计 算 机 工 程
Computer Engineering

文献标识码：A 中图分类号：TP391.4

基于 3D 卷积神经网络的人体动作识别

张 瑞[1,2]，李其申[2]，储 珺[2]
（南昌航空大学信息工程学院[1]，南昌 330063；江西省图像处理与模式识别重点实验室[2]，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，提出一种基于 3D 卷积神经网络结构的人体动作识别方法。以连续的 16 帧视频为一组，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，有效地训练网络参数，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，最终通过两层全连接与 softmax 分类器得到识别分类结果。通过与 iDT、P-CNN、LRCN 三种典型算法比较，实验结果表明，本文提出的方法识别准确率更高，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

Human Action recognition based on 3D convolution neural network

ZHANG Rui[1,2],LI Qishen[2],CHU Jun[2]
(School of Information Engineering, Nanchang Hangkong University[1], Nanchang 330063, P.R.China; Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition[2], Nanchang 330063, P.R.China)

Abstracts: Human action diversity, scene noise, the camera motion angle changes and other factors increase the difficulty of human action recognition. Thus, this paper proposes a 3D convolution neural network method to recognize human action . We use relevant techniques to identify the classification results. Firstly, successive 16 frames of the video are divided into a group. Secondly, we process the group data by the gray, gradient-x, gradient-y, optflow-x and optflow-y as multichannel. It effectively training network parameters. Thirdly, the extracted features are obtained using 5 layers 3D convolution, 5 layers 3D pooling to increase time dimension information, Finally, the recognition results are obtained by 2 layers full connection and softmax classifier. Compared with the other three typical algorithms: iDT, P-CNN, LRCN, this algorithm effectively improves the accuracy of human action recognition and has faster running speed.

Keywords:human action recognition; multichannel;3D convolution; 3D pooling; time dimension

---

基金项目：国家自然科学基金（61663031）；江西省自然科学基金项目（20132BAB201046）；南昌航空大学研究生专项创新资金（YC2016009）

作者简历：张瑞，（1993-），女，河南省洛阳市人，南昌航空大学计算机应用技术硕士研究生，研究方向为图像处理与模式识别，E-mail：546029094@qq.com；李其申，（1975-），男，河北省衡水市人，博士，南昌航空大学副教授，研究方向为图像处理与模式识别；储珺，（1967-），女，教授、博士生导师，博士，江西省图像处理与模式识别重点实验室副主任，江西省中青年学科带头人

通迅地址：330063 江西省南昌市丰和南大道 696 号南昌航空大学信息工程学院

0 概述

人体动作识别在计算机视觉中是极具挑战性的课题之一，涉及模式识别、图像处理、计算机视觉、人工智能等多个学科领域。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中的参数膨胀问题，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作为输出，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很好的效果[4]。卷积神经网络（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，包括分类、检测、分割等任务[5]。

目前基于卷积神经网络的人体动作识别方法主要分为两大类：一类是将动作视频帧通过二维卷积神经网络或与其他神经网络相结合提取特征及动作分类，例如长时递归卷积神经网络[6]（Long term Recurrent Convolutional Network,LRCN）等，这类方法一般都是针对图像进行处理，对基于视频分析的问题，通常忽略了时序信息。另一类则是将二维卷积拓展到三维卷积，加入时间维度，直接对视频进行分析捕获，这一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，多个连续帧数据进行卷积得到多个特征图[7-8]。三维卷积比二维卷积更好地表达视频中的有效运动信息，具有一定的优越性。Karpathy 等使用多分辨率的卷积神经网络，将视频数据流分为低分辨率和原始分辨率的数据流[9]。Simonyan 等同样使用两个数据流的卷积神经网络，将视频分为静态帧数据流和帧间数据流，均采用卷积神经网络提取特征后进行合并，通过全连接经 SVM 分类器做分类识别[10]，该算法虽然取得较好地效果，但是网络结构复杂，时间及空间复杂度较高。Cheron 等使用单帧数据和光流数据捕获运动信息，其卷积网络的第一层包含灰度数据、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[11]，该算法在机场视频监控中有较好的识别率，但在其他场合的视频中没有展现较好地识别准确率[1-5]。

本文基于 P-CNN 算法模型，设计 5 层 3D 卷积，5 层 3D 池化，两次全连接与 softmax 分类器得出最终结果，实验表明对于日常的人类动作有更好的识别准确率。

1 经典的 3D 卷积神经网络人体动作识别算法

2015 年 Cheron 等提出具有时间维度信息的 3D 卷积神经网络[11]，是人体动作识别方面经典的 3D 卷积神经网络，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中的数据进行连接，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二维特征图[12]，卷积过程可表示为：

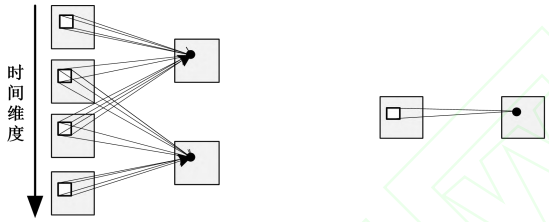
$$v_{ij}^{xy}=\tanh\left(b_{ij}+\sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}w_{ijm}^{pq}v_{(i-1)m}^{(x+p)(y+q)}\right) \tag{1}$$

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，$\tanh(\cdot)$是双曲正切函数，b_{ij} 是第 i 层卷积层的第 j 个特征图的偏差，m 是第$(i\text{-}1)$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$(i\text{-}1)$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，所以将二维卷积拓展到三维卷积，从空间维度和时间维度上计算特征。

3D 卷积是数据集通过三维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：

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

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，tanh(·)是双曲正切函数，b_{ij} 是第 i 层卷积层的第 j 个特征图的偏差，m 是第$(i\text{-}1)$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。三维卷积与二维卷积相比增加了时间维度，如下图 1(a)、(b)所示：



(a) 3D 卷积　　(b) 2D 卷积

图 1 3D 卷积(a)与 2D 卷积(b)比较

（注：(a)图箭头形状相同表示权值共享）

1.2 Cheron 3D 卷积神经网络

Cheron 依据 3D 卷积原理，基于机场视频监控 TRECVID 数据库，提出了一种 3D 卷积神经网络结构用于人体动作识别，具体结构如图 2：

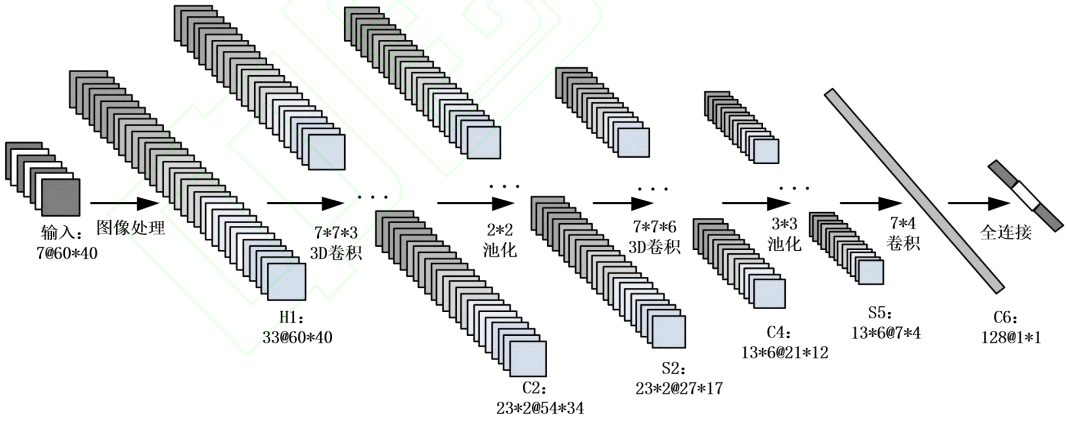


图 2 Cheron 算法的 3D 卷积神经网络结构

此网络结构以 7 帧 60×40 的视频图像为一组，作为 3D 卷积神经网络的初始输入：

（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，两个相邻连续帧计算光流，得到第一层共 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，为了增加其特征图的数量，即提取不同的特征信息，用两种不同的卷积方式进行卷积，得到两组共 46 个特征图，每组是 23 个特征图；

（3）对第二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

的特征图与第二层比，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，同样与第二层一样，为了增加特征图的数量，两组特征图都采用 3 种不同的卷积方式，即第四层卷积层得到 6 组不同的特征图，每组有 13 个特征图；

（5）对第四层的特征图进行窗口大小为 3×3 的池化，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：每个特征图与第五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 Cheron 3D 卷积神经网络算法的不足

在 Cheron 等提出的 3D 卷积神经网络算法中，3D 卷积神经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，其输入是连续的 7 帧为一组，但不是所有的人体动作都能在 7 帧之内表达完成，而 Cheron 等提出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，Cheron 等提出的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，仍丢失了部分时间维度信息。基于上述考虑，提出一种改进的 3D 卷积神经网络算法。

2 改进的 3D 卷积神经网络算法

2.1 3D 卷积、池化操作，以及卷积核参数的选择

Cheron 等提出的算法，其网络结构只在卷积层使用三维卷积，而在池化层依旧采用二维池化，不可避免导致一部分时间维度信息的丢失；与 3D 卷积和 2D 卷积相比类似，3D 池化同样增加了时间维度，3D 池化与 3D 卷积一样输出也是三维的。本文所提出的 3D 卷积神经网络结构的卷积层与池化层均采用三维。

卷积核大小会对图像处理有明显的影响，如果使用的卷积核过小，对于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；如果使用的卷积核过大，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 卷积神经网络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），一般的 3D 卷积神经网络中卷积核的时间维度均为 3，因此本文所提出的网络结构使用大小为 3×3×3 的 3D 卷积核。

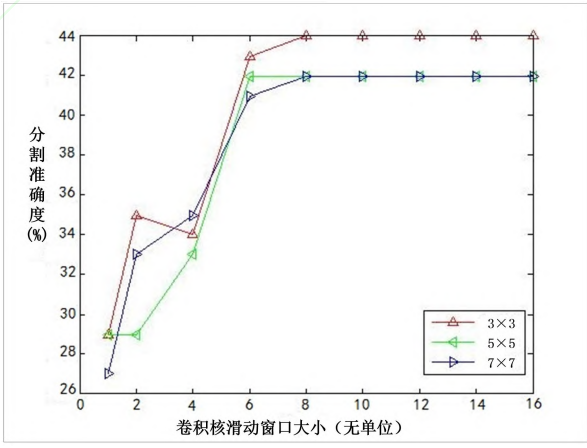


图 3 不同大小卷积核的动作识别分割准确度比较（卷积核大小 3×3、5×5、7×7）

2.2 改进的 3D 卷积神经网络算法结构

改进的 3D 卷积神经网络依旧采用 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 2.1 所述，卷积核均采用 3×3×3 的大小。本文设计的卷积网络除处理为 5 通道的第一层外，还含有 5 层卷积，5 层池化以及两层全连接 与 softmax 分类器的输出层；结构如下图 4 所示。

图 4 改进的 3D 卷积神经网络结构

本节所描述的 3D 卷积网络结构,训练网络结构参数所用的数据库是 UCF101，预测类标签属于 101 种不同的动作，所有的视频帧大小为 128×171，训练此 3D 卷积网络所输入的尺寸为 5×16×128×171（5 表示 5 通道：灰度、*x* 方向梯度、*y* 方向梯度、*x* 方向光流、*y* 方向光流，16 表示输入的是 16 帧无重叠划窗的视频帧数，128×171 表示输入视频帧的大小）；用于网络训练的输入是随机抽取处理的 16 帧视频。网络结构从第一层卷积层到第五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，使得各卷积层中输入与输出大小无变化。所有池化层（除第一层池化层外）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比池化前的输入信号大小降低了 8 倍，即经过池化降低了时间空间分辨率，第一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便获得更多的动作信息，最后经过两次全连接得到 2048 维特征向量通过 softmax 分类器得到最终结果。

3. 实验结果对比及分析

为了对本文所提的方法进行有效地评估，在公共基准的 UCF101 数据库上与文献所提出的方法进行对比试验；UCF101 数据库包含 101 类动作、共 6680 段视频，均为真实场景下拍摄，充分考虑多个场景类型、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，数据集中所有视频采用尺寸改为 112×112。实验结果与 iDT 算法、LRCN 算法、P-CNN 算法比较，如下表 1。

表 1 在数据库 UCF101 上的对比实验结果

人体动作识别算法	识别准确率（%）
iDT[12-13]	76.2
LRCN	82.9
P-CNN	88.5
本文算法	91.2

从上述表中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。

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[14-17]。LRCN 算法[18]是 2D 卷积神经网络(2D Convolutional Neural Network,2D CNN)、递归神经网络(Recurrent Neural Network,RNN)相结合的长时递归卷积神经网络(Long term Recurrent Convolutional Network,LRCN)解决了 RNN 中出现梯度消亡现象,此算法使用 LSTM（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 算法即为第二部分 Cheron 所提出的经典 3D 卷积网络算法，采用 3D 卷积与 2D 池化的卷积神经网络。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，通过两次全连接到 softmax 分类器取得了较好的实验结果。

实验均在 Linux14.04 操作系统、CPU（6 核，1.6GHZ）、GPU GTX1080（显存 8G）使用 Caffe，基于同一数据库 UCF101，LRCN 算法、P-CNN 算法以及本文算法的运行时间如下表 2：

表 2 四种算法的运行时间比较

算法名称	LRCN	P-CNN	本文算法
CPU/GPU	GPU	GPU	GPU
运行时间（小时）	12.2	6.7	2.2

注：iDT 算法未找到 GPU 版本不做对比，所有算法的运算速度均是在无重叠的视频帧下获得的。

从上表可以看出，本文所提出的算法在运行时间上更有优势。

4. 结束语

本文基于 Cheron 等提出的 3D 卷积神经网络算法，对人体动作识别问题，将视频通过预处理变为多通道作为网络输入，更加有效地训练了网络参数，再通过 5 层 3D 卷积、5 层 3D 池化使得提取的动作特征含有更多时间维度与空间维度的信息，最终经过两层全连接通过 softmax 分类器，提高了人体动作识别的准确率，较好地适应场景不同的复杂变化，具有较快的运行速度。下一步将通过研究更为复杂场景下的人体动作识别，对所提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，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]. 中国图像图形学报, 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]. 中国图像图形学报, 2016, 21(1):53-59.

[6] Donahue J, Hendricks L A, Guadarrama S, et al. Long-term Recurrent Convolutional Networks for Visual Recognition and Description[J]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015(6):2625-2634.

[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]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014:1725-1732.

[10] Simonyan K, Zisserman A. Two-Stream Convolutional Networks for Action Recognition in Videos[J]. Neural information processing systems, 2014:568-576.

[11] Cheron G, Laptev I, Schmid C. P-CNN: Posed-Based CNN Features for Action Recognition[J]. IEEE Interational Conference on Computer Vision (ICCV), 2015(10):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[J]. Computer Science, 2014.

[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]// Visual Communications and Image Processing. IEEE, 2016:1-4.

[17] Peng X, Zou C, Qiao Y, et al. Action Recognition with Stacked Fisher Vectors[C]// European Conference on Computer Vision. Springer International Publishing, 2014:581-595.

[18] Bezak P. Building recognition system based on deep learning[C]// International Conference on Artificial Intelligence and Pattern Recognition. IEEE, 2016:1-5.