面向特定场景的行为识别算法的研究

王华锋 张 鹏
（北方工业大学信息学院，北京 100144）

摘 要：本文对深度学习在行为识别中的应用进行了研究。本研究进行的创新主要体现为以下两个方面：首先提出了将骨架信息与卷积网络相融合的行为识别模型。其次在网络中设计了一个新颖的时空卷积模块并创新性地引入了注意力机制，用于对人体骨架运动情况进行时空相关性建模。

关键词：行为识别；深度学习；注意力机制
中图分类号：TP391 文献标识码：A 文章编号：2096-4390（2020）28-0012-02

1 方法

1.1 融合骨架信息的卷积神经网络

人体的骨架图是对人体轮廓的抽象，是一种包含了人体主要关节点信息和骨架连接结构信息的图形。与传统的卷积神经网络不同的是，本此研究提出的网络输入中包含了人体的骨架信息。其特征用如下公式来表示：

$$G=(V, E) \tag{1}$$

$$E=\left\{v_{ti}v_{tj} \mid (i, j) \in H\right\} \tag{2}$$

其中 V 为骨架图中所有关节点的集合，包含了两个子集，第一个子集描述每个帧的骨架内连接，第二个子集包含帧间边缘信息，代表了关节随着时间的运动轨迹。在此网络的传播规则如下：

$$f\left(D^i, I\right)=\sigma\left(ID^iW^i\right) \tag{3}$$

W^i 表示第 i 层的权重矩阵，σ 表示 ReLU 激活函数，I 为输入邻接矩阵，它是某一关节点和其他关节点连接特征的数学表示。首先在邻接矩阵左侧乘以特征矩阵，完成了特征的融合操作，然后再在右侧乘以权重矩阵，实现了传统卷积神经网络加权操作。具体到 x 层卷积其输出为：

$$f_{\text{out}}\left(v_{ti}\right)=\sum_{v_{tj}\in B(v_{ti})}\frac{1}{Z_{ti}\left(v_{tj}\right)}f_{in}\left(\mathbf{p}\left(v_{ti}, v_{tj}\right)\right)\cdot \mathrm{w}\left(v_{ti}, v_{tj}\right) \tag{4}$$

$$\mathbf{p}\left(v_{ti}, v_{tj}\right)=v_{tj} \tag{5}$$

$$\mathbf{w}\left(v_{ti}, v_{tj}\right)=\mathbf{w}'\left(l_{ti}\left(v_{tj}\right)\right) \tag{6}$$

其中，p 为采样函数，表示对输入信息的特征提取，列举了位置 x 的近邻区域，K 为卷积核的维度，加权函数 w 提供了一个权重向量，来计算其与经过采样的输入向量的内积。在网络中将骨骼节点的近邻划分为固定的 K 个子集，W 通过索引一个 K 维的张量来实现。则上述公式被重新定义如下：

$$f_{\text{out}}\left(v_{ti}\right)=\sum_{v_{tj}\in B(v_{ti})}\frac{1}{Z_{ti}\left(v_{tj}\right)}f_{in}\left(\mathbf{p}\left(v_{ti}, v_{tj}\right)\right)\cdot \mathbf{w}\left(v_{ti}, v_{tj}\right) \tag{7}$$

$$f_{\text{out}}\left(v_{ti}\right)=\sum_{v_{tj}\in B(v_{ti})}\frac{1}{Z_{ti}\left(v_{tj}\right)}f_{in}\left(v_{tj}\right)\cdot \mathbf{w}\left(l_{ti}\left(v_{tj}\right)\right) \tag{8}$$

1.2 注意力机制的设计

本次研究在上述提到的骨架网络结构里创新性地引入了注意力模块，模块中包括了两种注意力机制：时间注意力机制和空间注意力机制。在空间维度上不同关节点之间存在连通，具有很强的动态性，可以用注意力机制去自适应调节空间维度上节点之间的动态相关性：

$$S=V_s*\sigma\left(\left(X_h^{r-1}W_1\right)W_2\left(W_3X_h^{r-1}\right)^T+b_s\right) \tag{9}$$

$$S'_{i,j}=\frac{S_{i,j}}{\sum_{j=1}^{N}\exp\left(S_{i,j}\right)} \tag{10}$$

其中为前一层的输入，V 是骨架关节点的集合，W 为需要进行学习的参数。在 S 中代表了索引 i 的点和索引 j 的点的相关性。在进行卷积操作时将使用邻接矩阵 A 与空间注意力矩阵 S 一起对两点之间的动态权重值进行更新。

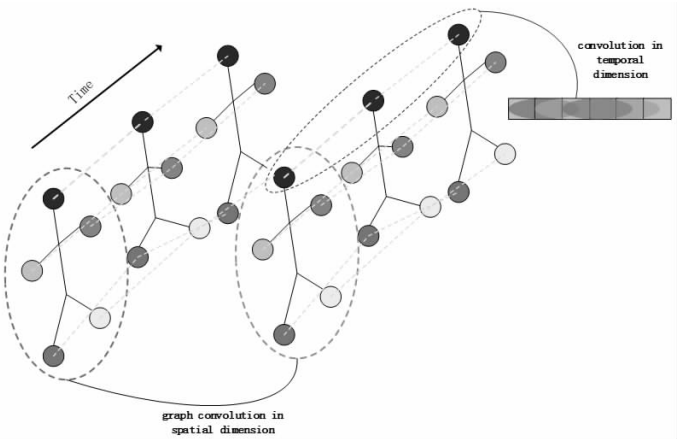

图 1 时空注意力机制示意图

时空注意力模块能够使网络对不同时刻的信息赋予不同的权重。在时间维度上，不同的帧之间的行为特征也存在着相关性，这些相关性在不同点之间也是时刻发生着变化，因此，网络使用一个时间注意力机制去调整数据在各个时间的权重：

$$E=V_e*\sigma\left(\left(X_h^{r-1}U_1\right)U_2\left(U_3X_h^{r-1}\right)^T+b_e\right) \tag{11}$$

作者简介：张鹏（1994–），男，汉族，河北省衡水市人，硕士，研究方向：机器学习。

$$E'_{i,j} = \frac{E_{i,j}}{\sum_{j=1}^{N} \exp\left(E_{i,j}\right)} \tag{12}$$

1.3 数据集的建立

本研究中构建了自己的行为识别骨架数据集。数据集中主要有五类日常的动作：跌倒，坐下，站立，喝水，太极，其中每个动作有 30-40 个视频剪辑，共计 170 个。其中一部分视频片段在实验室拍摄完成，一部分从网络的视频中提取。首先通过 ffmpeg 将视频的帧率转到 30FPS，并将视频的分辨率进行统一调整为 340*256，然后开始对视频中的人体骨架信息进行提取，每个信息元组中包含三组数据包括人体关节点的 2D 坐标和置信度，最终将其保存为 json 文件。

2 实验

本文从传统的卷积神经网络模型出发，引入了骨架网络模型，并对日常中常见的动作进行了分类和识别。实验完成了行为动作的分类任务，但当检测目标在快速运动下的采样会出现模糊，导致骨架信息丢失。

图 2 行为检测结果图(1)

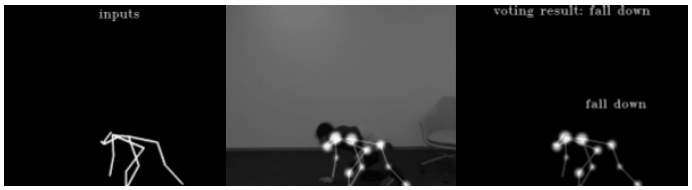

图 3 行为检测结果图(2)

如图所示，骨架关节点的亮度半径代表了其权重的大小，不同关节点在不同动作中的权重有高有低，其中的空间注意力模块决定了不同关节点在行为构成中占的权重，时间注意力决定了不同帧在行为识别中占的比重。

具体而言在跌倒中的肩部关节点的权重较高，关节点亮度范围较大；喝水的动作中手部和胳膊的关节点权重更高，而下肢的权重则相对较低。

在行为识别任务上与四种不同的算法进行了对比，Two Stream Networks[1]以堆叠的光流矢量的形式对运动特征进行了建模，然后通过两个单独的网络进行训练。C3D[2]利用 3x3x3 的三维卷积网络在大规模有监督的数据集上进行训练。LRCN 网络[3]输入图片后先使用传统 CNN 来提取输入图片的特征，然后送入后续的 LSTM 网络进行处理。TSN 网络[4]对整个视频进行稀疏采样，而不是采用单帧分析的方法，然后网络对各个片段进行动作类别的初步预测。可以看出，本文提出的模型在识别率上更有优势。

3 总结与展望

本文从传统的卷积神经网络模型出发，引入了骨架网络模型，并对日常中常见的动作进行了分类和识别。主要的研究成果主要体现在如下方面：

表 1 实验结果

算法模型	Top1 准确率
Two Stream Networks	83.5%
C3D	82.3%
LRCN	82.92%
TSN	86.1%
本文网络	88.6%

(1) 本文提出了一种基于骨架信息的神经网络模型，并且在其中引入了注意力模块，对时间和空间的特征进行了研究。

(2) 构建了行为识别的数据集。对生活中常见的行为如跌倒，坐下，站起等进行了拍摄剪辑，最后利用相关算法对骨架信息进行了采集，得到了经过预处理的行为识别数据集。

参考文献

[1]Christoph R P W, Pinz F A. Spatiotemporal residual networks for video action recognition [J]. Advances in Neural Information Processing Systems, 2016.

[2]Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks [C]//Proceedings of the IEEE international conference on computer vision. 2015.

[3]Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[4]Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition [C]//European conference on computer vision. Springer, Cham, 2016.