第 58 卷 第 2 期/2021 年 1 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

结合时序动态图和双流卷积网络的人体行为识别

张文强，王增强，张良*

中国民航大学天津市智能信号与图像处理重点实验室，天津 300300

摘要 为了更好地对人体动作的长时时域信息进行建模，提出了一种结合时序动态图和双流卷积网络的人体行为识别算法。首先，利用双向顺序池化算法来构建时序动态图，实现视频从三维空间到二维空间的映射，用来提取动作的表观和长时时序信息；然后提出了基于 inceptionV3 的双流卷积网络，包含表观及长时运动流和短时运动流，分别以时序动态图和堆叠的光流帧序列作为输入，且结合数据增强、模态预训练、稀疏采样等方式；最后将各支流输出的类别判定分数通过平均池化的方式进行分数融合。在 UCF101 和 HMDB51 数据集的实验结果表明：与传统双流卷积网络相比，该方法可以有效利用动作的时空信息，识别率得到较大的提升，具有有效性和鲁棒性。

关键词 图像处理；双流卷积网络；人体行为识别；时序动态图；数据增强

中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP202158.0210007

Human Action Recognition Combining Sequential Dynamic Images and Two-Stream Convolutional Network

Zhang Wenqiang, Wang Zengqiang, Zhang Liang*

Tianjin Key Laboratory of Advanced Signal and Image Processing, Civil Aviation University of China, Tianjin 300300, China

Abstract In order to well model the long-term time-domain information of human action, a human action recognition algorithm based on sequential dynamic images and two-stream convolution network is proposed. First of all, the sequential dynamic images are constructed by using sequential pooling algorithm to realize the mapping of video from three-dimensional space to two-dimensional space, which is used to extract the apparent and long-term sequential information of actions. Then, a two-stream convolution network based on inceptionV3 is proposed, which includes apparent and long-time motion flow and short-time motion flow. The input of the network is sequential dynamic images and stacked frame sequence of optical flow, and it combines data augmentation, pre-trained model, and sparse sampling. Finally, the classification judgment scores output by each branch is fused by average pooling. Experimental results on UCF101 and HMDB51 datasets show that, compared with the traditional two-stream convolution network, this method can effective use the temporal and spatial information of the action, and the recognition rate can be improved greatly, which shows effectiveness and robustness.

Key words image processing; two-stream convolutional network; human action recognition; sequential dynamic images; data augmentation

OCIS codes 100.4996; 100.5010; 110.4153

1 引 言

伴随着海量视频数据的涌现，人体行为识别已经成为计算机视觉领域研究的热点，在监控安防、人工智能交互、辅助医疗、虚拟现实等领域具有广泛的应用前景[1-6]。利用计算机可以获取视频中人体的

收稿日期：2020-06-05；**修回日期**：2020-06-23；**录用日期**：2020-07-07

基金项目：国家自然科学基金(61179045)

* **E-mail**：l-zhang@cauc.edu.cn

行为特征，并且建立起与人体动作之间的映射关系，从而实现视频底层数据和高层语义之间的自动关联。受背景光照视角等方面影响，人体行为识别仍面临着巨大挑战。

已有的人体行为识别方式主要包括基于人工特征提取的方法和基于深度学习的方法。基于人工特征提取的方法模型设计简单，表现出较好的鲁棒性，但也存在提取特征的预处理成本较高、准确率较低的缺陷[7-8]。而随着卷积神经网络仿照生物神经元的工作机制表现出对光照、背景和噪声的鲁棒性，深度学习应用于动作识别成为了研究的热点之一[9]。现有的方法主要将注意力集中在有效的动作特征描述和人体行为识别模型的改进[10-15]。Karpathy 等[16]将堆叠的 RGB 视频帧序列作为深度卷积网络的输入，来表述人体的行为特征。但序列中存在着色彩、光照、复杂背景等冗余信息以及对动态嘈杂场景的不鲁棒性均会影响识别率；Simonyan 等[17]提出将多帧堆叠的光流帧序列应用于双流网络中的时间流，利用光流记录瞬时位移的特性来提取人体的短时运动信息，但受限于光流的高成本和对长时间动作的识别表现一般；Zhang 等[10]提出将视频解码过后获得的运动矢量作为视频动作描述子，可以缓解光流的高成本，但受限于图片精度和噪声的影响；上述描述动作特征的方法仍然是对视频中人体表观或运动信息的浅层次描述，无法捕获到深层次的特征信息。在网络模型方面，依据输入流的数目可以分为单流、双流和多流的网络模型。Tran 等[18]尝试将应用于图像的二维卷积网络扩展到三维（3D）空间，提出了 3D 卷积网络，可以直接处理视频，加快了视频的处理速度；Simonyan 等[17]参考了 RGB 视频帧和光流等的输入方式，提出了包含时间流和空间流的双流卷积网络用于提取视频中的表观和时序特征，使用支持向量机（SVM）分类器对最终结果进行处理；Wang 等[19]采用稀疏采样的方式基于双流网络对长时间范围结构进行了建模，将时间流和空间流处理后的结果进行融合，取得了不错的效果；Lan 等[20]进一步研究了双流网络时空融合的方式，提出了时序线性编码层来对视频中不同位置的特征进行融合编码；Shi 等[12]提出了深度轨迹描述符（SDTD），搭配 RGB 帧图片和光流序列作为输入，提出一个包含连续深度轨迹描述符流、空间流和时间流的三流网络。

已有的研究表明，对视频特征进行深层次的描述，充分提取视频中所包含的空间表观信息和时间运动信息对于行为识别具有重要意义。针对现有方法存在的问题，本文提出了一种结合时序动态图和双流卷积网络的行为识别算法。利用双向顺序池化（BRP）算法对视频三维特征进行压缩，在表观和长时时域进行建模，且结合堆叠光流（SOF）提取人体动作的短时序信息，提出了一种由表观和长时序卷积网络以及短时序卷积网络组成的双流卷积神经网络（TS-CNN）模型，进行人体行为识别。

2 动作表征方式

视频具有连贯性，可以看作是多帧静态图像（SI）按照时间顺序的排列。视频特征可以从空间和时间两个角度来进行表述。空间角度表现为连续的多帧静态图像序列，用来描述视频中的人物和场景等表观信息。但光照、遮挡以及背景等复杂冗余信息会给行为识别带来挑战。时间角度表现为帧与帧之间的运动变化信息，用来描述视频中物体的运动状态。但复杂的动作类别往往需要上百帧的静态图像序列来进行呈现。因此去除冗余信息和提取视频的长时时域信息对于人体行为识别十分关键。

而目前传统的双流网络模型只能处理不超过十帧光流的输入，限制了网络提取视频的长时时域信息。为了解决这些问题，本文提出采用双向顺序池化算法将多帧视频图像序列组成的三维信息压缩到二维空间中，同时去除掉光照、背景等冗余信息，得到时序动态图（SDI），用来表征视频中人体的表观信息和长时序信息；采用工具 denseflow 提取堆叠光流，光流中包含着动作瞬时的运动信息。分别以上述方式作为双流卷积网络的输入，将各支流输出的类别判定分数通过平均池化的方式进行分数融合，并进行行为识别。动作表征的具体流程如图 1 所示。

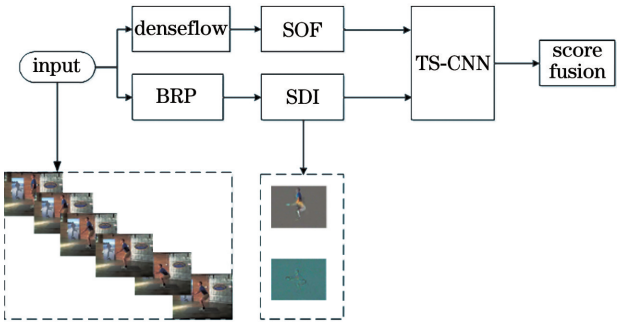

图 1 动作表征整体流程示意图

Fig. 1 Overall flow diagram of action representation

2.1 双向顺序池化算法

双向顺序池化算法是一种时间编码的过程，通过对视频序列进行编码来获取视频帧随时间变化的

动态特性。

给定一个 k 帧连续的视频图像序列，描述为

$$\boldsymbol{D}=[x_1,x_2,\cdots,x_t,\cdots,x_{k-1},x_k], \tag{1}$$

式中：从 x_1 到 x_k 按照时间顺序来进行排列，x_t 表示第 t 帧图像。

从序列中 $\boldsymbol{D}$ 的每一帧图像中提取特征向量 $\boldsymbol{C}(x)\in\mathbf{R}^d$，取其前 t 帧的均值 $\boldsymbol{m}_t$ 进行平滑操作得到新的序列描述为 $\boldsymbol{V}$，计算方法分别为

$$\boldsymbol{m}_t=\frac{1}{t}\sum_{\tau=1}^{t}\boldsymbol{C}(x), \tag{2}$$

$$\boldsymbol{v}_t=\frac{\boldsymbol{m}_t}{\left\|\boldsymbol{m}_{\frac{t}{2}}\right\|}, \tag{3}$$

$$\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_t,\cdots,\boldsymbol{v}_{k-1},\boldsymbol{v}_k], \tag{4}$$

式中：$\boldsymbol{v}_t$ 是对于均值 $\boldsymbol{m}_t$ 进行平滑操作的公式，这样可以降低特征向量间的偏差和噪声带来的影响；序列 $\boldsymbol{V}$ 包含着视频帧在时间段$[0,k]$内的时序变化信息。

将在时间段$[t,t+1]$内的视频帧时序变化设为 $\boldsymbol{E}$，对 $\boldsymbol{E}$ 进行编码。同时序列 $\boldsymbol{V}$ 在经过平滑操作后足够平滑，假设其达到理想的平滑条件，可以通过参数 $\boldsymbol{u}$ 的时序线性函数 $\theta(\boldsymbol{u})=\theta(\boldsymbol{V};\boldsymbol{u})$ 来编码绝对平滑条件下视频帧的时序变化。采用时序函数 $\theta(\boldsymbol{u})$ 来无限逼近于时序变化 $\boldsymbol{E}$，该过程描述为

$$\underset{\boldsymbol{u}}{\operatorname{argmin}}\left\|\boldsymbol{E}-\theta(\boldsymbol{u})\right\|。 \tag{5}$$

池化的主要目的是将视频帧序列提取到的特征进行压缩，经降维处理后映射到二维空间中，即学习到参数 $\boldsymbol{u}$。为了获得视频帧的顺序关系，本文引进了得分函数 $r_t=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{v}_t$ 来区分视频帧的先后顺序。一般情况下，时序越靠后，其得分函数越大。记$[\boldsymbol{v}_1\cdots\boldsymbol{v}_t\cdots\boldsymbol{v}_k]$为视频帧的顺序关系，则得分函数满足帧顺序的约束条件 $i<j,r_i<r_j$。这就是顺序池化的主要思想，即在满足视频帧顺序的约束条件下，从视频中提取特征，通过提取到的视频特征学习参数 $\boldsymbol{u}$，学习的过程可以通过 Ranksvm 算法来进行表述，即

$$\underset{\boldsymbol{u}}{\operatorname{argmin}}\frac{1}{2}\|\boldsymbol{u}\|^2+a\sum_{i<j}\boldsymbol{\xi}_i,\boldsymbol{\zeta}_j,$$
$$\text{s.t. }\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{v}_t-\boldsymbol{v}_i)\geqslant 1-\boldsymbol{\zeta}_{i,j},\boldsymbol{\zeta}_{i,j}\geqslant 0, \tag{6}$$

式中：a 为一个常数，满足约束条件 $a>0$；$\boldsymbol{\zeta}_{i,j}$ 表示松弛向量，是一个较小的非负数。

2.2　时序动态图

顺序池化算法的主要思想是在满足时序关系的约束条件下学习到参数 $\boldsymbol{u}$，作为该视频的特征描述子，用来表征整个视频的运动信息，因此将 $\boldsymbol{u}$ 定义为时序动态图。而该算法在进行池化操作时，提取到的特征更偏向于图像序列的起始帧，故采用双向顺序池化来降低偏差。按照帧顺序的约束条件从起始帧到结束帧所学习到的参数是正向顺序池化的过程，生成正向时序动态图(FSDI)。反之，当约束条件发生反转时，生成的是反向时序动态图(BSDI)。正向时序动态图和反向时序动态图主要获取视频帧随时间的动态变化，在表观性上并无本质区别，因此在实验中将其组合作为表观和长时运动流的输入。图 2 为静态视频帧以及对应的时序动态图示例。其中，图 2(a)为部分动作的单帧静态图像，图 2(b)为这些动作所生成的时序动态图，图 2(c)为这些动作对应的光流图。对比图 2(a)～(c)可以看出，时序动态图在去除背景冗余信息的同时，可以更好地表

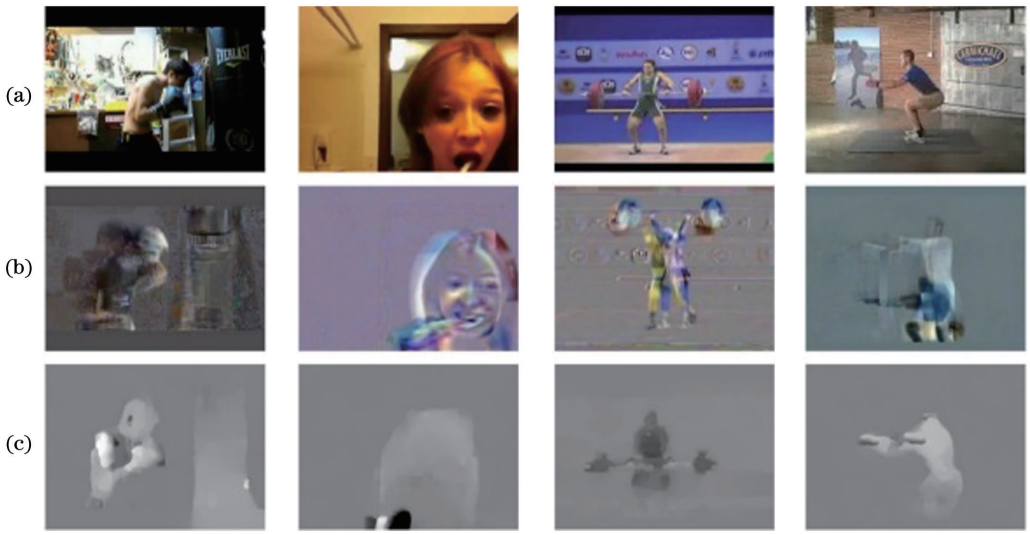

图 2　静态视频帧以及对应的时序动态图。(a)静态图像；(b)时序动态图；(c)光流图

Fig. 2　Static video frames and corresponding timing dynamic diagrams. (a) Static images; (b) timing dynamic diagrams; (c) optical flow diagrams

征长时间动作的运动信息；而光流更注重动作的瞬时特征。

3　TS-CNN

为了解决传统双流卷积网络难以提取长时序信息的问题，更好适配时序动态图作为输入方式，本文提出一种新的双流人体行为识别模型，即 TS-CNN，分别包含表观和长时序卷积网络以及短时序卷积网络，如图 3 所示。在此基础上采用稀疏采样的方式将视频平均分为四段，将每个视频片段通过 BRP 算法生成时序动态图作为表观和长时序卷积网络的输入，采用 inceptionV3 网络模型来提取视频中人体的表观和长时运动流信息。将每段视频片段随机抽取一帧光流（x 方向和 y 方向共 2 张图片）作为短时序卷积网络的输入，采用 inceptionV3 网络模型来提取动作的帧间运动信息。将各支流输出的类别判定分数通过平均池化的方式进行分数融合，得到最终的预测结果。

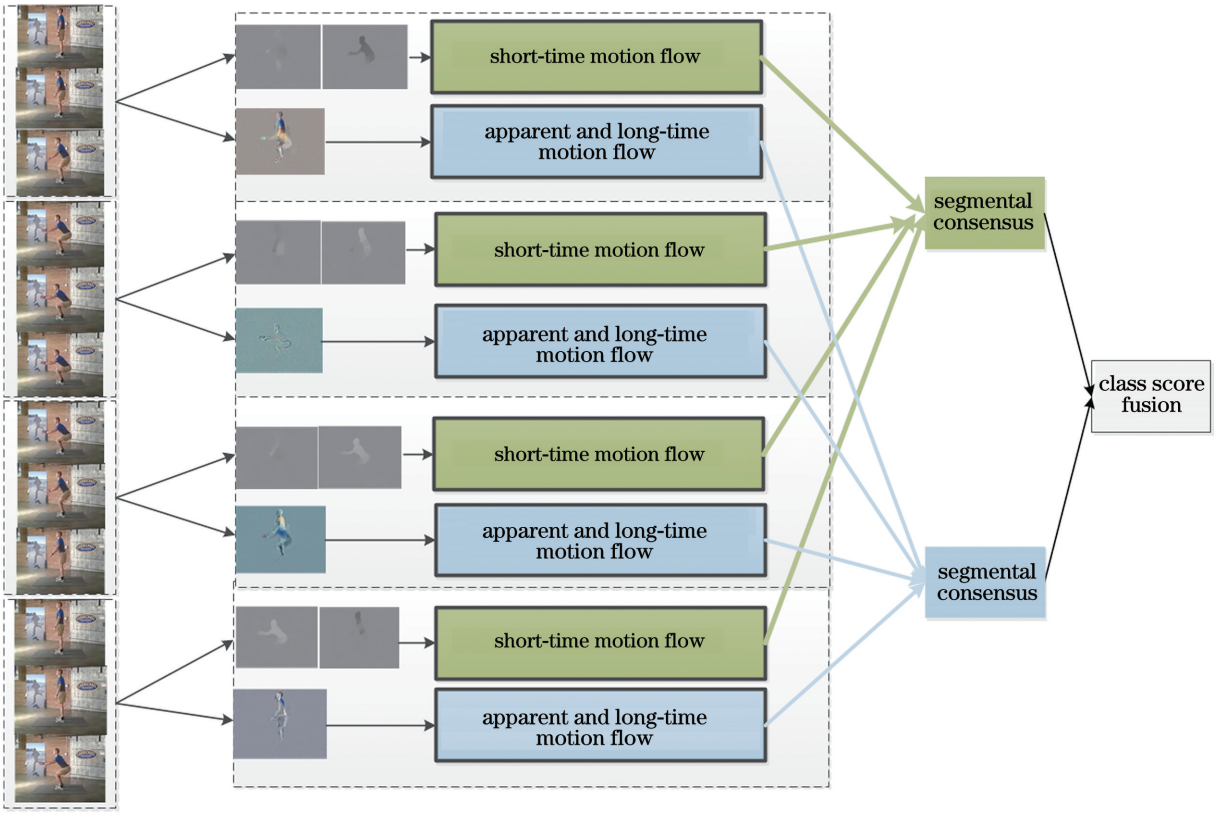

图 3　TS-CNN 网络框架

Fig. 3　TS-CNN network framework

3.1　表观和长时序卷积网络

该网络的输入方式时序动态图本质上是静态 RGB 图片，故可以直接利用二维卷积来进行图片特征的提取。之前的许多工作[16-17]表明，网络结构的深度可以提高人体行为识别的性能，但同时需要考虑到深度带来的计算成本。通过对比常用的几种网络结构，决定采用 inceptionV3 作为表观和长时序卷积网络的结构。该网络相比于其他网络结构具有很好的平衡性，在保证网络深度的同时，加速了计算过程、减弱了网络的非线性、降低了过拟合的概率。

inceptionV3 网络共有 42 层，其中包括 6 个卷积层、2 个池化层、3 个 inception 模块、1 个全连接层以及 1 个 Softmax 输出层。所有卷积核的尺寸均为 3×3，步长分别为 1 和 2。网络结构如表 1 所示。输入图片的尺寸为 299×299，通过卷积、池化和 inception 模块，最终经过全连接层得到 2048 维向量，通过 Softmax 输出层输出 1000 类。在 UCF101 和 HMDB51 数据集中分别加入新的 Softmax 输出层，输出类别数分别为 101 和 51。

时序动态图本质上是对视频特征的压缩，而压缩的视频帧数过多会丢失掉部分运动信息。因此对于每段行为视频通过稀疏采样进行分割，再通过顺序池化算法生成若干张时序动态图。首先，将行为视频平均分割成 n 个 w 帧的视频片段，再将每个片段进行特征压缩得到由 n 张时序动态图组成的图像序列。这些图像序列包含着整个行为视频的长时

时序信息，将图像序列的尺寸调整为 299×299，采用稀疏采样的方式将整个序列平分为四段，每段抽取一张图片作为 inceptionV3 网络的输入，将得到的类别分数通过平均池化进行段共识融合，得到最终的预测结果。而生成的时序动态图在数据量上与静态图像存在较大差距，在训练过程中容易导致网络产生过拟合的情况，影响识别效果。因此采用双向顺序池化以及数据增强的方式来对其数据量进行扩增，增强泛化性。

3.2　短时序卷积网络

短时序卷积网络以 inceptionV3 网络作为特征提取器来提取视频的帧间时序信息。相对于其他深度网络模型，inceptionV3 网络在保证网络深度的同时减少了参量个数，加速了计算过程，增加了网络的非线性，减小了模型在训练过程中过拟合的概率。网络提取到的视频帧间时序信息实际上是关于动作的短时运动信息。其表征方式即网络输入的是 x 方向和 y 方向的堆叠光流帧。光流利用图像序列中像素在时间域上的变化可以捕捉到场景中目标在帧前后的位移，且可以屏蔽掉相机运动和场景带来的影响。

在模型的训练过程中因数据集训练样本不足而导致模型的泛化能力较弱，容易出现过拟合的情况。为了避免此类风险，拟采用数据增强的方式对于标注样本进行 10 倍的数据增强。其中包括角点剪裁和颜色增强。角点剪裁首先将训练样本的尺寸从 299×299 缩放到 256×256，分别从中心和四个边角裁剪出尺寸为 224×224 的样本。在实验中发现颜色增强的训练样本有助于提高动作识别效果，因此在角点剪裁的数据基础上对其进行颜色增强，最终将进行角点剪裁和颜色增强的数据应用于短时序网络的训练，从而实现了数据的 10 倍增强。

4　实验结果与分析

4.1　实验环境设置

实验所采用的计算机操作系统为 Ubuntu16.04，图形处理器(GPU)为 NVIDIA RTX 2080Ti ×3。实验所采用的人体行为识别双流网络模型基于深度学习平台 Pytorch 结构搭建。采用小批量随机梯度下降算法来训练网络，动量设置为 0.9，批训练大小为 64，初始学习率为 0.001，网络在基于 ImageNet 的预训练模型上进行参数初始化。时序动态图的生成采用顺序池化算法，并利用 LibSVM 工具包和 Matlab2016b 对预处理后的 UCF101 和 HMDB51 标准行为数据集的视频特征进行压缩。光流采用 OpenCV 视觉库中 TVL1 算法，利用 denseflow 工具和 GPU 编译进行计算。

4.2　动作数据集

实验在 UCF101 和 HMDB51 视频行为数据集进行测试，验证所提出的人体行为识别方法。UCF101 数据集是由佛罗里达大学收集 Youtube 网站上的视频数据所建立的。该数据集包含 101 类动作，每个动作包括 25 组，每组包含 4～7 段视频，视频的空间分辨率为 320×240。在动作多样性、场景复杂度、背景扰动等方面都给人体行为识别带来巨大的挑战，是当前最为主流的动作识别数据集。HMDB51 数据集，共包含 51 类动作，6766 段视频，视频的空间分辨率为 320×240，均来源于 Youtube 网站和数字电影。视频的场景变化更为复杂多样，动作受到相机运动、背景、光照变化影响较大。

上述两个数据集的训练和测试均将数据集分为三组，每组的实验数据集和测试数据集的视频量比值为 3∶1，通过计算三组数据集的平均准确率来评估算法的优劣。

4.3　输入表征方式的对比

时序动态图的生成本质上是对于视频帧序列的有效压缩，而压缩的视频帧数过多会导致动作的部分关键信息丢失，从而影响识别率。因此在计算时序动态图时，需要对视频帧序列进行相应的预处理。首先将行为视频的帧序列重叠分割成若干个单位长度为 x、重叠部分长度为 $x/2$ 的子序列，对每个子序列片段进行压缩生成时序动态图。保留重叠部分的目的是防止行为视频在分割过程中丢失掉关键帧，影响动作识别的准确率。选择合适的子序列长度 x 对于保留动作的关键信息、达到长时时域建模等具有重要作用。图 4 所示为单独使用表观和长时序动

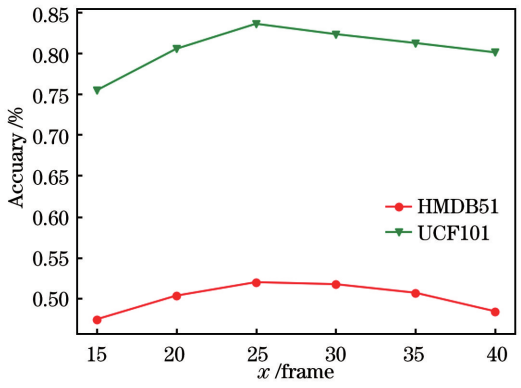

图 4　不同子序列长度的识别结果

Fig. 4　Recognition results of different subsequence lengths

作流进行行为识别时，子序列 x 的长度对于两个数据集结果的影响。而 HMDB51 数据集和 UCF101 数据集中视频的帧率为 25 frame/s，实验中发现，当子序列长度 x 值的设定小于 25 时，在动作幅度较大的行为特征压缩过程中容易出现时序动态图模糊的情况，对识别准确率造成一定的影响。因此实验选择的 x 最小值设置为 25。由图 4 可知，当 $x=25$ 时，在 HMDB51 数据集和 UCF101 数据集中取得最高识别率。因此，后续实验过程中选择的子序列长度为 25 frame。

生成时序动态图后，在 inceptionV3 框架下进行了多组对照实验，对比传统双流网络常见的输入 SI、SOF 和本文所提特征描述方式 FSDI、BSDI、SDI、经过数据增强后的时序动态图(ESDI)及其融合的情况，如表 1、2 所示。通过实验发现：1)时序动态图更注重视频特征中的逻辑关系，因此正向与反向对于实验结果无影响；2)由于时序动态图在一定程度上对视频进行了压缩，数据量的减少容易影响实验结果。本文提出的ESDI在UCF101数据集中相比 SI 的识别率提高了 2.4%，在 HMDB51 数据集中识别率提高了 3.3%。实验结果表明：时序动态图是一种高效的视频特征的描述方式，可以在表观和长时时域范围进行建模，应用于行为识别中可以提高准确率。

表 1 不同输入方式下 UCF101 数据集识别准确率

Table 1 Recognition accuracy of UCF101 dataset with different input modes unit: %

Method	Split1	Split2	Split3	Accuracy
SI	84.6	84.9	85.0	84.8
SOF	87.3	89.9	91.0	89.4
FSDI	83.9	83.8	83.1	83.6
BSDI	84.1	83.3	84.3	83.9
SDI	85.7	86.2	85.5	85.8
ESDI	87.2	86.8	87.6	87.2
SI+SOF	93.2	94.0	94.2	93.8
ESDI+SOF	94.8	94.6	95.3	94.9

表 2 不同输入方式下 HMDB51 数据集识别准确率

Table 2 Recognition accuracy of HMDB51 dataset with different input modes unit: %

Method	Split1	Split2	Split3	Accuracy
SI	54.8	50.4	49.6	51.6
SOF	64.2	63.6	62.7	63.5
FSDI	50.7	51.4	53.6	51.9
BSDI	51.6	51.5	54.1	52.4
SDI	54.5	52.9	53.7	53.7
ESDI	53.6	55.5	55.6	54.9
SI+SOF	68.7	67.5	68.4	68.2
ESDI+SOF	69.6	71.2	71.6	70.8

4.4 改进的双流卷积网络

本文的双流人体行为识别模型可分为表观和长时时序运动流以及短时时序运动流，输入分别为时序动态图、堆叠的光流帧序列，并且结合稀疏采样、跨模态预训练、数据增强等方式，两个支流得到的结果通过分数融合得到最终识别率。采用不同的融合方式在各个数据集中的识别准确率如表 3 所示。实验结果表明：采用平均池化作为分数融合的方式，实验结果最好。在后续实验中均采用平均池化作为两条支流融合的方式。

表 3 不同融合方式在数据集中识别准确率

Table 3 Recognition accuracy of different fusion methods on dataset unit: %

Consensus function	UCF101	HMDB51
Max	93.0	69.1
Average	94.9	70.8
Weighted average	93.8	69.7

在网络架构方面，本文选取了当前主流的网络模型 Resnet101、Bn-inception 以及 inceptionV3 作为两条支流的主要网络结构，分别测试其对于人体行为识别准确率的影响。实验结果如表 4 所示，通过对比发现以 inceptionV3 为主要网络结构的双流模型在两个数据集的识别率方面相比 Resnet101、Bn-inception 均有不同程度的提高。因此选择 inceptionV3 作为本文的主要网络结构。

表 4 不同网络模型在数据集中的识别准确率

Table 4 Recognition accuracy of different network models on dataset unit: %

Network structure	UCF101	HMDB51
Resnet101	93.6	68.4
Bn-inception	94.2	68.2
InceptionV3	94.9	70.8

为了评价本文提出双流网络的性能，分别测试表观和长时序运动流、短时序运动流和经过融合后 TS-CNN 在 UCF101、HMDB51 两个数据集的识别准确率，将其结果与文献[17]的原始双流卷积网络、

文献[19]提出的网络以及其他主流人体行为识别模型进行对比，实验结果如表 5 所示。结果表明：本文提出的双流网络融合后的结果比其他两个网络在 UCF101 数据集上分别提高 6.9%、0.9%，在 HMDB511 数据集上分别提高 11.4%、1.4%，在准确率上均有不同程度的提高。

表 5 不同人体行为识别模型的识别准确率

Table 5 Recognition accuracy of different human behavior recognition models unit：%

Network	UCF101	HMDB51
Spatial stream	84.8	51.4
Temproral stream	89.4	63.5
Original two-stream	88.0	59.4
Ref. [19]	94.0	69.4
Appearance and long-sequential stream	87.2	54.9
Short sequential stream	89.9	64
TS-CNN	94.9	70.8

4.5 运算速度

人体行为识别模型在训练阶段涉及到神经网络的反向传播过程，需要对网络的各个参数进行梯度计算，并且利用梯度下降的方法进行参数更新，上述过程一般需要多次迭代运算；在测试阶段，每个输入视频仅涉及到一个前向运算过程，即可得到输出结果，在时间复杂度和资源消耗上远小于训练阶段。

为了测试本文算法在识别速度上的性能，在训练好网络模型后，对验证数据集进行测试时，需要统计视频预处理和前向运算两者所需要的时间，再统计测试视频的总帧数，以处理帧率作为评价指标。在 UCF101 和 HMDB51 数据集中对本文算法进行了行为识别速度的测试，取得了分别为 77 frame/s 和 89 frame/s 的运算速度。在同等条件下，对于文献[19]的网络进行测试取得了 31 frame/s 和 37 frame/s 的结果。实验结果表明：基于时序动态图对于视频特征的有效压缩和 inceptionV3 网络结构在运算效率上的优势，本文算法在识别速度上优于传统双流卷积网络，且满足实时性的要求。

4.6 算法对比

为了对本文算法的性能做出客观的评价，针对 HMDB51 和 UCF101 数据集，以动作的平均识别准确率作为评价指标，与现有文献中基于传统的特征提取以及深度学习的算法进行比较，各个算法的识别结果如表 6 所示。

表 6 不同算法的识别准确率

Table 6 Recognition accuracy of different algorithms unit：%

Feature extraction	Method	UCF101	HMDB51
Tradition	Ref. [7]	84.8	57.2
	Ref. [8]	87.9	61.1
Deep learning	Ref. [17]	88.0	59.4
	Ref. [21]	88.6	–
	Ref. [22]	91.5	65.9
	Ref. [23]	93.1	63.3
	Ref. [24]	93.4	66.4
	Ref. [19]	94.0	69.4
	Proposed	94.9	70.8

由表 6 对比可知，基于深度学习的特征提取算法对比传统算法更容易学习到深层次的信息，更好地提取到视频中的动作特征，因此识别准确率较高。本文提出的结合表观和长时运动流以及短时运动流的双流算法相比其他深度卷积神经网络可以更好地提取视频的长时运动信息，在长时时域范围对于复杂动作进行建模，可以有效提高识别准确率。

5 结 论

本文提出了一种结合时序动态图和双流卷积神经网络的人体行为识别算法。算法利用时序动态图来提取视频中动作的表观和长时时域信息，并且通过 inceptionV3 网络构造了一种包含表观和长时运动流以及短时运动流的双流卷积网络，分别以时序动态图和堆叠的光流帧序列作为输入，通过平均池化的方式进行融合得到识别结果。实验结果表明，以时序动态图作为新的视频特征描述方式，可以构建视频的表观和长时时域结构，结合以 inceptionV3 为网络结构的双流卷积神经网络分别在 HMDB51 和 UCF101 数据集上与其他算法进行比较，显著提高了识别的准确率，验证了本文算法的有效性和鲁棒性。

参 考 文 献

[1] Zhu Y, Zhao J K, Wang Y N, et al. Areview of human action recognition based on deep learning[J]. Acta Automatica Sinica, 2016, 42(6): 848-857.
朱煜, 赵江坤, 王逸宁, 等. 基于深度学习的人体行为识别算法综述[J]. 自动化学报, 2016, 42(6):

848-857.

[2] Li Y P, Liu T T, Zhang L. Human action recognition based on deep learning[J]. Application Research of Computers, 2020, 37(1): 304-307, 316.
李玉鹏, 刘婷婷, 张良. 基于深度学习的人体动作识别方法[J]. 计算机应用研究, 2020, 37(1): 304-307, 316.

[3] Luo H L, Tong K, Kong F S. Theprogress of human action recognition in videos based on deep learning: a review[J]. Acta Electronica Sinica, 2019, 47(5): 1162-1173.
罗会兰, 童康, 孔繁胜. 基于深度学习的视频中人体动作识别进展综述[J]. 电子学报, 2019, 47(5): 1162-1173.

[4] Li Q H, Li A H, Wang T, et al. Double-stream convolutional networks with sequential optical flow image for action recognition[J]. Acta Optica Sinica, 2018, 38(6): 0615002.
李庆辉, 李艾华, 王涛, 等. 结合有序光流图和双流卷积网络的行为识别[J]. 光学学报, 2018, 38(6): 0615002.

[5] Liu F, Yu F Q. Humanaction recognition based on global and local features[J]. Laser & Optoelectronics Progress, 2020, 57(2): 021004.
刘帆, 于凤芹. 基于全局和局部特征的人体行为识别[J]. 激光与光电子学进展, 2020, 57(2): 021004.

[6] Huang Y W, Wan C L, Feng H. Multi-feature fusion human behavior recognition algorithm based on convolutional neural network and long short term memory neural network[J]. Laser & Optoelectronics Progress, 2019, 56(7): 071505.
黄友文, 万超伦, 冯恒. 基于卷积神经网络与长短期记忆神经网络的多特征融合人体行为识别算法[J]. 激光与光电子学进展, 2019, 56(7): 071505.

[7] Wang H, Kläser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International Journal of Computer Vision, 2013, 103(1): 60-79.

[8] Wang H, Schmid C. Actionrecognition with improved trajectories[C]//2013 IEEE International Conference on Computer Vision, December 1-8, 2013, Sydney, NSW, Australia. New York: IEEE Press, 2013: 3551-3558.

[9] Sun S Y, Kuang Z H, Sheng L, et al. Optical flow guided feature: a fast and robust motion representation for video action recognition[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1390-1399.

[10] Zhang B W, Wang L M, Wang Z, et al. Real-time action recognition with enhanced motion vector CNNs[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2718-2726.

[11] Wang L L, Ge L Z, Li R F, et al. Three-stream CNNs for action recognition[J]. Pattern Recognition Letters, 2017, 92: 33-40.

[12] Shi Y M, Tian Y H, Wang Y W, et al. Sequential deep trajectory descriptor for action recognition with three-stream CNN[J]. IEEE Transactions on Multimedia, 2017, 19(7): 1510-1520.

[13] Chen S H, Chen Z Z. On human behavior recognition with deep learning and IR spectral signal restoration technologies in a natural classroom[J]. Infrared Physics & Technology, 2020, 105: 103167.

[14] Arivazhagan S, Shebiah R N, Harini R, et al. Human action recognition from RGB-D data using complete local binary pattern[J]. Cognitive Systems Research, 2019, 58: 94-104.

[15] Fernando B, Gavves E, Oramas M J, et al. Rank pooling for action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 773-787.

[16] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 1725-1732.

[17] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[EB/OL]. (2014-11-12)[2020-07-07]. https://arxiv.org/abs/1406.2199.

[18] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 4489-4497.

[19] Wang L M, Xiong Y J, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition[EB/OL]. (2016-08-02)[2020-07-07]. https://arxiv.org/abs/1608.00859.

[20] Lan Z Z, Zhu Y, Hauptmann A G, et al. Deep local video feature for action recognition[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1219-1225.

[21] Ng Y H, Hausknecht M, Vijayanarasimhan S, et al.

Beyond short snippets: deep networks for video classification[J]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 4694-4702.

[22] Wang L M, Qiao Y, Tang X O. Action recognition with trajectory-pooled deep-convolutional descriptors [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 4305-4314.

[23] Zhu W J, Hu J, Sun G, et al. A key volume mining deep framework for action recognition [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 1991-1999.

[24] Carreira J, Zisserman A. Quovadis, action recognition? A new model and the kinetics dataset [J]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 4724-4733.