ANN）三个单分类器，利用模糊积分对其进行融合[22]。Rahmadwati 等利用 *k*-means 聚类和颜色分割算法对宫颈癌细胞进行分类[23]。赵英红、洪雅玲等利用 *k*-means 聚类提取细胞边缘，并与种子区域生长算法相结合，成功分割出癌变细胞核和细胞浆[24]。关涛等利用图像平均灰度作为特征参数利用 *k*-means 聚类得出一种自适应阈值分割的算法[25]。余婷婷提出了一种水平集分割算法，首先通过 *k*-means 聚类找出宫颈区域所在位置，再进一步实现对多生暗区的分割[26]。

图论是近年来新的研究热点，该方法涉及到了大量的数学几何知识[27]。这其中包括 Normalized Cut、最小生成树（Minimum Spanning Tree）以及基于主集的方法[28-30]。

而将图论与 *k*-means 聚类相结合的方法应用于宫颈癌病理图像的研究目前尚未有研究结果发表出来。因此，本文提出一种基于图论的研究方法，利用图来寻找点的关系并利用其特征提取不同组织的结构信息，然后用聚类的方法对点进行划分。

1.3 研究意义与贡献

宫颈癌是危害女性身体健康与宝贵生命最恶劣的恶性肿瘤之一，且是最为常见的癌症之一。患者平均年龄在 52.2 岁，然而在 30-35 岁患病的女性也占据着相当大的比例。面对如此高的患癌率，早期检测宫颈癌有助于更好的对患者的治疗进行规划。而如果能够及时发现及早治疗，宫颈癌病人的 5 年存活率将提高近 50%[31]。由于宫颈癌发病率虽然很高，但是治愈率和存活率也很高的特性，值得医生和医院投入更多精力开展治疗，具有较好的研究意义。

组织病理学图像的分析是诊断宫颈癌的重要手段，然而每张组织影像内包含的细胞又多又密，仅仅靠医生肉眼进行癌细胞与正常细胞间的分辨不仅任务繁重，还非常容易出错。没有经验的医生很有可能诊断失误。因此对组织病理学影像进行自动检测和分析并对医生起到辅助诊断，非常具有研究意义。

当女性宫颈发生癌变时，其细胞形态发生变化，拓扑结构也发生变化，因此可以将拓扑结构的改变作为宫颈癌变的判别标准。而目前上没有相关研究利用图论和 *k*-means 聚类对宫颈癌细胞进行深入研究。本研究的主要创新点与贡献在于将最小生成树算法和 *k*-means 聚类应用于宫颈癌研究，为医生提供辅助诊断依据，为病理医生学习提供丰富的学习数据库，并提供手动修正的功能给医生，医生可以用鼠标对图像进行再标注。

1.4 研究内容与主要工作

本文对计算机辅助诊断在宫颈癌病理组织图像应用展开了深入研究。首先对病理图像进行颜色特征提取，再利用 *k*-means 第一次对图像进行处理，然后利用骨架化的方法生成节点，近似为宫颈细胞核的分布。随后，基于所生成的节点构造最小生成树并提取几何特征。然后，再次应用 *k*-means 算法聚类。最后，提供手动修正的功能给医生，医生可以用鼠标对图像进行标注。具体研究内容如图 1.1 流程图所示：

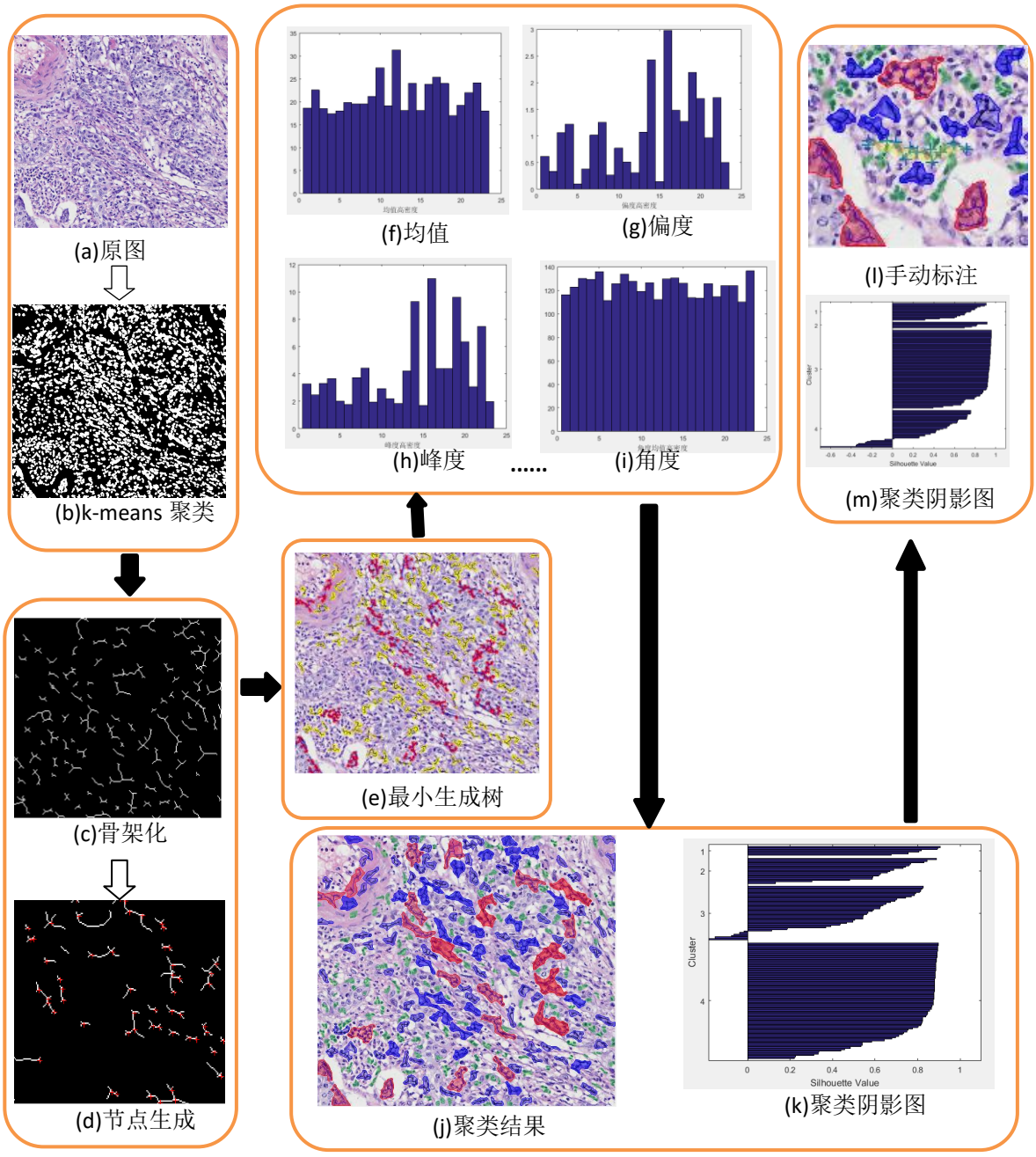


图 1.1 研究流程示意图

(a)为原始宫颈癌病理图像。(b)为提取颜色特征利用 *k*-means 做出第一次聚类的结果。(c)为在(b)基础上骨架化的结果。(d)是在骨架化的基础之上生成节点。细胞核的分布可用生成的骨架节点来表示。(e)是利用所生成节点构建的最小生成树图像，我们利用最小生成树提取各节点之间的邻近关系和空间排列。(f)、(g)、(h)、(i)为计算好的一些图的统计值，并将这些统计值作为图的特征，描述图像的拓扑空间结构。(j)、(k)利用图形特征，再次使用 *k*-means 聚类，得到了更明显的结果。(l)、(m)手动标注细胞核形成新的区域，并将其归为已知聚类中。

1.5 论文结构介绍

本文共分五章，结构框架介绍如下：

第 1 章：绪论。该章节主要介绍了宫颈癌症的研究意义，宫颈癌组织病理学相关的国内外研究现状，并简要描述了本文工作的主要研究内容，最后梳理脉络，对各章节做出安排。

第 2 章：基础知识。该章节主要介绍了研究所用到的主要的图像处理和分类的基本算法。

第 3 章：基于最小生成树的聚类研究。该章节主要阐述了本研究的核心内容和方法。本研究提出了一种针对宫颈癌病理图像的聚类方法。

第 4 章：实验及实验结果分析。该章节主要列出了不同图论方法对宫颈癌病理学图像聚类效果的比较，并且列出了实验结果。

第 5 章：总结与展望。该章节对本文的主要工作作出了总结，并对下一步工作方向进行展望。

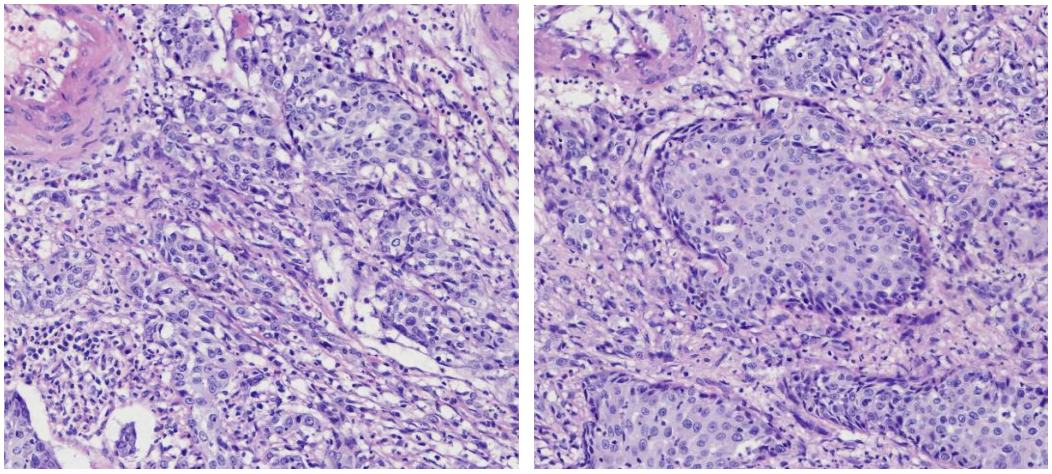
2 图论及机器学习基础知识

本章节将着重介绍宫颈癌、特征提取、机器学习、图论等理论方面的基础知识，方便读者理解文章内容。

2.1 宫颈癌组织病理学图像

在肿瘤的研究中，组织病理学的诊断是非常重要的。医生利用显微镜观察宫颈癌肿瘤切片通过分析细胞分布来进行诊断。对细胞进行染色处理有助于医生观察细胞。目前在宫颈癌细胞中主要应用的染色方法有以下三种：

HE 染色又叫做苏木精-伊红染色，苏木精可以将易于和碱性物质作用的物质蓝紫色。伊红则作用于嗜酸性物质，将其染成粉红色。即细胞核会被染为蓝紫色，细胞浆则是紫红色。巴氏染色用到的染料为苏木精和橘黄。橘黄能够渗透至小分子结构中去，适用于细胞浆。瑞氏染色是较为简单的染色方法，用到了酸性的伊红和碱性的美蓝，经相互作用氧化后变为一种中性染料[32]。



(a) (b)

图 2.1 宫颈癌组织病理学图像示意图

本研究所用的宫颈癌病理学图像像素为 976×881，染色方法为苏木精-伊红染色。可以看到图像中被染为紫色的细胞核呈现较为理想的染色效果。同时，在本研究中在第二次聚类中将黏连细胞分为高、中、低三种密度，并标注不同颜色。

2.2 特征提取

图像特征提取是计算机识别图像的重要步骤，特征提取的效果和最终图像识别的效果息息相关。如何从原图像中提取最能描述图像的特征是图像处理的研究热点。

2.2.1 颜色特征

颜色特征是较为稳定的特征，图像自身的方向和角度还有尺寸对颜色特征的影响较小。以下简要介绍三种主要的颜色特征的表示方法 [33]。

（1）颜色直方图

由 Swam 和 Ballard 最早应用的颜色直方图是应用较为广泛的颜色特征。其需要将空间划分为多个块，通过记录每个块中的像素点得到颜色直方图[34]。它所统计的是各个颜色在影像中所占的比例，是一种全局分布描述方法。这就使得它比较适用于忽略空间位置而难以自动分割的图像。

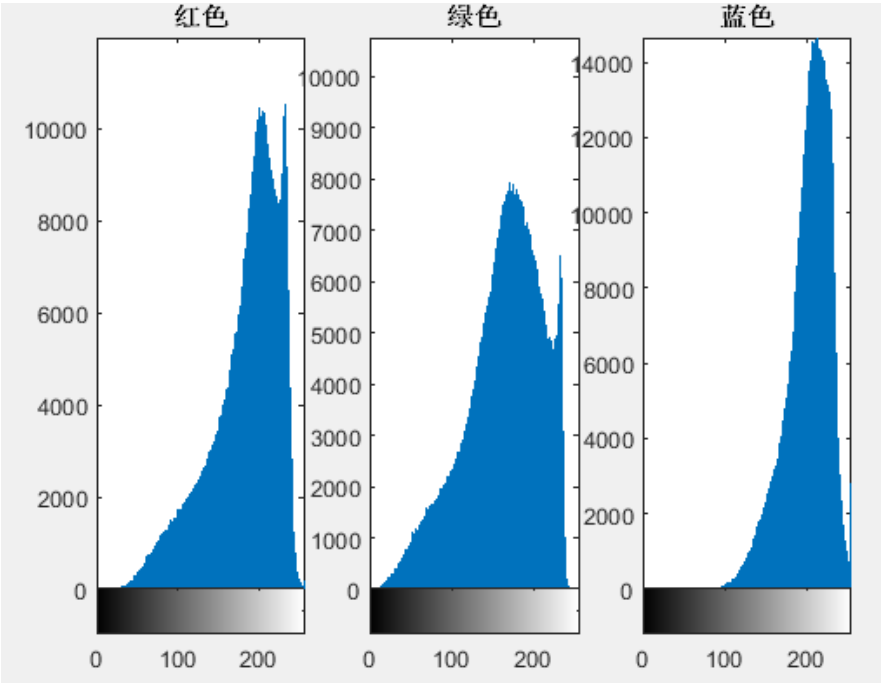


图 2.2 病理图像颜色直方图

设 $S(x_i)$ 为图像 I 的在某一特征值 x_i 的像素个数，对 $S(x_i)$ 做归一化处理得:

$$h(x_i)=\frac{S(x_i)}{\sum_i S(x_i)} \tag{2.1}$$

设 n 为某一 特征取值的个数，则图像 I 的颜色直方图表示为:

$$H(I)=[h(x_1),h(x_2),...,h(x_n)] \tag{2.2}$$

（2）颜色集

鉴于颜色直方图是一种全局特征，对局部颜色信息描述不够准确，Smith 和 Chang 提出的一种新的描述方法[35]。颜色集用来描述图像颜色特征的方法规避掉了一部分颜色直方图的缺点。并且这种描述方法方便使用二分树进行检索，能够更快速的查找图像。

颜色集方法需要转换颜色空间，将图像从常用的 RGB 空间转换为 HSV 空间，之后，利用不同颜色区域进行图像分割。分割为数个小块之后，每个小空间用各自的某个分量来表示，从而建立一个索引[36]。

（3）颜色矩

颜色矩起源于概率密度矩阵。最早是由 Stricker 和 Orengo 提出的。它是一种简单有效的表示方法[37]。图像的信息分布可以转换为多阶矩来描述，不仅仅是颜色信息，多阶矩中还蕴藏着更多图像的信息。

但是事实上，颜色信息主要存在低阶矩中，所以颜色矩主要用到的主要是前三阶矩。其优点在于不需要像颜色直方图一样，需要统计空间内各个像素值，同时特征向量的维数也较低。

设图像共有 *N* 个像素数量，颜色矩数学定义如下：

一阶矩（均值，mean），用来衡量颜色的平均强度，计算方法如下：

$$\mu_i=\frac{1}{N}\sum_{j=1}^{N}p_{i,j} \tag{2.3}$$

二阶矩（方差，variance），用来衡量图像颜色分布的均匀性，计算方法如下：

$$\sigma_i=\left(\frac{1}{N}\sum_{j=1}^{N}(p_{i,j}-\mu_i)^2\right)^{\frac{1}{2}} \tag{2.4}$$

三阶矩（斜度，skewness），描述的是图像总体分布的对称性和扭曲度，计算方法如下：

$$s_i=\left(\frac{1}{N}\sum_{j=1}^{N}(p_{i,j}-\mu_i)^3\right)^{\frac{1}{3}} \tag{2.5}$$

四阶矩（峰度，kurtosis），用来描述图像颜色分布形态的陡峭程度，计算方法如下：

$$\gamma_i = \left(\frac{1}{N} \sum_{j=1}^{N} (p_{i,j} - \mu_i)^4 \right)^{\frac{1}{4}} \tag{2.6}$$

其中 $p_{i,j}$ 是图像中第 j 个像素的第 i 个颜色分量。i 可以取 1,2,3，分别表示 R、G、B 颜色分量。由于颜色矩的多阶矩除颜色信息外还有更多的几何意义，所以也可以用其提取出的均值、斜度、峰度等特征作为几何特征。同时，多阶矩阵在颜色分量为 1 的灰度图像中也依旧可以使用。

本研究中首先应用了颜色直方图提取颜色特征，进行第一次聚类分离出稀疏细胞核黏连细胞核。随后在最小生成树建立时应用了四阶矩提取更多的图像信息，进行第二次聚类。因此颜色矩对于本研究不仅仅只是颜色特征提取，更可以用其高阶矩进行深层次的几何特征提取。

2.2.1 形状特征

形状特征提取主要有两种，一种是利用边缘特征，即形状的轮廓；另一种是利用整个形状的区域[38]。主要的方法有边界特征法、傅里叶描述符、形状无关矩、几何参数法。

边界特征法通过对边界的特征进行提取所需参数，其中最为经典的是 Hough 变换直线检测方法。这种方法将图形的外沿连接起来组成闭合形状，并将图像变至参数空间进行参数提取。

傅里叶描述符方法具有良好的几何不变性。最早由 Zahn 和 Roskies 提出，是一种应用广泛的基于边缘的描述方法[39]。其对图形的外沿进行变换，将图像变换到频域来提取特征。由傅里叶描述符可得出曲率函数、质心距离和复坐标函数。

几何参数法更加方便和直接的描述区域特征。它更多采用的是形状定量测度，而可以得出形状面积、周长、圆度。偏心率等几何参数[40]。需要注意的是，形状参数的提取需要对图像做出预处理，即进行图像处理和分割。因此几何参数法还受到图像处理效果的影响。

在本文中首先提取了各细胞的质心以及骨架，随后又应用了区域周长和区域面积的形状特征结合上文颜色特征进行第二次聚类。

2.3 机器学习理论

机器学习旨在让计算机模拟人类的学习行为，是人工智能的核心技术。按学习形式分类可分为有监督学习、无监督学习和半监督学习[41]。

2.3.1 有监督学习

有监督学习的训练集需要提前的标注，使计算机学习已知数据集的特征。通过已有的数据和对应的结果去训练最终建立起良好的模型来预测未知数据的结果[42]。通过这个模型可以将之后的输入对应到相应的输出中，也就具有了将未知数据分类的功能，从而实现分类的目的。因此它往往是让计算机去学习人为创建好的分类模型。有监督学习常用于训练神经网络和决策树。这两种技术都是比较依赖于预先的标注，这样会使得分类系统被给予足够的分类信息。对于神经网络，分类系统首先由系统提供的信息来判断网络的错误，然后不断调整参数。对于决策树，分类系统使用它来确定哪些属性提供的信息最多。有监督学习典型的算法是 KNN 和 SVM。

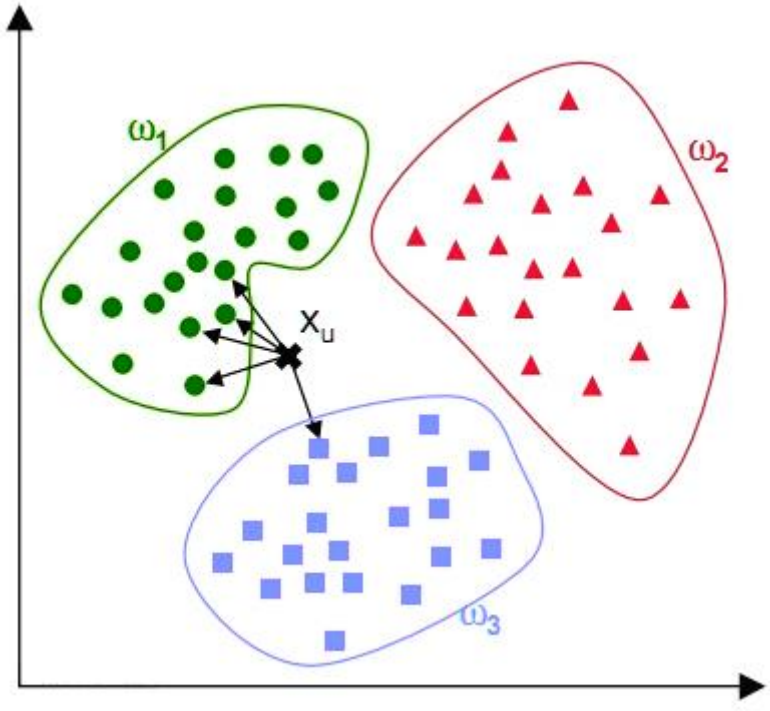


图 2.3 KNN 算法示意图[43]

所谓 *k*-临近算法（KNN），在特征空间中假设存在样本 *k*，它附近的其他样本已经做出人为标注，即已经正确分类，那么其最接近的样本大部分属于哪一类，它也属

于该类。在分类决策中，该方法只根据一个或多个最近样本的类别来确定要划分的样本的类别，因此只涉及极少数相邻样本。由于 k-临近算法这个特性，比其它方法更适合于类域重叠较多的样本集。k-临近算法的结果和 k 的选择有直接关系。同时，k-临近通过依据 k 个对象中最占优势的类别，而不是对单一对象进行决策，这也是 k-临近算法的优势所在。

2.3.2 无监督学习

无监督学习常用于关联、聚类和降维。其输入数据不需要被标记，也没有确定结果，且样本数据类别未知，这是无监督学习不同于有监督学习的地方。由于无监督学习在训练过程中的样本不需要预先标注类别，而事实上，在实际应用的过程中，不少情况下无法知道样本的标签。甚至训练样本对应的类别可能也没有，因此无监督学习比有监督学习应用方向更宽阔。常用的无监督学习算法主要有主成分分析法，等距映射、局部线性嵌入等[44]。

聚类是其中比较典型的例子，用于把相似的东西聚在一起，且并不需要关心这一类是什么。聚类算法最主要的是划分方法和层次方法两种。典型的划分方法有 k-means 算法、k-medoids 算法、CLARANS 算法。划分聚类算法需要把数据集分割为 k 个部分，并且需要 k 作为参数输入进去。典型的层次聚类方法是 BIRCH 算法、DBSCAN 算法。其中，由于 k-means 收敛速度快，易于实现，常得到广泛应用[45]。

2.3.3 k-means 算法

k-means 算法最先由 Macqueen 提出。这种方法会在所有样本中选取若干个点作为初始的聚类中心，而这种选取过程是随机的。然后该方法会计算其余的点与初始选取的点的距离，并把每个样本分配给距离最近的聚类中心[46]。每一次聚类时，都会根据所有聚类中的点重新计算，更新更为合适的聚类中心，而所有划分来的样本和聚类中心组成一个聚类。

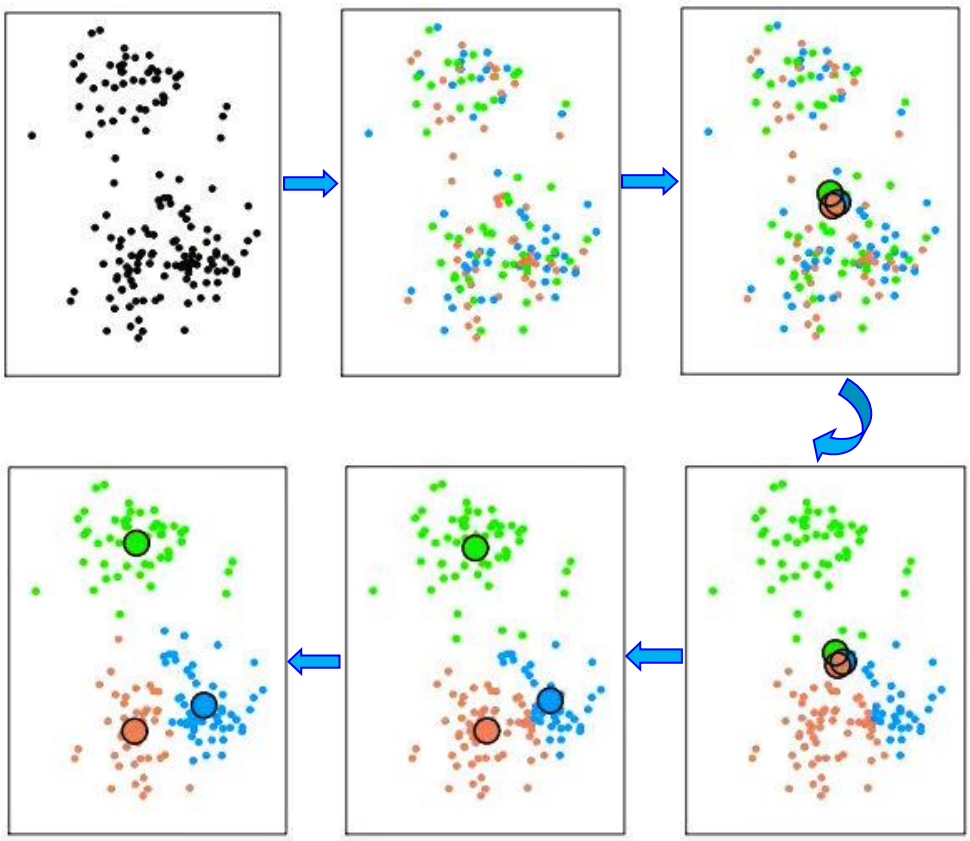


图 2.4 *k*-means 算法示意图[47]

表 2.1 *k*-means 算法

***k*-means** 算法
输入：
***k*:**簇的数目
D:包含 m 个对象的数据集
输出：**C**：***k*** 个簇的集合
方法：**1**. 从数据集 $D=\{x_1,x_2,...,x_m\}$ 中随机选择 *k* 个样本作为初始簇中心，其质心向量为 $\mu_1,\mu_2,...,\mu_k$
2. 对于 i=1, 2, . . . , m，计算样本 $\mathbf{x}_i$ 和各质心向量 μ_j (j=1, 2, . . . , *k*)的欧几里得距离 $d_{ij}=\|x_i-\mu_j\|_2^2$，将求出的各 d_{ij} 进行比较，并给 $\mathbf{x}_i$ 标记其中最小的 d_{ij} 所对应的类别 λ_i，分配到最相似的簇。
3. 此时重新计算簇的集合 $C_{\lambda_i}=C_{\lambda_i}\cup\{x_i\}$
4. 重新计算新的质心向量直到不再发生变化
5. 输出簇 $C=\{C_1,C_2,...,C_k\}$

在本研究中，第一次聚类的结果的细胞核被分为两类：一类是稀疏的细胞核，另一类是黏连细胞核，这种聚类处理可以用来区分不同类型的组织结构。由于稀疏细胞在第一次聚类中已经被成功区分出，下一阶段的聚类重点主要放在黏连细胞核之上。通过第二次聚类得到更为详细的结果，可以更好的预测组织患癌症风险。

2.4 图论

图论是近年来新的研究热点，涉及了大量的几何理论知识，是数学的一个分支。它以图为研究对象来研究图的性质，图是多个节点 V 和连接节点的边 E 的集合，可以表示为 $G=(V,E)$ 的有序对[48]。换句话说，图由若干给定的点和连接这两个点的线条组成，这些线条通常用来描述某些事物之间的特殊关系，事物以一个点表示，并以连接这两个点的线条表示之间的关系。关于图的几个概念定义如下：

无向图：一个图中所有边都是无向的，则称为无向图。

连通图：在无向图中，任意两个顶点都有路径相通的图。

连通网：在连通图中，图的边具有一定的意义，每一条边对应着一个权值。

图论的算法提供了一种简单和系统的模型，可以去除特定边缘，并将图分成若干子图进行图像分割，在计算机领域中起着重要作用[49]。许多问题都可以转化为图论问题，再通过算法加以解决。Normalized cut 方法、基于主集的方法以及最小生成树（MST）都是常用的图论方法。

生成树是是一种边的数量比顶点数量恰巧少 1 的连通图。假设其顶点数为 m 则其只有 $m-1$ 条边。而最小生成树是所有生成树中，所有边的代价和最小的生成树。

最小生成树问题是图论中最常见的网络拓扑问题。其往往涉及到识别单个核心的空间位置，然后将图构造为一组连接的核心节点，这样就可以描述图像的拓扑结构或空间信息[50]。此外，可以从这些单个细胞核相对彼此的密度和空间排列得到更多关于图信息的描述[51]。Graham 和 Hell 对 MST 的各个方面进行了全面的分析，对于稠密图应用普里姆(Prim)算法较为合适，其时间复杂度为 $O(n^2)$；而对于稀疏图应用克鲁斯卡尔算法(Kruskal)更为合适，其时间复杂度为 $O(e\,loge)$，其中 n 和 e 分别是图的顶点数和边数[52-54]。

表 2.2 Prim 算法

Prim 算法
输入：连通网 ***G***，其中顶点集合 ***V***，边集合 ***E***
方法：**1**. 初始化：$V_{new}=\{x\}$，其中 x 为集合 V 中的任一起始点 $E_{new}=\{\ \}$为空 **2**. 在集合 ***E*** 中选取权值最小的边<*u*, *v*>，其中节点 $u\in V_{new}$，而节点 $v\notin V_{new}$ 且 $v\in V$，如果有多条相同权值的边则任选其一 **3**. 将 *v* 加入集合 V_{new} 中，并将<*u*, *v*>边加入集合 E_{new} 中 **4.**重复 2, 3 操作直到 $V_{new}=V$
输出：用集合 V_{new} 和 E_{new} 来描述所得到的最小生成树

表 2.3 Kruskal 算法

Kruskal 算法
输入：连通网 ***G***，其中顶点集合 ***V***，边集合 ***E***
方法：**1**. 初始化：G_{new}，拥有原图 G 中相同的节点，但没有边 $L_{new}=\{\ \}$为图 G_{new} 中节点的连通分量 **2**. 在集合 **E** 中选取权值最小的边<*u*, *v*>，其中节点 u, v 不在图 G_{new} 的同一个连通分量中 **3**. 将<*u*,*v*>作为新的连通分量加入集合 L_{new} 中，更新 L_{new} 中的连通分量 **4.**重复 2, 3 操作直到图 G_{new} 中所有节点都在同一个连通分量里
输出：由 G_{new} 得到的最小生成树

通常生成图的方法分为全局和局部两种类型，全局图有 Voronoi 图、Delaunay 三角网和最小生成树。全局图倾向于研究图像中所有节点的结构。而局部图，例如单元簇图和最近邻图，倾向于观察局部邻域中的节点结构。本研究在 Cruz-Roa 等人对于 Delaunay 三角剖分的最小连通子图进行了研究，其包含了原始图中的所有节点，权重之和最小[55]。本研究在其研究基础之上，选择了最小生成树方法并应用了 Prim 算法。

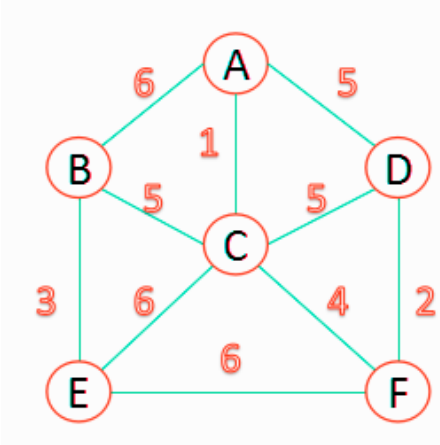


图 2.5 连通网示意图

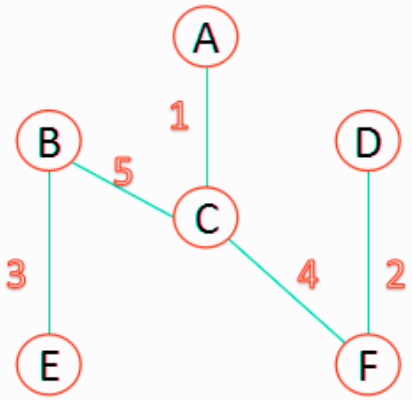


图 2.6 MST 示意图

2.5 鼠标取点

本研究的开发工具为 MATLAB 软件，其中 MATLAB 中有自带的函数可以调用操作面板利用鼠标进行标记，并记录下标记坐标。常用的鼠标标记函数主要有 ginput() 和 getpts()。考虑到宫颈癌病理图像细胞稠密，本研究选择了标记更为精确的 ginput() 函数进行标记，帮助医生手动修正图像。

2.6 本章小结

本章主要介绍了关于图论、机器学习、特征提取等实验中所用到的相关基础知识，以便于读者更好的理解本文所做的工作。在下一章节中本文将具体阐述实验所用到的方法。

3　基于最小生成树的病理图像聚类方法

本文的主要研究内容为基于最小生成树的病理图像聚类方法，将宫颈癌组织病理图像中不同密度的组织结构通过计算机智能标注，为医生快速定位细胞异常部位，辅助医生诊断和分析。

3.1　预处理

在特征提取之前，本研究首先尝试使用了大津阈值分割、*k*-means 图像分割方法对图像进行预处理。由于经过分割后的图像效果并不是十分理想，图像信息损失较大，对于之后的研究在特征提取方面提取更多有意义的信息有一定的影响。因此本文仅将病理图像的预处理作为一种尝试。

3.1.1　大津阈值分割

大津阈值分割通过阈值将图像分为两部分，当阈值达到最佳时，前景与后景的差别达到最大。在大津阈值算法中衡量差别的标准是最大类间方差。然而由于该方法以类间方差为主要判别，对噪声和图像都比较敏感，得到如图 3.2 结果：

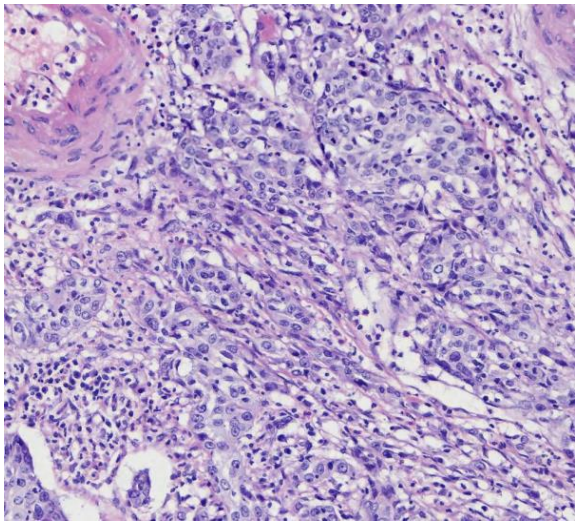


图 3.1 宫颈癌病理图像原图

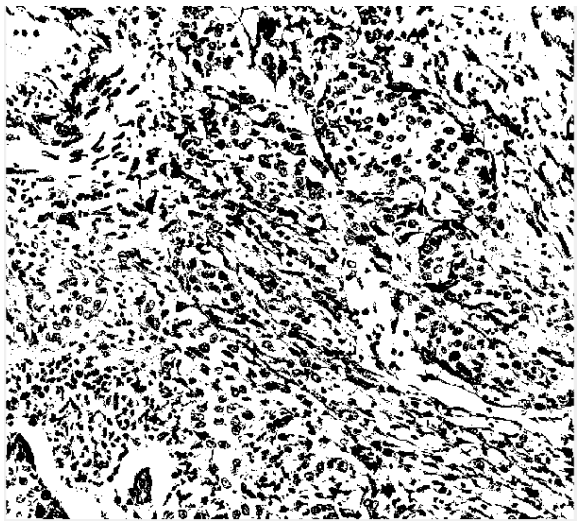
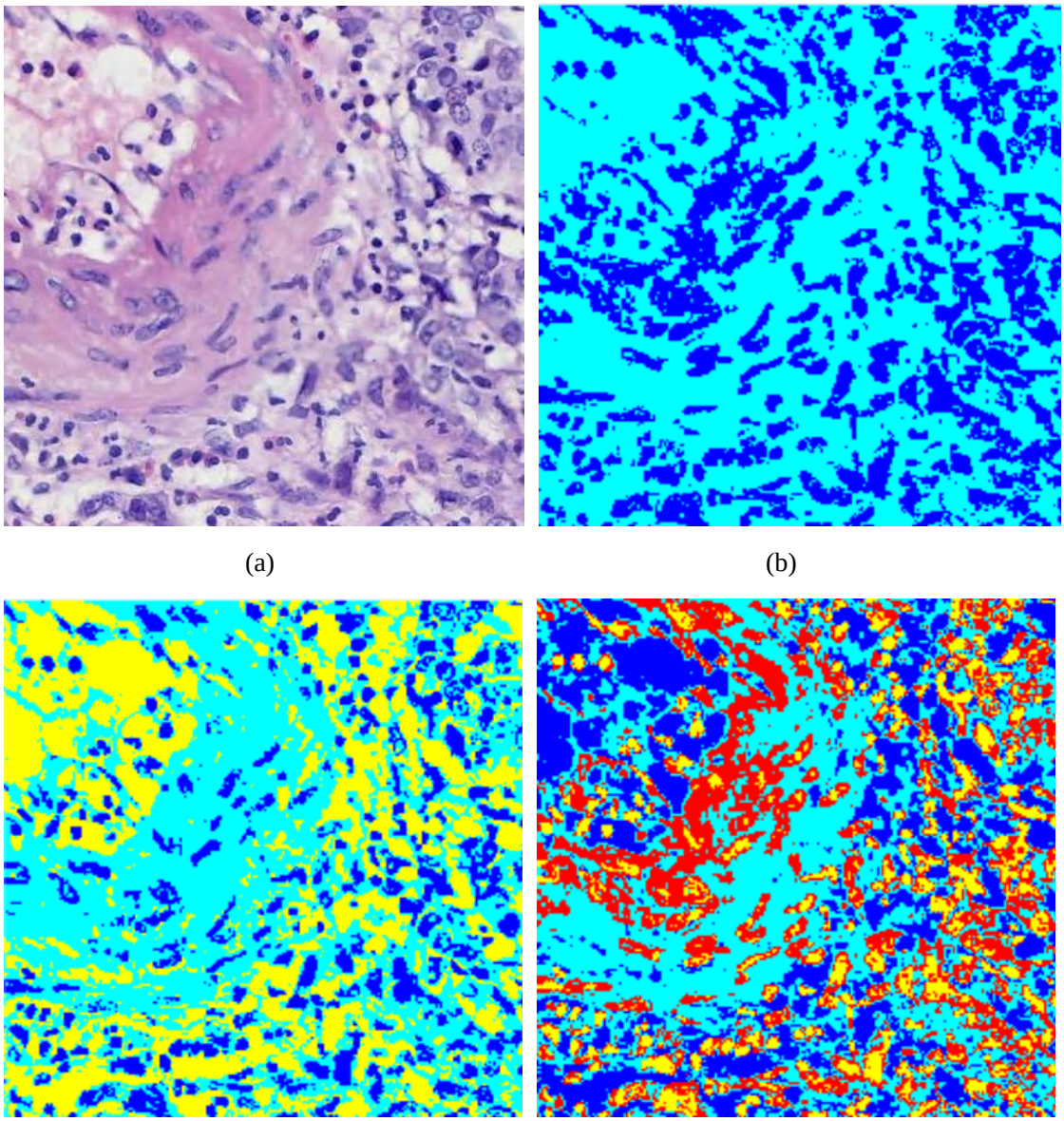


图 3.2 大津阈值分割后结果

可以看出大津阈值在图像分割之后对细胞核的分割效果并不是十分友好，并且丧失了很多组织结构信息。

3.1.2 *k*-means 图像分割

在 *k*-means 进行图像分割时，RGB 色彩空间的图像中的每个像素点都被当做是一个三维向量。在 *k*-means 聚类时，聚类的类数 *k* 和最大迭代次数 N 都对图像分割结果有一定影响。本文在实验中尝试了 *k*=2, *k*=3, *k*=4 等不同簇数，最大迭代次数为 50 次，为了使结果看起来更加明显，本文将不同簇的像素用不同的颜色表示，得到如图 3.3 所示结果：



(a)　　(b)

(c)　　(d)

图 3.3 *k*-means 在 *k*=2,3,4 时的分割结果

本研究所用显微病理图像由苏木精-伊红染色，细胞核呈蓝紫色，细胞浆呈现粉红色，而不含有蛋白质和核酸的组织液则显示出白色。由于本研究倾向于通过细胞核的结构分布来提取特征检测病灶点，所以并不需要将细胞浆和组织液分割出来。由图像可以看出在 k=2 时聚类效果很好，而 k=3 时组织液的结构分布也被分割出来，而 k=4 时的结构信息比较混乱，受噪声干扰较大。经更改簇数和最大迭代数得出的多个结果进行比较，最终本文选择簇数 k=2 的图像分割结果，并进行简单的形态学处理，转化为二值图像得到图 3.4 所示结果：

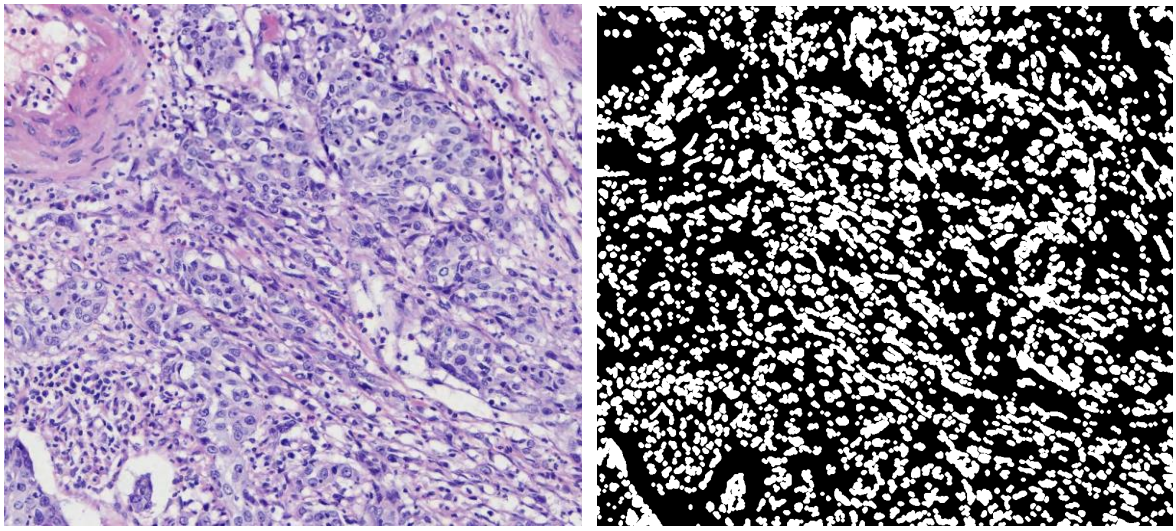


图 3.4 形态学处理后的二值图像

从图 3.4 中可以看出，经过形态学处理后的二值图像分割效果更为明显，填充了在分割时细胞中出现的细小孔洞，并转化为二值图像来进行后续的骨架化提取等一系列操作。

3.2 最小生成树构建

3.2.1 颜色直方图特征提取

本文中对宫颈癌病理组织图像提取了颜色直方图，下图将 RGB 颜色空间三通道的颜色分布纳入了一张图中，得到直观的直方图结果。下图中黄色区域、浅蓝色区域、深蓝色区域分别表示 R、G、B 三颜色通道。提取的颜色直方图如图 3.5 所示：

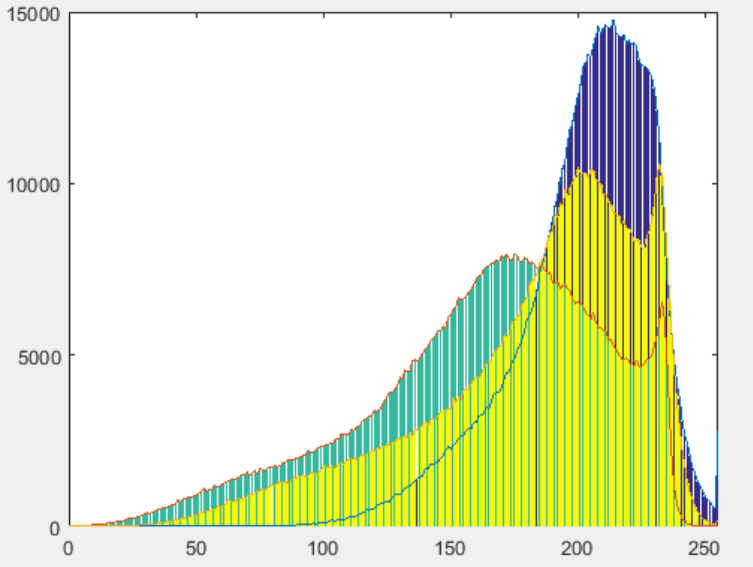


图 3.5 颜色直方图提取结果

通过提取 RGB 像素值作为颜色特征，设置簇数 k=2，应用 k-means 算法将宫颈癌组织病理学图像分为两组，第一组为稀疏的细胞，他们可以很好的聚类。第二组为黏连细胞，这些细胞重叠在一起往往难以识别。稀疏细胞多为形状结构正常的细胞，而黏连细胞中有很大一部分具有成为病灶点的潜力，因此本文接下来的研究重点聚焦在黏连细胞之上。

3.2.2 连通区域筛选

本节在黏连细胞中，人为的将其分为高密度、中密度、低密度三类，并通过筛选黏连细胞区域面积进行分组。设 L、M、N均是在处理好的二值图像上标记了四连通区域的矩阵。设标记区域内像素点个数为 $S(i)$

(1) 在 L 中剔除 $S(i)>500$的连通区域作为低密度区域

(2) 在 M中剔除 $S(i)<500 \| S(i)>1700$的连通区域作为中密度区域

(3) 在 N中剔除 $S(i)<1700$的连通区域作为高密度区域

3.2.3 骨架化

本节对二值图像进行骨架化处理，并保证骨架结构不断裂的基础之上进行了细化操作。

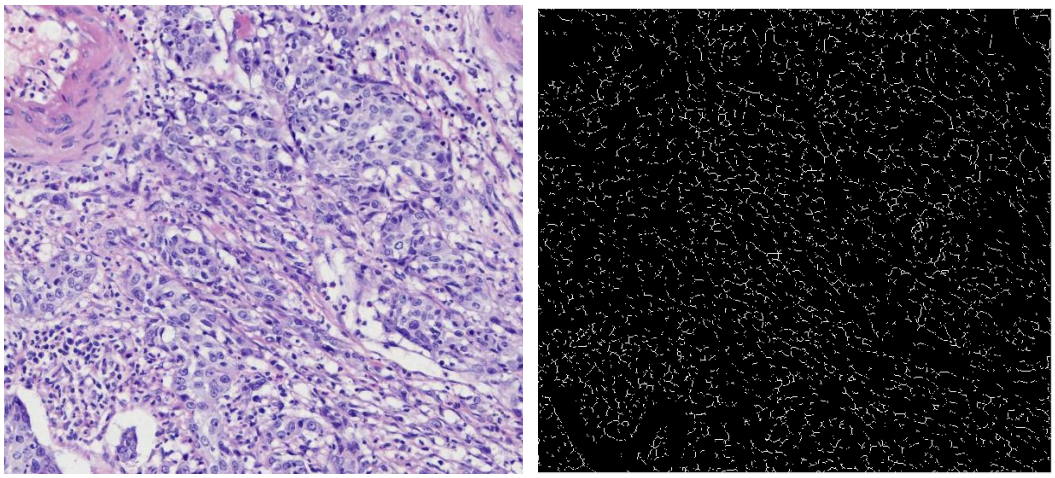
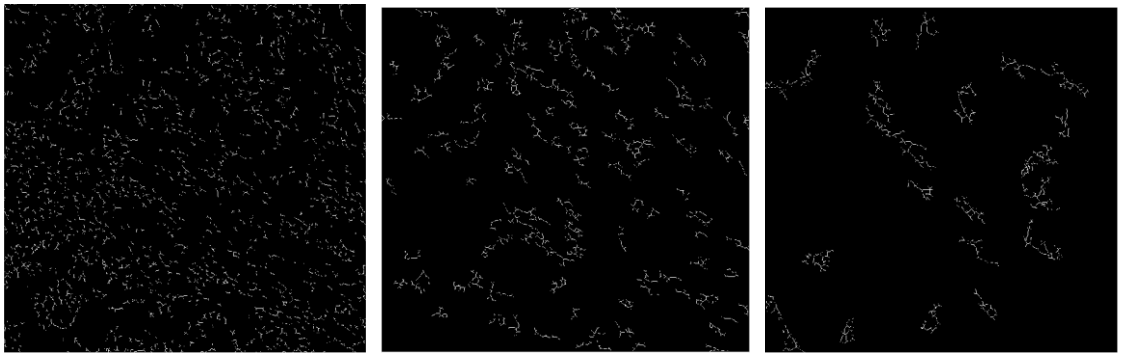


图 3.6 骨架提取结果

除此之外，本节还对 3.2.2 的三个连通区域矩阵也进行了骨架化处理。提取的细胞骨架来进行后续操作，图 3.7 三张图像分别为 L、M、N连通区域矩阵所提取的细胞骨架：



(a) **(b)** **(c)**

图 3.7 剔除部分结构的骨架提取结果

3.2.4 节点生成

将提取的骨架化图像附在病理学细胞的二值图像上，由图 3.8 可观察到其分叉点与细胞核位置近似，可以用来代替细胞核的位置，而且分叉点具有一定的特点可以进行提取。

本文在已经提取好的骨架中遍历，寻找其八邻域中也为骨架中的点，当其八邻域中有超过三个点也为骨架中的点，则这个点必然是一个分叉点，即为节点。图 3.9 为节点生成的结果，红色标注为节点所在位置。

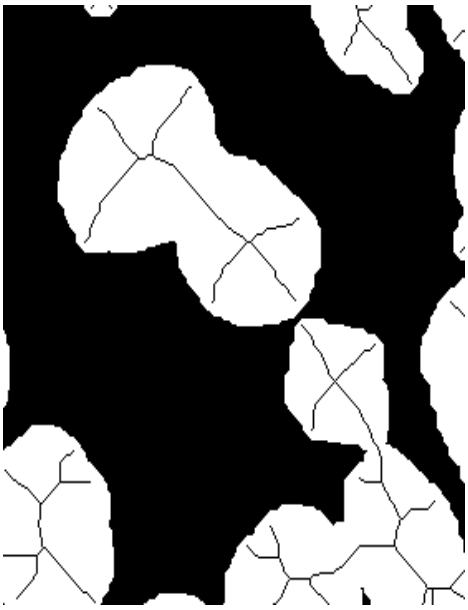


图 3.8 骨架化结构示意图

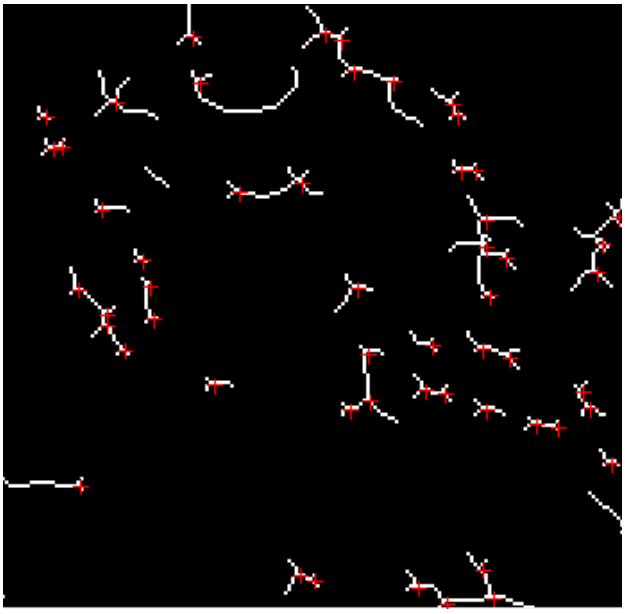


图 3.9 节点生成结果图

同时本文在寻找到节点之后对节点位置进行了计算，删除了部分多余节点，排除了相邻分叉点的干扰，确保了一个节点与一个细胞核一一对应的关系。

3.2.5 最小生成树构建

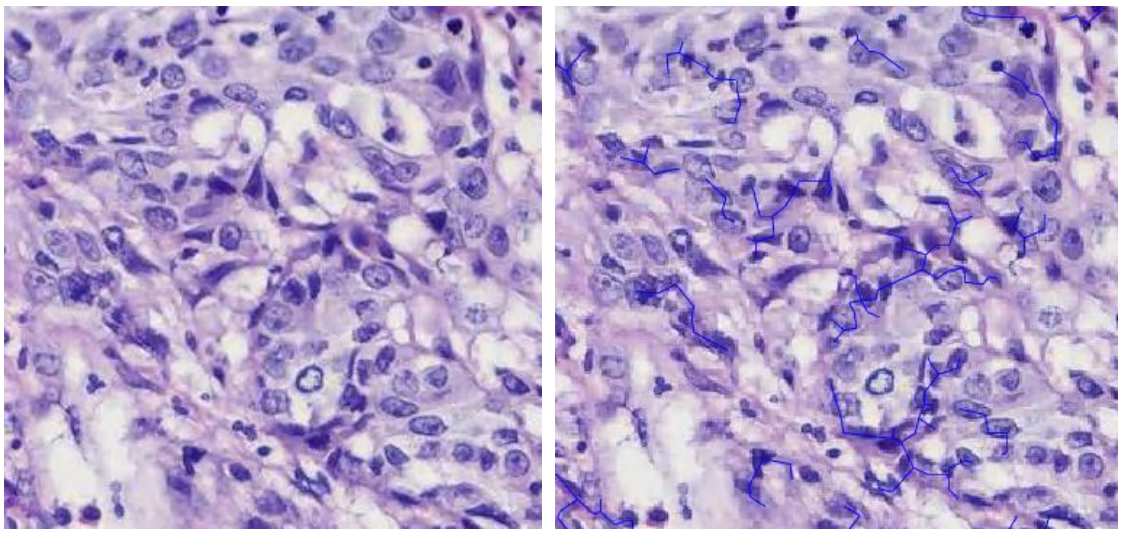


图 3.10 最小生成树生成结果图

由于筛选出的低密度黏连细胞多为少量的细胞黏连，本节中仅对中密度、高密度中的大片区域黏连进行了最小生成树的构建，图 3.10 为最小生成树构建的可视化结果。本节分别对高密度、中密度区域进行最小生成树构建，这种分离处理、分别构建的方法可以帮助区分不同类型的组织，为了使构建结果更加直观可辨，本研究最终用

两种不同颜色的连线来表示。黑色连线为中密度区域构建最小生成树，蓝色连线为高密度区域构建最小生成树。

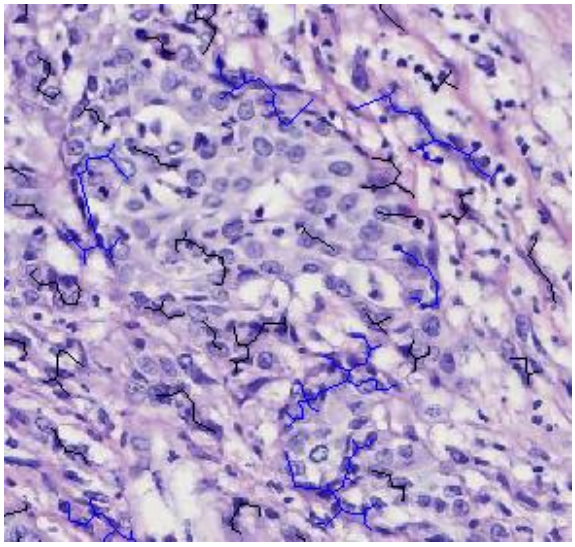


图 3.11 最小生成树生成结果图

3.3 *k*-means 聚类

3.3.1 特征值计算

本节根据最小生成树生成的图形，计算出各种统计值，并将这些统计值作为形状和几何特征来描述不同的组织。本研究提取出来的形状特征包括边缘长度的均值、方差、偏度和峰度，以及每个图形的角度。几何特征则包括组织的周长，和每个组织内节点的独立排列。

均值是一组数据集的趋势，计算方法如下：

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.1}$$

方差是测量一组数据时评判数据集分散程度的度量，计算方法如下：

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\mathrm{n}-1} \tag{3.2}$$

偏度是统计数据偏态和程度的度量，计算方法如下：

$$S_k=\frac{\mu_2}{\mu_2^{\frac{3}{2}}}=\frac{\mu_3}{\sigma^3}=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{3}{2}}} \tag{3.3}$$

其中 μ_3 是三阶中心矩，σ 是标准差。

峰度是统计数据变化陡峭的程度的度量，通常以四阶中心矩除以概率分布方差的平方再减去 3 来定义峰度，计算方法如下：

$$\gamma_2=\frac{\mu_4}{\sigma^4}-3=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}-3 \tag{3.4}$$

其中 μ_4 表示四阶中心矩。

3.3.2 *k*-means 聚类

稀疏细胞在第一次聚类中被提取出来，在黏连细胞的后续处理上，通过最小生成树提取的图形特征和集合特征，本节再次使用 *k*-means 聚类，分别为两种类型的组织设置 *k* 值，然后通过第二阶段聚类得到更为详细的结果，来预测组织的癌症风险，快速发现病灶部位。根据最小生成树所提取的图形特征所聚类出来结果，本研究簇数选择 *k*=2，将高密度和中密度分别聚类，结果如图 3.12 所示。

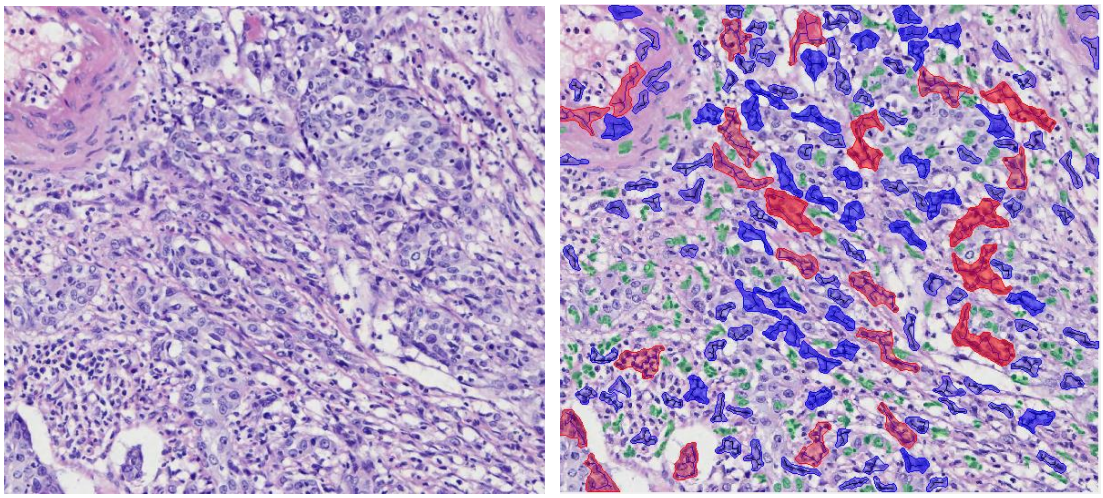


图 3.12 *k*-means 聚类结果图

3.4 手动标注

图 3.12 所示聚类结果辅助医生快速发现病灶点，同时可以来预测癌症风险。除此之外，本文还提供给医生手动标注细胞核的功能，来标记密集区域或可疑病灶点。图 3.13 所示黄色区域为手动标注区域显示结果。医生点出细胞核后，将生成一片带颜色的标记区域，同时其细胞核会用线相连接构成图形结构，以便后续提取特征，加入黏连细胞聚类之中。

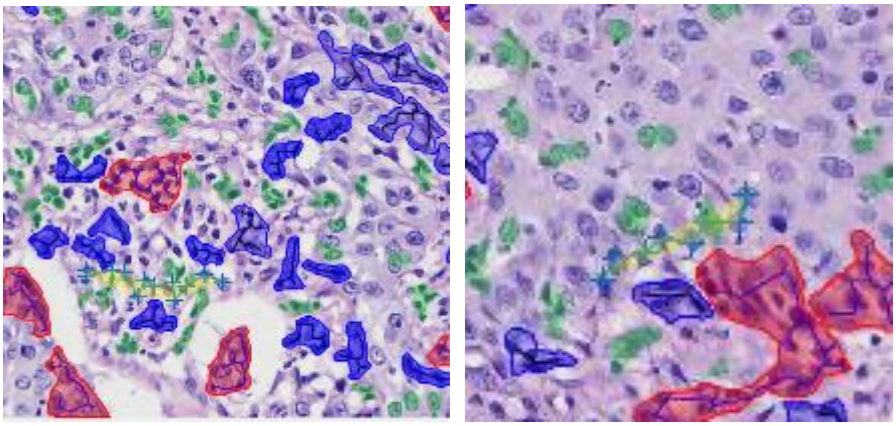


图 3.13 手动标注病灶区域

3.5 本章小结

本章中详细介绍了研究工作中所用到的一些方法，工作内容的创新点在于利用骨架化生成节点代替细胞核分布在图像中做出标注，并可以利用鼠标进行操作。在下一章节中，本文将详细讨论实验所得结果并加以分析。

4 实验结果与分析

本文在上一章中具体介绍了基于最小生成树的病理图像聚类方法，本章节将对实验所产生的结果进行总结、分析和评价。

4.1 实验设置

4.1.1 数据来源简介

本实验所用组织病理学图像为宫颈癌组织病理学图像。该宫颈癌组织切片经制备和拍摄，由辽宁肿瘤医院研究所的两位组织病理学家提供。实验中采用苏木精伊红染色法对细胞进行上色，随后人工封片，每张切片由奥林巴斯数字扫描显微镜进行图像采集。

4.1.2 数据设置

经奥林巴斯数字扫描显微镜图像采集所得图像为 10 幅放大 40 倍的图像，每张图像大小在 200M 到 2024M 之间，像素数在30000×30000到60000×60000之间。由于每张扫描图像非常大，且尺寸不一，本研究在实验开始之前将每个图像裁剪成多个子图像，并使其每张子图像大小在 1M 到 2M 之间，像素值都为976×881，以方便研究。同时，本研究在多张子图像中选取较为代表性的十张图像作为主要研究对象，来对聚类结果进行检验和评价。图 4.1 为实验中选取的十张子图像，图 4.2 为 12 张子图像进过第一阶段聚类后得到的图像。

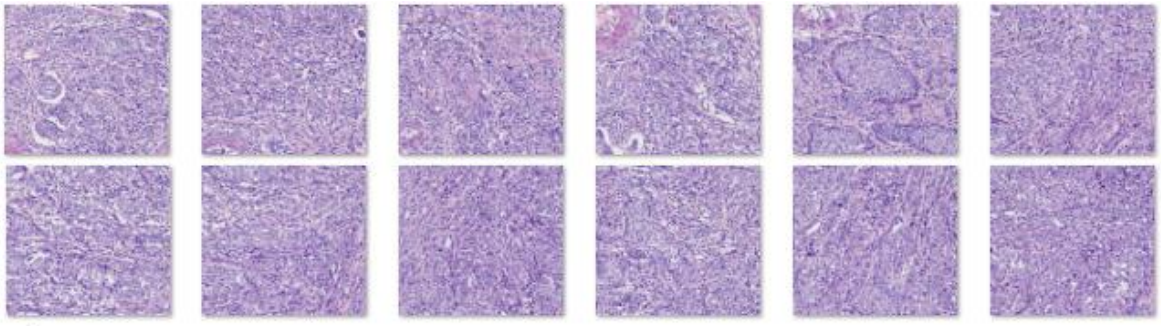


图 4.1 选取的 12 张子图像

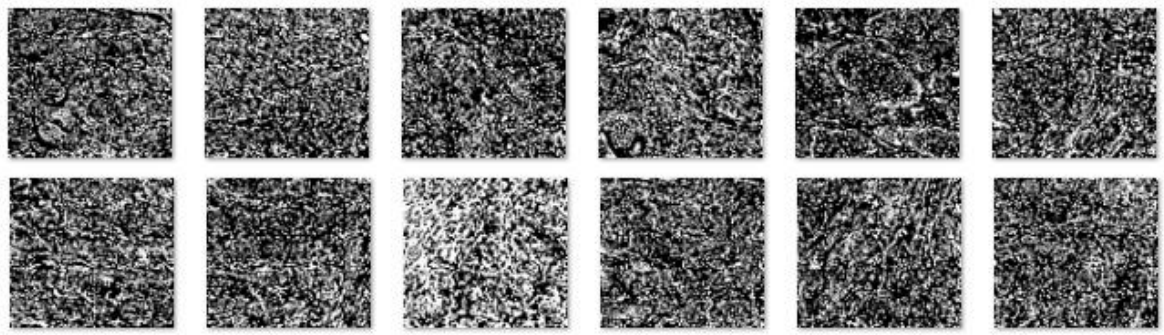


图 4.2 第一阶段聚类结果

4.1.3 实验环境

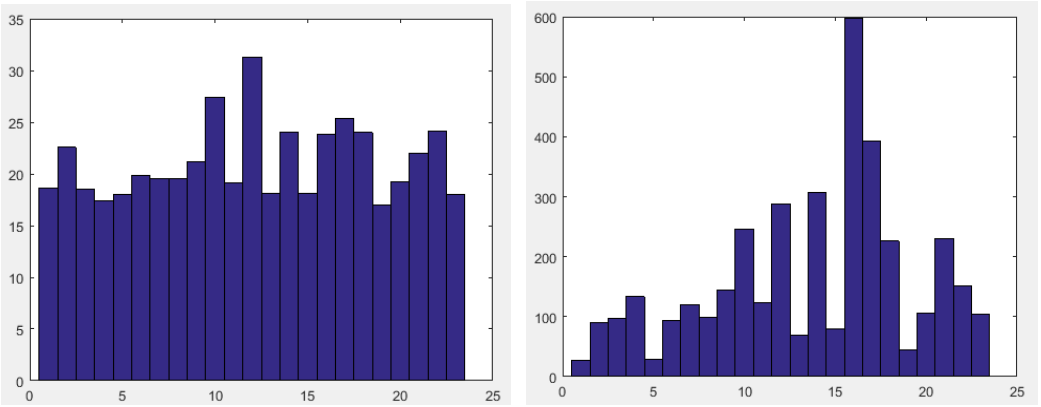
本文采用的编程软件是 MATLAB，操作系统采用是 Windows10 系统。

4.2 实验结果及分析

本节将对前文所使用的方法所产生的实验结果进行分析，并将用直观的轮廓图的可视化的阴影图来评价聚类结果。

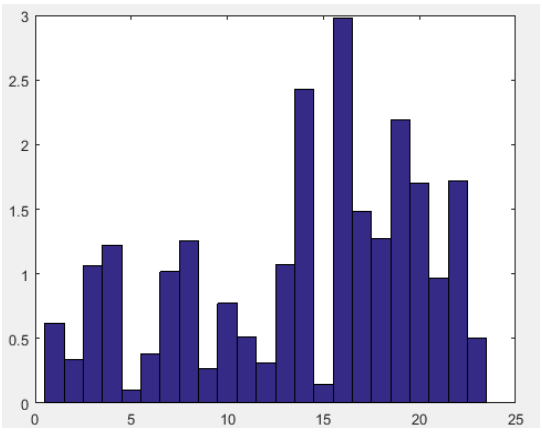
4.2.1 基于最小生成树的病理图像聚类研究结果与分析

在 3.3.1 节中介绍了图形特征提取的主要计算公式。本节中由计算公式算出统计数据，并作出可视化的结果呈现。提取的图形特征可视化结果由图 4.3-4.10 所示，从图表可看出，在同一幅宫颈癌组织病理学图像中，高密度黏连的细胞核组织结构信息更加明显，其数量比中密度细胞核更少。

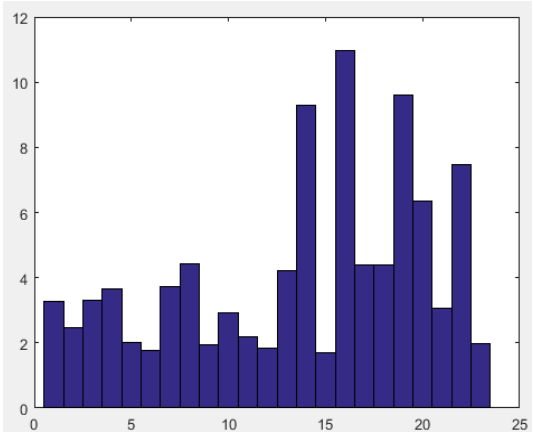


(a) (b)

图 4.3 高密度组织长度的均值和方差

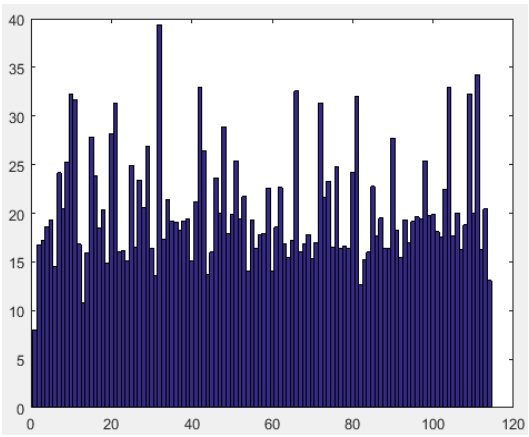


(c)

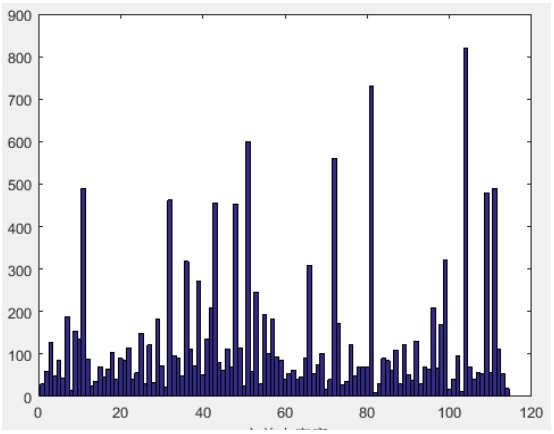


(d)

图 4.4 高密度组织长度的偏度和峰度

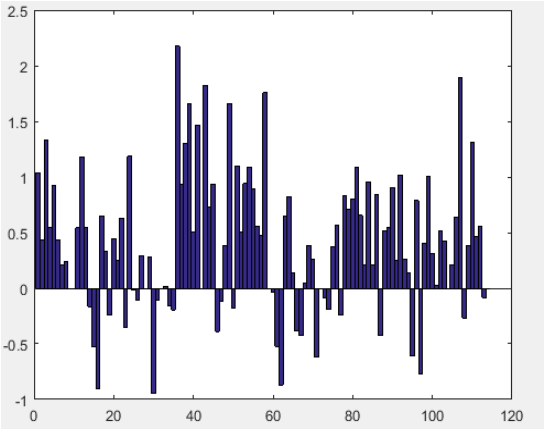


(e)

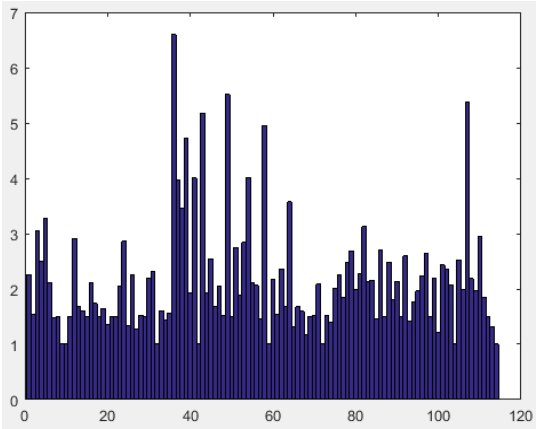


(f)

图 4.5 中密度组织长度的均值和方差

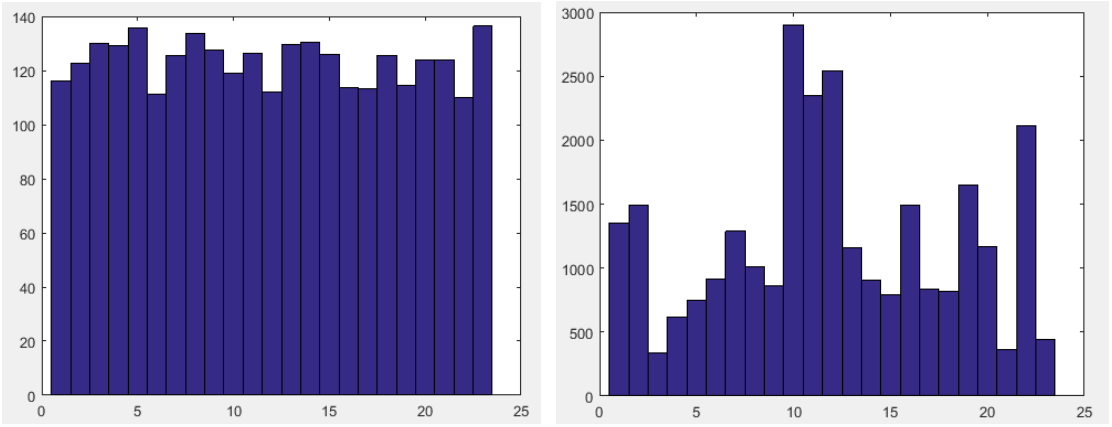


(g)



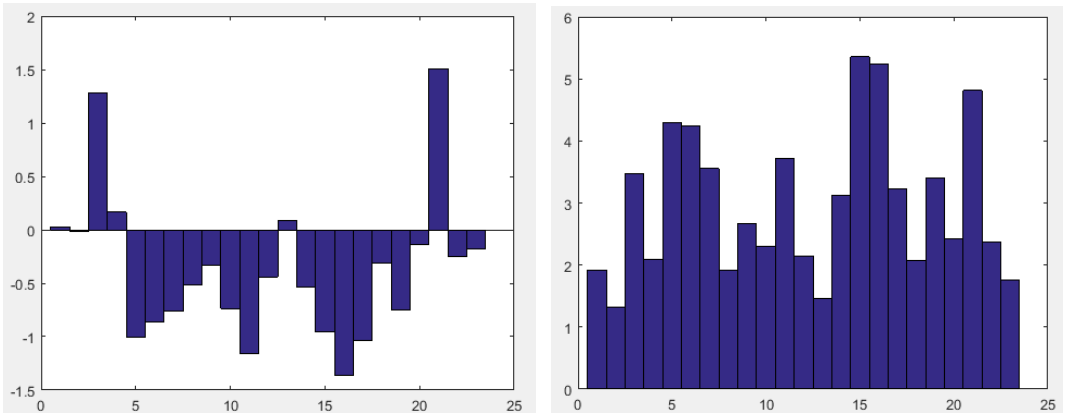
(h)

图 4.6 中密度组织长度的偏度和峰度



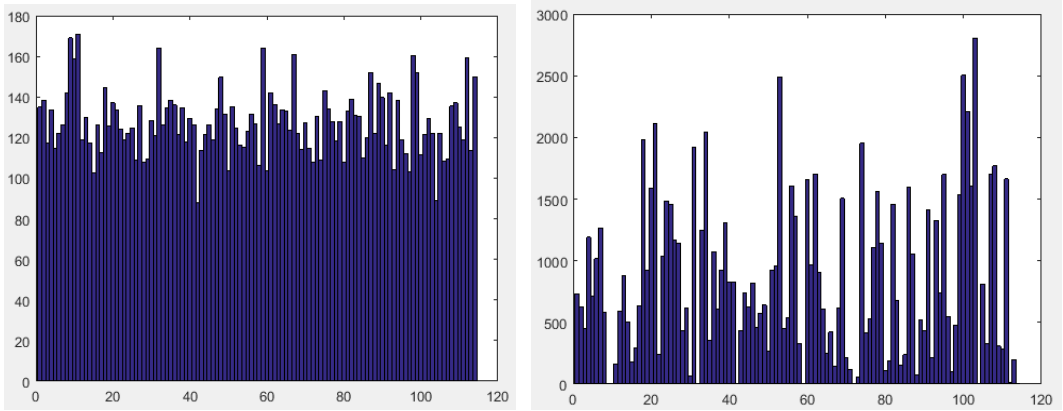
(i) (j)

图 4.7 高密度组织角度的均值和方差



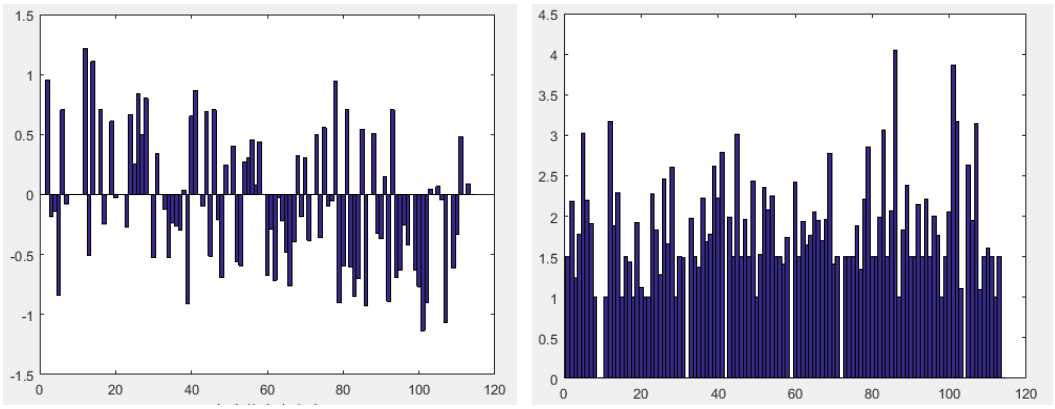
(k) (l)

图 4.8 高密度组织角度的偏度和峰度



(m) (n)

图 4.9 中密度组织角度的均值和方差

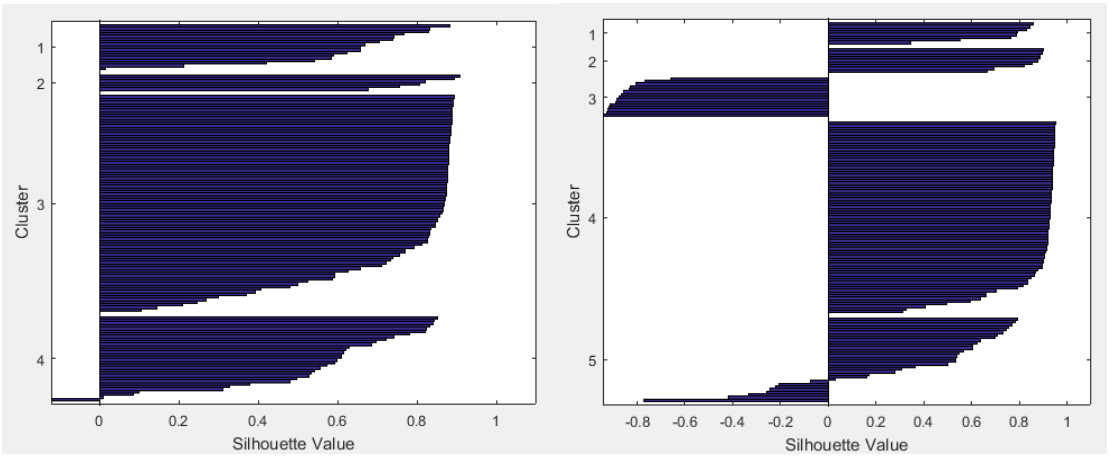


(o) (p)

图 4.10 高密度组织角度的均值和方差

从图 4.3-4.10 可以看出，高密度组织的长度和角度的平均值相对稳定，没有显著差异，但方差、偏度和峰度的差异明显。因此，高密度组织可以用 *k*-means 分为两到三类。而中密度组织的长度和角度的平均值都呈现出一定波动，其中方差、偏度和峰度波动更为明显。中密度组织可以用 *k*-means 分为两到三类。

此外，研究中使用了一种有效的评价方法，进一步对比分析了聚类结果。阴影图用于说明一个集群中每个数据点与相邻集群中的其他数据点之间的距离。若这个平均值很高，则说明整体性能较好。阴影图有效的评估了聚类的效果。



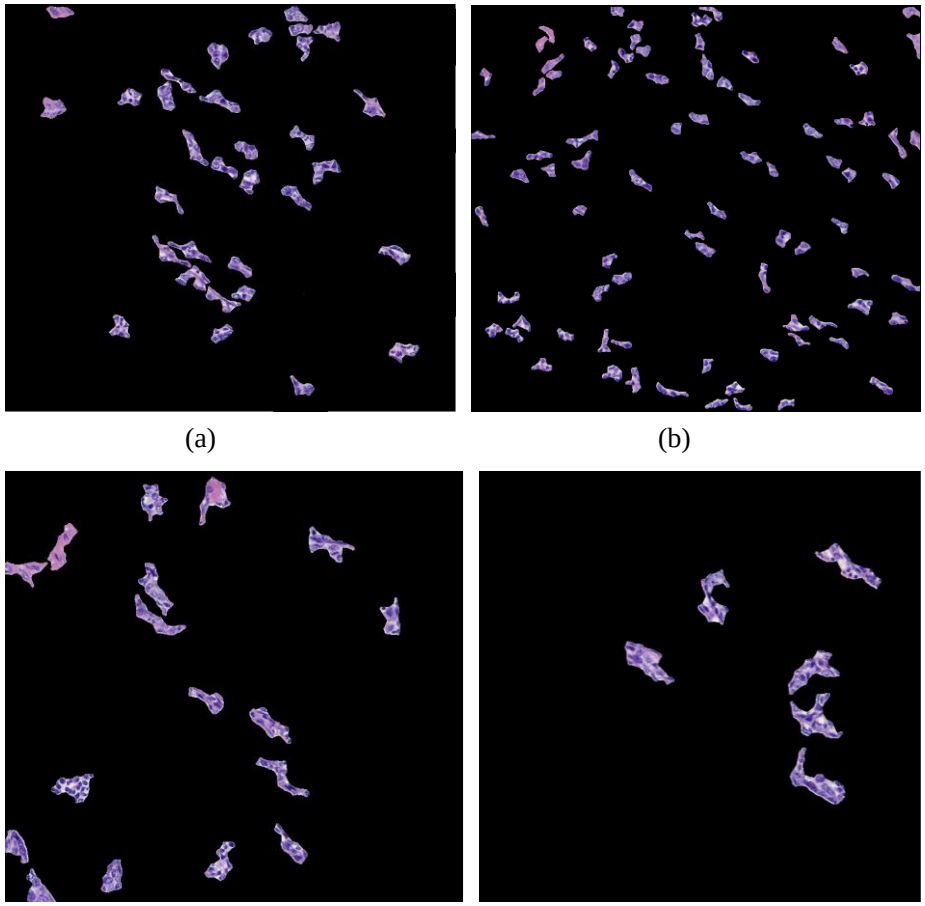
(a) {b)

图 4.11 *k*=4 和 *k*=6 的聚类结果阴影图

图 4.11 对比了簇数为 *k*=4 和 *k*=6 的聚类阴影图。(a)是 *k*=4 产生的结果，(b)是 *k*=6 产生的结果。横轴代表轮廓值，纵轴代表簇号。(a)中，纵轴上的 1/2 分别代表两种高密度组织，3、4 代表两种中密度组织。(b)中，纵轴上的 1、2 和 3 分别代表三种类型的高密度组织，4、5 和 6 代表三种类型的中密度组织。

从图 4.11 可以看出，当 *k*=4 时，聚类的效果比较好，随着 *k* 数目的增加，聚类

变得多而乱。因此实验中，本文选择以 k=4 来进行聚类。同时，为了更好的观察聚类出的结果，实验中将聚类结果的组织结构分别提取出来，并分别显示出来。图 4.12 中(a)、(b)是两种中密度细胞结构的聚类结果，(c)、(d)是两种高密度细胞结构的聚类结果。从图中可以更清晰的看出哪些组织属于同种聚类，方便医生进行对比。



(a) (b)

(c) (d)

图 4.12 高密度中密度结构提取结果图

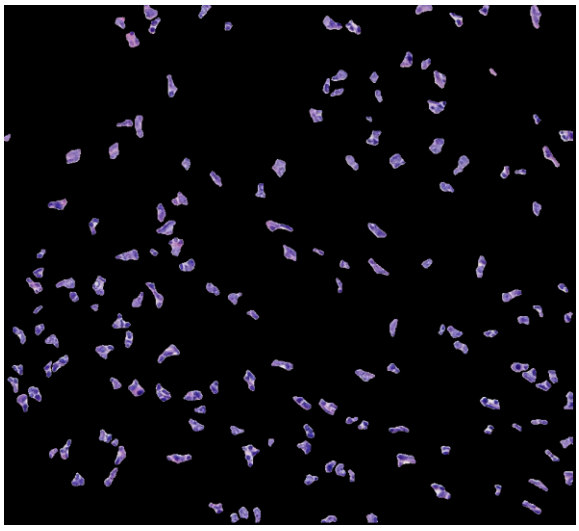


图 4.13 低密度结构提取结果图

除此之外，研究中还将低密度组织细胞结构也提取出来。本文在 3.2.5 中曾提到，低密度黏连细胞仅有 2-3 个细胞黏连，结构并不复杂，因此研究中没有生成低密度的最小生成树，但是其也有也成为病灶点的潜力，因此也为医生做出标注。

为了区分高密度、中密度、低密度，在最终的图像显示上，研究中用不同颜色以及深浅标记了聚类结果。如图 4.14，高密度用红色标记，中密度用蓝色标记，低密度用绿色标记，同时区域颜色越深代表其组织的拓扑结构越复杂。可以看到图 4.14 中多数密集细胞都能很好的聚类，进行区域标注和最小生成树构建，但仍有小部分没有被识别出来。

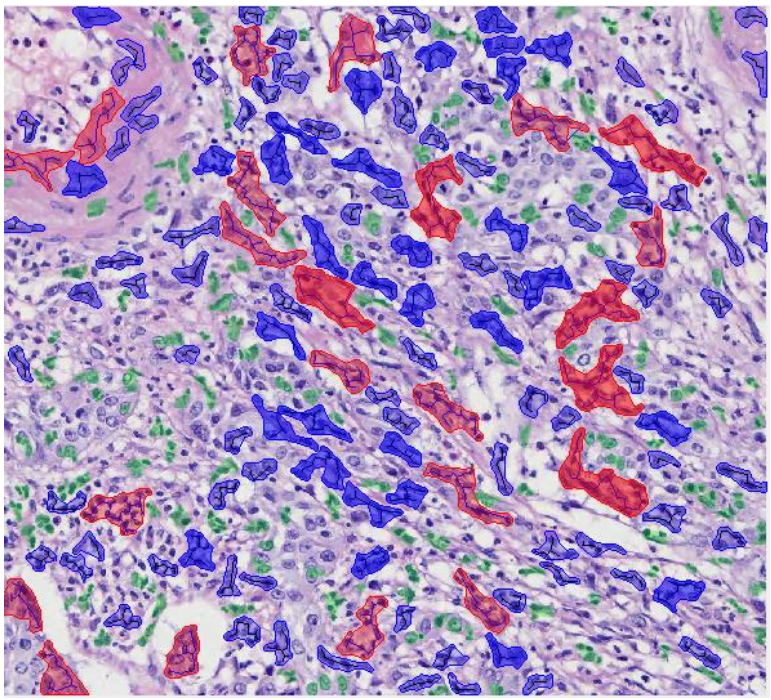


图 4.14 宫颈癌病理图像标注结果图

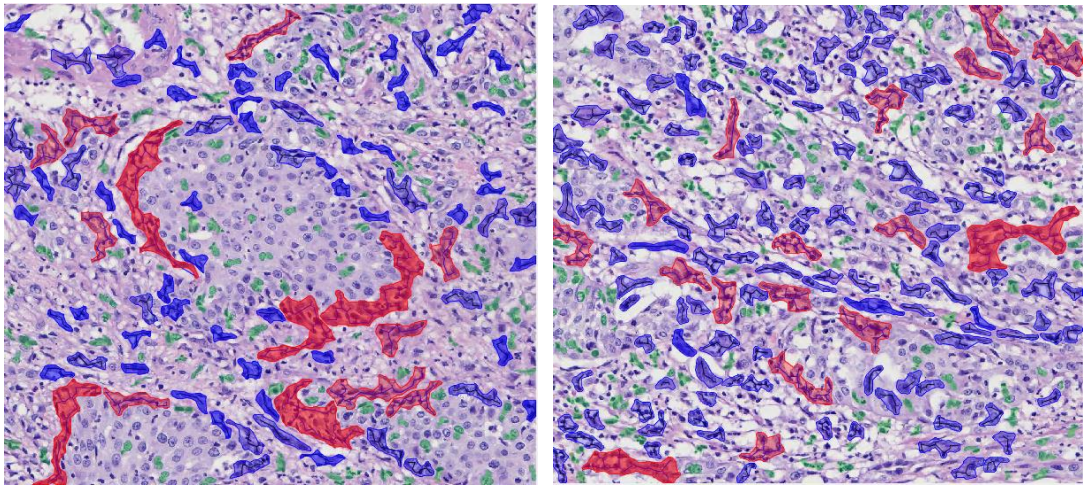
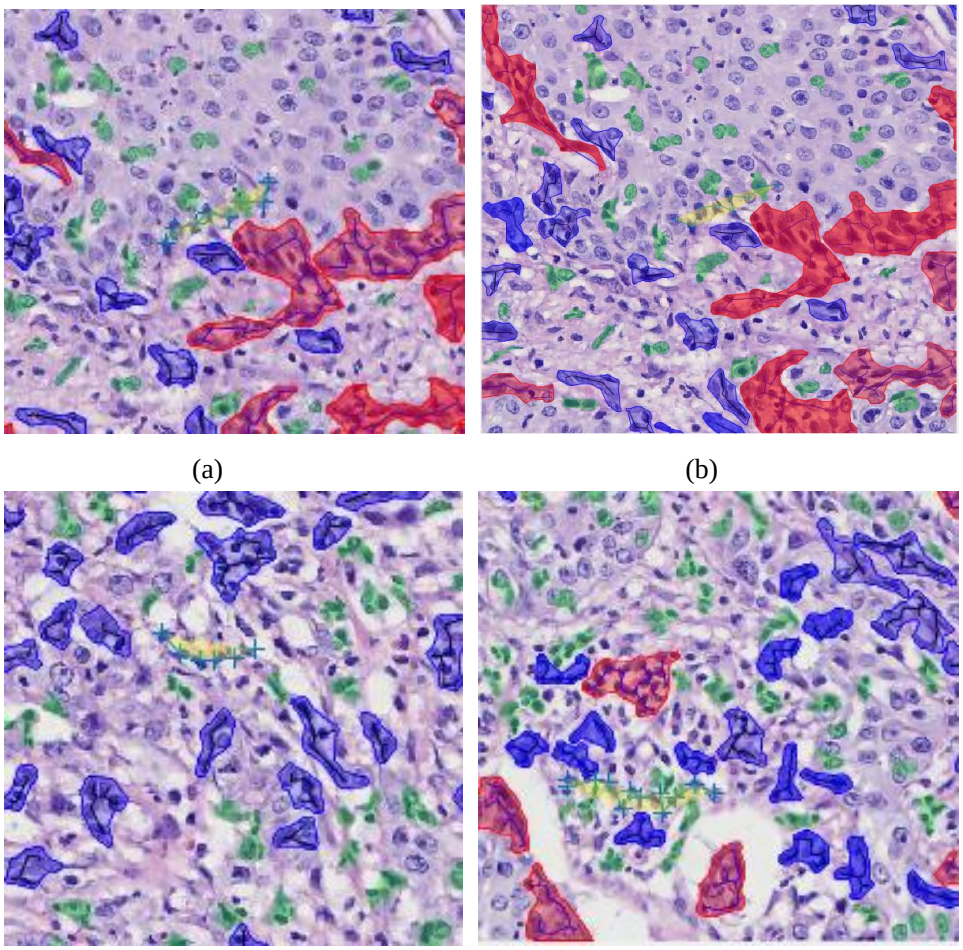


图 4.15 宫颈癌病理图像标注结果图

由于本研究使用的是实际医学数据，不同类型组织的结构和数目差异很大，为了

检验本实验所用方法的普遍适用性，研究中还使用了多张宫颈癌病理图像进行聚类查看聚类效果。

在聚类结果可视化的图像显示出后，医生查看组织病理学图像，可以通过面板点取细胞核进行手动标注。图 4.16 为手动标注的几处区域实例图，模拟医生对某处组织结构的重点标注和修改。



(a) (b)

(c) (d)

图 4.16 手动标注结果实例图

在进行手动标注后，本研究将对手动标注的区域进行图结构特征提取，将其归类为较近的聚类中去。图 4.17-4.20 为图 4.16 中医生标注前后聚类结果的对比。

由图 4.17-4.20 可以看出在标注后重新聚类的效果并不是十分稳定，与医生手动标注的结构有较大关系，因此获得的阴影图每次都具有差异性。因而有时，聚类获得的阴影图会出现不太好的效果。

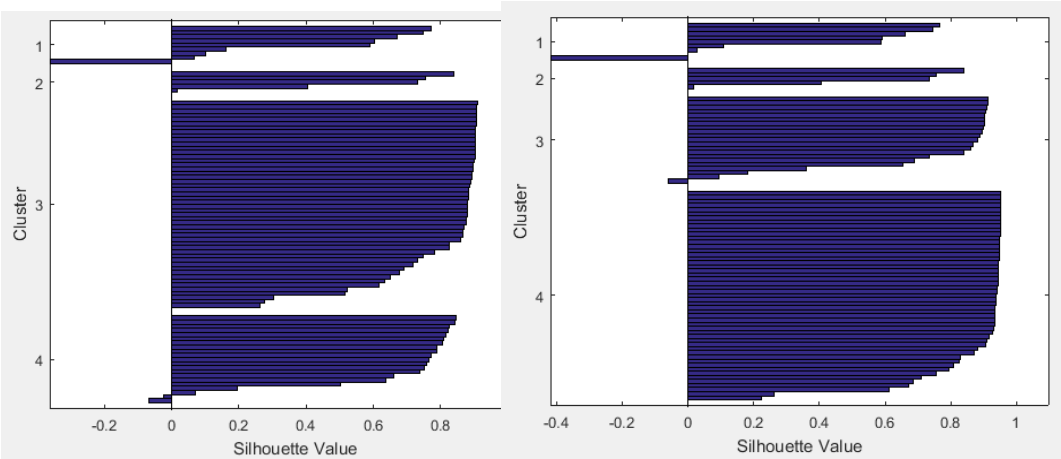


图 4.17 图(a)标注下的前后聚类对比

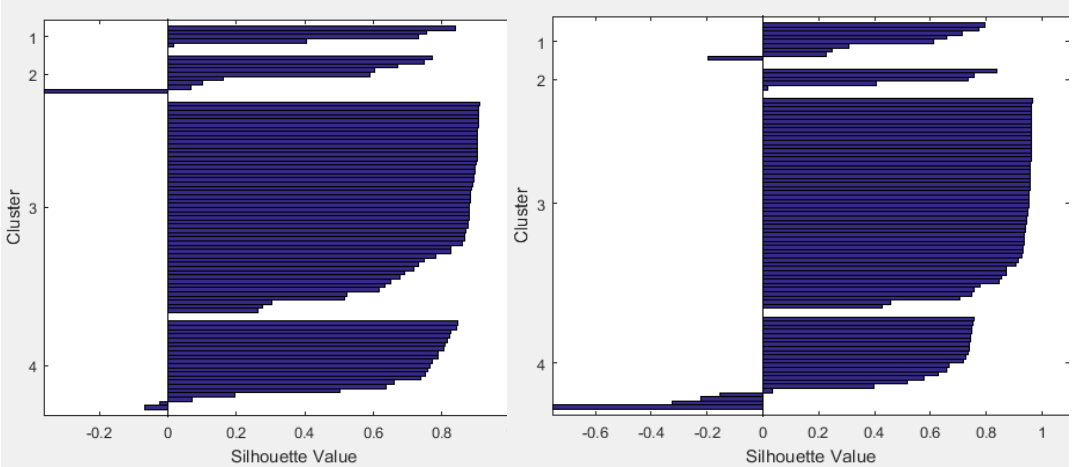


图 4.18 图(b)标注下的前后聚类对比

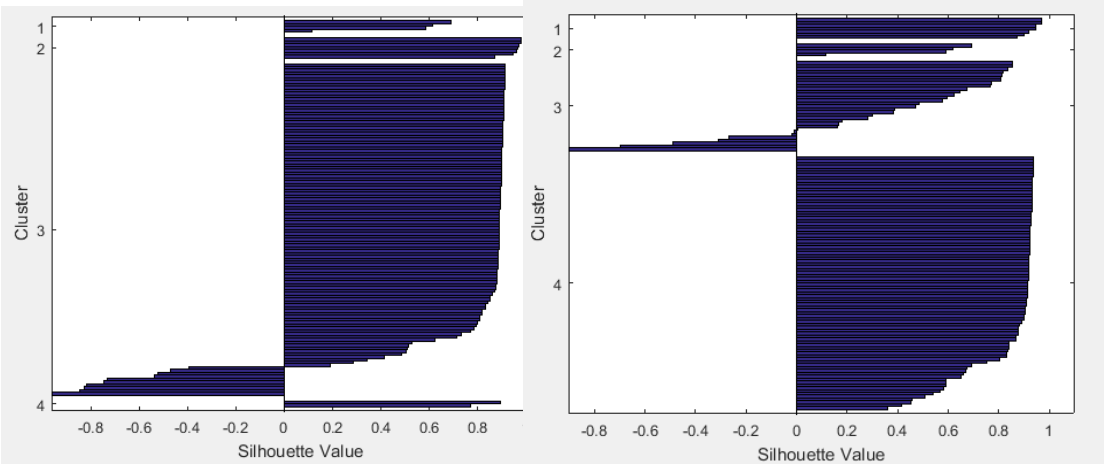


图 4.19 图(c)标注下的前后聚类对比

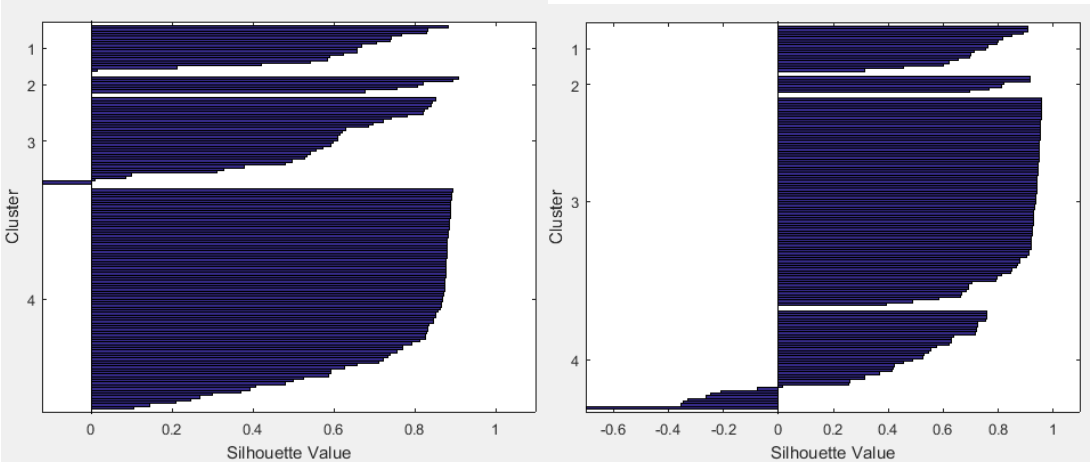


图 4.20 图(d)标注下的前后聚类对比

考虑到医生在观察阴影图时会浪费更多的时间，并且可能受限于专业知识，对于聚类和阴影图并不是十分了解，本文为医生给出了更简单的结果：在操作面板中直接显示所标注的区域更近似哪种密度。

图 4.21 操作面板弹出提示

4.2.2 错误结果分析

研究中，由于宫颈癌组织病理学图像组织结构多样，差异性大，结构和数目不定，在聚类时 k 值的选择可能会有所变化。除此之外由于高密度黏连细胞结构复杂，在构建最小生成树时不容易被计算机识别。

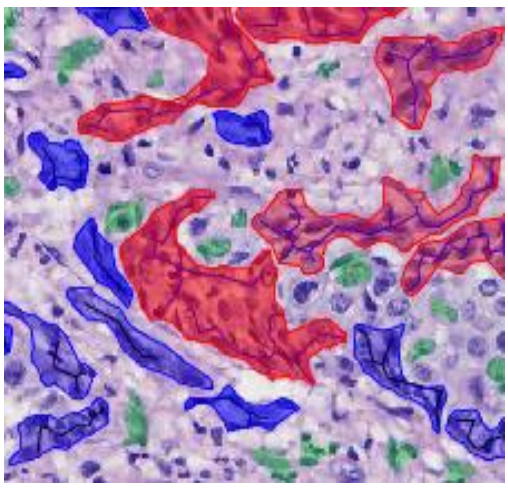


图 4.22 未被识别的细胞

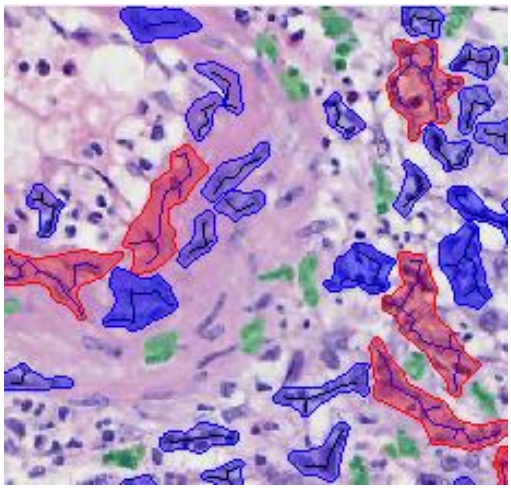


图 4.23 错误识别的组织

图 4.22 中深红色区域中部分细胞没有被构建进入到最小生成树中，造成这种错误的原因有可能是在图像分割时并不是十分精确，导致图像信息丢失，计算机不能正确识别细胞。因此，研究在预处理方面还需要更加精确完美的图像分割。

图 4.23 中红色区域为错误分割的组织，实际的组织结构中该部位并没有密集细胞，错误原因可能是图像分割时未能将白色的组织液分割出去，而被计算机误认为是大块的黏连细胞。

4.3 本章小结

本章给出了实验的具体结果和详尽分析，并提出了一些错误结果产生的可能原因。实验结果显示标注的颜色清晰，更有助于医生迅速发现病灶区域，并对可能产生病变的组织作出有效预防。在下一章节中，本文将提出对于实验的总结和后续的发展思路。

5 总结与展望

5.1 总结

如今在电子技术日新月异的时代，各种电子元器件的制造和开发越发先进高效，高处理能力的计算机系统被开发出来，进而计算机辅助诊断技术得到发展。近年来计算机辅助诊断技术越发火热。

本文提出了一种基于图论的无监督学习方法针对宫颈癌组织病理学显微图像进行聚类研究，着重在显微细胞图像的细胞黏连程度加以研究，旨在解决癌症风险预测和快速发现病灶点的问题。在本文实验研究中，也进行了很多尝试，具体工作内容总结如下：

1. 本文详细的研究了宫颈癌组织病理学、图论、无监督学习的相关文献，对于计算机辅助诊断组织病理学显微图像的研究背景和研究现状进行了总结，强调了其作用的贡献和意义。之后，本文对，颜色特征提取、机器学习理论、图论等研究中所用的基础知识进行了简要的介绍。

2. 本文提出了一种基于骨架化节点生成的方法来近似表示图像中的细胞核分布，并删除了多余节点。这是在面对高度黏连细胞较为有效的处理办法。

3. 本文提出了一种基于最小生成树的宫颈癌组织病理学图像聚类方法，该方法利用 k-means 聚类进行图像分割，并利用最小生成树进行图形和几何特征提取，并将颜色、图形、几何信息相结合，利用这些特征向量进行聚类，并在原始显微病理图像上进行标注，使聚类结果清晰可见。

4. 本文利用了一种阴影图来评价聚类结果的效果，结果显示在 k=4 时效果最佳，成功完成聚类任务。在实验中，通过对一个实际的图像数据集展开研究，验证了该方法的有效性和发展潜力。

5. 本文实现了一种手动标注细胞核结构的功能，利用了操作面板进行人工修改，并对此重新聚类，弥补了 2 中部分密集细胞没有被计算机识别的不足。

5.2 展望

本文的研究工作和成果在计算机辅助诊断在组织病理学图像研究中具有一定的发展潜力和应用价值，并且实现了标注病理图像的聚类功能。但受到专业知识的有限性，本文的研究仍需要继续学习和实践，本文的工作还需要不断完善和改进。根据组织病理学显微图像信息量大、结构复杂多样的特性，研究在很多方面仍具有深入研究的潜力。在今后的学习中，可以在以下几个方面进行研究：

1. 考虑更优的图像分割方法，以确保图像信息尽量不流失或较少流失。通过上述实验错误结果分析可看出，部分错误原因是由于图像分割方法并不完善导致最小生成树构建不够全面，以至于后续部分密集细胞没有被正确标注。因此，提高图像分割方法是记下来较为重要的研究方向。

2. 使用更多图论的方法进行提取特征，并进行多方面比较。在本文研究中，对于最小生成树方法展开详细的研究，但是在图论的其他方法中，也十分有研究前景和潜力。

3. 本文仅使用 MATLAB 实现了病理组织的聚类功能，在后续的研究工作中可以将软件封装集成为一个易于医生使用的软件，来方便医生使用。

参考文献

[1] 周晖,王东雁,罗铭,et al.《FIGO 2018 妇癌报告》——子宫颈癌指南解读[J].中国实用妇科与产科杂志,2019,35(01):99-107.

[2] Siegel, R., Miller, K., Jemal, A. Cancer Statistics. CA: A Cancer Journal for Clinicians, 2017,67(1), 7 - 30 (2017).

[3] 巩姣梅.宫颈癌筛查新技术研究[D].郑州：郑州大学,2013.

[4] 冯玲,杨湖珍. TCT、HPV-DNA 分型及阴道镜在宫颈癌筛查中的应用[J].当代医学,2013,25:19-21.

[5] 丁海艳,孙允高,叶大田.计算机自动识别宫颈细胞涂片技术[J].国外医学.生物医学工程分册,2000,02:85-90.

[6] Sukumarand, P., Gnanamurthy, R.: Computer Aided Detection of Cervical Cancer Using Pap Smear Images Based on Adaptive Neuro Fuzzy Inference System Classififier, Journal of Medical Imaging and Health Informatics 6(2), 312–319 (2016).

[7] 杨玉涛,张帆,曾莉等.细胞 DNA 定量分析技术在宫颈癌筛查中的回顾性研究[J],中国妇幼保健,2013，28（14）：2304-2306.

[8] 刘新梅. 探讨毛细式液基细胞学薄层染色技术对宫颈癌筛查应用价值研究[J].中外医疗,2016(1):54-55.

[9] 林钻娣,彭奕琼,刘永珠,等. 叶酸受体介导的宫颈特殊染色法和液基细胞学在宫颈癌筛查中的应用比较[J]. 慢性病学杂志, 2017(04):116-118.

[10] 李晓琳,师俊梅,刘涛,等. 细胞 DNA 定量分析法与巴氏染色细胞学诊断宫颈癌及癌前病变的对比分析[J]. 广西医学, 2015(4):497-499.

[11] 郑晓娟,胡新荣,唐泽立,等. p16 与 eIF4E 在宫颈癌中基因改变情况的初步研究[J]. 实用癌症杂志, 2011(1):13-15.

[12] 科学网. 专家呼吁利用生物标志物提高宫颈癌分级准确性[J].上海：上海医药,2015(6):13-13.

[13] 王志瑞,闫彩良.图像特征提取方法的综述[J].吉首大学学报（自然科学版），2011,05:43-47.

[14] 董小丽,张立和，米晓莉.基于颜色索引相关统计的彩色图像特征提取[J].光电子.激光，2011,04:623-628.

[15] 巩艳华，朱爱红,代凌云.基于颜色直方图的颜色特征提取[J].福建电脑，2007(5):96-97.

[16] 许宗敬，胡平．显微图像纹理特征提取方法综述[J]．微计算机应用 2009,30(6):6-13.

[17] Anousouya Devi, M & Subban, Ravi&Vaishnavi, J & Punitha, S. (2016). Classification of Cervical Cancer Using Artificial Neural Networks [J]. Procedia Computer Science, 89. 465- 472. 10.1016/j.procs.2016.06.105.

[18] Mustafa N,Isa N A M,Mashor M Y, et al.Colour Contrast Enhancement on Preselected Cervical Cell for ThinPrep® Images[C]// International Conference on International Information Hiding and Multimedia Signal Processing. IEEE Computer Society, 2007:209-212.

[19] 董建军．基于人工神经网络的子宫颈癌细胞识别[D]．沈阳工业大学，2006.

[20] 马瑾,曹阳．人工神经网络模型在子宫颈癌细胞检测方面的研究[J]．电子世界,2013(8).

[21] 鲁武警．基于 Snake 分割和 SVM 的宫颈细胞识别研究[D]．2015.

[22] 李文杰．一种多分类器融合的单个宫颈细胞图像分割、特征提取和分类识别方法研究[D]．2016.

[23] Rahmadwati R, Naghdy G, Todd C. Computer aided decision support system for cervical cancer classification[C]// SPIE Optical Engineering + Applications. International Society for Optics and Photonics, 2012:105-121.

[24] 赵英红，洪雅玲，孙存杰．基于 K 均值聚类算法的宫颈癌细胞分割方法[J].临床医学工程，2014，21(9):1089-1090.

[25] 关涛，周东翔，刘云辉,等．基于自适应阈值分割的宫颈细胞图像分类算法[J].信号处理，2012，28(9).

[26] 余婷婷．荧光宫颈图像多生暗区的配准与分割[D]，2018.

[27] Bondy J A , Murty U S R . Graph Theory With Application[J]. The Mathematical Gazette, 1978, 62(419):237-238.

[28] SHIJ,MALJKJ. Normalized cut and image segmentation [J]. IEEE Trans on Pattern

Anlysis and Machine Intelligence,2000,22(8):26-3.

[29]]FELZEN SZWALB P F,HUTTENLOCHER D P. Efficiet graph-based image segmentation [J]. Intemational Joural of Computer Vision,2004,59(2):167-181.

[30] PAVAN M,PeL LLO M. A new geaph-theoretic approach to clusterin and segmentation[C]. Proceedings of 2003 IEEE Computer Society Conference on CVPR,2003,1:145-152.

[31] Radke R J , Andra S , Al-Kofahi O , et al. Image change detection algorithms: a systematic survey[J]. IEEE Transactions on Image Processing, 2005, 14(3):294-307.

[32] 魏策，侯向萍．两种不同染色方法在液基细胞学(TCT)检查中的应用价值[J]．中外女性健康研究，2015(13)．

[33] 宋艳．综合颜色和形状特征的图像检索关键问题研究[D]．山东师范大学.

[34] Swain M,J,Ballard DH.Color indexing [J].International Journal of Computer Vision, 1991,7(1):11-32.

[35] John R,Smith,S,Fu Chang.Tools and techniques for color image retrieval [C],In Prac.of SPIE:Storage and Retrveval for Image and Video Database.vol 2670,1995, 2670:426-437.

[36] 刘铝．基于内容的图像检索方法的研究与实现[D]．湖南大学，2011.

[37] Stricker M,A,Orengo M,Similarity of color images [J].In:Proc of SPIE:Storage and retrieval for Image and Video databases III,San Jose,CA,1995,2420:381-392.

[38] Haralick R M , Shanmugam K , Dinstein I . Textural Features for Image Classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2007, SMC-3(6):610-621.

[39] Zahn C T , Roskies R Z . Fourier Descriptors for Plane Closed Curves[J]. IEEE Transactions on Computers, 1972, C-21(3):269-281.

[40]张跃．融合颜色和纹理特征的建材图像检索技术研究[D]．武汉理工大学，2012.

[41] 韩波．网格环境下的分布式离群数据挖掘方法研究[D]．太原科技大学.

[42] 陈凯，朱钰．机器学习及其相关算法综述[J]．统计与信息论坛，2007(5):105-112.

[43]百度百科“临近算法”词条[EB]/[OL]．https://baike.baidu.com/item/%E9%82%BB%E8%BF%91%E7%AE%97%E6%B3%95/1151153?fr=aladdin.2019.5.25.

[44] 殷瑞刚，魏帅，李晗,等.深度学习中的无监督学习方法综述[J].计算机系统应用，25(8).

[45] Peng,Y.,Park,M., Xu,M.,et al.:Clustering Nuclei Using Machine Learning Techniques. In: Proc. of IEEE/ICME International Conference 2010. pp. 52–57 (2010) .

[46] Lee H B,Macqueen J B. A K-Means Cluster Analysis Computer Program With Cross-Tabulations and Next-Nearest-Neighbor Analysis[J]. Educational and Psychological Measurement, 1980, 40(1):133-138.

[47] 真依然很拉风.聚类分析：k-means 和层次聚类[EB]/[OL].https://www.jianshu.com/p/794e91f60170.2019.5.26.

[48] 殷剑宏，吴开亚. 图论及其算法[M]. 中国科学技术大学出版社，2003.

[49] 陈杏，李军. 基于图论的分割算法研究综述[J]. 计算机与数字工程，2016,44(10).

[50] Miranda G H B , Barrera J , Soares E G , et al. Structural Analysis of Histological Images to Aid Diagnosis of Cervical Cancer[C]// Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference on. IEEE Computer Society, 2012.

[51] Park M , Jin J S , Xu M , et al. Microscopic image segmentation based on color pixels classification[C]// Proceedings of the First International Conference on Internet Multimedia Computing and Service. ACM, 2009.

[52] R I.Graham,P Hell. On the history of the minimum spanning tree problrm[J].Annals of the History of Computering,1985,7(1):43-57.

[53] R C PrimmShortest connection networks and some generalizations[J].Bell systems Technology Journal,1957,36:1389-1401.

[54] J B Kruskal.On the shortest spanning tree of a graph and the traveling salesman problem[J].Proceedings of the American Mathematical Society,1956,7:48-50.

[55] Cruz-Roa A , Xu J , Madabhushi A . A note on the stability and discriminability of graph-based features for classification problems in digital pathology[C]// 10th International Symposium on Medical Information Processing and Analysis. International Society for Optics and Photonics, 2015.

致谢

首先，非常感谢我的指导老师李晨对我的耐心指导。在这次毕业设计中遇到了很多问题，正是因为老师的指导，教授了我很多专业的知识，让我学到了很多东西，才能逐一把这些问题解决掉。非常感谢老师的帮助，让我少走了许多弯路，在选题时帮我选了基于最小生成树的组织病理学图像聚类研究这一个课题，并且还提供了许多专业书籍和文献资料给我。在实验中，也遇到了许多困难，是老师的指导让我坚持下来。并且在最后的论文撰写中也给我了许多帮助。在此衷心的感谢我的老师这一年来的无私指导。

同时，要非常感谢影像系的所有老师，感谢大学四年来教导过我们的所有老师。是你们的无私指导，用心教授，辛勤付出才使得知识能够一代代传承下去。

除此之外，要非常感谢辽宁肿瘤医院各位医生的大力支持，是他们制作了宝贵的病理学切片，并提供了大量的、详尽的图像和数据作为我的毕业设计进行研究。如果没有他们的帮助，我也无法完成这份毕业设计。

感谢在毕业设计中给予我帮助的同学张家伟、贺良子、卢渤霖、胡志杰学长以及我的室友乌日娜、刘逸如同学。是他们一直陪伴着我完成了这次的毕业设计，期间我有许多问题向他们请教，他们都及时给予了帮助，帮我度过了难关。

特别要感谢的是我的父母和家人，他们在我读大学时的提供了帮助，并鼓励我，支持我，是我最坚强的后盾。

最后要特别感谢东北大学四年来对我的培养，给了我一个宽阔的学习平台，让我不断获取新的知识，充实自己。

个人简历

姓名：尚麟静　　年龄：23 岁　　籍贯：辽宁

民族：汉族　　学历：本科　　专业：生物医学工程

政治面貌：共青团员

教育背景：

沈阳市雨田实验中学　　2010-2012

沈阳市第十一中学　　2012-2015

东北大学　　2015-2019

所获荣誉情况：

中荷学院青马优秀学员

科研成果：

专利：

李晨，陈昊，尚麟静，许宁，孙洪赞，张乐，胡志杰，薛丹. 基于点式阵列条件随机场的宫颈癌组织病理图像诊断方法[P].中国专利，2018：201811552817.4.（实审）