·图形图像处理· 文章编号：1000-3428(2019)01-0259-05 文献标志码：A 中图分类号：TP391

基于 3D 卷积神经网络的人体动作识别算法

张 瑞[1,2]，李其申[2]，储 珺[2]

（1. 南昌航空大学 信息工程学院，南昌 330063；2. 江西省图像处理与模式识别重点实验室，南昌 330063）

摘 要： 由于人体动作的多样性、场景嘈杂、摄像机运动视角多变等特性，导致人体动作识别的难度增加。为此，基于 3D 卷积神经网络，提出一种新的人体动作识别算法。以连续的 16 帧视频为一组输入，采用视频图像的灰度、*x* 方向梯度、*y* 方向梯度、*x* 方向光流、*y* 方向光流做多通道处理，训练网络参数，经过 5 层 3D 卷积、5 层 3D 池化增加提取特征中时间维度的动作信息，最终通过 2 层全连接与 softmax 分类器得到识别分类结果。在 UCF101 数据库上进行实验，结果表明，相比 iDT、P-CNN、LRCN 算法，该算法具有较高的识别准确率，且运行速度更快。

关键词： 人体动作识别；多通道；3D 卷积；3D 池化；时间维度

中文引用格式： 张瑞，李其申，储珺. 基于 3D 卷积神经网络的人体动作识别算法[J]. 计算机工程，2019，45(1)：259-263.

英文引用格式： ZHANG Rui，LI Qishen，CHU Jun. Human action recognition algorithm based on 3D convolution neural network[J]. Computer Engineering，2019，45(1)：259-263.

Human Action Recognition Algorithm Based on 3D Convolution Neural Network

ZHANG Rui[1,2]，LI Qishen[2]，CHU Jun[2]

（1. School of Information Engineering，Nanchang Hangkong University，Nanchang 330063，China；
2. Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition，Nanchang 330063，China）

【Abstract】 Human action diversity，scene noise，the camera motion angle changes and other factors increase the difficulty of human action recognition. This paper proposes a human action recognition algorithm based on 3D convolution neural network. Firstly，successive 16 frames of the video are divided into a group as the input. Secondly，the input data is multi-channel processed using the gray，gradient-x，gradient-y，optflow-x and optflow-y，which effectively trains the network parameters. Thirdly，the extracted features are obtained using 5-layer 3D convolution and 5-layer 3D pooling to increase time dimension information，Finally，the recognition results are obtained by two full connection layers and the softmax classifier. Experiment is made on the UCF101 database，and the results show that compared with iDT，P-CNN，LRCN algorithms，the proposed algorithm has a higher accuracy of human action recognition and a faster running speed.

【Key words】 human action recognition；multi-channel；3D convolution；3D pooling；time dimension

DOI： 10.19678/j.issn.1000-3428.0048978

0 概述

人体动作识别是计算机视觉领域的一个极具挑战性的课题，涉及模式识别、图像处理、人工智能等多个学科领域，并广泛应用在人机交互、运动捕捉分析、视频监控和安全、环境控制检测与预测[1-2]。当前人体动作识别主要受到个体差异、视角变化、摄像机运动、光照角度的影响[3]。

卷积神经网络（Convolutional Neural Network，CNN）是基于深度学习理论的一种人工神经网络，利用权值共享减小普通神经网络的参数膨胀问题，在前向计算过程中使用卷积核对输入数据进行卷积操作，最终通过一个非线性函数输出结果。其通过在原始图像上交替运用滤波与局部近邻操作获取复杂特征的层次结构，在视觉物体识别中取得较好的效果[4]。CNN 由于其特殊的网络结构对复杂背景、光照、角度变化等不敏感的特性，近些年来被广泛应用于计算机视觉中的分类、检测、分割等任务[5]。

目前基于 CNN 的人体动作识别方法主要分为 2 类。一类将动作视频帧通过 2D CNN 与其他神经网络结合提取特征及动作分类，例如长时递归卷积神经网络（Long term Recurrent Convolutional Network，LRCN）[6]等。这类方法一般都是针对图像进行处理，对视频分析通常忽略了时序信息。另一类将 2D 卷积拓展到 3D 卷积，加入时间维度，直接

基金项目： 国家自然科学基金(61663031)；江西省自然科学基金(20132BAB201046)；南昌航空大学研究生创新专项资金(YC2016009)。
作者简介： 张 瑞(1993—)，女，硕士研究生，主研方向为图像处理、模式识别；李其申，副教授、博士；储 珺，教授、博士、博士生导师。
收稿日期： 2017-10-16 **修回日期：** 2017-12-09 **E-mail：** 546029094@qq.com

对视频进行分析捕获。这类方法是对视频数据的时间维度和空间维度进行特征计算,对多个连续帧数据进行卷积得到多个特征图[7-8]。3D 卷积比 2D 卷积能更好地表达在视频中的有效运动信息,具有一定的优越性。文献[9]使用多分辨率的 CNN,将视频数据流分为低分辨率和原始分辨率的数据流。文献[10]将视频分为静态帧数据流和帧间数据流,采用 CNN 提取特征后进行合并,通过全连接经支持向量机(Support Vector Machine,SVM)分类器做分类识别。该算法虽然取得较好的效果,但是网络结构复杂,时间及空间复杂度较高。文献[11]使用单帧数据和光流数据捕获运动信息,其卷积网络的第1 层包含灰度数据、x 与 y 方向的梯度、x 与 y 方向的光流,网络结构共有 3 个卷积层,2 个池化层和 1 个全连接层以及 softmax 分类器的输出层。该算法(下文简称 P-CNN 算法)在机场视频监控中有较好的识别率,但在其他场合监控视频中没有展现较好的识别准确率[1,3,5]。

本文基于 P-CNN 算法模型,采用 5 通道的处理方式,将输入图像经 5 层 3D 卷积、5 层 3D 池化获取更多时间维度信息,并将所有信息通过 2 次全连接与 softmax 分类器得出最终结果。

1 P-CNN 算法

文献[11]提出的具有时间维度信息的P-CNN算法,采用的网络被认为是用于人体动作识别的经典 3D CNN。该网络使用单帧数据和光流数据,在卷积过程中与多个连续帧的数据进行连接,采用已标记的机场监控视频库 TRECVID 训练得到模型参数,结合 softmax 分类器得到最终识别结果。

1.1 3D 卷积

2D 卷积的实质是从前一层特征映射中提取局部邻域特征,在空间维度上卷积得到 2D 特征图[12],卷积过程可表示为:

$$v_{ij}^{xy}=\tanh\left(b_{ij}+\sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}w_{ijm}^{pq}v_{(i-1)m}^{(x+p)(y+q)}\right) \quad (1)$$

其中,v_{ij}^{xy} 代表第 i 层第 j 个特征图像素点 (x,y) 的结果值,b_{ij} 是第 i 层第 j 个特征图的偏差,m 是第 $i-1$ 层的特征图个数,P_i、Q_i 是第 i 层 2D 卷积核的空间维度大小,w_{ijm}^{pq} 是第 $i-1$ 层第 m 个特征图连接的卷积核权重。在视频分析问题中,需要获取的运动信息数据在多个连续帧中,所以将 2D 卷积拓展到 3D 卷积,从空间维度和时间维度上计算特征。

3D 卷积是数据集通过 3D 卷积核由多个连续帧叠加在一起形成的。多个连续帧依次通过卷积层,卷积层中每个特征图都与上一层的多个相邻连续帧相连,从而获取一定的运动信息[11],可表示为:

$$v_{ij}^{xyz}=\tanh\left(b_{ij}+\sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1}w_{ijm}^{pqr}v_{(i-1)m}^{(x+p)(v+q)(z+r)}\right) \quad (2)$$

其中,v_{ij}^{xyz} 代表第 i 层第 j 个特征图像素点 (x,y,z) 的结果值,b_{ij} 是第 i 层卷积层的第 j 个特征图的偏差,m 是第 $i-1$ 层的特征图个数,P_i、Q_i、R_i 是第 i 层 3D 卷积核的空间维度与时间维度大小,w_{ijm}^{pqr} 是前一层第 m 个特征图连接的卷积核权重。3D 卷积与 2D 卷积相比,增加了时间维度,其框架结构如图 1 所示。

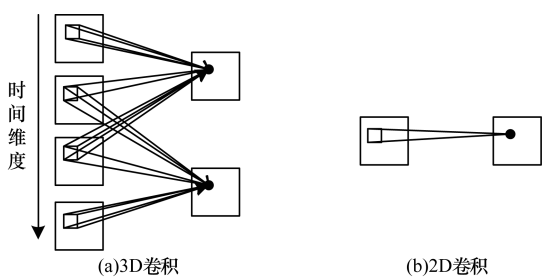

图 1 3D 卷积与 2D 卷积框架结构

1.2 P-CNN 网络结构

文献[11]提出的 P-CNN 算法网络结构如图 2 所示。

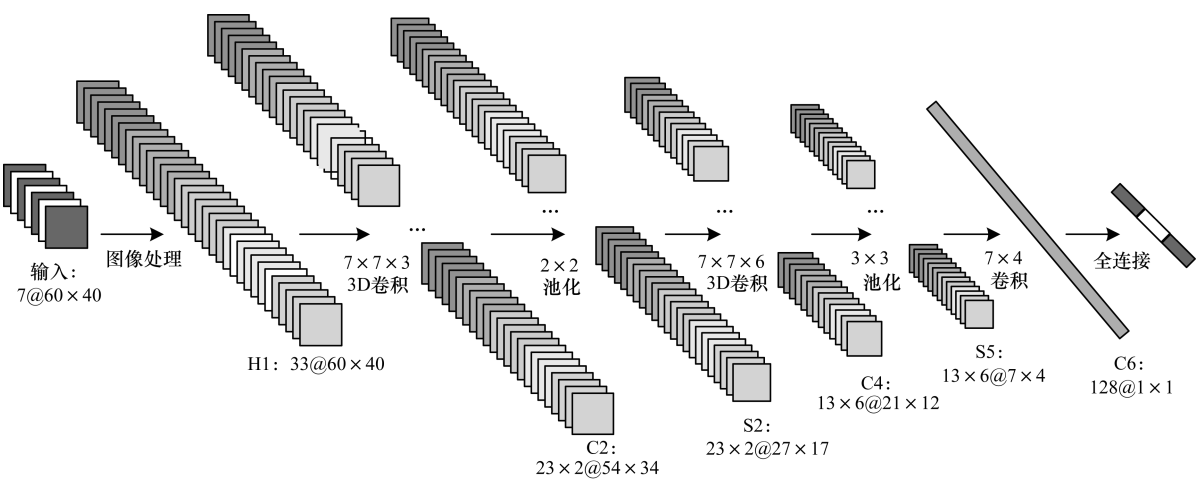

图 2 P-CNN 网络结构

此网络结构以 7 帧 60 ×40 的视频为一组,作为初始输入,其卷积过程如下:

1)对图像进行处理,分为 5 个通道,即灰度、x 方向梯度、y 方向梯度、x 方向光流、y 方向光流。每帧图像计算灰度、梯度,2 个相邻连续帧计算光流,得到第 1 层共 33 个 60 ×40 的特征图。这一层是基于先验知识对图像进行处理,比随机初始化,更利于有效地训练网络结构参数。

2)分别在 5 个通道上进行卷积核大小为 7 ×7 × 3(7 ×7 为空间维度,3 为时间维度)的 3D 卷积。为了增加特征图的数量,提取不同的特征信息,用 2 种不同的卷积方式进行卷积,得到 2 组共 46 个特征图,每组是 23 个特征图。

3)对第 2 层的特征图进行窗口大小为 2 ×2 的池化,得到相同数目 23 ×2 的特征图。此层的特征图与第 2 层比,降低了空间分辨率。

4)对 5 个通道的特征图分别进行卷积核大小为 7 ×6 ×3 的 3D 卷积。与第 2 层一样,为了增加特征图的数量,2 组特征图都采用 3 种不同的卷积方式,第 4 层得到 6 组不同的特征图,每组有 13 个特征图。

5)对第 4 层的特征图进行窗口大小为 3 ×3 地池化,得到同样数目但空间分辨率降低的特征图。

6)此时,时间维上帧的个数已经很小,所以此层只在空间维度卷积,卷积核大小为 7 ×4,即输出的特征图大小为 1 ×1。此层共有 128 个特征图,每个特征图与第 5 层中所有的特征图进行全连接,即得到最终的 128 维特征向量。

经过多层卷积池化后,每连续的 7 帧视频均转化为 128 维的特征向量,采用 softmax 分类器对其特征向量进行分类,实现行为识别。

1.3　P-CNN 算法的不足

P-CNN 算法的网络结构是基于机场视频监控 TRECVID 数据库[11-12]提出的。为降低网络参数训练的复杂度,将连续 7 帧视频为一组作为输入,但不是所有的人体动作都能在 7 帧视频之内表达完成。而 P-CNN 算法的网络结构在一定程度上忽略了这种高层的运动信息,导致提取的特征向量无法包含部分信息,使得识别准确率有所降低。同时,该网络结构只在卷积层中运用 3D 卷积,而在池化层中依旧采用 2D 池化操作,丢失了部分时间维度信息。针对上述缺陷,本文提出一种改进的 3D CNN 算法。

2　改进的 3D CNN 算法

2.1　卷积核参数选择

P-CNN 算法的网络结构只在卷积层使用 3D 卷积,而在池化层依旧采用 2D 池化,不可避免地丢失了一部分时间维度信息。与 3D 卷积类似,3D 池化同样增加了时间维度,3D 池化与 3D 卷积一样输出也是 3D 的。本文提出的 3D CNN 结构的卷积层与池化层均采用 3D。

卷积核大小会对图像处理有明显的影响。如果使用的卷积核较小,对图像感兴趣的动作信息增强效果不明显,反而可能加大同频带噪声,掩盖图像原有的一些细节信息;如果使用的卷积核较大,卷积的计算量也随之增大,计算复杂度较高。所以,合适的卷积核大小对于整体的视频图像来说,至关重要。根据大量的 2D CNN 结构在人体动作识别上的应用发现,卷积核大小为 3 ×3 卷积网络结构的效果最好[13](如图 3 所示),一般 3D CNN 的卷积核的时间维度均为 3,因此本文提出的网络结构使用大小为 3 ×3 ×3 的 3D 卷积核。

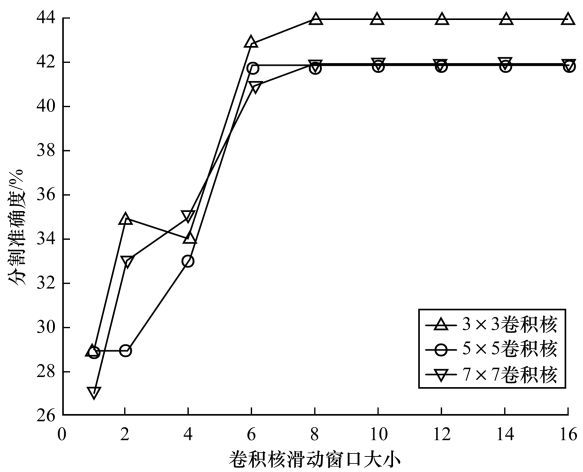

图 3　分割准确度与卷积核滑动窗口大小之间的关系

2.2　改进的 3D CNN 算法网络结构

改进的 3D CNN 采用 5 通道的处理方式,即灰度、x 方向梯度、y 方向梯度、x 方向光流、y 方向光流,并依据第 2.1 节所述,卷积核均采用3 ×3 ×3 的大小。本文设计的卷积网络除了 5 个通道的第 1 层外,还含有 5 层卷积、5 层池化以及 2 层全连接与 softmax 分类器的输出层,结构如图 4 所示。

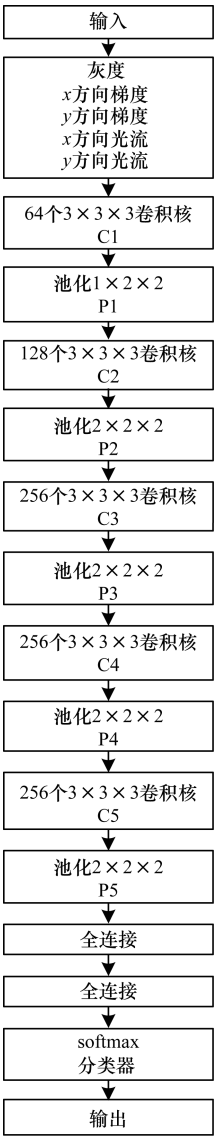

图 4 改进的 3D CNN 网络结构

本文采用 UCF101 数据库训练本节的 3D CNN 网络结构参数,输入尺寸为 5×16×128×171,其中,5 表示 5 个通道,16 表示输入 16 帧无重叠划窗的视频,128×171 表示输入视频帧的大小。网络结构从第 1 层卷积层到第 5 层卷积层中的滤波器数量即卷积核数量分别为 64、128、256、256、256,所有卷积层中的卷积核大小为 3×3×3 以适当的 *padding* = true(卷积核类型,以便处理图像边缘)及步长 *stride* = 1(卷积滑动窗口)运算,使各卷积层中输入与输出大小无变化。所有池化层(除第 1 层池化层外)的池化核大小均为 2×2×2 以步长即池化滑动窗口 *stride* = 1 进行最大值池化,每一层池化后的输出信号均是池化前的输入信号大小的 1/8,即经过池化降低了时间空间分辨率,第 1 层池化层的池化核大小为 1×2×2,目的是不过早地缩减视频信息的长度以获得更多的动作信息,最后经过 2 次全连接得到 2 048 维特征向量,将特征向量通过 softmax 分类器可得最终结果。

3 实验与结果分析

为了对本文所提方法进行有效评估,在公共基准的 UCF101 数据库上与 iDT 算法[12-13]、LRCN 算法、P-CNN 算法进行对比实验。UCF101 数据库包含 101 类动作、共 6 680 段在真实场景中拍摄的视频。数据集包括了多个场景类型、视角区域以及光照、背景变化与摄像头移动。为了更好地进行实验,将数据集中所有视频帧尺寸改为 112×112。实验结果见表 1,可以看出本文提出的算法在数据库 UCF101 上有较高的识别准确率。

表 1 4 种算法对比实验结果 %

算法	识别准确率
iDT 算法	76.2
LRCN 算法	82.9
P-CNN 算法	88.5
本文算法	91.2

iDT 算法是非深度学习算法在动作识别领域中效果最佳的算法,此算法依据先验知识提取动作识别的低层特征稠密光流轨迹对运动场景进行分割,并使用运动边界编码,利用 SVM 分类器进行识别分类[14-16]。LRCN 算法[17]通过 2D CNN 和递归神经网络(Recurrent Neural Network,RNN)结合解决了在 RNN 中出现梯度消亡现象,此算法使用长短记忆网络(Long Short Term Memory,LSTM)对视频建模,将上一时刻视频帧通过 CNN 的输出作为下一时刻 RNN 的输入。P-CNN 算法的详细描述见第 1 节。本文算法利用先验知识计算灰度、*x* 方向梯度、*y* 方向梯度、*x* 方向光流、*y* 方向光流较好地训练了网络参数,设计 5 层 3D 卷积、5 层 3D 池化使动作视频的时间维度信息得到有效的表征,通过 2 次全连接到 softmax 分类器取得了较好的实验结果。

对比实验均在 Linux 14.04 操作系统、CPU(6 核,1.6 GHz)、GPU GTX1080(显存 8 GB)下进行。使用 Caffe 框架,基于同一 UCF101 数据库,LRCN 算法、P-CNN 算法以及本文算法的运行时间如表 2 所示。

表 2 3 种算法的运行时间比较

算法	处理器	运行时间/h
LRCN 算法	GPU	12.2
P-CNN 算法	GPU	6.7
本文算法	GPU	2.2

因为 iDT 算法未找到 GPU 版本,所以没有进行

对比。所有算法的运算速度均是在无重叠的视频帧下获得的。由表 2 可以看出,本文算法在运行时间上更有优势。

4　结束语

本文基于 P-CNN 算法,对人体动作进行识别,将视频通过预处理变为多通道作为网络输入,有效地训练了网络参数,通过 5 层 3D 卷积、5 层 3D 池化提取出含更多时间维度与空间维度的动作特征,并经过 2 层全连接并通过 softmax 分类器输出,提高人体动作识别的准确率。实验结果表明,该算法能较好地适应复杂场景,并具有较快的运行速度。下一步将对网络进行补充与调整,使其能够更好地适应复杂场景下的人体动作识别问题,进一步提高识别准确率。

参考文献

[1] 郑胤,陈权崎,章毓晋. 深度学习及其在目标和行为识别中的新进展[J]. 中国图象图形学报,2014,19(2):175-184.

[2] 徐勤军,吴镇扬. 视频序列中的行为识别研究进展[J]. 电子测量与仪器学报,2014,28(4):343-351.

[3] 雷庆,陈锻生,李绍滋. 复杂场景下的人体行为识别研究新进展[J]. 计算机科学,2014,41(12):1-7.

[4] 杜友田,陈峰,徐文立,等. 基于视觉的人的运动识别综述[J]. 电子学报,2007,35(1):84-90.

[5] 李岳云,许悦雷,马时平,等. 深度卷积神经网络的显著性检测[J]. 中国图象图形学报,2016,21(1):53-59.

[6] DONAHUE J, HENDRICKS L A, GUADARRAMAS, et al. Long-term recurrent convolutional networks for visual recognition and description[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 39(4):677-691.

[7] 李瑞峰,王亮亮,王珂. 人体动作行为识别综述[J]. 模式识别与人工智能,2014,27(1):35-48.

[8] 单言虎,张彰,黄凯奇. 人的视觉行为识别研究回顾、现状及展望[J]. 计算机研究与发展,2016,53(1):93-112.

[9] KARPATHY A,TODERICI G,SHETTY S,et al. Large-scale video classification with convolutional neural networks[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C. ,USA:IEEE Press,2014:1725-1732.

[10] SIMONYAN K,ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Neural Information Processing Systems,2014(1):568-576.

[11] CHERON G,LAPTEV I, SCHMID C. P-CNN: Posed-based CNN features for action recognition[C]// Proceedings of IEEE Interational Conference on Computer Vision. Washington D. C. ,USA:IEEE Press, 2016:3218-3226.

[12] JI S,XU W,YANG M,et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(1):221-231.

[13] CHAQUET J M, CARMONA E J, FERNANDEZ-CABALLERO A. A survey of video datasets for human action and activity recognition[J]. Computer Vision and Image Understanding,2013,117(6):633-659.

[14] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2017-09-01]. http://cn. arxiv. org/pdf/1409. 1556v6.

[15] 徐渊,许晓亮,李才年,等. 结合 SVM 分类器与 HOG 特征提取的行人检测[J]. 计算机工程,2016,42(1):56-60.

[16] ZHANG B,WANG H. Encoding scale into fisher vector for human action recognition[C]//Proceedings of Visual Communications and Image Processing. Washington D. C. ,USA:IEEE Press,2016:1-4.

[17] BEZAK P. Building recognition system based on deep learning[C]//Proceedings of International Conference on Artificial Intelligence and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2016:1-5.

编辑　刘盛龄

(上接第 258 页)

[9] LIU C,KANG S,LI F,et al. Canopy leaf area index for apple tree using hemispherical photography in arid region[J]. Scientia Horticulturae,2013,164:610-615.

[10] NILSON T. A theoretical analysis of the frequency of gaps in plant stands[J]. Agricultural Meteorology,1971, 8:25-38.

[11] OTSU N. A thresholding selection method from gray-level histogram[J]. IEEE Transactions on Systems Man and Cybernetics,1979,9(1):62-66.

[12] 傅隆生,孙世鹏,MANUEL V A,等. 基于果萼图像的猕猴桃果实夜间识别方法[J]. 农业工程学报,2017,33(2):199-204.

[13] 桂媛. 基于 SVM 的植物叶片类别识别研究[J]. 信息技术与信息化,2015(9):148-150.

[14] 张善文,张云龙,尚怡君. 一种基于 Otsu 算法的植物病害叶片图像分割方法[J]. 江苏农业科学,2014,42(4):337-339.

[15] 刘柯楠,杨启良,郭浩,等. 基于不同滴灌方式的苹果幼树根系生长的三维可视化模拟[J]. 中国生态农业学报,2013,21(6):744-751.

[16] 刘健庄,栗文青. 灰度图象的二维 Otsu 自动阈值分割法[J]. 自动化学报,1993,19(1):101-105.

[17] 景晓军,李剑峰,刘郁林. 一种基于三维最大类间方差的图像分割算法[J]. 电子学报,2003,31(9):1281-1285.

编辑　吴云芳