应用研究

DOI:10.19695/j.cnki.cn12-1369.2020.09.24

基于深度学习的视频场景下的人体动作识别研究

李轩 孙美鑫
(沈阳航空航天大学电子信息工程学院,辽宁沈阳 110136)

摘要:动作识别是通过计算机处理分析连续的视频流来自动识别一些人类的行为动作。本文首先在vgg传统深度学习模型的基础上改进网络结构提出了一种用于动作识别的新型网络结构。然后建立3dcnn,利用目标检测中的光流提取法创新性的进行模型融合。并由实验结果证明本文提出的模型在动作识别上有着良好的表现效果。

关键词:深度学习;3D卷积;光流法;动作识别
中图分类号:TP391.41 **文献标识码**:A **文章编号**:1007-9416(2020)09-0062-03

0 引言

人体行为多种多言,从简单到复杂可以分为四种方面:(1)由人体部位简单运动形成的动作,如鼓掌、踢腿等;(2)由简单动作形成的个人行为:如快走、慢跑、蹦;(3)交互行为:一种小群体的相互行为。如打球、看书等;(4)群体行为:类似于集体活动,开会等[1]。当前人体行为识别的研究主要分为两个子任务:行为分类和时序行为检测。行为分类一般是对分割好的视频片段给出一个行为的类别标签,每一个视频片段仅包含一个行为实例。然而,现实生活中大部分视频都是未分割的长视频,要明确的是,时序行为检测是行为分类更复杂一级的研究任务,正如图像识别任务中目标检测与图像分类的关系一样。行为分类是时序行为检测的基础[2]。

为提升识别的准确性越来越多的人开始利用目标检测中的光流法结合3D-CNN来进行实验。3D卷积提取出的时间维度图像中的不同像素之间的特征关系,光流是反应不同像素间形成差异的运动信息,利用卷积提取出特征,再输入到网络中去[3]。

1 数据集的介绍

本次设计用的数据集是KTH数据集。KTH数据集是动作识别领域的传统数据集,目前在现有的数据库中包含了6种不同行为动作,分别是拳击,握手、挥手、慢跑、快跑和走路。是由25个人在实际生活中的室内外不同场景进行展示,还包括同一个人在不同情景下的不同着装。

这些视频是由固定的静态摄像机在不同场景以25fps每帧的速度进行拍摄的,这些采集到的时间序列的分辨率是160×120,一共包含了2391个序列。视频的平均长度为4秒。本次实验将所有采集到的时间序列以8,8,9分配给训练集,验证集和测试集。由上述的6个动作在4个变换的场景中完成。下载后的数据集以文件形式存储,每个文件包含4个子序列,序列以avi格式存储。

2 实验所用的模型

自从,深度学习在目标检测上进行应用,不同类型的卷积结构开始被陆续提出。例如反卷积和转置卷积等,但这些都是对数据中的单张图片进行操作。其中,3D-CNN是由Tran等人研究提出的。它主要被用于视频中的动作识

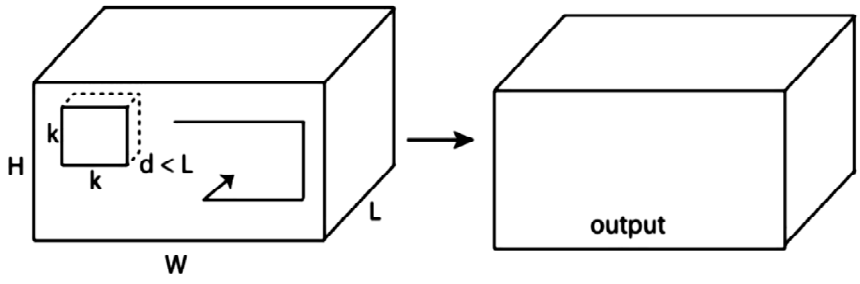

图1 3D卷积示意图

收稿日期:2020-06-24

作者简介:李轩(1967—),男,吉林辽源人,硕士,副教授,研究方向:信号与信息处理;孙美鑫(1995—),女,辽宁本溪人,研究生,研究方向:无线与移动通信。

别。如图1所示，为3D卷积的卷积过程。如果采用不同大小的卷积核对该立方体进行卷积操作，就会得到多种相对丰富的时间和空间特征。

图1展示了3D的完整结构图以及它与2D卷积的不同之处，它是利用连续三帧的图像信息，并且卷积层中的每一个特征图都会与上一层结构相连接，以此来充分提取到视频场景下的运动信息。

其中，卷积中的具体卷积层的特征图的位置(x，y，z)的值v由下面的式(1)计算得到：

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

由上式所示，其中R_i表示在时间维度上的大小，用w_{ijm}^{pqr}来表示权值的大小，

本次研究是在此基础上进行改动，输入的是采集到的尺寸大小，本次实验是60×80×1，然后经过第一层卷积层，用了32个卷积核，大小为3×3。通过堆叠3×3卷积核来增强感受野的大小，并且与传统的5×5和7×7相比减少了一些参数，还可以一定程度上增加非线性映射，尺寸不发生改变。然后经过bp处理。输入是一个小批量的值：B={x_i…m}要学习的参数：Y，β。

具体的算法如下式所示：

$$\mu_B\rightarrow\frac{1}{m}\sum_{i=1}^{m}x_i \tag{2}$$

$$6^2\beta\rightarrow\frac{1}{m}\sum_{i=1}^{m}(x_i-\mu\ \beta)^2 \tag{3}$$

$$\hat{x}_i\rightarrow\frac{x_i-\mu_B}{\sqrt{6^2\beta+C}} \tag{4}$$

$$y_i\rightarrow Y\hat{X}_i+\beta \tag{5}$$

使用的小卷积核和常见的大卷积核相比有着以下两点优势。能够增强神经网络的容量和模型复杂度，可以减少卷积的参数个数。在网络全连接层的训练阶段Dropout设置为0.5，可以防止模型的过拟合。激活函数使用的是Relu，其在正向部分为1，可以避免梯度爆炸等问题。并且，Relu函数计算速度快，加快了网络的训练。以下为其具体的公式：

$$\text{Relu}=\max(x,0)=\begin{cases}0, & x<0\\ x, & x>0\end{cases} \tag{6}$$

3 模型融合

光流为optical flow，它通过输入的连续的图像序列来检测物体微小的动作变化。光流在图像中的含义就是动作向量(motion vector)(u，v)，分别表示位移在x和y方向上的变化率光流常用于计算图像中各个位置的瞬时移位，是依据视频中连续两帧之间的各像素相关性而计算得到的。记录视频中的两帧对应的像素点的灰度值，然后进行两帧相减，取绝对值为最终的计算结果。使得移位过后，保持各像素点的位置在下一时刻的对应值一致。

实验中首先通过目标检测方法提取出光流，本次实验使用opencv中的光流法对视频中的动作进行识别。相关的算法设计思路如下：首先采集到的第i帧的图像，然后获得它后面的图像。接下来分别对两个图像进行去噪声处理，利用去后的结果得到光流场。接下来计算出局部动能场，然后利用边缘检测的方法，由于视频中使运动的目标，所以要尽可能的提取出运动部分的图像，计算出其对应的中心位置。

输入的是video length 30，经过计算得出光流为29。光流和RGB的3个通道不一样，光流是沿着x方向和y方向各有一个光流图，这两个方向单独通过网络层，这个训练的过程可以用来刻画视频中流动的运动信息。单个方向的光流输入就是batchsize，这里设置为64，29，w，h，1。Conv1−1是一个3D卷积核，大小为1×3×3，输出大小为32，进行了一次下采样。然后对特征值做归一化处理，这里不影响输出。Conv1−2和conv1−1的原理相同，此时的输出由30×40变为28×38，产生了微小的变化是因为使用了(3×3)的卷积核在特征图上做了对应的滑动平均计算。由于这里没有使用padding，padding可以有效保留边界信息和保持尺寸不发生改变。卷积在特征图上进行了28次运算，所以得出了28×38的输出。然后是pool1−1，这是个MaxPooling3D pooling，用2×2×2的立方块大小划分出对应区域，取出窗口里面的最大值。这样由(29，28，38)到(14，14，19)，这里是因为使用了padding=valid，所以最后一次计算不用了。以下的两个卷积层中的操作原理与第一层卷积层的原理相同。然后通过三个全连接层，最后把两个3dcnn进行融合起来，通过分类器，实现对数据集中6类动作的识别。

4 模型的训练过程及结果

使用到的KTH数据库作为本次实验的数据集，由于硬件配置的缺陷，有的是谷歌的GPU作为服务器进行模型的训练。用的是keras作为神经网络的框架来搭建的模型。Keras的核心数据结构是模型。模型是用来组织网络层的方式。模型有两种，一种叫Sequential模型，另一种叫Model模型。Sequential模型是一系列网络层按顺序构成的一种结构，是单输入和单输出的，层与层之间只有相邻关系，是最简单的一种模型。Model模型是用来建立更复杂的模型的。这里先介绍简单的Sequential模型的使用，上面第三章

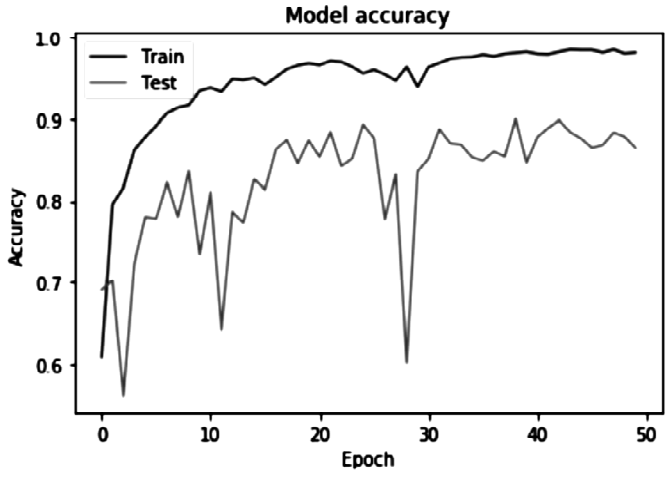

图2 模型的准确率示意图

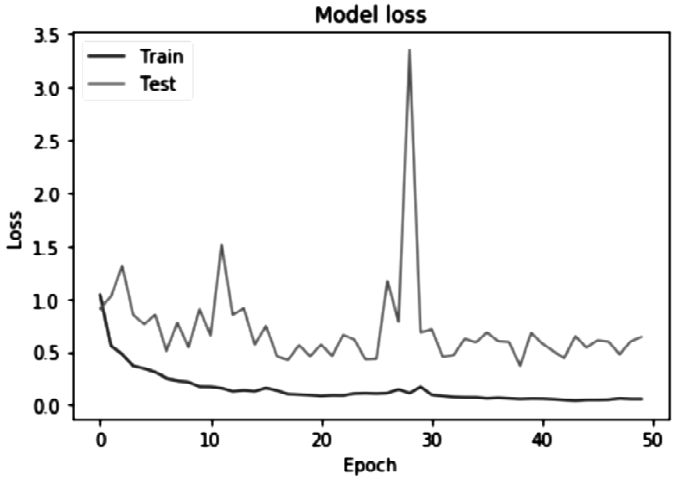

图3 模型的 loss 示意图

的网络模型就是用model搭建的。

Batch-size设置为64，一共训练了50轮。用matplotlib绘制出模型的loss和准确率图像，如图2所示。

由图3所示，本次实验设计的网络模型在经过多次的训练得到了良好的效果。在训练了50轮后，模型在训练集和测试集的准确率均在85%以上，通过框架的内置函数对模型进行评估得到最大的预测精度达到了0.912037037037037。而loss值这边在50轮训练过后也能很好的控制在0.5左右，说明模型具有有效性。

5 结语

在本次研究中我们通过以传统的VGG为基础进行了模型上的改进，并融合了目标检测中的光流提取法。经过实验验证了这个双流卷积模型在数据集的训练集和测试集上都有着良好的表现。通过学习率和dropout的设置，有效地避免了模型出现过拟合情况。在数据集Kth的训练过程最大的准确率可以达到90%以上，并且能够准确识别出6个动作。

参考文献

[1] 郑潇,彭晓东,王嘉璇.基于姿态时空特征的人体行为识别方法[J].计算机辅助设计与图形学学报,2018,30(9):1615-1624.

[2] 张承玺.固定场景下的人体姿态识别[D].哈尔滨:哈尔滨工业大学,2014.

[3] 张苗辉.基于视觉系统的行人检测与跟踪方法研究[D].上海:上海交通大学,2013.

Research on Human Action Recognition in Monitoring Scenes Based on Depth Learning

LI Xuan,SUN Mei-xin

(School of Electronic Information Engineering, Shenyang Aerospace University, Shenyang Liaoning 110136)

Abstract: Action recognition is to automatically recognize some human actions through computer processing and analysis of continuous video streams. Firstly, this paper improves the network structure based on vgg's traditional deep learning model and proposes a new network structure for motion recognition. Then 3dcnn is established, and the model fusion is carried out by using the optical flow extraction method in target detection. The experimental results show that the model proposed in this paper has good performance in motion recognition

Key words:deep learning; 3Dconvolution; optical flow method; motion recognition