计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201100185

基于场景先验知识的室内人体行为识别方法

刘 昕[1] 袁家斌[1,2] 王天星[1]
1 南京航空航天大学计算机科学与技术学院 南京 211106
2 南京航空航天大学信息化处(信息化技术中心) 南京 211106
(liuxinx@nuaa.edu.cn)

摘 要 目前,室内人体行为识别技术被广泛应用于视频内容理解、居家养老、医疗护理等领域,现有研究方法更多的是对人体行为进行建模,忽略了视频中场景与人体行为间的联系。为了充分利用场景信息与室内人体运动的关联性,文中对基于场景先验知识的室内人体行为识别方法进行了研究,提出了一种基于场景先验知识的双流膨胀3D行为识别网络(Scene-Prior Knowledge Inflated 3D ConvNet,SPI3D)。首先使用ResNet152网络提取场景特征进行场景分类,再基于场景分类的结果,引入量化后的场景先验知识,通过对权值进行约束来优化总体目标函数。另外,针对现有数据集多聚焦于人体行为特征、场景复杂且场景特征不明显的问题,自建了室内场景-行为识别数据集(Scene-Action DataBase,SADB)。实验结果表明,在SADB数据集上,SPI3D网络的识别准确率为87.9%,比直接利用I3D网络的识别准确率高6%。由此可见,引入场景先验知识后的室内人体行为识别模型具有更好的表现。

关键词:场景识别;深度学习;先验知识;行为识别
中图法分类号 TP391

Interior Human Action Recognition Method Based on Prior Knowledge of Scene

LIU Xin[1],YUAN Jia-bin[1,2] and WANG Tian-xing[1]
1 School of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 211106,China
2 Information Department(Informationization Technology Center),Nanjing University of Aeronautics and Astronautics,Nanjing 211106,China

Abstract Currently,the recognition technology targeted at human action in an interior scene is widely used in video content understanding,home-based care,medical care and other fields,and existing researches pay more heed to the modelling of human action,while ignoring the connection between interior scene and human action in videos. With a view to making full use of the relevance between the scene information and the human motion,this paper studies the recognition approaches for human action in an interior scene based on scene-prior knowledge. Yet,the paper proposes scene-prior knowledge inflated 3D ConvNet(SPI3D). Firstly,the ResNet152 network is adopted to extract scene features for scene classification. Then,based on the results,combined with scene-prior knowledge,this paper introduces quantified scene prior knowledge,optimizes the overall objective function by constraining the weights. Additionally,aiming at the problem that most of the existing data sets focus on the characteristics of human action,whereas the scene information remains complex and plain,an interior scene-action database(SADB) is established. It is shown in experimental results,on the SADB,the recognition accuracy rate of SPI3D reaches 87.9%,6% higher than the recognition accuracy of I3D directly. It can be seen that the modelling for the recognition on human action in interior scene is featured by better performance after introducing the prior knowledge of the scene.

Keywords Scene recognition,Deep learning,Prior knowledge,Action Recognition

1 引言

随着移动网络和视频平台的快速发展,视频已经成为数据量最大、传播最广泛的数据之一,短视频平台快手、抖音的日均活跃用户数在2020年上半年分别达到3亿和4亿,庞大的用户数量以及每日产生的海量视频在网络上传播,对视频平台的审核机制提出了挑战,而人工审核的方法需要很大的成本,并且人工审核的方式无法及时处理违规视频。2020

到稿日期:2020-11-26 返修日期:2021-04-01
基金项目:国家重点研发计划课题(2017YFB0802303);国家自然科学基金(62076127,61571226)
This work was supported by the National Key Research and Development Program of China(2017YFB0802303) and National Natural Science Foundation of China(62076127,61571226).
通信作者:袁家斌(jbyuan@nuaa.edu.cn)

年，我国公共监控摄像头的数量超过 6 亿，传统的数据管理方法已不适用于管理海量的视频数据。另一方面，智能家居系统的发展也对视频内容的识别任务提出了新的要求。视频中含有大量可用于视频内容识别的信息，多数视频的主体是人，因此研究视频人体行为识别可以有效促进视频内容理解的研究。如何利用视频中的有效信息，将视频中的各类信息进行关联，进而解决视频中人体行为识别的问题成为一项具有挑战性的任务。

现实世界中，人体行为与其所处的场景密切相关，场景所含元素（如场景内的物体、环境结构、场景属性等）都会影响主体的行为。虽然某些动作是相对独立于场景的，如微笑、哭泣等；但在特定场景下，主体仅可能完成特定的行为，例如：刷牙通常发生在卫生间，做饭通常发生在厨房，吃饭在餐厅发生的概率要大于在其他场景发生的概率。现有的研究方法主要是对人体行为识别网络、特征提取方法以及特征处理方法进行创新，这些方法更关注视频中的行为本身，而视频中的行为信息有限，视频的场景信息以及场景与行为间的关联性被忽略，此时，引入场景信息能够提升对主体行为的理解。此外，目前的公共行为识别数据集多聚焦于视频主体，视频中的人体占据了较大比例的画面，导致视频的场景被主体遮挡或无法通过视频内容得到该视频场景的类别，且这些数据集中包含的场景、行为大多是相对独立的，忽视了场景与行为间的关联。例如，Kinetics-400[1]数据集中标签为“洗头发”的视频中，视频主体所处场景有“公园”“厨房”“浴室”等，可对场景关联人体行为的识别方法产生较大影响。

本文首先针对当前公共数据集的不足，建立了场景-行为识别数据集（Scene-Action DataBase，SADB），然后根据室内人体行为受场景限制这一特性，提出一种基于场景先验知识的双流膨胀 3D 行为识别网络模型（SPI3D）。该模型首先识别视频中人物所处的场景，而后将场景对应的先验知识融入到行为识别模型中，通过将场景先验知识转换成行为识别模型中对权值的约束，优化目标函数，从而提高行为识别的效果。本文使用典型行为识别网络模型在 SADB 数据集上进行了基准测评实验，并将其与通用数据集 HMDB51[2]，UCF-101[3]，Kinetics 进行对比。对比 I3D 模型与本文提出的 SPI3D 模型，实验结果证明，将场景先验知识（prior knowledge）加入到行为识别中可以有效提高行为识别的准确率。

2 相关工作

根据特征提取方法的不同，将视频人体行为识别方法分为基于手工特征的行为识别方法和基于深度学习的方法。基于手工特征的方法利用特征提取符提取视频的行为特征，然后利用主成分分析和白化对底层特征进行预处理，再将特征编码成定长的特征向量，输入到行为分类器中进行训练，使用分类器完成行为分类。基于深度学习的方法中，特征提取部分由深度学习框架自动处理，数据进入训练网络后不受人为干预。这些基于深度学习的模型多为端到端的网络结构，可被训练的深度模型对输入到网络中的特征向量进行学习后输出在每个分类标签下的打分，深度模型再根据数据的类别标签利用反向传播调整网络模型参数，优化网络达到较好的分类效果。基于深度学习的方法比基于手工特征的方法更具普适性，学习到的特征更有效。

目前，基于深度学习的人体行为识别方法主要有双流网络（Two-Stream Network）和 3D 卷积网络（3D Convolutional Network）两种。1）双流网络。Simonyan 等[4]提出包含空间流和时间流的经典双流网络模型 Two-Stream Convolutional Networks，空间流通道对单帧图像学习空间（spatial）特征，时间流通道对光流图像学习运动信息，最后将两通道经过 softmax 的分数融合得到行为分类。Wang 等[5]为解决双流网络对长范围时间结构建模能力不足的问题，提出对视频采用稀疏采样方式的时域分割网络（Temporal Segment Networks，TSN）。2）3D 卷积网络[6]。Tran 等[7]提出了经典的 C3D 网络证明了 3D 卷积比 2D 卷积更适用于时空特征学习以及使用 $3\times3\times3$ 的卷积核可以达到最佳性能。针对 3D 卷积网络模型参数量大、对计算机资源需求高等问题，Qiu 等[8]提出一种基于残差网络结构[9]的伪 3D 残差网络（Pseudo-3D Residual Net，P3D ResNet），将 3D 卷积核拆分成两个串联的 2D 卷积核，并且可以将这种方式扩展到在图像数据训练集中训练好的 2D 卷积神经网络。Carreira 等[10]提出了一种基于 2D 卷积网络膨胀的双流膨胀 3D 卷积网络（Two-Stream Inflated 3D ConvNet，I3D），卷积核与池化核扩张到 3D，从视频中无缝学习时空特征，利用在图像数据集中训练好的 2D 卷积神经网络的参数初始化膨胀网络，该方法在基准数据集 HMDB51 和 UCF-101 上获得了当时最好的效果。Kim 等[11]同样利用预训练的 2D 卷积网络进行时间流的学习，提出一种双流卷积神经网络，在空间流和时间流均使用预训练的 2D 卷积网络，将从视频中选择的 3 个灰度图像分配给 RGB 三通道，形成一个堆叠的灰度三通道图像（SG3I），分别通过 RGB 图像和 SG3I 对 CNN 进行微调，避免计算光流造成的高花费。Yang 等[12]和 Yan 等[13]的研究也推动了深度学习模型的发展。

一些研究者在深度学习的框架下探索了场景上下文信息对行为识别的影响。Marszalek 等[14]利用文本挖掘方法从电影脚本中发现行为和场景的共现关系，发现场景上下文信息可以改善行为识别分类的结果。Zhang 等[15]将场景建模为中级的“隐藏层”，以桥接动作描述符和动作类别，在提取包含时空动作特征的混合视觉描述符和场景描述符后，通过朴素贝叶斯近邻算法学习场景与行为之间的联合概率分布情况，以推断动作类别。Dong 等[16]提出一种基于字典学习的场景和行为的联合学习模型，利用字典学习的方法对联合特征进行稀疏表示，提取出更具有解释性的特征信息，提高了传统单一行为识别网络的效果。Monteiro 等[17]建立了一个双流神经网络，其中一个流进行动作识别，另一个流通过识别动作发生的上下文来改善最终的行为识别结果。Vu 等[18]在一个自建的大规模图像数据集上分析发现，大多数场景动作的关联性很强，他们将室内、室外动作类别中的典型动作标签与场景类别关联，使用基于场景的动作注释进行分类学习，利用这种关联性完成静态图像中的人体行为识别任务，通过实验证明了不同场景类别具有不同的行为模式。Peng 等[19]提出一种

基于运动场景交互约束的无监督视频动作聚类方法，通过约束多视图的互补性，规范场景和运动子空间表示的分歧，约束多场景上下文的一致性，优化构造的目标函数，基于此辅助聚类过程，进而完成行为识别。Park 等[20]提出一种可以同时识别视频中场景和动作的网络，基于预测的结果对场景和动作的关联程度打分，通过注意加权特征进行分类，优化视频分类的结果。Ding 等[21]通过建立室内数据集和视频私有帧数据集，将室内数据集中的所有类别生成均值图像，利用这种颜色知识以及由场景中对象出现频率组成的场景知识，确定场景权重，最终通过这两种先验知识提高室内物体识别准确率。

与上述工作不同的是，本文将场景与行为间的关联转换成场景先验知识，并依赖深度学习网络学习到的知识克服先验知识中的缺点，通过先验知识在模型中对权值进行约束，优化目标函数。实验结果表明，融合了场景先验知识的行为识别网络具有更高的识别准确率。

3　基于场景先验知识的双流膨胀 3D 行为识别网络模型

图 1 为本文提出的 SPI3D 模型的结构图。本文设计模型的思路如下：1）为了获取视频所属场景类型，我们使用公布在 Places 上经过预训练的神经网络 ResNet152[22]，并在实验数据集上进行微调；2）将通过观察样本和围绕场景上下文得到的先验知识进行量化，并将其融合到行为识别网络中；3）使用稀疏采样策略对视频帧进行采样，人体行为识别网络在对行为建模具有较好效果的 I3D 网络的基础上进行改进。

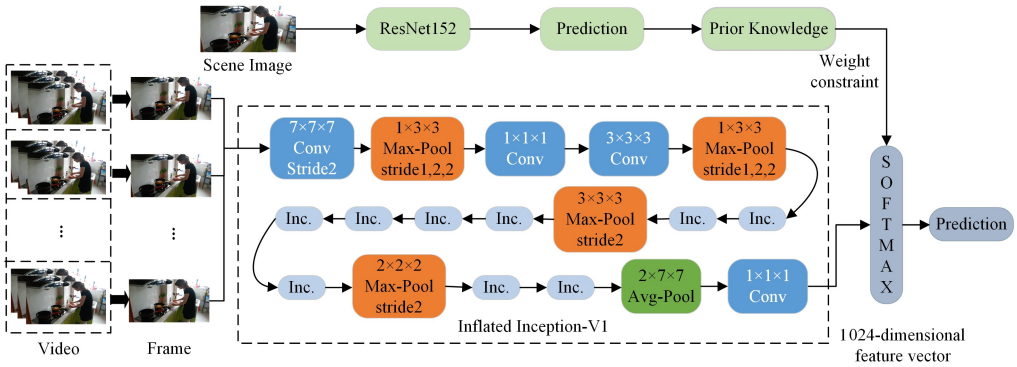

图 1　SPI3D 的模型结构图

Fig. 1　Structure diagram of SPI3D model

3.1　视频场景识别

为了探究场景对人体行为的影响，首先需要得到视频中的场景内容，因此将视频场景识别部分作为数据的预处理。

卷积神经网络在行为识别方面不断发展的同时，在场景分类方面也取得了巨大成功，因此我们使用现有的网络模型提取视频场景内容。麻省理工学院 Zhou[22]团队提出一个名为 Places 的场景数据集，其中包含了 700 万张带有标签的场景图片，205 个场景分类，其随后又发布了 Places2[23]数据集。本文使用该团队公开的、在 Places2 的子集 Places365 上预训练后的 ResNet152-places365 模型，然后在实验数据集上进行微调。Places365 包含 365 个场景类别，共有约 180 万张图像，每个场景类别最多有 5000 张图像。

利用 ResNet152-places365 提取场景特征进行分类学习，该网络作为 ImageNet ILSVRC 2015 大赛的冠军模型，解决了网络退化问题，相比传统卷积神经网络，它使用了由两个 1×1 和一个 3×3 卷积组成的 Deeper Bottleneck 结构，使网络参数量大幅缩减，训练难度降低，因此适合作为数据预处理部分的网络。

提取视频场景特征后在全连接层进行分类，场景类别共 8 个，第 i 个视频的场景识别结果表示为：

$$\boldsymbol{s}_i=(x_{i1},x_{i2},\cdots,x_{ij})(1\leqslant j\leqslant 8) \tag{1}$$

其中，x_{ij} 表示第 i 个视频属于第 j 个场景分类的概率值。

为了使行为识别模型更全面地利用场景信息，需要保留所有视频在所有场景中的概率值，假设视频个数为 N，所有视频的场景识别结果为：

$$S=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_N)^{\mathrm{T}} \tag{2}$$

3.2　先验知识的定义和量化

将先验知识融入深度神经网络中可以使其具有与人脑更加相似的学习方式，而人脑中高水平的知识和低水平的感官输入在学习中起着共同的作用[24]。对于多分类任务，先验知识的应用有明显的优势，判断样本类别时可以排除一些样本不可能属于的分类，这样可以使神经元之间连接的权重绝对值降低，使神经网络快速收敛到局部最优解[25]。

在视频人体行为识别问题中，场景是视频中的重要元素，根据场景属性可以判断视频中不可能或是发生概率很小的行为，例如在卫生间内不会做饭、在卧室通常不会洗衣服等，这些知识都属于先验知识。

本文提出的行为识别网络利用 softmax 函数进行分类，行为类别共 21 个，第 i 个视频的人体行为识别结果表示为：

$$\boldsymbol{A}_i=(a_{i1},a_{i2},\cdots,a_{ik})(1\leqslant k\leqslant 21) \tag{3}$$

其中，a_{ik} 表示第 i 个视频属于第 k 个行为分类的概率值。假设视频个数为 N，所有视频经过神经网络识别的行为分类结果为：

$$\boldsymbol{A}=(\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_N)^{\mathrm{T}} \tag{4}$$

不同的场景可以得到不同的先验知识，我们将先验知识定义为视频中人体在不同场景下完成的行为属于各类别的概率。在基于场景先验知识的人体行为分类问题中，建立从场景到行为的映射关系，将先验知识进行量化，从场景到行为的

所有映射关系表示为：

$$\boldsymbol{G}_S=(\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_j)^{\mathrm{T}}(1\leqslant j\leqslant 8) \tag{5}$$

其中，$\boldsymbol{G}_j$ 表示第 j 个场景下的先验知识，$\boldsymbol{G}_j$ 的公式为：

$$\boldsymbol{G}_j=(G_{ja_1},G_{ja_2},\cdots,G_{ja_k})(1\leqslant j\leqslant 8,1\leqslant k\leqslant 21) \tag{6}$$

其中，G_{ja_k} 表示第 j 个场景下发生第 k 个动作的先验概率，$0\leqslant G_{ja_k}<1$。

由于对视频中场景的判断是由神经网络作出的，而神经网络对场景的分类结果是由概率表示的，因此对于先验知识的选择，取场景结果的 TOP3，将场景的预测概率用作场景先验知识的权值。第 i 个视频经过场景识别后的先验知识可表示为：

$$\boldsymbol{G}_i^p=\sum_{m=1}^{3}x_{ij}^m\boldsymbol{G}_j(1\leqslant j\leqslant 8) \tag{7}$$

其中，x_{ij}^m 表示第 i 个视频在第 j 个场景下概率的 TOP3，$\boldsymbol{G}_i^p=(g_{i1}^p,g_{i2}^p,\cdots,g_{ik}^p)$，$g_{ik}^p$ 为每个视频对应先验知识的行为发生概率。同时，深度神经网络对样本的学习是通过分析数据的分布进行的，为了不使先验知识的引入对最终结果产生决定性影响，设 μ 为先验知识的影响权值，$0<\mu<1$。第 i 个视频结合了先验知识的行为识别结果，表示为：

$$\boldsymbol{y}_i=\mu\boldsymbol{G}_i^p+(1-\mu)\boldsymbol{A}_i \tag{8}$$

3.3 融合先验知识的人体行为识别模型

SPI3D 的人体行为识别部分的主体框架采用对人体行为建模具有较好效果的 I3D 网络，I3D 网络是将 2D 滤波器扩充为 3D 的网络，2D 神经网络使用的是 Inception-V1[26] 网络，并且扩充后的 3D 网络可以使用 2D 网络在 ImageNet[27] 预训练的参数，其方法是对 2D 卷积核的参数沿时间复制后除以 3D 卷积核时间维度的大小。本文受 Stewart[28] 的思路的启发，使用约束学习监督神经网络在计算机视觉中的任务，将量化后的场景先验知识融合到损失函数中，通过神经网络的反向传播，优化网络学习以及先验知识的影响权值。

在特征提取部分，本文将视频平均分成 N 个片段，并从每个视频片段中随机采样一张图片，这样的稀疏采样可以使采集到的图片尽可能地覆盖整个视频，并且使特征提取的效果不受视频长度的影响，这种采样策略可以用在任意的卷积神经网络中[29-30]。行为识别网络由卷积层、最大池化层、Inception 模块、平均池化层和 softmax 组成，4 个卷积层中，最后一个卷积层的卷积核为 $1\times1\times1$，用来生成分类分数，其他卷积层后均有 BN 操作和 Relu 激活，dropout 层仅放在平均池化层后，这种结构可以提高学习速率。利用 softmax 训练分类时，softmax 的输出向量为神经网络的预测结果，k 个分类的训练集可以表示为 $(\boldsymbol{y},\boldsymbol{Y})=[(y_1,Y_1),(y_2,Y_2),\cdots,(y_k,Y_k)]$，$\boldsymbol{y}$ 表示预测值，$\boldsymbol{Y}$ 表示标签。此时，单个训练样本的损失函数即为：

$$\boldsymbol{L}(\boldsymbol{Y},\boldsymbol{y})=-\sum_{j=1}^{k}Y_j\log y_j \tag{9}$$

训练时引入场景先验知识 $\boldsymbol{G}_P$。在训练过程中，如果先验知识与预测结果呈正相关，则增大先验知识的权重 μ 值；若先验知识不准确，则降低 μ 值，将整个数据集的损失函数表示如下：

$$\boldsymbol{J}(\boldsymbol{\omega}_i,\boldsymbol{b}_i,\boldsymbol{\mu})=\frac{1}{N}\sum_{1}^{N}\boldsymbol{L}(\boldsymbol{Y}_i,\boldsymbol{y}_i) \tag{10}$$

其中，$\boldsymbol{\omega}_i$，$\boldsymbol{b}_i$ 分别为 softmax 需要学习的权重、偏置值，$\boldsymbol{\mu}$ 为先验知识权重的集合。

4 实验结果与评估分析

本文设置了几组对比实验，首先使用 ResNet152 网络进行场景识别实验，然后使用几种典型的神经网络方法对自建的数据集 SADB 进行测评，最后测试 SPI3D 模型的行为识别能力。

4.1 实验环境

本文中场景识别实验所用实验环境为 Windows 10，GPU 使用 RTX 2060，CUDA10.0。

行为识别实验均使用 4 块 Nvidia Tesla K40m 的 GPU 计算完成，操作系统为 Linux，使用的深度学习框架根据网络模型的预训练框架决定，将数据分为 32 个批量(batch size)在 GPU 上并行计算。

4.2 实验数据集

现有的行为识别数据集大多聚焦于人体行为本身，缺乏具有室内场景特征的视频数据集。为了探究场景先验知识对人体行为识别准确性的影响，本文建立了一个场景-行为识别数据集(SADB)，该数据集关注室内场景与人体行为的关联关系，视频背景比较清晰明确。8 个场景类别分别是：浴室、卧室、餐厅、厨房、客厅、办公室、楼梯、电脑房。人体行为的 21 个类别涉及日常生活的常见行为以及一些特定场景下的特定行为，包括摔倒、躺下、坐下、摔东西、肚子疼、咳嗽、撞头、哭泣、刷牙、洗头、洗手、洗衣服、清洁厕所、铺床、做饭、吃东西、喝水、倒水、阅读、扫地、使用计算机。数据集没有很深的层次结构，而是在每个场景下包含了可能出现在此场景中的所有行为分类，如客厅(摔倒、躺下、吃东西、喝水、阅读…)，浴室(刷牙、洗头、洗手、洗衣服…)。图 2 给出了来自 SADB 的剪辑。

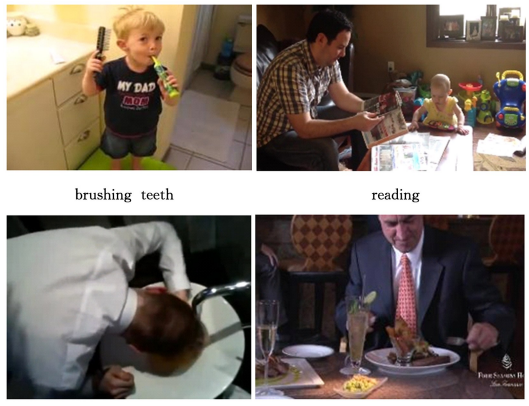

图 2 SADB 数据集的剪辑

Fig. 2 Clip of SADB

SADB 中每个类别的行为都有 180～250 个剪辑，每个剪辑持续约 10 s，21 类行为共包含 4 577 个视频。图 3 给出了 SADB 的行为类别及视频数量。除此之外，为了保证场景识别的准确率，在视频剪辑过程中通过人工抽取的场景最为清晰，将视频所有帧中对场景识别负影响最小的一帧作为该剪辑的场景图。

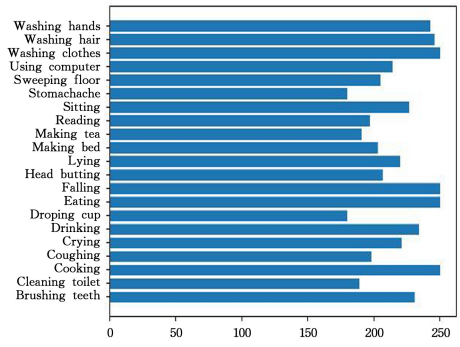

图 3　SADB 行为类别及视频数量

Fig. 3　SADB action category and number

4.3　场景识别实验

本节实验使用在公共场景数据集 Places365[23] 上预训练后的 ResNet152-places 网络，然后在 SADB 数据集上进行微调训练。

进行训练时将场景图片小边的分辨率转换为 256，输入时随机截取分辨率为 224×224 的图像，同时使用水平翻转、随机旋转等方式进行数据增强。

初始学习率为 0.01，每迭代 1 000 次进行一次验证，当验证集的预测准确率稳定后，将学习率下降为原来的 1/10，降低学习率后准确率不变化则停止训练。Batch Size 设为 32，使用深度学习框架为 caffe。训练过程中使用 K 折交叉验证，将训练集均分为 10 组训练样本集，10 组子样本集分别做一次验证集后计算平均错误率，进行 10 次 10 折交叉验证后取最佳性能的模型使用全部训练样本进行一次训练，得到最后的识别结果。表 1 列出了 ResNet152 在 Places365 上的准确率，以及在 SADB 数据集上 8 类场景识别结果的平均准确率。

表 1　ResNet152 在 Places365 和 SADB 数据集上的场景识别结果

Table 1　ResNet152 scene recognition results on Places365 and SADB

（单位：%）

Dataset	Top-1	Top-3	Top-5
Places365	56.65	—	85.07
SADB	78.46	90.84	92.73

图 4 给出了 ResNet152-places 网络在 SADB 数据集上预测的示例，结果显示其中大部分场景的 Top-3 能够与标签正确对应。由于 SADB 数据集中的场景类别只有 8 类，且基本不存在容易混淆的场景类别，所以本节实验能够取得一个比较高的识别准确率。因此，本文在后续的行为识别实验中采用本节实验结果的 Top-3，同时将 Top-3 的识别结果按 100% 的比例重新进行计算。

图 4　ResNet152-places 网络在 SADB 数据集上预测的示例

Fig. 4　Example of ResNet152-places prediction on SADB dataset

4.4　SADB 数据集测试实验

由于深层神经网络在人体行为识别中具有良好的表现，本文选择 3 种比较具有代表性的神经网络对 SADB 数据集进行基准测试，分别为 C3D 网络、TSN 网络以及 I3D 网络。

为了对比 3 种网络在原实验数据集与 SADB 数据集中的表现，文中保留了大部分训练策略，包括数据预处理方法和数据增强方法，实验选择原实验中表现较优的参数。

4.4.1　C3D 网络实验

本节实验使用的 C3D 网络是在 Sports-1M 数据集上预训练后的网络，在 SADB 数据集上进行微调训练。

实验训练初始学习率设置为 0.001，*epoch* 为 20，学习率在每 5 个周期后除以 10。最终测试结果由 5 次测试结果平均得到。C3D 网络[7] 在 UCF-101 以及 SADB 上得到的行为识别准确率结果如表 2 所列。实验采用 RGB 帧作为 C3D 网络的输入，结果表明，C3D 网络在 SADB 数据集上的整体表现差于在 UCF-101 数据集上的表现，但 SADB 的行为分类数要小于 UCF-101 的分类数，主要原因是 SADB 中存在一些不连续的片段，提取到的帧间信息不完整或不连续，同时 SADB 中视频的视角不断变化，背景更为复杂，室内动作具有特殊性，部分动作幅度较小，更难识别。

表 2　C3D 网络在 UCF-101 和 SADB 上的准确率

Table 2　C3D accuracy on split 1 of UCF-101 and SADB

（单位：%）

Modality	UCF-101	SADB
RGB	82.3	68.7

4.4.2　TSN 网络实验

本节实验使用的 TSN 网络是在 UCF-101 数据集上预训练后的网络，在 SADB 数据集上进行微调训练。将视频均分为 7 段，每个片段内随机抽取一帧，空间流和时间流两个分支的融合策略与原网络保持一致。

实验中空间流网络的初始学习率为 0.001，迭代次数为 2 400（*epoch*=20），在 *epoch* 为 8 和 16 时学习率减小 1/10。

时间流网络的初始化学习率设为 0.005，迭代次数为12000（*epoch*=100），在 *epoch* 为 40 和 80 时学习率减小 1/10，Dropout 的概率设置为 0.7，优化函数选用随机梯度下降算法，动量设置为 0.9。最终的识别准确率取 5 次测试的平均值。表 3 列出了 TSN[5] 使用不同模态特征在不同数据集上的分类准确率。

表 3 TSN 在 UCF-101，HMDB-51 和 SADB 上的分类准确率

Table 3 Classification accuracy of TSN on UCF-101, HMDB-51 and SADB

（单位：%）

Modalities	UCF-101	HMDB-51	SADB
RGB	86.0	53.5	72.4
Flow	88.1	66.5	73.3
RGB+Flow	94.0	68.5	76.7

在使用 TSN 网络测试 SADB 数据集时，使用光流作为模型输入与使用 RGB 作为模型输入相比，识别准确率没有明显提升，而在 HMDB-51 数据集上，光流的表现要明显优于 RGB，部分原因是 SADB 数据集中有较多由运动的摄像镜头拍摄的视频，而光流是利用帧间相关性得到的，当相邻帧间的亮度相差较大或相邻帧间的运动幅度较大时，提取到的光流特征不能有效表示视频中包含的运动信息。将 RGB 和光流融合后，因稀疏采样造成的帧跨度较大的不足与光流特征的不足相互弥补，并且使模型学习到更多的运动信息，所以模型的效果得到了一定的提升。

4.4.3 I3D 网络模型实验

本节使用的网络是在 ImageNet 和 Kinetics 数据集上进行预训练的 I3D 网络模型，在 SADB 数据集上进行微调训练。本文视频输入大小为 112×112，测试阶段随机抽取视频的 64 帧作为模型输入。

实验中训练使用的优化函数为标准 SGD，动量设置为 0.9，初始学习率为 0.1，在验证集的准确率稳定后再降低至原学习率的 1/10(0.01)，Dropout 率设为 0.5。表 4 对比了不同输入特征下 I3D 模型在不同数据集上的识别效果。

表 4 I3D 网络在 Kinetics 和 SADB 上 Top-1 和 Top-5 的行为识别准确率

Table 4 I3D Top-1 and Top-5 action recognition accuracy on Kinetics and SADB

（单位：%）

Modalities	Kinetics		SADB	
	Top-1	Top-5	Top-1	Top-5
RGB	71.1	89.3	77.6	87.2
Flow	63.4	84.9	72.4	82.5
RGB+Flow	74.2	91.3	81.7	89.9

从表 4 可以看出，光流在两个数据集上的表现均差于 RGB 的表现，但 I3D 模型在 SADB 数据集上的整体表现是不错的。需要注意的是，I3D 模型在原文[10]中使用 32 块 GPU 并行训练，3D 卷积网络使用了 64 块 GPU，输入图像大小为 3×64×224×224，而本节实验受实验环境限制，使用 4 块 GPU 进行训练和测试，输入图像大小为 3×64×112×112，GPU 数量减少、输入图像的分辨率降低、单次训练样本的输入量减少等因素都会影响网络训练的效果，无法完全发挥 I3D 网络的性能。

4.5 SPI3D 性能测试实验

本节对 SPI3D 模型的性能进行测试，特征模态使用 RGB、光流以及 RGB+光流的双流特征做多组对比实验。实验使用在 ImageNet 和 Kinetics 上预训练后的 I3D 模型，将先验知识融合到 I3D 模型后在 SADB 数据集上进行端到端训练，微调 SPI3D 网络的全部参数。实验以 4.3 节的场景识别结果为基础，结合先验知识对行为识别结果进行优化，同时在反向优化的过程中对先验知识及其所占比例进行优化。

为了将迁移学习的效果更好地表现出来，同时由于实验环境的限制，将 I3D 模型的输入大小由原文的 3×64×224×224 修改为 3×64×112×112；训练使用 64 帧，使 SPI3D 模型能从预训练中获得更大的收益；本实验在测试时使用 64 帧输入。

本节实验使用 TSN 网络中提到的稀疏采样策略，将每条视频平均分成 64 段后，从每一段剪辑中随机采样一帧作为 SPI3D 模型的输入，实验中训练使用的优化函数为标准 SGD，动量设置为 0.9，初始学习率为 0.01，在验证集的准确率稳定后再降低至原参数的 1/10(0.001)，Dropout 率设为 0.5，训练使用的深度学习框架为 pytorch，光流图像使用 OpenCV 的 TV-L1 算法提取，训练阶段的数据增强使用随机裁剪、左右翻转的方式，测试阶段使用中心裁剪的方式进行数据增强。对每个视频剪辑进行了场景识别后，建立视频-场景-先验知识-行为识别模型输入的映射关系，将先验知识融入神经网络的训练过程中，每条先验知识对应于不同的 μ 值，且需要在训练中不断调整此参数。

实验使用与 TSN 相同的稀疏采样策略，保证了整段视频信息的学习，使得更新参数时模型可用的信息更多，对识别长时间的动作有更好的效果。表 5 列出了 I3D 和 SPI3D 在 SADB 上使用不同模态的行为识别结果。可以看出，单独对比 3 种模态，SPI3D 模型已经取得了优于 I3D 模型的效果，可以证明当先验知识可用时，SPI3D 在行为识别准确率上有很好的效果。同时使用 RGB 和光流两种模态作为 SPI3D 模型的输入，准确率由 84.2% 提升至 88.3%，证明 SPI3D 在 SADB 数据集上融合光流后可以更好地学习到运动信息。但由于场景识别准确率无法达到理想值，对行为识别结果会产生一定的影响，因此后续提升场景识别准确率也可以进一步提升行为识别的准确率。

表 5 I3D 和 SPI3D 在 SADB 上的行为识别准确率

Table 5 Action recognition accuracy of I3D and SPI3D on SADB

（单位：%）

Modalities	I3D		SPI3D	
	Top-1	Top-5	Top-1	Top-5
RGB	77.6	87.2	83.8	90.9
Flow	72.4	82.5	75.9	87.0
RGB+Flow	81.7	89.9	87.9	94.8

表 6 列出了 I3D 和 SPI3D 在 SADB 中部分行为识别结果的对比。从实验结果可以看出，SPI3D 在对做饭、刷牙、摔倒这几类行为的识别中取得了明显高于 I3D 的识别准确率，在

对哭泣、咳嗽、坐下的识别中，两个网络的识别准确率相差不大，分析原因是SADB中包含做饭、刷牙和摔倒这类行为的视频与主体场景有着较大的关联，场景先验知识在这个过程中起到比较重要的作用。而哭泣、咳嗽、坐下这类行为在两种模型中识别准确率相差不大，其中哭泣和咳嗽两类行为的识别准确率较低，原因是这两类动作的变化幅度较小，容易与其他行为发生混淆。对整体的数据进行对比可以看出，SPI3D在对某些与场景关联性较强的行为进行识别时具有很好的效果，而对某些与场景关联性较弱的行为则没有较好的识别能力。

表6　I3D和SPI3D在SADB数据集上对部分行为的识别准确率对比

Table 6　Comparison of recognition accuracy of some categories of I3D and SPI3D on SADB dataset

（单位：%）

Action	I3D	SPI3D
Falling	90.1	98.7
Cooking	81.6	87.3
Brushing teeth	88.7	94.0
Coughing	57.8	60.2
Crying	49.4	53.2
Sitting	73.5	72.8

表7列出了不同模型在公开数据集以及SADB数据集上的基准表现。SPI3D在SADB数据集上的表现均优于其他模型，部分原因是其在训练时使用了64帧输入，单次训练样本数提升，更密集的时间特征使模型学习到了更多的运动信息，并且在迁移学习后从预训练中获得了更大的收益。与C3D网络相比，SPI3D网络添加了光流特征，获得了从ImageNet中预训练的好处，同时SPI3D继承了TSN网络的稀疏采样策略，但SADB数据集包含的是室内人体行为，SPI3D充分利用了室内场景信息与人体运动的关联性和场景对人体运动的限制，使其比其他网络更适用于对室内行为的识别。SADB数据集的规模要小于Kinetics数据集，但SPI3D模型在SADB数据集上仍然取得了很好的表现，这也得益于其在ImageNet和Kinetics两个大型数据集上的预训练。

表7　对比不同模型在UCF-101，HMDB-51，Kinetics和SADB数据集上的识别准确率

Table 7　Comparison of recognition accuracy of different models on UCF-101，HMDB-51，Kinetics and SADB

（单位：%）

Model	UCF-101	HMDB-51	Kinetics	SADB
C3D	82.3	—	56.1	68.7
RGB-TSN	86.0	53.5	—	72.4
Flow-TSN	88.1	66.5	—	73.3
Two-Stream TSN	94.0	68.5	—	76.7
RGB-I3D	95.1	74.3	71.1	77.6
Flow-I3D	96.5	77.3	63.4	72.4
Two-Stream I3D	97.8	80.9	74.2	81.7
RGB-SPI3D	—	—	—	83.8
Flow-SPI3D	—	—	—	75.9
Two-Stream SPI3D	—	—	—	87.9

结束语　本文首先针对当前行为识别数据集出现的问题，建立一个场景-行为识别数据集（SADB）。然后针对目前的行为识别模型中没有有效利用场景与行为间的关联性，提出一种基于场景先验知识的双流膨胀3D行为识别网络模型（SPI3D），引入场景先验知识，对损失函数进行改进。实验表明，本文提出的SPI3D模型优于其他行为识别模型，在解决特定场景中的人体行为识别问题上有着很好的表现。后续的工作将持续优化场景识别结果，进一步提升行为识别的效果，同时引入不同先验知识，提升模型在不同数据集上的泛化能力。

参 考 文 献

[1] KAY W，CARREIRA J，SIMONYAN K，et al. The kinetics human action video dataset[J]. arXiv：1705.06950，2017.

[2] KUEHNE H，JHUANG H，GARROTE E，et al. HMDB：A Large Video Database for Human Motion Recognition[C]// 2011 International Conference on Computer Vision. Barcelona：IEEE，2011：2556-2563.

[3] SOOMRO K，ZAMIR A R，SHAH M. UCF101：A Dataset of 101 Human Actions Classes From Videos in The Wild[J]. arXiv：1212.0402，2012.

[4] SIMONYAN K，ZISSERMAN A. Two-Stream Convolutional Networks for Action Recognition in Videos[M]. Advances in Neural Information Processing Systems. Berlin：Springer，2014：568-576.

[5] WANG L，XIONG Y，WANG Z，et al. Temporal Segment Networks：Towards Good Practices for Deep Action Recognition[C]// European Conference on Computer Vision. Cham：Springer，2016：20-36.

[6] JI S，XU W，YANG M，et al. 3D Convolutional Neural Networks for Human Action Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2013，35(1)：221-231.

[7] TRAN D，BOURDEV L，FERGUS R，et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]// 2015 IEEE International Conference on Computer Vision (ICCV). Santiago：IEEE，2015：4489-4497.

[8] QIU Z，YAO T，MEI T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks[C]// 2017 IEEE International Conference on Computer Vision(ICCV). Venice：IEEE，2017：5534-5542.

[9] HE K，ZHANG X，REN S，et al. Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas：IEEE，2016：770-778.

[10] CARREIRA J，ZISSERMAN A. Quo vadis，action recognition? a new model and the kinetics dataset[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Honolulu：IEEE，2017：6299-6308.

[11] KIM J H，WON C S. Action Recognition in Videos Using Pretrained 2D Convolutional Neural Networks[J]. IEEE Access，2020，8：60179-60188.

[12] YANG W B，YANG H C，LU C，et al. Gesture Recognition Based on Skin Color Features and Convolutional Neural Network[J]. Journal of Chongqing Technology and Business University(Natural Science Edition)，2018，35(4)：75-81.

[13] YAN H，LUO C，LI H，et al. Gait Recognition Method Based on Gait Energy Map Combined with VGG[J]. Journal of Chongqing

University of Technology(Natural Science),2020,34(5):166-172.
[14] MARSZALEK M,LAPTEV I,SCHMID C. Actions in context [C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami:IEEE,2009:2929-2936.
[15] ZHANG H B,LEI Q,CHEN D S,et al. Probability-based method for boosting human action recognition using scene context [J]. IET Computer Vision,2016,10(6):528-536.
[16] DONG X,TAN L,ZHOU L N,et al. Short Video Behavior Recognition Combining Scene and Behavior Features[J]. Journal of Frontiers of Computer Science and Technology,2020,14(10):1754-1761.
[17] MONTEIRO J,GRANADA R,MENEGUZZI F,et al. Using Scene Context to Improve Action Recognition[C]// 23rd Iberoamerican Congress(CIARP 2018). Madrid,2018:954-961.
[18] VU T H,OLSSON C,LAPTEV I,et al. Predicting actions from static scenes[C]// European Conference on Computer Vision. Cham:Springer,2014:421-436.
[19] PENG B,LEI J,FU H,et al. Unsupervised Video Action Clustering via Motion-Scene Interaction Constraint[J]. IEEE Transactions on Circuits and Systems for Video Technology,2018,30(1):131-144.
[20] PARK J,LEE J,JEON S,et al. Video Summarization by Learning Relationships between Action and Scene[C]// 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul:IEEE,2019:1545-1552.
[21] DING X,LUO Y,LI Q,et al. Prior knowledge-based deep learning method for indoor object recognition and application[J]. Systems Science & Control Engineering,2018,6(1):249-257.
[22] ZHOU B,GARCIA A L,XIAO J,et al. Learning Deep Features for Scene Recognition using Places Database[J]. Advances in Neural Information Processing Systems,2015,1:487-495.
[23] ZHOU B,LAPEDRIZA A,KHOSLA A,et al. Places:A 10 Million Image Database for Scene Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2018,40(6):1452-1464.
[24] DILIGENTI M,ROYCHOWDHURY S,GORI M. Integrating Prior Knowledge into Deep Learning[C]// IEEE International Conference on Machine Learning & Applications. IEEE,2017:920-923.
[25] XUAN D M,WANG J Y,YU H,et al. Application of prior knowledge in deep learning[J]. Computer Engineering and Design,2015,36(11):3087-3091.
[26] SZEGEDY C,LIU W,JIA Y Q,et al. Going deeper with convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Boston:IEEE,2015:1-9.
[27] DENG J,DONG W,SOCHER R,et al. ImageNet:A large-scale hierarchical image database[C]// IEEE Conference on Computer Vision & Pattern Recognition. Miami:IEEE,2009:248-255.
[28] STEWART R,ERMON S. Label-Free Supervision of Neural Networks with Physics and Domain Knowledge[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. California:AAAI,2017:2576-2582.
[29] SCHLOSSER P,DAVID M,ARENS M. Investigation on Combining 3D Convolution of Image Data and Optical Flow to Generate Temporal Action Proposals[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach:IEEE,2019:2448-2456.
[30] YANG C,XU Y,SHI J,et al. Temporal Pyramid Network for Action Recognition[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle:IEEE,2020:588-597.

LIU Xin, born in 1995, postgraduate. His main research interests include deep learning and action recognition.

YUAN Jia-bin, born in 1968, Ph.D, professor, is a senior member of China Computer Federation. His main research interests include deep learning, high performance computing and information security, etc.

(责任编辑:柯颖)