学号：？？？？？

**数据分析**

**课程论文**

**基于机器学习对三维神经元图像的胞体检测**

|  |  |
| --- | --- |
| **姓 名** | 李琦 |
| **学 号** | 18721802 |
| **论文评分** |  |

**2019年 2月18日**

目录

[**1、** **引言。** 2](#_Toc1400862)

[**2、胞体自动化检测算法** 3](#_Toc1400863)

[**2.1数据集** 3](#_Toc1400864)

[**2.2数据预处理** 3](#_Toc1400866)

[**2.3胞体特征** 5](#_Toc1400869)

[**2.4机器学习算法检测胞体** 5](#_Toc1400870)

[**3、实验结果** 7](#_Toc1400871)

[**4、结论与展望** 9](#_Toc1400872)

基于机器学习对三维神经元图像的胞体检测

[摘要]神经元的计算和分析中两个重要的任务是寻找神经元的胞体和重建它的结构。其中寻找胞体更为重要，因为我们需要它来展开图像追踪，从而重建结构。本文描述了一个基于机器学习的自动神经元胞体检测技术。测试图像是随机选择的包含噪声和部分神经元结构的原始图像，图像中的胞体也没有提前标注，实验表明该方法是非常高效的。

[关键字]神经元图像，胞体，机器学习

**Soma Detection in 3D Images of Neurons using Machine**

**Learning Technique**

**Abstract:** Two important tasks in the calculation and analysis of neurons are to find the soma of the neuron and reconstruct its structure. It is more important to look for the soma because we need it to expand the image tracking to reconstruct the structure. This paper describes an automated neuron soma detection technique based on machine learning. The test image is a randomly selected original image containing noise and part of the neuron structure, and the soma in the image is not labeled in advance. Experiments show that the method is very efficient.

**Key words:** Images of Neurons; Soma; Machine learning

1. **引言。**

神经系统是人体内起主导作用的功能调节系统。人体的结构与功能均极为复杂，体内各器官、系统的功能和各种生理过程都不是各自孤立地进行，而是在神经系统的直接或间接调节控制下，互相联系、相互影响、密切配合，使人体成为一个完整统一的有机体，实现和维持正常的生命活动。同时，人体又是生活在经常变化的环境中，神经系统能感受到外部环境的变化对体内各种功能不断进行迅速而完善的调整，使人体适应体内外环境的变化。可见，神经系统在人体生命活动中起着主导的调节作用，人类的神经系统高度发展，特别是大脑皮层不仅进化成为调节控制的最高中枢，而且进化成为能进行思维活动的器官。因此，人类不但能适应环境，还能认识和改造世界。

神经元是大脑中的重要神经细胞，也是构成神经系统结构和功能的基本单位。神经神经元的胞体是神经元的重要组成部分。 神经递质和神经元肽都是在这里合成并运送到神经末梢。关于胞体的研究对神经元分割至关重要并且是神经系统研究中必不可少的一步。

之前已经出现了手动设置胞体检测参数的方法，来帮助专家们寻找胞体位置，以及基于目标样本的自动化方法。Pawley在2006年提出直接将图像阈值来作为二进制掩码进行胞体检测的方法，彭汉川在2011年提出将最大距离变换，2015年Zhou提出了把强度和差异作为阈值，这些方法在分割后的荧光灯图像中都非常有效。

本体提出了一种针对果蝇神经元细胞的基于机器学习的全自动神经元胞体检测算法，相比较和之前他人方法的检测结果，本文的方法更加有效。

**2、胞体自动化检测算法**

**2.1数据集**

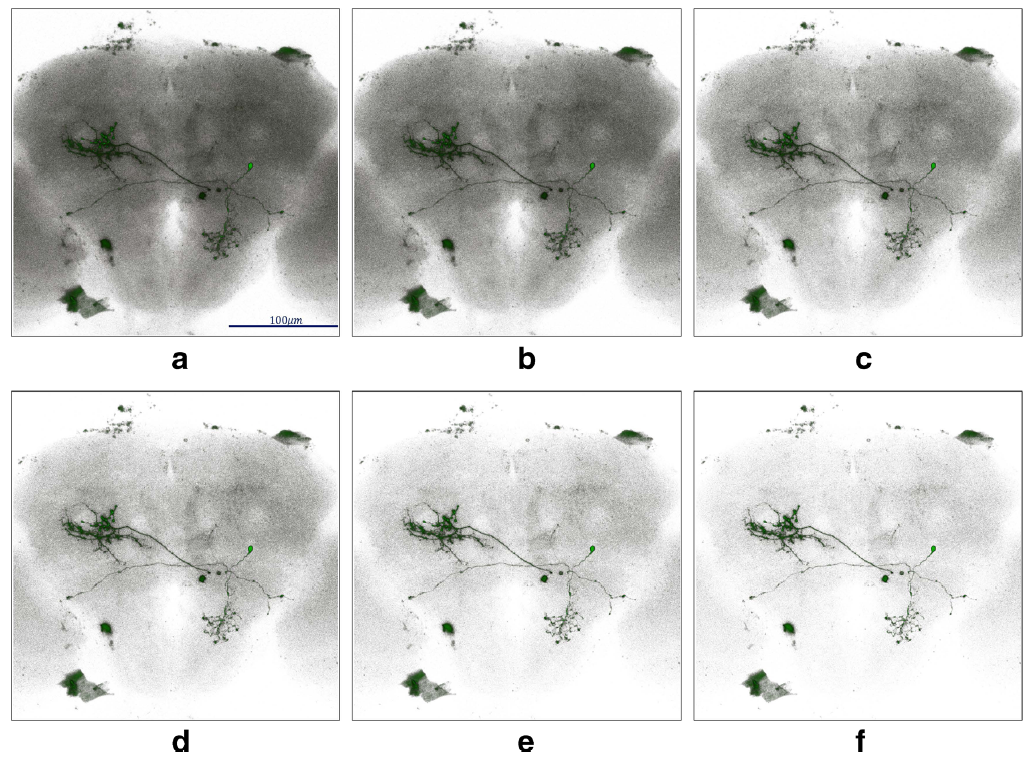
本实验的数据库使用了FlyCircuit的图像数据库。每个图像都包含使用共聚焦显微镜获得的三维体数据。数据集被分为两部分，第一部分是未经处理的包含噪声和一些不需要的神经元结构的图像（如图1a），第二部分是经过人为处理的完整的神经元图像（去除了噪声，只保留了干净的神经元信号，如图1b和1c）。

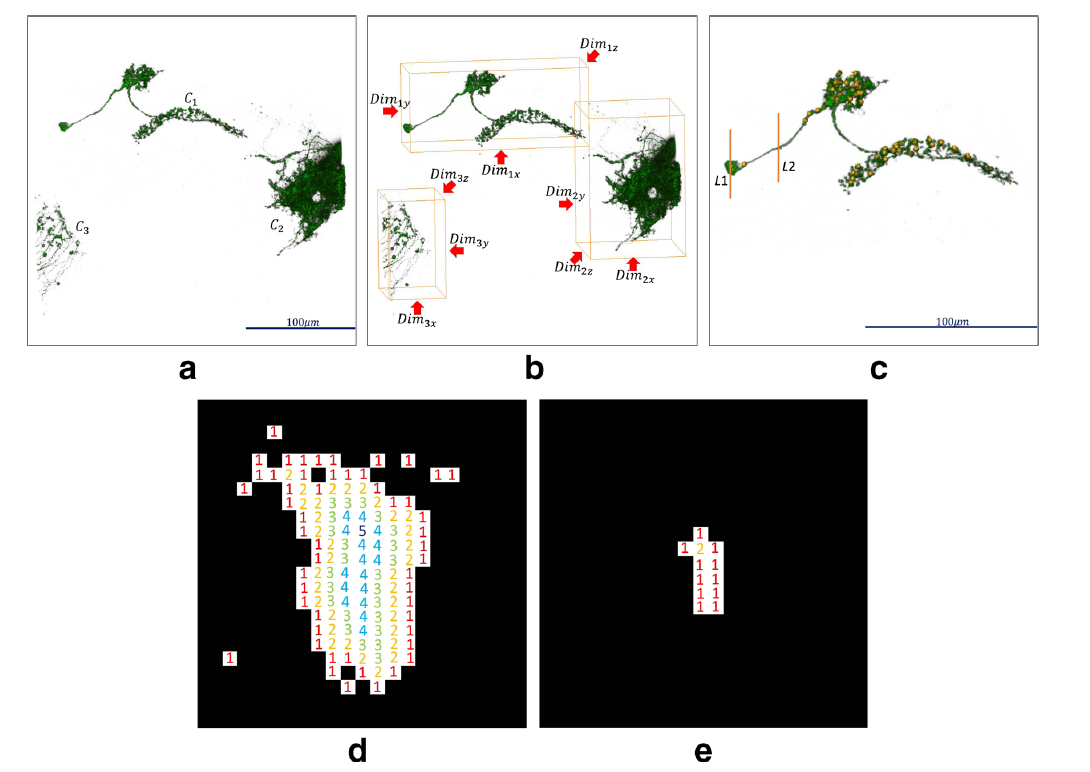
****

Figure

**2.2数据预处理**

首先我们把数据按照阈值分别为整个图像强度的60%，70%，80%，90%，95%来进行处理保留，然后进行二值化处理（体素值为0/1）。整个图像此时有n个连通分量，每个连通分量C都被一个Bounding Box（B）所包围。和，分别为Bounding Box的x,y,z分量。如果连通分量有超过一百个体素在Bounding Box边缘，那我们认为它是一个不完整的神经元（因为神经元可能正好在图像切割的边缘）。值为图像到连通分量边缘的距离，我们设置的阈值为5，当时，我们才认为它可能是Soma。如果两个Soma距离小于3个体素，我们就把它们融合成一个Soma。



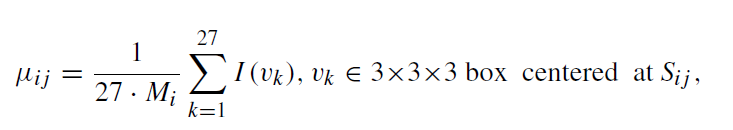


**2.3胞体特征**

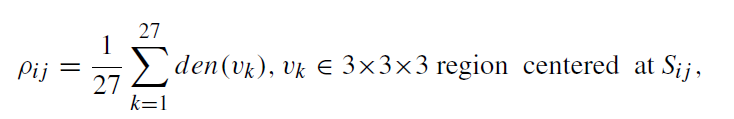
胞体候选者需满足以下特征才可能是一个神经元的中心点。

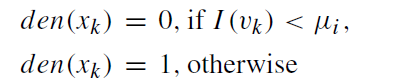
让和分别为第i个连通分量体素的平均值和最大值，为第i个连通分量的第j个Soma候选者，其Bf值为。我们提出了以下三个特征，这些特征均被归一化，其值越接近1，表明为Soma的可能性越大。

1. Soma比神经元其他部分具有更高的强度值。归一化强度值是以候选者为中心的立方体的平均强度值。

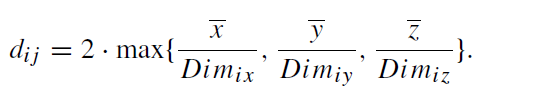


1. Soma除了体素值外，其外形更接近于球体，并且体素值更加紧凑。我们定义紧凑值为候选者周围的体素超过平均值的数量





1. Soma的位置通常也是一个特征，其一般情况不会在Bounding Box的中心位置，所以我们可以定义一个特征，x，y，z分别是候选者在三个维度到Bounding Box中心点的位置



1. Soma的外形类似于球体，所以我们可以为之确立一个均衡性特征。

设置一个平面，穿过候选者体素，左边点个数为f，右边为r，则均衡性计算方法为

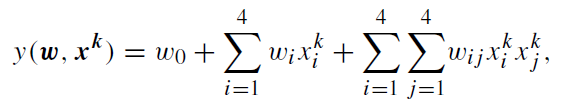


（若r<f，则r为分子）

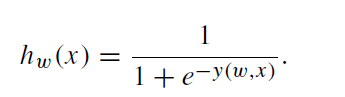
**2.4机器学习算法检测胞体**

我们从清理过的数据集中选择训练样本，其中包括正向和反向样本。我们选择连通分量中具有最大Bf值的体素，如果其和真实位置相差小于6微米，那么其为正向样本，否则为反向样本。这个过程被重复5次，来创建5个例子，如下图，红色为正向样本，其余为反向，但是红色并不是第一个被选择的体素值。

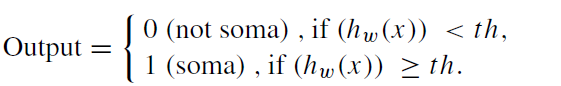
上节我们提到了四个特征，让Xi分别代表这四个特征，我们可以创建出如下的回归方程学习模型：



其中w为权重，k代表第k个训练实例，y函数结果为0/1，表明其是否为一个Soma。应用逻辑回归，将连续目标转换为分类目标，h函数定义如下：



针对h函数，需要一个阈值来确立输出：

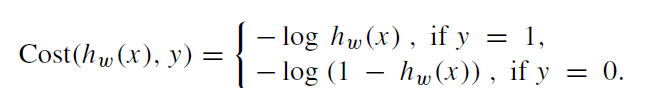


本文中设th为0.5

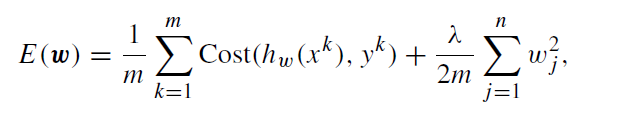
我们的目标是计算权重w，来使之匹配y输出，使得判断结果为最佳。

以下提出两个函数，代价函数和误差函数，每次迭代计算之后，结果都需要减去一个误差函数。

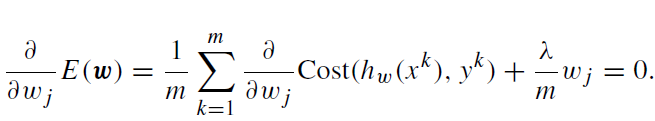
代价函数：



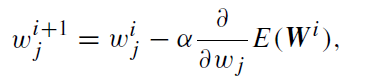
误差函数：



其中m是训练实例数，n是权重个数。等式右侧第二个表达式是为了避免过度拟合。误差函数是为了累积对目标预测的误差，我们的目标是找到w的集合，使得代价函数最小化，可以通过求解误差函数的偏导来实现。



联立之前等式最终可以得到w的更新式：

. 

α = 0.1，为学习率，学习的停止条件为误差函数提高小于0.0001或者迭代次数大于5000。最终可以获得方程中的权重。

**3、实验结果**

从FlyCircuit数据库中随机选择200个原始图像，由人手工提前标注。（但有些Soma没有被人标注，也可能被算法检测出来）

为了分析我们算法的效果，我们定义了几个变量：

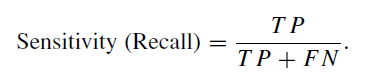
TP:220个中选择正确的Soma个数

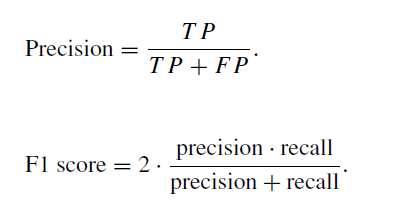
TS:220个之外选择正确的Soma个数

FN：220个中选择没有被检测出的Soma个数

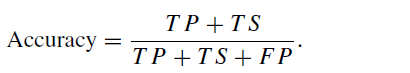
FP：选择错误的Soma个数

我们使用灵敏度，F1分数和准确度来估算算法表现：



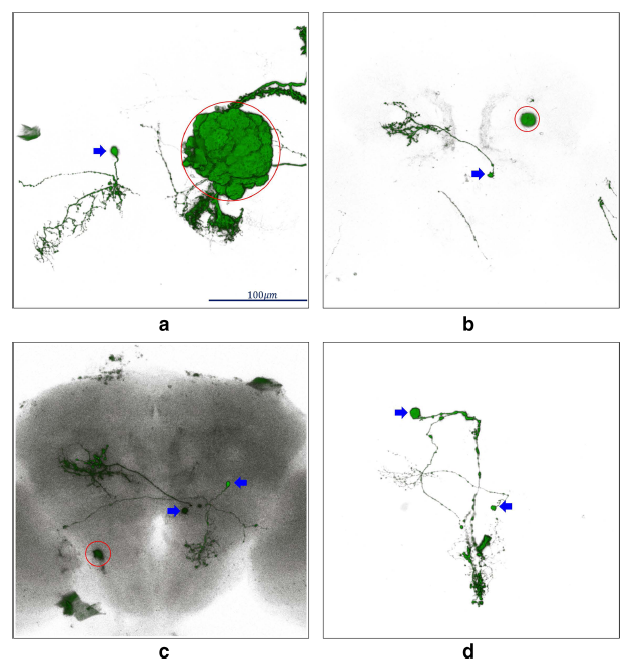


因为图像中包含的Soma并没有提前被标注出来，而且数量未知，所以灵敏度和F1分数没有包含TS。而又因为算法识别出了不在数据库中的Soma，所以我们将精度测量值修改为：



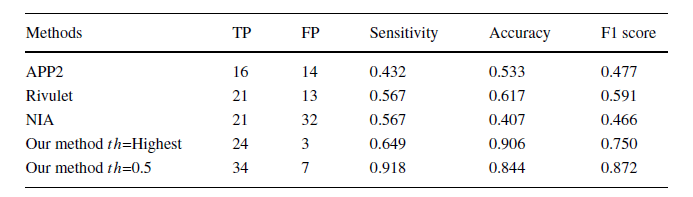
灵敏度和准确度取决于方程式中的阈值th。如果只有一个soma，则将th设置为最高hw（x）。 在这种情况下，我们得到TP = 156，T S = 20，FP = 21，FN = 64。 敏感度为70.9％，F1评分为78.6％，准确率为89.3％。 考虑到图像中可能存在多个soma，th被设置为0.5作为hw（x）的阈值。 然后我们得到TP = 195，T S = 55，FP = 57，FN = 25.灵敏度为88.6％，F1得分为82.6％，准确度为81.4％。

下图展示了大多数情况下的四个疑难案例体细胞检测算法。 图7a-c包含明亮的物体可导致体细胞检测到错误的位置。 在图7c-d中有两个体细胞。 这个图中的四个图像中的somas用蓝色箭头标记。本文的方法正确确定了所有的胞体。



本文还调用了其他的方法针对上图进行检测，包括APP2，Rivulet，SomaExtraction，NIA

对比结果如下



Table

图8显示了获得的结果提出的方法和其他三种方法。 在图8中，

通过所提出的方法检测到的体细胞用蓝色标记箭头; APP2，NIA和Rivulet获得的结果是分别用黄色，红色和蓝色圆圈标记。该其他三种方法被吸引到图1中的错误位置。在图8a中，仅检测到一个体细胞，并且在图8d中仅检测到一个体细胞。

除了困难的情况，我们随机选择33来自200个测试数据集的体积数据集。 然后我们将方法APP2，Rivulet和NIA应用于33一组测试数据。 有37名专家选择的神经元图片。 表1列出了所有的方法。本文提出的方法显然更加有效。

**4、结论与展望**

我们描述了一种机器学习方法来检测果蝇大脑的共聚焦显微图像上的体细胞。 所提出的方法是有效的并且可以应用于原始原始图像，其中噪声未被去除，并且胞体的数量是未知的。根据不同的显微镜，为获取图像，操作员和图像特征而设置的参数可能不同。为了获得最佳结果，可以更改功能，并且应该重新训练模型。

在这项研究中，一旦机器训练完成，胞体识别任务就会完全自动执行。 随机选择200个测试数据集，灵敏度和准确度均令人满意。 该算法具有鲁棒性和实用性。

**参考文献**

[1]彭昱忠，王谦，元昌安，林开平.数据挖掘技术在气象预报研究中的应用[J].干旱气象,2015,33(1):19-27.

[2]金龙,吴建生.基于遗传算法的神经网络短期预报预测模型[J].灾害学,2004,8(1): 15-16．

[3]段文广,周晓军,石永炜.数据挖掘技术在精细化温度预报中的应用[J].干旱气象，2012,30(1):130-135．

[4]Enireddy Vamsidhar，Varma K V S RP,Sankara RaoP，et al. Prediction of Rainfall Using Back propagation Neural Network Mode[J].International Journal on Computer Science and Engi- neering,2010,2(4):1119-1121．

[5]Sendonaris A, Erkip E, Aazhang B.User cooperation diversity-Part I and II [J]. IEEE Transactions on Communications, 2003 ,51(11):1927-1948.

[6]江涛，张玉芳，王银辉.一种改进的粒子群算法在BP网络中的应用研究[J].计算机科学，2006，33（9）：164-166.

[7]Peter J. Angeline Evolutionary Optimization Versus Particle Swarm Optimization, Philosophy and Performance Difference[J]. Lecture Notes In Computer Science, 1998, 1447:601-610.

[8]Yang Shangdong. Li Xiang. A New Ann Optimized By Improved PSO Algorithm Combined With Chaos And Its Application In Short-term Load Forecasting[J]. IEEE Transaction on Computational Intelligence and Security, 2006 International Conference, 2006, 2:945-948.

[9]周岩,王盛,高传善,孙慰迟.基于改进粒子群算法的模糊神经网络及其在短时天气预报中的应用[J].计算机应用与软件,2010,27(5):234-237.