应用科技
Applied Science and Technology
中国科技核心期刊
ISSN 1009-671X
CN 23-1191/U

基于多特征融合及Transformer的人体跌倒动作检测算法

刘文龙, 陈春雨

引用本文:
刘文龙,陈春雨. 基于多特征融合及Transformer的人体跌倒动作检测算法[J]. 应用科技, 2022, 49(2): 49–54,62.
LIU Wenlong,CHEN Chunyu. A human falling motion detection algorithm based on multi–feature fusion and Transformer[J]. *Applied science and technology*, 2022, 49(2): 49–54,62.

在线阅读 View online: https://dx.doi.org/10.11991/yykj.202104017

DOI: 10.11991/yykj.202104017
网络出版地址:https://kns.cnki.net/kcms/detail/23.1191.U.20211126.1403.010.html

基于多特征融合及 Transformer 的人体跌倒动作检测算法

刘文龙，陈春雨

哈尔滨工程大学 信息与通信工程学院，黑龙江 哈尔滨 150001

摘 要: 为解决跌倒动作的检测和空间定位问题，本文以 YOLOv3 目标检测算法为基础，提出了一种全新的用于人跌倒动作识别的检测架构。本算法将视频拆分成一系列的图片序列，并在图片序列中指定关键帧。通过 3D 卷积神经网络提取视频序列中的时间维度特征，2D 卷积神经网络提取关键帧中的空间维度特征，经通道融合机制在不同尺度的预测特征层进行通道融合，融合后的特征经过特征金字塔 Transformer 进行深层次特征提取和融合。该算法实现了端到端的训练。通过在自己制作的跌倒动作数据集上进行训练和测试，证明了多特征融合和 Transformer 结构在人体跌倒动作检测中的有效性。

关键词: 动作识别；Transformer 结构；特征融合；空间注意力机制；通道注意力机制；卷积神经网络；YOLOv3；预选框
中图分类号: TP18 **文献标志码:** A **文章编号:** 1009−671X(2022)02−0049−07

A human falling motion detection algorithm based on multi-feature fusion and Transformer

LIU Wenlong, CHEN Chunyu

College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China

Abstract: In order to solve the problems of detecting and locating falling motions, a new detection architecture for human falling motions recognition is proposed in this paper, based on YOLOv3 object detection algorithm. The algorithm divides the video into a series of image sequences, and specifies key frames in the image sequences. 3D convolutional neural network (CNN) is used to extract temporal dimension features in video sequences, and 2D convolutional neural network is used to extract spatial dimension features in key frames. The channel fusion mechanism is used to fuse the prediction features of different scales, and then the fused features are extracted and fused in depth through the feature pyramid Transformer. The algorithm can achieve end-to-end training. The effectiveness of multi-feature fusion and the Transformer structure in human falling action detection is proved through training and testing on the self-made falling action data set.

Keywords: action recognition; transformer structure; feature fusion; spatial attention mechanism; channel attention mechanism; convolutional neural networks; YOLOv3; anchor

当前，人工智能和高性能计算设备的迅速发展，为大规模深度神经网络的训练和部署提供了可行性保障。动作识别作为计算机视觉的研究方向之一，在各个领域具有良好的应用前景。在医疗领域，可以通过动作识别来判断病人的状态，减轻病房看护者的负担。病人一旦发生摔倒等行为可以通过动作识别来进行判断，并提醒看护人员进行及时救助[1]。在安防领域，可以利用动作识别实现对打架斗殴等违法犯罪行为进行识别并记录，为刑事案件的侦破提供有利帮助[2]。在监考任务中，可以通过行为识别来判断应试者是否存在扭头、转身以及东张西望等作弊行为[3]。可见动作识别在现实场景中具有重要的应用价值。传统的视频动作行为识别常常通过人工肉眼观察来完成，耗费了大量的人力和时间，长期工作还会对视力造成损伤，这项任务被计算机取代是大势所趋。基于深度学习的动作检测方法相对于传统方法可以实现无人化、智能化[4]。

陈永彬等[5] 采用了 OpenPose 网络提取人体关

收稿日期: 2021−04−19. **网络出版日期:** 2021−11−29.
基金项目: 国家自然科学基金项目(61871142)；基于人工智能架构的多传感器信息融合与决策系统的研究与实现(KY10800180032)；中央高校基本科研业务费项目(3072020CFT0803).
作者简介: 刘文龙，男，硕士研究生.
陈春雨，男，副教授，博士.
通信作者: 陈春雨，E-mail:springrain@hrbeu.edu.cn.

键点信息，并融合了场景的语义信息，通过判断头部以下关键点的下移量进行判断是否存在摔倒行为。乌民雨等[6]采用了包含长短期记忆（long short-term memory, LSTM）的双流的网络结构，2个流分别输入RGB图片和光流图片。刘峰等[7]通过目标检测器检测到运动的行人，对行人提取了头部运动特征，通过深度森林算法对特征进行分类，来判断行人是否发生摔倒行为。魏振刚等[8]通过对视频中的人物的前景提取得到目标，对前景目标进行形态学处理得到了最小外接矩形，通过分析外接矩形的宽高比筛选出可能存在跌倒行为的样本，再运用统计学方法对异常目标绘制椭圆边界作为特征进行分类，判断是否发生跌倒行为。Tran 等[9]提出了3D卷积（3D convolution, C3D）方法，可以通过训练学习，简单有效地对动作的时空特征进行提取，这项方法常用于对动作进行分类，但单独的C3D方法无法完成对执行动作的目标进行空间定位，对此本文吸收了C3D方法可以提取时空特征的思想，实现了跌倒动作的时空定位。

1　网络结构

得益于GPU性能的高速发展，大规模神经网络的训练及部署成为了可能。动作识别作为计算机视觉领域的重要研究分支，可以通过深度学习方法来解决；动作的定位可以通过借鉴目标检测的方法，采用边界框的回归来实现。对于动作的分类不仅需要空间特征信息，还需要时间维度上的特征信息，因为动作是连贯的，所以必须通过一定的手段来提取时间维度上的特征。因此本文借鉴了YOLOv3目标检测网络[10]和双流网络[11]的设计思想。

1.1　动作识别网络整体结构

本文提出的网络整体结构如图1所示。

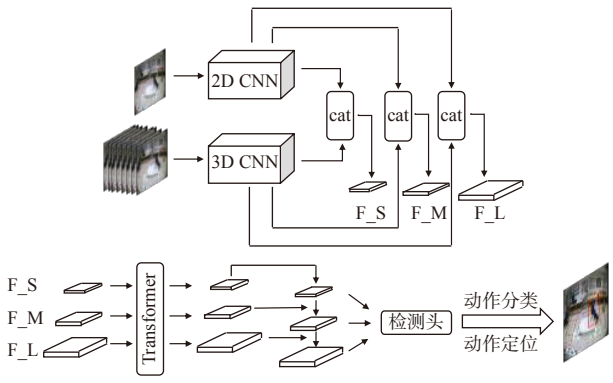

图1　整体网络结构

本文设计了可以实现动作识别与定位的算法，通过在2D卷积神经网络（convolutional neural network, CNN）并联了一路3D卷积神经网络（3D CNN）用于提取时间维度特征，并通过通道融合机制对空间维度特征和时间维度特征在通道维度上进行拼接（cat）。为了提高模型对小目标的检测能力，本算法采用特征金字塔结构，对骨干网络提取3个不同尺度的特征（分别为F_S、F_M、F_L），提升对小目标的检测能力，对这3个尺度的特征采用Transformer结构进行深层次特征融合，提取语义信息。最终经过分类网络和回归网络进行动作分类和动作定位。

1.1.1　特征提取网络

由于动作具有连续性，几乎不能通过单张图片来进行判断动作类型，所以必须获得动作的连续信息，也就是时间维度上的特征。这种时间维度上的特征采用3D卷积来进行提取，他的计算过程可以通过图2来表示。图2中h_{3D}、w_{3D}和l分别为特征图空间维度的高度、空间维度的宽度、时间维度的长度，k_{3D}为卷积核空间维度的宽度和高度，d为卷积核时间维度的长度。

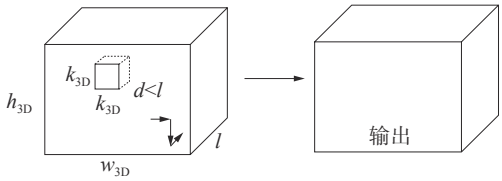

图2　3D卷积计算示意

3D卷积的输入是由视频拆分出来的一系列连续的图片帧。本文中采用的是连续的16帧图片，将其组成一个五维张量，并由B表示批次大小、C表示图片通道数目、N表示图片数目、W表示图片的宽度、H表示图片的高度。3D卷积的卷积核包含3个维度，可以在图2中的长方体数据中的3个方向滑动，最终堆叠到同一个输出，这样就可以提取多个相邻图片之间的特征。本文采用的3D卷积网络为3D ResNet101[12]。

图3所示的是2D卷积示意图，图中h_{2D},w_{2D}分别代表特征图空间维度的高度和宽度。k_{2D}代表卷积核空间维度的宽度和高度。在3D卷积输入的多个连续图片序列中抽取最后一帧作为关键帧，将其输入到2D卷积网络中。2D卷积的卷积核包含2个维度，可以在图3中的长方形数据中的2个方向进行滑动，得到特征图，提取输入图像空间上的信息。本文采用ResNeXt101作为2D网络，该网络具有良好的特征提取能力。

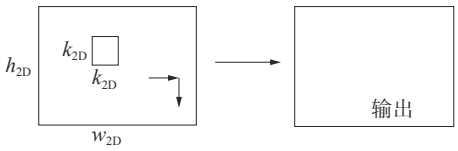

图3　2D 卷积计算示意

1.1.2　特征金字塔 Transformer 结构

为了更好地提取语义特征，提升模型对动作的专注度，本文采用了 Transformer 结构。Transformer 最早应用于自然语言处理任务中，因其优秀的表现被迁移到计算机视觉中[13]。而本文则将其应用到动作检测任务中，将其作为注意力机制与传统注意力机制进行对比。本文应用的 Transformer 结构如图 4 所示。

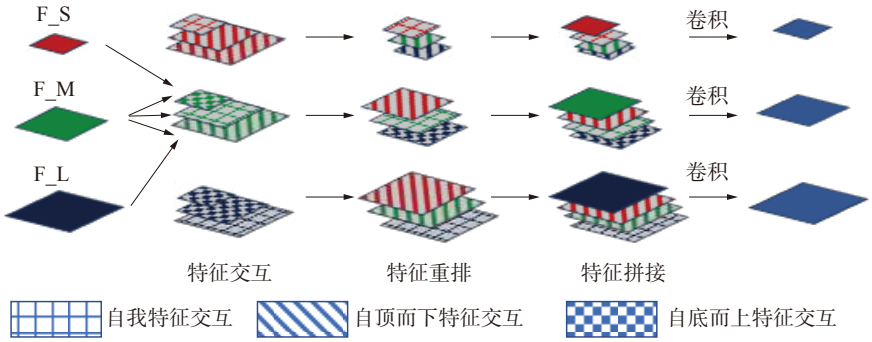

图4　特征金字塔 Transformer 结构

本文对特征提取网络的 3 个预测特征层构成的特征金字塔使用了 Transformer 结构，Transformer 结构可以使得特征可以跨空间和尺度进行交互。文中应用的特征金字塔 Transformer 包含了 3 个子模块，分别是自我特征交互模块、自顶而下特征交互模块和自底而上特征交互模块。经过该结构输出的特征图与输入特征图尺寸相同，但蕴含了丰富的上下文信息。

自我特征交互模块的作用在于捕捉一个特征图上同时出现的目标特征，它是 Non-Local 算法的改进，它的输出特征图$\hat{X}$的大小与输入特征图X的大小一致。两者主要的区别在于本文的自我特征交互模块改进了标准化函数F_{mos}，将标准的 Softmax 函数改进为 mos 函数，其计算如式（1）所示。设$\boldsymbol{q}_i = f_{\text{q}}(X_i) \in \boldsymbol{Q}$是第$i$个查询向量，$\boldsymbol{k}_j = f_{\text{k}}(X_j) \in \boldsymbol{K}$和$\boldsymbol{v}_j = f_{\text{v}}(X_j) \in \boldsymbol{V}$分别为第$j$个键向量和值向量；$f_{\text{q}}$、$f_{\text{k}}$和$f_{\text{v}}$分别代表查询向量、键向量和值向量的 Transformer 函数[14]。

$$F_{\text{mos}}\left(s_{i,j}^{n}\right) = \sum_{n=1}^{N} \pi_n \frac{\exp\left(s_{i,j}^{n}\right)}{\sum_j \exp\left(s_{i,j}^{n}\right)} \tag{1}$$

式中：N表示将$\boldsymbol{q}_i$和$\boldsymbol{k}_j$划分成了N部分，$s_{i,j}^{n}$为对这N部分中的每一部分进行了点积F_{sim}计算后得出的相似度分数，π_n为可学习的参数。

自我特征交互模块在单个特征图X内对查询向量$\boldsymbol{Q}$、键向量$\boldsymbol{K}$和值向量$\boldsymbol{V}$进行运算，自我特征交互模块的计算过程如式 (2)：

$$\begin{aligned} &\text{Input}: \boldsymbol{q}_i, \boldsymbol{k}_j, \boldsymbol{v}_j, N \\ &\text{Similarity}: s_{i,j}^{n} = F_{\text{sim}}(\boldsymbol{q}_{i,n}, \boldsymbol{k}_{j,n}) \\ &\text{Weight}: w_{i,j} = F_{\text{mos}}(s_{i,j}^{n}) \\ &\text{Output}: \hat{X}_i = F_{\text{mul}}(w_{i,j}, \boldsymbol{v}_j) \end{aligned} \tag{2}$$

自顶而下特征交互模块是特征金字塔自顶而下的交互，它将相对高层次的特征图X^c与相对低层次的特征图X^f建立联系。设$\boldsymbol{q}_i = f_{\text{q}}(X_i^f) \in \boldsymbol{Q}$是第$i$个查询向量，$\boldsymbol{k}_j = f_{\text{k}}(X_j^c) \in \boldsymbol{K}$和$\boldsymbol{v}_j = f_{\text{v}}(X_j^c) \in \boldsymbol{V}$分别为第$j$个键向量和值向量；$f_{\text{q}}$、$f_{\text{k}}$和$f_{\text{v}}$分别代表查询向量、键向量和值向量的 Transformer 函数。自顶而下特征交互模块的输出特征图$\hat{X}^f$大小与输入特征图X^f的大小一致。经验表明，当 2 个特征图的语义信息不同时，欧氏距离的负值在计算相似性方面比点积运算更有效，所以使用式（3）所示的F_{eud}来计算相似性：

$$F_{\text{eud}}(\boldsymbol{q}_i, \boldsymbol{k}_i) = -\|\boldsymbol{q}_i - \boldsymbol{k}_i\|^2 \tag{3}$$

自顶而下特征交互模块的计算过程如式 (4)：

$$\begin{aligned} &\text{Input}: \boldsymbol{q}_i, \boldsymbol{k}_j, \boldsymbol{v}_j, N \\ &\text{Similarity}: s_{i,j}^{n} = F_{\text{eud}}(\boldsymbol{q}_{i,n}, \boldsymbol{k}_{j,n}) \\ &\text{Weight}: w_{i,j} = F_{\text{mos}}(s_{i,j}^{n}) \\ &\text{Output}: \hat{X}_i = F_{\text{mul}}(w_{i,j}, \boldsymbol{v}_j) \end{aligned} \tag{4}$$

自底而上特征交互模块是特征金字塔自底而上的交互，旨在通过合并相对低层次的特征图用以呈现高层次特征。自底而上特征交互模块不是对像素的操作，而是对整个特征图进行操作。高层次的特征图被定义为$\boldsymbol{Q}$，低层次的特征被定义为$\boldsymbol{K}$和$\boldsymbol{V}$。$\boldsymbol{Q}$和$\boldsymbol{K}$是通过通道注意力机制来相互作用的，$\boldsymbol{K}$通过全局平均池化算法来计算$\boldsymbol{Q}$的权重，

加权后的$\boldsymbol{Q}$经过3×3卷积与通过3×3卷积进行下采样的$\boldsymbol{V}$相加，相加后的结果经过3×3卷积运算后得到输出特征图。计算过程如式 (5)：

$$\begin{aligned}&\text{Input}:\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\\&\text{Weight}:w=\text{GAP}(\boldsymbol{K})\\&\text{Weight - Query}:\boldsymbol{Q}_{\text{att}}=F_{\text{att}}(\boldsymbol{Q},w)\\&\text{Down - sample}:\boldsymbol{V}_{\text{dow}}=F_{\text{sconv}}(\boldsymbol{V})\\&\text{Output}:\hat{X}^{c}=F_{\text{add}}(F_{\text{conv}}(\boldsymbol{Q}_{\text{add}}),\boldsymbol{V}_{\text{dow}})\end{aligned}\tag{5}$$

式中：GAP 为全局平均池化函数；F_{att}为外积函数；F_{sconv}和F_{conv}为两个不同的3×3的卷积，F_{sconv}用于特征图下采样，F_{conv}用于提取特征；F_{add}为特征图像素大小相加函数，$\hat{X}^{c}$为输出特征图。

如图 4 所示的网络结构图，3 个不同尺度的预测特征层在经过上文所述的 3 种模块之后得到的特征图，按照尺寸大小进行重新排列，再将原始预测特征层与排列后的特征图在通道维度进行拼接，经过卷积得到与原始预测特征层尺寸相同的特征图。可以看到，输出的特征图与输入特征图相比，增加了空间和尺度信息的交互，包含了更为丰富的语义信息。

1.2　数据处理及损失函数

本文考虑到小目标检测的问题，采用了 YOLOv3 算法的检测方案。通过 K-means 算法对数据集的边界框大小进行统计，最终聚类出 9 个预选框。选取 3 个预测特征层，在 3 个尺度上对目标进行检测，每个预测特征层对应 3 个聚类出来的预选框。在经过 Transformer 和卷积之后，最终的每个预测特征层通道为$3\times(A+5)$。其中：3 为该预测特征层预选框的数目；A 为数据集中动作类别数目；每个预选框对应$A+5$个参数，5 为边界框的中心坐标和边界框的宽高回归所需要的 4 个参数和 1 个置信度参数。

损失函数共包括 3 部分，分别是边界框回归损失、分类损失和置信度损失。其中边界框回归损失采用完整交并比 (complete intersection-over-union, CIoU) 损失函数[15]，分类和置信度损失采用 Focal Loss 损失函数[16]。由于 YOLOv3 的正样本负样本存在严重的不均衡问题，大量的负样本大部分为简单易于区分的负样本，会造成对真样本的淹没，影响检测效果，针对样本不均衡问题，本文采用式(6)作为分类和置信度损失函数。

$$L_{\text{FocalLoss}}=\begin{cases}-\alpha(1-y')^{\gamma}\log y', & y=1\\-(1-\alpha)y'^{\gamma}\log(1-y'), & y=0\end{cases}\tag{6}$$

边界框回归损失函数通常采用 L2 范数损失函数，由于 L2 范数损失函数不能很好地反应预测框与真实位置的重合程度，所以在本文中采用了 CIoU 损失函数，它能够很好地反应预测框与真实位置的重叠程度，并具有尺度不变性，还充分考虑了预测框与真实位置的中心距离、预测框的长宽比等因素，比 L2 范数损失函数具有更大的优势。图 5 为 CIoU 函数计算所需参数示意图。其中绿色矩形框为目标的真实边界框，它的宽高分别为w^{gt}和h^{gt}；黑色方框为网络预测的边界框，它的宽高分别为w^{p}和h^{p}；d为真实框和预测框的中心距离，c为真实框与预测框的最小外接矩形的对角距离。

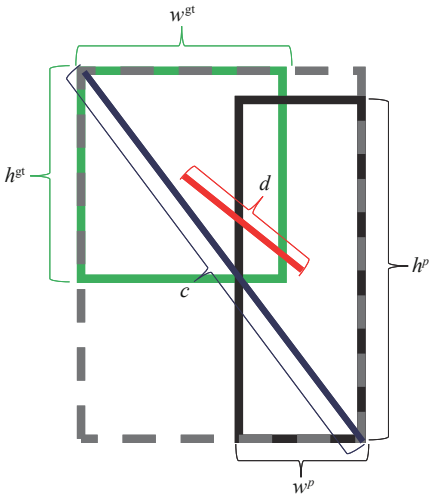

图5　CIoU 函数计算所需参数示意

式(7)为 CIoU 损失的计算过程：

$$L_{\text{CIoU}}=1-R_{\text{CIoU}}\tag{7}$$

式中：$R_{\text{CIoU}}=R_{\text{IoU}}-\left(\dfrac{d^2}{c^2}+\alpha v\right)$，$R_{\text{IoU}}$代表真实框和预测框的面积交并比；$d^2/c^2$用于修正预测框中心；$\alpha v$用于修正预测框的宽高，$v$衡量了预测框和真实框的宽高一致性，$v=\dfrac{4}{\pi^2}\left(\arctan\dfrac{w^{\text{gt}}}{h^{\text{gt}}}-\arctan\dfrac{w}{h}\right)^2$，$\alpha$为权衡参数，$\alpha=\dfrac{v}{1-R_{\text{IoU}}+v}$。

2　实验

2.1　跌倒动作数据集

目前，跌倒动作的公开数据集数目不多，很少有提供跌倒动作的人物边界框数据。本文从 Muticam 数据集中选取部分视频，共包含 6 个场景，每个场景由 8 个摄像头在不同角度进行拍摄，分辨率为 320×240。对拍摄的视频进行了分帧处理，并进行了手工标注，一共标记了 7083 张图片。其中 6373 张图片作为训练集，710 张图片作为验证集和测试集。每次输入到网络进行训练和推理的图片数据均来自于同一个摄像头，8 个摄像头的数据最终都会独立地被送入网络进行训练和推理。

2.2　跌倒动作识别实验

2.2.1　实验环境与训练策略

本文中所有实验的硬件平台均在 Ubuntu 18.04 操作系统下，GPU 采用 Nvidia RTX2080Ti，GPU 显存大小为 11 GB，CPU 采用 I7-8700K，内存大小为 32 GB。软件环境如下：深度学习框架为 Pytorch-1.7.1、CUDA 版本为 10.1。

每次实验训练集迭代训练 10 次，一次训练输入到网络的样本数目为 2，共计训练 31 870 次。基本学习率（R_{learning}）为 0.002，学习率的动态变化如图 6 所示，其中包含了 2 种策略：对于前 1 000 次训练采用线性预热；预热完成后采用余弦衰减策略直到训练结束。优化器采用带有动量参数的随机梯度下降算法。

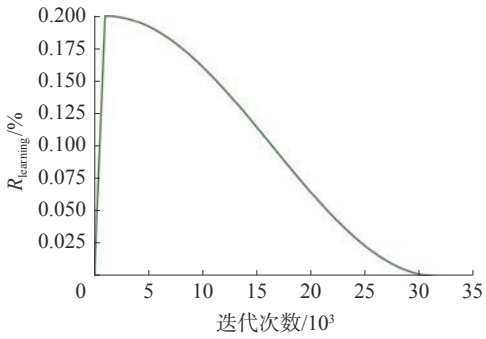

图6　学习率变化曲线

2.2.2　预选框的设置

本文的算法需要预先设定好预选框，合适的预选框设定可以加速模型的收敛。预选框的设定原则是尽可能与数据集中的目标宽高相似，所以需要对数据集进行分析。由于本文采用了 3 个尺度的预测特征层，每个预测特征层对应 3 个预选框，所以一共需要设置 9 个预选框。预选框通过 K-means 算法得到。图 7 中黄色的圆点代表数据集中的目标的宽高信息，其中横坐标为目标的宽度像素值，纵坐标为目标的高度像素值，其余颜色的 9 个点为聚类中心，红色的 3 个聚类中心是设置在特征金字塔网络输出的最大尺寸特征图上的预选框，绿色的 3 个聚类中心是设置在特征金字塔网络输出的中等尺寸特征图上的预选框，蓝色的 3 个聚类中心是设置在特征金字塔网络输出的最小尺寸特征图上的预选框。因为特征金字塔网络中尺寸相对较大的特征图用于更好地预测小目标，尺寸相对较小的特征图用于预测相对较大的目标，这里通过对聚类中心面积的大小进行排序将这 9 个预选框分配给特征金字塔网络的 3 个特征层。

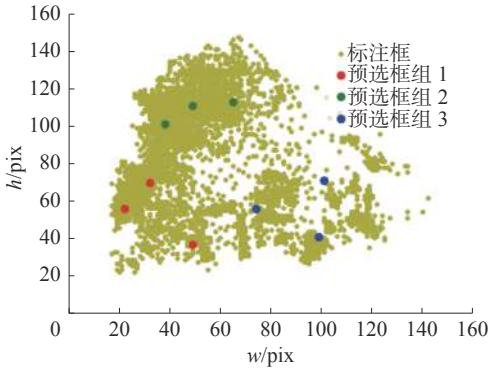

图7　聚类算法设定预选框

2.2.3　骨干网路结构有效性实验

该部分实验旨在判断不同的骨干网络结构对检测精度的影响，从而确定最佳的网络结构。对比实验中除了骨干网络的结构不同外其余的参数均相同，该部分的注意力机制均采用了特征金字塔 Transformer，输入图像像素尺寸均为 320×256，3D 卷积神经网络均输入连续 16 张图片，采用 IoU 阈值为 0.5 的均值平均精度 (mean average precision, mAP) δ_{mAP} 作为评价指标。

对比单独使用 2D 网络、单独使用 3D 网络、同时使用 2D 和 3D 网络的实验精度结果，从表 1 数据可以看出，同时使用 2D 和 3D 网络具有最好的 mAP 精度。

表 1　不同网络结构实验精度结果　　%

骨干网络结构	$\delta_{\text{AP}_{0.5}}$			δ_{mAP}
	行走	跌倒	躺下	
2D	83.13	59.54	86.40	76.36
3D	94.61	84.68	90.40	89.90
2D+3D	94.18	**93.80**	97.61	**95.20**

分析发现，单独使用 2D 网络，无法提取时间维度上的特征。从跌倒这个动作中来看，这个动作无法从单个图片中确定，所以单独使用 2D 网络得到该动作的 AP 指标相比于其余 2 组对比实验的实验结果低了很多；单独使用 3D 网络相比于单独使用 2D 网络的实验结果平均精度（average precision, AP）值 δ_{AP} 提升很大，但与同时使用 2D 和 3D 网络的 mAP 精度上还是有一定差距。这是因为 3D 网络注重于提取时间维度的特征而缺乏空间维度的特征，从而导致边界框的定位不准。可以看出，同时使用 2D 和 3D 网络在动作时空定位任务上具有充分的有效性。

2.2.4　3D 网络图片输入数目实验

在 2.2.3 节中对 3D 网络用于跌倒动作检测的有效性进行了验证，由于动作具有连续性的特点，人类无法通过一帧图片判定一个动作的类

别，同理计算机也如此。但是观察一个动作时间的或长或短，会对动作类别的预测产生影响。对于算法来说就是网络一次性该输入多少张连续图片这一问题值得深入探究。对此本文算法的3D网络部分，作者通过实验进行了探究，发现输入图片数目为16时具有最好的效果。分析其原因是因为当输入图片数目较少时，时间维度上提取的特征没有经过长时间的建模，包含的时间特征不够全面；而输入相对较多的图片时，很多动作持续时间较短，输入的图片序列中包含了其他动作的特征，造成网络学习混乱，效果也不佳。表2为3D网络一次性输入不同数目的图片时所对应的mAP结果。

表2　3D网络输入不同图片数目实验精度结果　%

输入图片数目	$\delta_{AP_{0.5}}$			δ_{mAP}
	行走	跌倒	躺下	
8	93.43	84.48	96.71	91.54
16	94.18	**93.80**	97.61	**95.20**
32	87.06	85.39	94.67	89.04

2.2.5　Transformer 有效性实验

本文对骨干网络获得的特征采用了通道拼接的方法进行特征融合，融合后的2D和3D特征缺乏深层次的融合交互，本文为解决这个问题使用了Transformer结构作为注意力机制，该结构应用在特征金字塔之后，可以融合不同预测特征层的特征，达到了特征跨空间、时间、尺度融合的目的。这部分实验将本文算法同通道注意力机制算法（Cam）和空间注意力算法（Sam）进行对比。表3的结果表明特征金字塔Transformer结构对检测性能具有很大的提升。

表3　特征金字塔Transformer结构有效性实验　%

结构	$\delta_{AP_{0.5}}$			δ_{mAP}
	行走	跌倒	躺下	
—	90.45	88.13	96.84	91.81
Cam	95.59	86.42	95.64	92.55
Sam	96.66	84.19	97.01	92.62
Cam + Sam	96.30	88.30	96.67	93.76
Transformer	94.18	**93.80**	97.61	**95.20**

3　结论

针对跌倒动作检测任务，本文提出了一种基于多特征融合及Transformer结构的时空定位检测架构，通过Transformer结构深度融合2D和3D网络提取的空间和时间维度以及不同尺度的特征，并通过一系列的对比实验证明了本文算法的有效性。本文算法相比于C3D等算法不仅完成了动作的分类还实现了动作的时空定位。与仅使用2D卷积网络相比，mAP提升了14.84%；与仅使用3D卷积网络相比，mAP提升5.3%。加入Transformer结构相比于传统注意力机制mAP最少提升了1.44%，可见将本方法应用于跌倒动作的时空定位，具有相当大的理论和经济价值。

参考文献：

[1] 张庆宾, 丁娜娜, 吴海波. 基于BP神经网络的摔倒动作识别方法 [J]. 指挥信息系统与技术, 2021, 12(1): 60–64.

[2] 吴松伟, 刘军, 范长军, 等. 面向刑侦视频的异常行为检测系统的设计与实现 [J]. 计算机时代, 2020(9): 67–71,75.

[3] 窦刚, 刘荣华, 范诚. 基于卷积神经网络的考场不当行为识别 [J]. 中国考试, 2021(2): 56–62,94.

[4] 张晓平, 纪佳慧, 王力, 等. 基于视频的人体异常行为识别与检测方法综述 [J]. 控制与决策, 2022, 37(1): 14–27.

[5] 陈永彬, 何汉武, 王国桢, 等. 基于机器视觉的老年人摔倒检测系统 [J]. 自动化与信息工程, 2019, 40(5): 37–41.

[6] 乌民雨, 吴宇豪, 陈晓辉. 一种基于双流网络的跌倒检测算法 [J]. 信息通信, 2020, 33(7): 32–34.

[7] 刘峰, 徐壮, 干宗良, 等. 一种基于时序运动特征的RGB-D视频跌倒行为检测算法 [J]. 南京邮电大学学报(自然科学版), 2020, 40(5): 117–124.

[8] 魏振钢, 孔勇强, 魏兆强, 等. 基于多摄像头监控的人体跌倒检测算法 [J]. 中国海洋大学学报(自然科学版), 2019, 49(7): 142–148.

[9] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 4489–4497.

[10] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. (2018–04–08)[2021–01–08]. https://arxiv.org/abs/1804.02767

[11] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[EB/OL]. (2014–06–09)[2021–01–08]. https://arxiv.org/abs/1406.2199.

[12] HARA K, KATAOKA H, SATOH Y. Learning spatio-temporal features with 3D residual networks for action recognition[C]//2017 IEEE International Conference on Computer Vision Workshops. Venice: IEEE, 2017: 3154–3160.

[13] ZHANG Dong, ZHANG Hanwang, TANG Jinhui, et al. Feature pyramid transformer[C]//Computer Vision–ECCV 2020. Cham: Springer International Publishing, 2020: 323–339.

(下转第 62 页)

[5] 薛文暄. 基于 yolov3-tiny 的小尺度行人检测算法研究 [D]. 北京邮电大学, 2020.
[6] 余珮嘉. 增强深度特征表示的行人检测研究 [D]. 贵阳: 贵州大学, 2020.
[7] ZHOU Chunluan, YUAN Junsong. Bi-box regression for pedestrian detection and occlusion estimation[C]// 15th European Conference on Computer Vision. Cham: Springer International Publishing, 2018: 138–154.
[8] WANG Xinlong, XIAO Tete, JIANG Yuning, et al. Repulsion loss: detecting pedestrians in a crowd[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7774–7783.
[9] MARÍN J, VÁZQUEZ D, LÓPEZ A M, et al. Occlusion handling via random subspace classifiers for human detection[J]. IEEE transactions on cybernetics, 2014, 44(3): 342–354.
[10] ZHANG Shifeng, WEN Longyin, BIAN Xiao, et al. Occlusion-aware R-CNN: detecting pedestrians in a crowd[C]//15th European Conference on Computer Vision. Cham: Springer International Publishing, 2018: 657–674.
[11] PANG Yanwei, XIE Jin, KHAN M H, et al. Mask-guided attention network for occluded pedestrian detection[C]// 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 4966–4974.
[12] LIU Wei, LIAO Shengcai, HU Weidong, et al. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting[C]//15th European Conference on Computer Vision. Cham: Springer International Publishing, 2018: 618–634.
[13] LIU Wei, LIAO Shengcai, REN Weiqiang, et al. High-level semantic feature detection: a new perspective for pedestrian detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 5182–5191.
[14] LAW H, DENG Jia. CornerNet: detecting objects as paired keypoints[J]. International journal of computer vision, 2020, 128(3): 642–656.
[15] DUAN Kaiwen, BAI Song, XIE Lingxi, et al. CenterNet: keypoint triplets for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 6568–6577.
[16] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//15th European Conference on Computer Vision. Cham: Springer International Publishing, 2018: 3–19.
[17] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6): 1137–1149.

本文引用格式:

叶正喆, 苍岩. 基于卷积神经网络的行人检测方法 [J]. 应用科技, 2022, 49(2): 55–62.
YE Zhengzhe, CANG Yan. A pedestrian detection method based on convolutional neural network[J]. Applied science and technology, 2022, 49(2): 55–62.

(上接第 54 页)

[14] VASWANI A, SHAZEER N, PARAMAR N, et al. Attention is all you need[EB/OL]. (2017–06–12) [2021–01–08]. https://arxiv.org/abs/1706.03762.
[15] ZHENG Zhaohui, WANG Ping, LIU Wei, et al. Distance-IoU loss: faster and better learning for bounding box regression[J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(7): 12993–13000.
[16] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[EB/OL]. (2017–08–07)[2021–01–08]. https://arxiv.org/abs/1708.02002.

本文引用格式:

刘文龙, 陈春雨. 基于多特征融合及 Transformer 的人体跌倒动作检测算法 [J]. 应用科技, 2022, 49(2): 49–54, 62.
LIU Wenlong, CHEN Chunyu. A human falling motion detection algorithm based on multi-feature fusion and Transformer[J]. Applied science and technology, 2022, 49(2): 49–54, 62.