结合有序光流图和双流卷积网络的行为识别

李庆辉, 李艾华, 王涛, 崔智高
火箭军工程大学作战保障学院, 陕西 西安 710025

摘要 为有效利用行为视频的长时时域信息,提高行为识别准确率,提出一种结合有序光流图和双流卷积神经网络的行为识别算法。首先利用 Rank 支持向量机(SVM)算法将连续光流序列压缩总结成单幅有序光流图,实现对视频长时时域结构的建模;然后设计一个包含表观和短时运动流与长时运动流的双流卷积网络,分别以堆叠 RGB 帧、有序光流图为输入提取视频的表观和短时运动信息与长时运动信息;最后将双流网络的 C3D 描述子和 VGG 描述子融合后输入线性 SVM 进行行为识别。在 HMDB51 和 UCF101 两个数据集的实验结果表明,该算法能够有效利用空域表观信息和时域运动信息,具有较高的行为视频识别准确率。

关键词 机器视觉; 行为识别; 有序光流图; 卷积神经网络; 支持向量机
中图分类号 TP391 **文献标识码** A **doi**: 10.3788/AOS201838.0615002

Double-Stream Convolutional Networks with Sequential Optical Flow Image for Action Recognition

Li Qinghui, Li Aihua, Wang Tao, Cui Zhigao
Academy of Operational Support, Rocket Force Engineering University, Xi'an, Shaanxi 710025, *China*

Abstract In order to effectively utilize the long-term temporal information of video for improving the accuracy of action recognition, a new recognition approach is proposed based on the sequential optical flow image and double-stream convolutional neural networks. Firstly, the Rank support vector machine (SVM) algorithm is used to compress the continuous optical flow frames into a single sequential optical flow image to realize the modeling of the long-term temporal structure of video. Secondly, we design a double-stream convolutional networks containing appearance and short-term motion stream and long-term motion stream. It takes the stacked RGB frames and the sequential optical flow images as input to extract the appearance and short-time motion information and the long-time motion information of the video. Finally, the linear SVM is adopted to integrate C3D descriptor and VGG descriptor for action recognition. The experimental results on HMDB51 and UCF101 datasets show that the proposed approach improves the action recognition accuracy effectively by using the spatial information and the temporal motion information.

Key words machine vision; action recognition; sequential optical flow image; convolutional neural network; support vector machine
OCIS codes 150.0155; 100.4996; 100.2960; 200.4260

1 引 言

人体行为识别(HAR)已经成为机器视觉领域的研究热点之一[1-2]。利用行为识别技术,计算机可以自动地理解和描述视频中的人体行为,从而将底层视频数据与高层语义自动关联起来,所以行为识别在视频监控、人机交互、运动分析和基于内容的视频检索等领域具有巨大的应用价值。由于视频拍摄视角、尺度、背景的复杂多变,以及行为的类内差异性和类间相似性,行为识别研究面临巨大的挑战。

行为识别方法主要分为基于人工设计特征的方法和基于深度学习特征的方法两类[3],其中后者是

收稿日期: 2017-11-27; **收到修改稿日期**: 2018-01-04
基金项目: 国家自然科学基金(61501470)、陕西省重点研发计划(2017GY-075)
作者简介: 李庆辉(1989—),男,博士研究生,主要从事机器视觉方面的研究。E-mail: lqhui1212@126.com
导师简介: 李艾华(1966—),男,博士,教授,博士生导师,主要从事机器视觉和人工智能方面的研究。E-mail: ahli66@126.com

当前的研究热点和难点之一。Karpathy 等[4]首先在 Sports-1M 数据集上测试了深度卷积网络，将堆叠的连续 RGB 视频帧直接输入网络进行识别；Simonyan 等[5]设计了一种包含空域网络和时域网络的双流卷积神经网络(CNN)，分别将单帧 RGB 图像和堆叠光流位移场输入空域网络和时域网络以获取视频的表观和运动信息；Tran 等[6]将 2D 空间卷积核扩展到时域网络，并提出了 3D 卷积网络，以多帧图像为输入单元，通过多次交替卷积、池化操作来学习视频的时空特征。此外，研究者发现利用深度网络提取视频的长时时域信息可以有效地提高行为识别准确率。Donahue 等[7]根据递归神经网络编码卷积特征的时域关系，提出了融合卷积层和长时递归层的长时递归卷积网络(LRCN)；Wang 等[8]在稀疏时域片段上构建长时时域结构，利用稀疏采样抽取多个视频片段，分别在每个片段上建立双流卷积网络，最后融合所有网络的输出结果进行预测分类；Varol 等[9]通过扩展 3D 卷积网络输入的时间长度设计了一种长时卷积(LTC)网络，并研究了不同底层表示(原始像素值、光流等)对识别结果的影响。

行为视频作为连续的图像序列，其静态表观信息、短时时域信息，以及长时时域信息的有效利用对行为识别具有重要意义。针对上述问题，本文提出一种结合有序光流图和双流卷积网络的行为识别算法。该方法通过有序光流图建模视频的长时时域结构，利用一个包含表观和短时运动流、长时运动流的双流卷积网络提取行为视频的表观信息和长短时运动信息，结合线性支持向量机(SVM)对行为视频进行识别。

2　有序光流图

视频可看作多帧静态图像在时间维度的顺序排列，可以从空间和时间两个部分来分析视频信息。空间信息表现为视频的每帧图像，可以描述视频中的场景和物体；时间信息表现为帧与帧之间的运动变化，可以描述视频中的物体和相机的运动[5]。视频中的复杂行为往往由数十帧甚至数百帧来共同呈现，因此提取视频的时域运动信息，尤其是长时运动信息，对识别视频中的行为具有非常重要的作用。

通常利用光流序列来表示视频的时域运动信息，但受网络参数限制，现有深度模型难以处理超过 10 帧的光流序列输入，因此难以提取视频的长时时域信息。受文献[10]的启发，本文在保留次序信息的条件下，将光流序列压缩总结到单幅图像上，并将这个单幅图像作为深度网络的输入信息，从而提取更长时间的运动信息。

给定一个 n 帧连续光流序列 $\boldsymbol{F}=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n]$，其中 $\boldsymbol{f}_i\in\mathbf{R}^{d_1\times d_2\times 2}$，$d_1$、$d_2$ 分别为光流图的高度和宽度。每帧光流图均为双通道图像对应于光流的水平分量和垂直分量，表示为 $\boldsymbol{f}_i^x$，$\boldsymbol{f}_i^y$。定义第 t 帧光流图 $\boldsymbol{f}_t$ 对应的加权移动平均图为

$$\hat{\boldsymbol{f}}_t=\sum_{i=1}^{t}\frac{i}{\sum_{j=1}^{t}j}\boldsymbol{f}_i,\tag{1}$$

(1)式的加权平均方法可以同时降低光流估计的错误率和白噪声的影响。

在光流序列的加权移动平均图上计算有序光流图，计算公式如下：

$$\begin{aligned}&\min_{\boldsymbol{G}\in\mathbf{R}^{d_1\times d_2\times 2},\xi\geqslant 0}\|\boldsymbol{G}\|^2+C\sum_{i<j}\xi_{ij}\\&\text{s.t.}\langle\boldsymbol{G},\hat{\boldsymbol{f}}_j\rangle-\langle\boldsymbol{G},\hat{\boldsymbol{f}}_i\rangle\geqslant 1-\xi_{ij},\ \forall i<j\end{aligned},\tag{2}$$

式中$\langle\cdot,\cdot\rangle$为内积，$C$为边界大小与训练误差之间的折中参数，$\xi_{ij}$为松弛变量。(2)式来源于排序算法 Rank SVM，约束条件$\langle\boldsymbol{G},\boldsymbol{f}_j\rangle-\langle\boldsymbol{G},\boldsymbol{f}_i\rangle\geqslant 1-\xi_{ij}$($\forall i<j$)保留了光流帧的顺序信息。采用训练学习得到的参数 $\boldsymbol{G}\in\mathbf{R}^{d_1\times d_2\times 2}$ 表示光流序列，事实上其与光流图的大小是相同的，因此将 G 定义为有序光流图(SOFI)。(2)式的求解等价于无约束优化问题，即最小化 Hinge Loss 函数：

$$\min_{\boldsymbol{G}\in\mathbf{R}^{d_1\times d_2\times 2}}\sum_{i<j}[1-\langle\boldsymbol{G},\boldsymbol{f}_j\rangle+\langle\boldsymbol{G},\boldsymbol{f}_i\rangle]_+\lambda\|\boldsymbol{G}\|^2,\tag{3}$$

式中$[\cdot]_+$为函数 $\max(0,x)$，$\lambda=1/C$。

需要注意的是，光流图的两个通道不是图像的颜色通道，而是速度向量的通道，两者共同描述每个像素点位置的运动向量，因此两者是相关的。但是 Rank SVM 算法默认不同通道是独立的，解决办法是通过矩阵对角化对两个通道进行去相关。实验中发现这种去相关操作并不能明显地提升性能，因此本文忽略这种相关关系。假设 $\boldsymbol{G}_x$、$\boldsymbol{G}_y\in\mathbf{R}^{d_1\times d_2}$ 为有序光流图 $\boldsymbol{G}$，分别对应于光流的水平和垂直分量的两个通道，则(2)式可转化为

$$\begin{aligned}&\min_{\boldsymbol{G}_x,\boldsymbol{G}_y\in\mathbf{R}^{d_1\times d_2},\xi\geqslant 0}\|\boldsymbol{G}_x\|^2+\|\boldsymbol{G}_y\|^2+C\sum_{i<j}\xi_{ij}\\&\text{s.t.}\langle\boldsymbol{G}_x,\hat{\boldsymbol{f}}_j^x\rangle+\langle\boldsymbol{G}_y,\hat{\boldsymbol{f}}_j^y\rangle-\langle\boldsymbol{G}_x,\hat{\boldsymbol{f}}_i^x\rangle-\\&\langle G_y,\hat{f}_i^y\rangle\geqslant 1-\xi_{ij},\ \forall i<j\end{aligned},\tag{4}$$

将得到的 $\boldsymbol{G}_x$、$\boldsymbol{G}_y$ 利用最小-最大规范化转换到[0,255]范围内叠加生成有序光流图，并将其作为深度网络

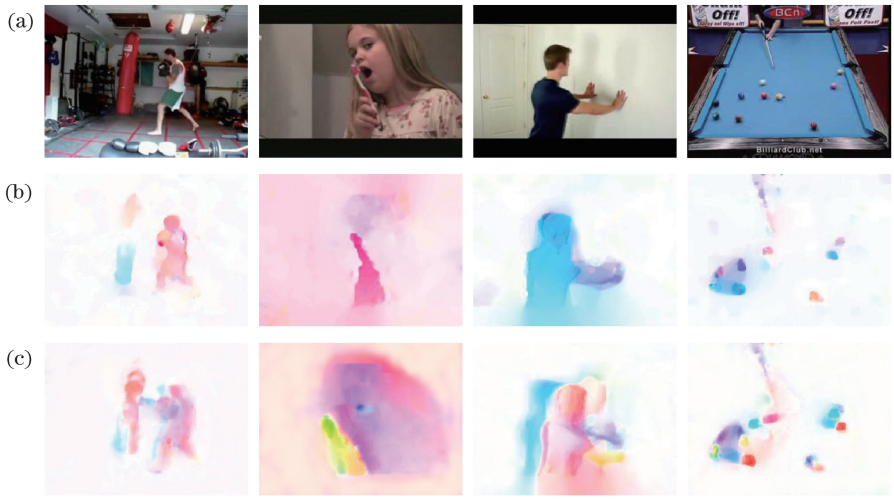

图 1　(a)原始视频帧；(b)光流图；(c)有序光流图

Fig. 1　(a) Original video frames; (b) optical flow images; (c) sequential optical flow images

的输入。通过以上过程实现从 n 帧光流序列到单幅有序光流图的映射，图 1 为有序光流图及其相应 RGB 视频帧、光流图示例，从图 1(c)可以看出，有序光流图可以表达多帧视频序列的运动信息。

3　双流卷积网络

为充分利用视频序列的表观信息、短时运动信息以及长时运动信息，提出一种双流卷积网络框架[11]。本文将双流卷积网络分成表观和短时运动流(A&STM stream)与长时运动流(LTM stream)，如图 2 所示。在表观和短时运动流中，以堆叠 RGB 帧序列为输入信息，采用 C3D net[6] 提取行为视频的表观和短时运动特征；在长时运动流中，以有序光流图为输入信息，采用 VGG-16 net[12] 提取行为视频的长时运动特征。最后融合两个网络 fc 6 层的输出响应对输入线性 SVM 进行分类识别。

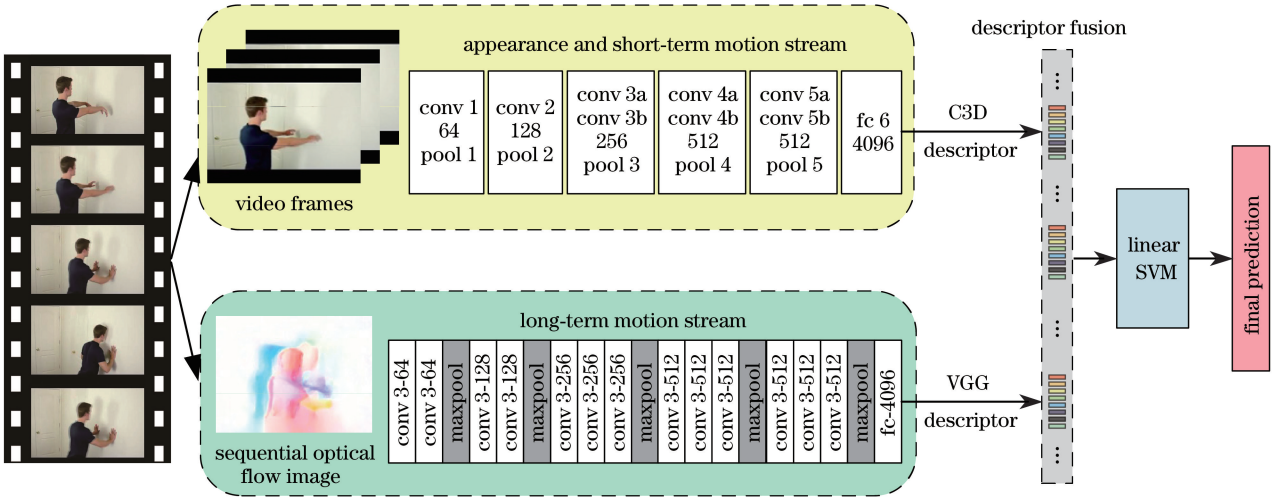

图 2　融合表观信息和运动信息的双流网络框架

Fig. 2　Double-stream network framework fusing appearance information and motion information

3.1　表观和短时运动流

表观和短时运动流以 3D 卷积网络 C3D net 作为特征提取器提取视频的表观和短时运动特征。C3D net 利用 3D 卷积核和池化核可以同时在时—空维度对多帧视频序列进行卷积和池化操作，能够提取空域表观信息和时域运动信息。这种时域运动信息实际上是一种短时运动信息。

C3D net 包含：8 个卷积层(conv x)，每层卷积核数如图 2 所示，尺寸为 3×3×3，步长为 1；5 个最大池化层(pool y)，除 pool 1 的池化核尺寸为 1×2×2 外，其余池化核尺寸均为 2×2×2；2 个全连接层(fc z)，每个全连接层的输出响应为 4096 维；1 个 softmax 输出层。网络以 16 帧的片段为输入单元，相邻片段重叠 8 帧，输入图片尺寸为 224 pixel×

224 pixel。将行为视频所有片段的 fc 6 层响应取平均并进行 L2 归一化，得到的 4096 维向量作为该视频的 C3D 特征。

3.2　长时运动流

有序光流图是单幅图像，可以直接利用 2D 卷积网络提取特征向量。VGG-16 net 包含：13 个卷积层，所有卷积核尺寸均为 3×3，步长为 1，每层卷积核数如图 2 所示，部分卷积层包含最大池化操作；3 个全连接层，输出响应分别为 4096、4096、1000 维；1 个 softmax 输出层。

在生成有序光流图时，为避免压缩的光流帧过多而导致信息丢失，对每段行为视频生成若干个有序光流图。首先，对于一段光流序列首先在时间维度上分成若干个以 w 帧为单位的子序列，间隔为 $w/2$，亦即相邻的子序列之间重叠 $w/2$ 帧。然后，在每个子序列上分别建立一个有序光流图，将这些有序光流图输入 VGG-16 net，输入图像尺寸同样调整为 224 pixel×224 pixel。将所有有序光流图的fc 6 层响应取平均并进行 L2 归一化得到 VGG 特征。

在训练深度网络时容易因标注样本不足而导致过拟合，降低了网络泛化能力[13]。为避免这种风险，采用两种策略对长时运动流的数据进行10 倍增强：角点裁剪和尺度抖动。在角点裁剪中，首先将图像尺寸缩放为 256 pixel×256 pixel，然后从中心和 4 个对角区域将图像裁剪为 5 个 224 pixel×224 pixel 的子图像，从而实现数据的 5 倍增强。尺度抖动是一种多尺度裁剪过程，首先将输入图像尺寸固定为 256 pixel×340 pixel，然后在角点裁剪的 5 个位置从{256，224，192，168}任选值作为宽和高对输入图像进行裁剪，最后将所有裁剪区域缩放为 224 pixel×224 pixel，这种方法同样实现了数据的 5 倍增强。

4　实验与分析

4.1　行为数据集

在常用的两个标准数据集 HMDB51 和 UCF101 上验证提出的行为识别方法。

HMDB51 数据集主要来源于视频网站 Youtube 和数字电影，包含 51 类共 6766 段行为视频，空间分辨率为 320 pixel×240 pixel。视频多数由真实场景下非固定相机拍摄，存在大量的面部动作、肢体动作以及交互行为。将该数据集分成三组，并以三组的平均准确率评估算法。

UCF101 数据集由中佛罗里达大学建立，包含 101 类共 13320 段行为视频。每类行为视频分成 25 组，每组至少包含 100 段视频，视频长度为 29～1776 帧，空间分辨率为 320 pixel×240 pixel。视频拍摄场景更为复杂，存在背景扰动、相机运动、尺度和光照变化。同样将该数据集分成三组，计算三组的平均准确率。

4.2　实验设置

实验计算机配置为 Intel Core i7-6700@3.4 GHz，NVIDA GeForce TITANX GPU，操作系统为 Ubuntu 15.10。实验中，双流卷积网络基于 Caffe 平台设计实现。网络训练采用小批量随机梯度下降法，动量为 0.9，权值衰减率为 0.0005，HMDB51 数据集的批大小为 64，UCF101 数据集的批大小为 128。对于表观和短时运动流，采用在 Sports-1M 行为库上预训练的 C3D net，初始学习率设为 0.005；对于长时运动流，采用在 ImageNet 图像库预训练的 VGG-16 net，初始学习率设为 0.001。光流的计算采用 OpenCV 视觉库中的 TVL1 算法，并利用 LibSVM 工具包在不同子序列计算有序光流图。

4.3　有序光流图

计算有序光流图时，首先将行为视频的光流序列有重叠地分割成若干个以 w 帧为单位的子序列，然后在每个子序列上计算有序光流图。如果子序列帧数过少，无法达到建模长时时域结构的目的，过多则可能会丢失部分运动信息，所以首先需要确定合理的子序列长度。图 3 所示为单独使用长时运动流进行行为识别时，不同子序列长度 w 在两个数据集上对应的识别结果。由图 3 可知，w 取 24 和 28 时，分别在 HMDB51 和 UCF101 上取得最高识别结果，因此，实验中子序列长度取中间值 26 帧。

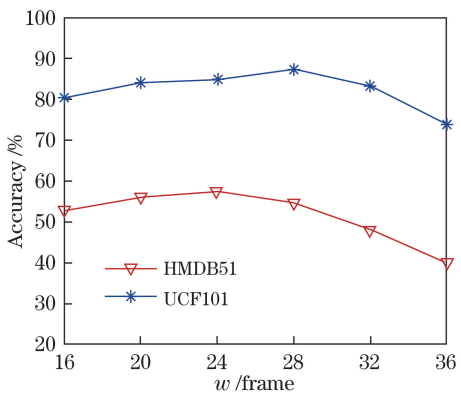

图 3　不同子序列长度的识别结果

Fig. 3　Results of different subsequence lengths

有序光流图实质上是对多帧光流图的有效压

缩，能够提取对识别行为有更重要意义的长时运动信息。在 VGG-16 net 框架下进行了多组验证实验，对比对象为卷积网络常用的输入：静态图像(SI)、堆叠光流场(SOF)、动态图(DI)[10]以及其组合。实验结果分别如表 1、2 所示，本文的 SOFI 对比 SI、SOF、DI 在 HMDB51 上识别准确率分别提高了 8%、3.4%、5.7%，在 UCF101 上识别准确率分别提高了 4.4%、5.6%、2.4%。在输入组合后，实验结果进一步提高，尤其是 SOFI＋SI 组合在两个数据集上分别取得最高识别结果 62.5%和 90.3%。实验结果表明：有序光流图是一种高效的视频表示，在应用到卷积网络后能够提高行为识别准确率。

表 1　HMDB51 识别准确率

Table 1　Recognition accuracy of HMDB51　%

Method	Split 1	Split 2	Split 3	Average
SI	49.1	50.6	49.6	49.8
SOF	55.2	53.4	54.7	54.4
DI	50.7	52.5	53.1	52.1
SOFI	57.8	58.4	57.2	**57.8**
SOFI＋DI	58.1	58.9	58.4	58.5
SOFI＋SI	63.3	61.8	62.5	**62.5**

表 2　UCF101 识别准确率

Table 2　Recognition accuracy of UCF101　%

Method	Split 1	Split 2	Split 3	Average
SI	81.4	80.5	81.9	81.3
SOF	79.3	81.5	79.5	80.1
DI	83.4	83.9	82.6	83.3
SOFI	85.8	86.1	85.2	**85.7**
SOFI＋DI	87.7	86.9	87.3	87.3
SOFI＋SI	89.6	90.9	90.3	**90.3**

4.4　双流卷积网络

本文的双流卷积网络分为表观和短时运动流与长时运动流，输入分别为堆叠 RGB 帧序列、有序光流图。为验证这种网络框架的有效性，分别测试两个支流网络以及融合后双流网络对 HMDB51、UCF101 数据集的识别结果。在测试支流网络时，取各自 fc 6 层响应作为描述子，经 L2 归一化后输入线性 SVM 分类器进行分类识别。实验对比方法为文献[5]的原始双流卷积网络(包括空间流和时间流)和文献[14]的 ST-ResNet(包括表观流和运动流)，实验结果如表 3 所示。由实验结果可知，融合后的双流网络识别结果比两个支流在 HMDB51 上分别提高了 7.7%、14.8%，在 UCF101 上分别提高了 4.7%、13.1%。对比三种双流网络，本文的双流网络比原始双流网络和 ST-ResNet 在两个数据集的识别结果均有不同程度的提高。实验结果表明：本文提出的双流卷积网络能够有效地融合行为视频的表观和长短时运动信息，识别准确率较高。

表 3　不同卷积网络的识别准确率

Table 3　Recognition accuracy of different convolutional networks　%

Network	HMDB51	UCF101
Spatial stream	41.6	81.2
Temporal stream	54.3	75.6
Original double-stream	59.4	88.0
Appearance stream	43.4	82.3
Motion stream	55.4	79.1
ST-ResNet	65.6	92.7
A&STM stream	64.9	90.1
LTM stream	57.8	81.7
Proposed double-stream	**72.6**	**94.8**

在两个数据集中，对比原始双流卷积网络，本文算法的准确率提高量排名前 10 位的行为类别如表 4 所示。HMDB51 中准确率提高量较大的行为类别为 Cartwheel、Climb_stairs、Swing_baseball 等；UCF101 中准确率提高量较大的行为类别为 IceDancing、Hammering、FloorGymnastics 等。这些行为相对复杂，延续时间长，而且在短时时域表现上和其他行为存在相似性，例如 Cartwheel 和 Handstand、Swing_baseball 和 Hit、IceDancing 和 ShakeHands 等。

表 4　准确率提高量排名前 10 的行为类别

Table 4　TOP10 categories of accuracy improvement　%

HMDB51		UCF101	
Action category	Increment	Action category	Increment
Cartwheel	35.6	IceDancing	24.6
Climb_stairs	32.3	Hammering	22.3
Swing_baseball	31.7	FloorGymnastics	17.8
Hit	30.0	JumpRope	17.2
Handstand	29.6	Fencing	16.4
Smoke	26.5	BrushingTeeth	13.9
Drink	24.3	Skiing	12.8
Draw_sword	19.8	Nunchucks	11.6
Shoot_ball	17.4	CricketBowling	11.2
Wave	16.9	HighJump	10.7

4.5 算法对比

为了体现算法的优势，针对 HMDB51 和 UCF101 两个数据集，将本文算法（SOFI＋Double-stream）与现有文献中的算法进行对比，各算法的识别结果如表 5 所示。

表 5　不同算法的识别准确率对比

Table 5　Comparison of recognition accuracy for different methods　%

	Method	HMDB51	UCF101
Shallow	IDT＋FV[15]	57.2	84.8
	IDT＋HSV[16]	61.1	87.9
	MoFAP[17]	61.7	88.3
Deep	CNN-hid6＋IDT[18]	—	89.6
	TDD＋IDT[19]	65.9	91.5
	TSN[8]	71.0	94.0
	I3D＋Double-stream[20]	66.4	93.4
	SOFI＋Double-stream	**72.6**	**94.8**

由实验对比结果可知，基于深度神经网络的方法能够学习得到行为视频的高层次语义信息，识别准确率高于只能获得浅层局部信息的人工设计特征方法；在基于深度神经网络的方法中，引入支流网络分别提取空域和时域信息的方法可以提高识别准确率（如 TSN、I3D＋Double-stream）。本文算法利用 C3D net 和 VGG-16 net 组成双流深度卷积网络分别提取表观和短时运动信息与长时运动信息，有效提高了识别准确率。实验中，本文算法在 HMDB51 和 UCF101 两个数据集上的运算速度分别为 53 frame·s^{-1} 和 71 frame·s^{-1}，满足实时性要求。

5　结　论

为有效利用行为视频的长时运动信息，提出一种结合有序光流图和双流卷积神经网络的行为识别算法。算法首先通过有序光流图建模视频的长时时域结构，并利用 C3D net 和 VGG-16 net 构造一个包含表观和短时运动流与长时运动流的双流卷积网络；然后分别以堆叠 RGB 帧、有序光流图为输入提取视频的表观和短时运动信息与长时运动信息；最后采用线性 SVM 对行为视频进行分类。在 HMDB51 和 UCF101 两个数据集上对本文算法的有序光流图和双流卷积网络分别进行了实验验证，并与几种先进算法进行了对比。多组实验结果表明，本文算法能够有效提高行为识别的准确率。

参　考　文　献

[1] Herath S, Harandi M, Porikli F. Going deeper into action recognition: A survey[J]. Image and Vision Computing, 2017, 60: 4-21.

[2] Ma M, Li Y B. Multi-level image sequences and convolutional neural networks based human action recognition method[J]. Journal of Jilin University Engineering and Technology Edition, 2017, 47(4): 1244-1252.
马淼, 李贻斌. 基于多级图像序列和卷积神经网络的人体行为识别[J]. 吉林大学学报(工学版), 2017, 47(4): 1244-1252.

[3] Subetha T, Chitrakala S. A survey on human activity recognition from videos[C]// Proceedings of IEEE International Conference on Information Communication and Embedded Systems, 2016: 1-7.

[4] Karpathy A, Toderici G, Shetty S, *et al*. Large-scale video classification with convolutional neural networks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1725-1732.

[5] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 1(4): 568-576.

[6] Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks[C]// Proceedings of IEEE International Conference on Computer Vision, 2015: 4489-4497.

[7] Donahue J, Hendricks L A, Rohrbach M, *et al*. Long-term recurrent convolutional networks for visual recognition and description[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 677-691.

[8] Wang L M, Xiong Y J, Wang Z, *et al*. Temporal segment networks: Towards good practices for deep action recognition[J]. ACM Transactions on Information Systems, 2016, 22(1): 20-36.

[9] Varol G, Laptev I, Schmid C. Long-term temporal convolutions for action recognition[J]. IEEE Trans Pattern Analysis and Machine Intelligence, 2018, 40(6): 1510-1517.

[10] Bilen H, Fernando B, Gavves E, *et al*. Dynamic image networks for action recognition[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016: 3034-3042.

[11] Lin S Z, Zheng Y, Lu X F, *et al*. Adaptive tracking algorithm for aerial small targets based on multi-domain convolutional neural networks and autoregression model[J]. Acta Optica Sinica, 2017, 37(12): 1215006.
蔺素珍, 郑瑶, 禄晓飞, 等. 基于多域卷积神经网络

与自回归模型的空中小目标自适应跟踪方法[J]. 光学学报, 2017, 37(12): 1215006.

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]// Proceedings of International Conference on Learning Representations, 2015: 1-14.

[13] Qu L, Wang K R, Chen L L, *et al*. Fast road detection based on RGBD images and convolutional neural network[J]. Acta Optica Sinica, 2017, 37(10): 1010003.
曲磊, 王康如, 陈利利, 等. 基于RGBD图像和卷积神经网络的快速道路检测[J]. 光学学报, 2017, 37(10): 1010003.

[14] Feichtenhofer C, Pinz A, Wildes R P. Spatiotemporal residual networks for video action recognition[C]// Proceedings of Neural Information Processing Systems, 2016: 3468-3476.

[15] Wang H, Schmid C. Action recognition with improved trajectories[C]// Proceedings of IEEE International Conference on Computer Vision, 2013: 3551-3558.

[16] Peng X J, Wang L M, Wang X X, *et al*. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice[J]. Computer Vision and Image Understanding, 2016, 150: 109-125.

[17] Wang L M, Qiao Y, Tang X O. MoFAP: A multi-level representation for action recognition[J]. International Journal of Computer Vision, 2016, 119(3): 254-271.

[18] Zha S X, Luisier F, Andrews W, *et al*. Exploiting image-trained CNN architectures for unconstrained video classification[J]. arXiv preprint arXiv: 1503.04144, 2015.

[19] Wang L M, Qiao Y, Tang X O. Action recognition with trajectory-pooled deep-convolutional descriptors[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015: 4305-4314.

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