视频行为识别综述

罗会兰，王婵娟，卢飞
（江西理工大学信息工程学院，江西 赣州 341000）

摘 要：目前行为识别发展迅速，许多基于深度网络自动学习特征的行为识别算法被提出。深度学习方法需要大量数据来训练，对电脑存储、运算能力要求较高。在回顾了当下流行的基于深度网络的行为识别方法的基础上，着重综述了基于手动提取特征的传统行为识别方法。传统行为识别方法通常遵循对视频提取特征并进行建模和预测分类的流程，并将识别流程细分为以下几个步骤进行综述：特征采样、特征描述符选取、特征预/后处理、描述符聚类、向量编码。同时，还对评价算法性能的基准数据集进行了归纳总结。

关键词：行为识别；手动提取；深度网络；数据集
中图分类号：TP391
文献标识码：A
doi: 10.11959/j.issn.1000-436x.2018107

Survey of video behavior recognition

LUO Huilan, WANG Chanjuan, LU Fei
School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China

Abstract: Behavior recognition is developing rapidly, and a number of behavior recognition algorithms based on deep network automatic learning features have been proposed. The deep learning method requires a large number of data to train, and requires higher computer storage and computing power. After a brief review of the current popular behavior recognition method based on deep network, it focused on the traditional behavior recognition methods. Traditional behavior recognition methods usually followed the processes of video feature extraction, modeling of features and classification. Following the basic process, the recognition process was overviewed according to the following steps, feature sampling, feature descriptors, feature processing, descriptor aggregation and vector coding. At the same time, the benchmark data set commonly used for evaluating the algorithm performance was also summarized.

Key words: behavior recognition, handcrafted, deep network, data set

1 引言

人体行为识别是指利用模式识别、机器学习等方法，从一段未知的视频中自动分析识别人体执行的行为。最简单的行为识别也称为行为分类，它可以将未知视频中的人体行为分类到预先定义的几种行为类别中。较为复杂的行为识别是指识别视频中多个人体正在交互进行的群体活动。行为识别的最终目标是自动分析视频中有什么人，在什么时刻、什么地方做了什么事情。人体行为识别在安防、交通管理、智能看护、娱乐休闲等现实生活中应用广泛。目前，行为识别的研究方法主要有 2 种：一种是基于手动提取特征的方法，另一种是基于深度网络学习特征的方法。2 种方法各有长短，基于手动提取特征的方法能够根据需要提取相应的特征，实现简单，但行为的表示能力也受所提取特征的限制；基于深度网络学习特征的方法能够自动学习特征，但需要大量数据支撑，不适于小型数据集处理，

收稿日期：2018-01-29；**修回日期**：2018-05-16
基金项目：国家自然科学基金资助项目（No.61105042, No.61462035）；江西省自然科学基金资助项目（No.20171BAB202014）
Foundation Items: The National Natural Science Foundation of China (No.61105042, No.61462035), The Natural Science Foundation of Jiangxi Province (No.20171BAB202014)

且整个过程是端到端的，像个黑盒子，不适于计算视觉领域的研究初学者熟悉图像、视频处理的基本技术和基本步骤。

Moeslund 等[1]按照行为的复杂程度将人体行为分为 3 个层级：基本动作、行为和活动。基本动作指的是能在肢体层次上描述的基本运动；行为指的是由基本动作构成，描述一个可能是周期性的全身运动；活动包含许多后续动作，并对正在执行的动作进行解释。例如，左腿向前是一个基本动作，跑步是一个行为，跨栏就是一个包括开始、跳跃和跑步动作的一个活动。与此类似，文献[2]认为行为识别可以分为 2 类：一类是低层动作的识别，另一类是高层行为的识别，其还认为前者是后者的基础，并依此将行为识别方法分为 2 类进行综述。

Ji 等[3]按行为识别的步骤将其分成 3 个子问题：人体检测、与视觉无关的姿势表示和估计、行为理解，并对其进行了综述。而 Dhamsania 等[4]按照视频场景中的目标人物数对识别方法进行了分类，将其区分为单人行为识别、双人或人与物互动的行为识别以及多人行为识别。Candamo 等[5]则讨论了交通监管视频场景中的行为识别问题：单人游荡识别、多人打架识别以及人与物体互动识别（如偷车、毁坏公共设施等）。Poppe 等[6]将视频行为识别的问题转化为图像序列的识别分类问题，并讨论了图像的各种表示及分类方法。

有些综述着眼于讨论某一特定动作类识别问题。Weinland 等[7]着眼于解决全身运动（如踢打、拳击等）识别问题的方法，并对这些方法按照如何表示动作的时空结构、如何对视频进行分割以及如何学习获得行为表示进行分类。Chaudhary 等[8]着眼于解决手势识别问题的方法，比较分析了当前一些流行方法的实验结果。

为了让初学者更好地理解传统视频行为识别方法的基本流程及其与最新深度网络模型方法的区别，本文分别综述了传统手动提取特征方法和深度网络学习方法，并重点论述了基于手动提取特征表示的行为识别方法，按照流程就每个相对独立的步骤进行了总结归纳，然后在此基础上综述了当前流行的用于行为识别的深度学习模型。主要贡献如下。

1) 对基于手动提取特征表示的行为识别方法进行了较为系统、全面的研究和分类，并对每类方法中的典型算法进行了阐述和分析。

2) 对 2012 年以来以卷积神经网络为代表的深度网络学习技术在行为识别中的应用进行了研究和阐述。

3) 对行为识别算法常用的基准数据集、算法性能评价指标进行了研究和介绍。

4) 讨论了行为识别中目前存在的、亟待解决的主要问题以及未来发展的趋势。

2 基于手动特征的行为识别方法

基于手动提取特征的行为识别方法一般包含如图 1 所示的处理流程，即首先对视频进行采样，然后对样本提取特征，接着对特征进行编码，再对编码得到的向量进行规范化，最后训练分类。

2.1 特征采样方法

一般而言，提取特征之前需要先对视频进行兴趣点采样，然后对采样兴趣点进行特征信息的提取。采样方式有基于兴趣区域的采样、基于轨迹的采样和基于身体部分的采样等。

2.1.1 基于兴趣区域的采样

基于兴趣区域的采样方法是指利用探测器检测视频的兴趣区域，从而提取特征描述的方法。这类方法不需要对行为视频进行前景背景分割，也不需要对发生行为的人体进行精确的定位跟踪。Laptev 等[9]提出对 Harris 角点检测方法[10]进行时空扩展，在行为视频中进行 Harris3D 兴趣点检测。Harris3D 检测空间维与时间维上都具有显著变化的点区域，并自适应地选择兴趣点的时间尺度与空间尺度。图 2 示例了对 UCF101[11]数据库中画眼妆这个动作的兴趣点采样截图。Oikonomopoulos 等[12]提出了一种基于时空显著点的行为表征方法：首先

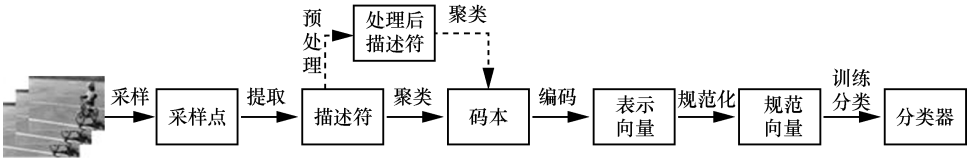

图 1 基于手动提取特征的行为识别流程

计算行为视频中每个像素点对应的时空邻域的信号直方图的熵，然后将取得 Shannon 熵的局部极大值的位置点视为时空显著点。以上 2 种方法检测到的采样点在空间尺度与时间尺度上都具有显著变化，但是视频中满足条件的采样点较少，这就导致采样得到的时空兴趣点比较稀疏，对后续的行为识别有一定的影响。针对这种问题，Dollar 等[13]提出了一种基于空间维上的高斯平滑滤波器与时间维上的 Gabor 滤波器的 Cuboid 检测方法，该方法检测出的时空兴趣点较为密集。Rapantzikos 等[14]提出使用离散小波变换，通过低通、高通滤波器的响应值来检测时空兴趣点。后来 Rapantzikos 等[15]又提出引入运动信息与颜色信息进行时空显著点检测。这些时空兴趣点检测方法均检测到了密集的时空兴趣点。Willems 等[16]提出将二维图像中的 Hessian 显著点检测方法扩展到三维视频中，这种方法被命名为 Hessian 时空兴趣点检测方法，它使用 3D Hessian 矩阵的行列式来评估视频中各位置点的显著性。Hessian 时空兴趣点检测方法以一种非迭代的方式，自动选择兴趣点的时空位置与尺度。这种方法能够检测到更为密集，且尺度不变的时空兴趣点。

(a) 原视频　　(b) Harris3D 兴趣点采样

图 2　原视频与兴趣点采样对比

2.1.2　基于轨迹的采样

伴随着人体运动的发生，会产生一条运动轨迹。Wang 等[17]提出沿着运动轨迹将轨迹邻域划分成细小的子空间，然后对每个子空间提取特征描述信息。基于轨迹的采样方法把时间信息也考虑进来了，通常来说，这种采样方法会比基于兴趣区域的采样方法对视频的表征能力更强。但是因为其沿着轨迹密集采样，所以采样得到的兴趣点数目较大，对于计算机的存储空间和运算速度的要求会更高。为了解决这个问题，文献[18-19]提出在稠密轨迹的基础上设置一些新的限制条件，从而减少稠密轨迹数。为了消除相机抖动对识别性能的影响，Wang 等[20]又提出了改进版的稠密轨迹提取方法，改进版中引入了对背景光流的消除方法，使特征更集中于对人体运动的描述。许多行为识别的研究工作[21-23]都是在改进稠密轨迹基础上进行的，在深度网络方法出现之前，该方法曾一度占据行为识别领域的领先位置。

2.1.3　基于身体部分的采样

基于身体部分的采样方法是通过姿态估计方法或深度图姿态估计方法，获取人体各部件的位置、关节点的位置以及关节点的运动信息来表征行为。这类方法一般需要先用前景背景分割、运动检测或行人检测跟踪算法对视频中的人体进行定位，然后对人体身体部位进行描述。通过这种采样方法所提取到的特征信息比较完整，对视频中的人体行为来说是一种良好的表征方式。Ali 等[24]利用人体头部与躯干的 5 个归一化节点的轨迹信息构建人体的行为。图 3 示例了演员表演 9 个不同动作时人体 5 个节点及其轨迹。Yilma 等[25]使用 13 个人体节点的轨迹信息进行行为识别。Jhuang 等[26]使用人工标记的 14 个关节点表达的姿态特征进行行为识别，并通过实验对比，发现了这种基于关节点的姿态特征表达比局部特征能获得更好的识别效果。Singh 等[27]使用 15 个链接点来表征人体行为的关键姿态，并利用跟踪信息进行行为识别。文献[28-29]则利用神经网络对自由度为 20 的人体行为骨架信息进行行为识别。

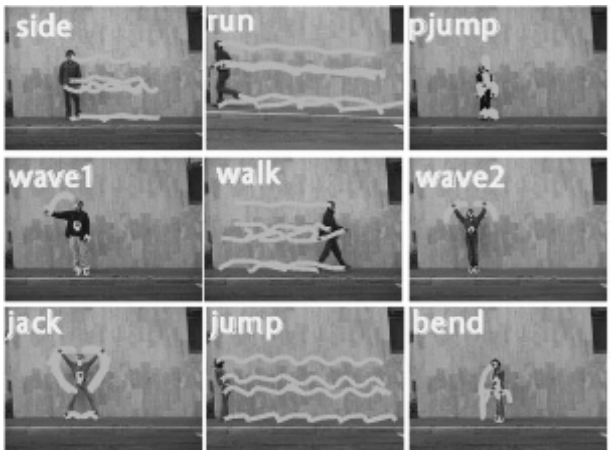

图 3　表演 9 个不同动作时人体 5 个节点的运动轨迹

2.2　描述符

特征提取的目的是收集通用的对背景变换顽健的视频描述信息。理想的特征应该是与尺度变化、旋转、仿射变化、光照变化、视角变化无关的。从全局来看，可以用外观、姿势或语境信息来描述视频中的人体行为。从局部来说，可以用方向梯度、

光流方向等来描述视频中的人体行为，本文将视频描述信息分为全局描述符和局部描述符来做进一步阐述。

2.2.1 全局描述符

全局描述符是对通过背景减图或跟踪的方法得到整个感兴趣的人体进行描述，通常采用的是人体的姿态、关节形状、剪影轮廓等信息。这些特征对噪声、部分遮挡、视角的变化比较敏感。

伴随着人体运动的发生，人体的姿态也会发生变化，因此，人体姿态也可以作为表征运动的一条线索。Wang 等[30]提出一种基于姿态的行为表示模型，用于描述人体姿势的时空结构。这类方法的处理流程如下：首先为每一帧估计 k 个最好的姿势，然后利用分段线索和时间约束推断最佳姿势。该方法在 UCF Sports 数据集和 MSR Action3D 数据集上分别获得了 90%和 90.22%的识别准确度，要优于同期其他方法。

众所周知，人体的运动是由关节带动发生的，因此，关节点的位置变化也能从侧面描述视频的运动信息。Jiang 等[31]提出了一种关节形状运动描述子，将光流场的运动模型和外观模型结合捕捉运动的不同性质。这种方法是将长视频看作基本动作的序列，然后利用关节形状运动描述子对基本动作进行匹配，从而实现视频的分类。文献[32]提出了一种基于关节点的元动作描述符，这种方法首先引入单关节点部位的动态聚类，采用关节点判别力来动态确定聚类中心个数。然后将判别力强的部位聚类个数增大，反之亦然。之后再引入判别力部位整体聚类，选出高识别率的判别力部位，将每个判别力部位内所有的关节点视为一个整体，串联特征后聚类，得到新的元动作，对于给定的样本，某个部位的元动作特征定义为该部位基础特征与各聚类中心归一化欧氏距离的串联。最后分别采用单关节点部位动态聚类和多判别力部位聚类的元动作特征来表示行为。

剪影表征的是人体的轮廓形象，做不同动作时人体的轮廓是不同的，例如，伸平双手和坐下，因此，行为视频中人体的剪影也可以作为人体运动的描述，Gorelick 等[33]使用背景差分法来提取人体的剪影信息，并据此将行为表征为时空形状。然后，基于泊松方程解的性质，利用提取的时空形状的方向、突出点、结构等特征的联合向量来表征行为。

2.2.2 局部描述符

局部描述符是指对提取出的局部兴趣点进行描述的方法，最常用的有梯度方向直方图（HOG, histogram of oriented gradient）、光流梯度方向直方图（HOF, histograms of oriented optical flow）、运动边界直方图（MBH, motion of boundary history）这 3 种方法。

HOG[34]描述的是静态外观信息，首先需要将图像分割成细小的子空间，然后统计每个子空间中各像素点的梯度方向，最后合并每个子空间的统计直方图并将其作为图像的 HOG 特征描述符。为了获得更好的光照、阴影等不变性，还可以先把这些子空间的局部直方图在图像中更大的区间内进行对比度归一化。

HOF[35]表达的是局部运动信息，首先是将光流图像分割成许多细小的子空间，然后加权统计每个子空间的光流方向，得到光流梯度直方图。由于视频中发生行为的人体的尺寸会随着时间发生变化，相应的光流特征描述子的维度也会变化。所以，光流的计算对背景噪声、尺度变化以及运动方向都比较敏感。为了使其对运动方向及尺度变化顽健，可以横轴为基准计算夹角并对得到的光流梯度直方图进行归一化。

MBH[36]表达的是相关运动信息。MBH 的计算方法是将 x 和 y 方向上的光流图像视作 2 张灰度图像，然后提取这些灰度图像的梯度直方图，即 MBH 特征是分别在图像的 x 和 y 方向的光流图像上计算 HOG 特征，实现对运动物体的边界信息的提取。

2.3 特征预/后处理技术

从视频中提取的底层特征以及编码后的特征向量需要经过一些处理技术防止数据过拟合的情况。本文将应用于从视频提取的底层特征上的处理方法称为预处理技术，将应用于编码后的特征向量上的处理方法称为后处理技术。有一些研究者会忽略对特征数据进行预处理而直接编码，但最近有研究[37]表明，对特征进行预处理能提升识别准确度。

常用的预处理技术分为 2 类，一类是降维处理，另一类是白化操作。主成份分析（PCA, principal component analysis）是一个常用的线性降维方法。PCA 把原先的 n 维特征用数目更少的 m 维特征取代，通过最大化样本方差，尽量使新的 m 个维度互不相关。白化的目的是去掉数据之间的相关度，是很多算法进行预处理的步骤。例

如，当训练图片数据时，因为图片中相邻像素值有一定的关联，所以很多信息是冗余的，这时就可以利用白化进行去相关操作。常见的白化操作有 PCA Whitening 和 ZCA Whitening。PCA Whitening 的操作流程是先通过 PCA 消除特征之间的相关性，然后利用缩放因子使特征具有相同的方差。ZCA Whitening 本质上是换一种方法实现特征的去相关及归一化，将经过 PCA Whitening 后的数据重新变换回原来的空间。对于卷积神经网络算法来说，因为它对自然图像的局部特征依赖较大，所以使用和原始数据同一空间表达的 ZCA Whitening 会比 PCA Whitening 的效果更好。但是对于大多数其他的机器学习算法来说，两者的效果相差不大。

编码后的特征向量往往需要经过后处理进行规范，常用的后处理技术有池化和归一化。池化分为最大池化、求和池化和平均池化。最大池化就是取这些描述符的编码系数中最大的值作为视频的全局表示。求和池化就是将所有描述符的编码系数求和并将得到的和值作为视频的全局表示。平均池化就是将所有描述符的编码系数求和之后再取平均值并将平均值作为视频的全局表示。常用的归一化方式有 4 种：L1 归一化、L2 归一化、Power Normalization 和 Intra Normalization。假设 $\boldsymbol{v}=\{x_1,\cdots,x_n\}$ 表示一个视频的编码向量，则各规范化策略计算式如下。

L1 归一化：

$$\boldsymbol{v}_{\mathrm{L1}}=\left[\frac{x_1}{|x_1|+\cdots+|x_n|},\cdots,\frac{x_n}{|x_1|+\cdots+|x_n|}\right] \tag{1}$$

L2 归一化：

$$\boldsymbol{v}_{\mathrm{L2}}=\left[\frac{x_1}{\sqrt{x_1^{\,2}+\cdots+x_n^{\,2}}},\cdots,\frac{x_n}{\sqrt{x_1^{\,2}+\cdots+x_n^{\,2}}}\right] \tag{2}$$

Power Normalization：

$$\boldsymbol{v}_{\mathrm{Power}}=sign(\boldsymbol{v}).*(abs(\boldsymbol{v}).\wedge\alpha) \tag{3}$$

Intra normalization：

$$\boldsymbol{v}_{\mathrm{Intra}}=\left[\frac{\boldsymbol{v}^1}{\|\boldsymbol{v}^1\|_2},\cdots,\frac{\boldsymbol{v}^k}{\|\boldsymbol{v}^k\|_2}\right] \tag{4}$$

式(3)中的 α 为规范参数，且满足条件 $0\leqslant\alpha\leqslant1$。式(4)中的 $\boldsymbol{v}^k$ 表示和第 k 个聚类中心或第 k 个高斯分量相关的单词向量。

2.4　聚类方法

在对视频进行特征提取得到视频的特征集之后，需要对视频的特征集进行聚类得到后面编码需要的码本。视频动作识别领域常用的聚类方式有 K 均值聚类和混合高斯模型（Gaussian mixed model）聚类。

K 均值聚类是依据特征点之间的相似性聚类。它初始时随机选择 K 个特征点作为 K 个簇的均值点或代表点，然后将每个特征点分配给离它最近的均值点代表的簇，分配完毕后重新计算各个簇的均值，这个过程不断重复，直到准则函数收敛。它的结果会受初始选择的 K 个均值点的影响。

混合高斯模型指的是多个高斯分布函数的线性组合，表示为

$$p(x)=\sum_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k) \tag{5}$$

其中，$N(x|\mu_k,\Sigma_k)$ 表示混合高斯分布中的第 k 个高斯分量，π_k 表示混合系数，且 π_k 满足条件：$\sum_{k=1}^{K}\pi_k=1$，$0\leqslant\pi_k\leqslant1$。

2.5　编码方法

采样视频兴趣点并描述得到训练特征集，然后通过聚类得到特征码本，还需要对每个视频的特征进行编码获取表示向量。常用的编码方法有矢量量化（VQ, vector quantization）、稀疏编码（SC, sparse coding）、费舍尔编码（FV, Fisher vector）和局部聚合描述符矢量（VLAD, vector of locally aggregated descriptor）。

VQ 是一种投票式的硬性编码方法，投票规则如下：给定一个 k 维的码本 $D=(d_1,\cdots,d_i,\cdots,d_k)$，对于视频的描述符集 $X=(x_1,\cdots,x_j,\cdots,x_n)$，其中，$x_j$ 表示视频的第 j 个描述符，则 x_j 对视觉词典中第 i 个视觉单词 d_i 的投票只有 2 个取值，1 或 0，如果 x_j 和 d_i 的距离最近，则投票值为 1，否则为 0。通过这种投票方法，第 j 个描述符就获得了一个 k 维的编码系数 s_j，且 s_j=[…000010000…]。类似地，可获得视频描述符集 X 中每个描述子的编码系数。这种编码方法是一种硬量化，容易导致信息损失。

SC 是一种重建型编码方法，它的目的是使编码系数 s 能依据聚类得到的字典最大可能重建描述子 x。给定一个大小为 K 的码本 $D=\{d_k,k=1,\cdots,K\}$，对于视频的描述符集 X 的编码系数 s，计算式为

$$s=\arg\min_{s}\left\|\chi-Ds\right\|_2^2+\lambda\left\|s\right\|_1 \tag{6}$$

其中，$\left\|s\right\|_1$ 表示对编码系数做 L1 正则化处理，保证编码系数具有稀疏性。

FV 是由 Perronnin 等[38]提出的用于大尺度图像分类的。因其在图像分类中的杰出表现，逐渐被引入视频的行为识别中。用描述符集 X 来描述一段视频，给定一个大小为 K 的混合高斯分布模型，视频描述符集的编码系数 s 可以表示为

$$s=\left[\xi_{\mu_1}^X,\cdots,\xi_{\mu_K}^X,\xi_{\sigma_1}^X,\cdots,\xi_{\sigma_K}^X\right] \tag{7}$$

其中，

$$\xi_{\mu_i}^X=\frac{1}{\sqrt{\pi_i}}q_i\left(\frac{X-\mu_i}{\sigma_i}\right) \tag{8}$$

$$\xi_{\sigma_i}^X=\frac{1}{\sqrt{2\pi_i}}q_i\left[\left(\frac{\left(X-\mu_i\right)^2}{\sigma_i^{\,2}}\right)-1\right] \tag{9}$$

$$q_i=\frac{\pi_i N\left(X;\mu_i,\Sigma_i\right)}{\sum_{j=1}^{K}\pi_j N\left(X;\mu_j,\Sigma_j\right)} \tag{10}$$

VLAD 是费舍尔编码的一种特殊形式，由 Jegou 等[39]在图像搜索中首次提出，这种编码方法的计算过程如下：假设在训练特征集上聚类得到大小为 K 的视觉词典 D，表示为 $D=\left\{d_k,k=1,\cdots,K\right\}$，其中，$d_k$ 表示码本中第 k 个视觉单词。假设一个视频的特征描述集为 X，则视频的编码系数 s 为

$$s=\left[0,\cdots,\left(X-d_k\right),\cdots,0\right] \tag{11}$$

其中，

$$k=\arg\min_{i}\left\|\chi-d_i\right\|_2^2 \tag{12}$$

3 基于深度学习的行为识别方法

行为识别方法的性能主要取决于视频特征的表达，与手动提取特征表示方法不同，基于深度网络学习特征表示的方法是从原始数据中自动学习特征。这种方法是端到端的，输入视频，输出分类结果。

深度学习中用于行为识别的深度网络主要有卷积神经网络（CNN, convolutional neural network）和递归神经网络（RNN, recurrent neural network）。卷积神经网络通常遵循 3 层体系结构，分别是卷积层、池化层和全连接层。比较经典的是 Simonyan 等[40]提出的用于行为识别的双流 CNN，其将视频看作一段图像序列，空间流计算图像帧的 CNN 特征，时间流计算若干图像帧间的光流 CNN 特征，最后再将两者进行融合。图 4 为双流 CNN 工作流程[40]。

这种方法虽然将立体的视频识别问题转化为平面的图像识别问题，但却丢失了动作的时间关联信息。为了弥补双流架构在时间信息上的丢失，Wang 等[41]提出了三流 CNN 架构。该架构在双流架构的基础上将时间流进一步细分，分为局部时间流和全局时间流。动作图像特征和光流特征分别作为空间流和局部时间流的输入，并将学习运动叠差图像（MSDI, motion stacked difference image）的 CNN 特征作为全局时间流的输入。在 UCF101 及 HMDB51[42]数据库上的实验表明，基于三流 CNN 架构的识别准确度比双流 CNN[40]方法分别提高了 1.7%和 1.9%。

还有一些研究者对 CNN 特征提取对象做了改进，例如，Gkioxari 等[43]提出不对整个图像帧学习

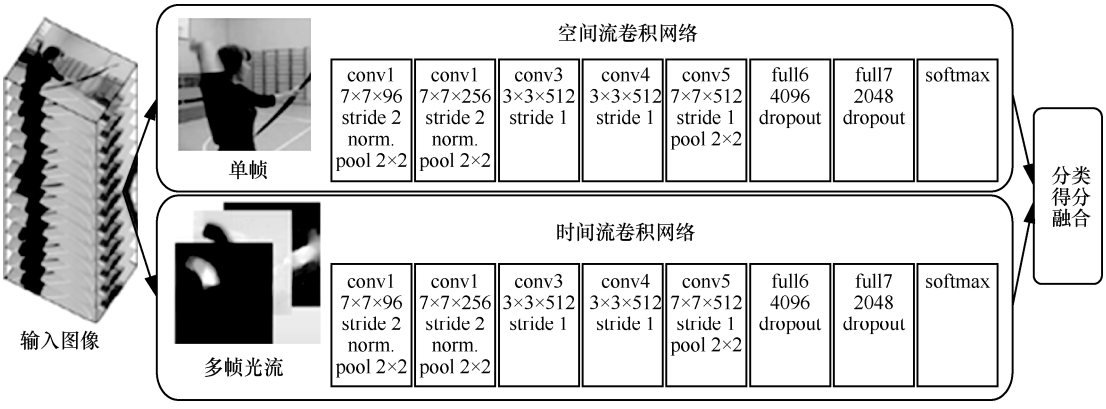

图 4 双流 CNN 工作流程

特征，而是在图像帧上先选择一个包含人体的包围盒作为主区域，然后根据主区域定义若干个次区域，利用最大值函数计算次区域包含的信息量并将其得分给主区域，再利用 RCNN（region-based convolutional network method）对主区域和次区域进行训练学习得到视频的特征表示。该方法在数据库 PASAL VOC Action dataset 上获得了 90.2%的平均准确度，超出同期其他方法[44-47]。Cheron 等[48]提出 P-CNN（pose based CNN）方法，该方法是先对输入的视频进行姿势估计，然后对身体不同部分提取 CNN 特征，再将各个部分的特征融合起来，该方法在数据库 JHMDB[26]和 MPII Cooking dataset[49]上均领先同期方法[26,49-51]。

递归神经网络[52]也常被用于深度学习模型中，它是将之前若干时刻的数据作为当前时刻的数据输入，从而保留了时间维度上的信息。长短时记忆[53]（LSTM, long short-term memory）类型的 RNN 是普通 RNN 的扩展，主要用于解决 RNN 中的梯度消亡现象。Niebles 等[54]提出了一种非监督式的 LSTM 模型来计算视频的表示信息。在文献[55-56]中，还提出将 CNN 和 RNN 结合起来识别视频中的人体行为。文献[57]在此基础上又提出了一种递归混合网络模型，该模型首先从彩色图像和光溜中提取空间特征和短时时间特征，然后对相邻的 P 帧特征进行池化并将池化结果输入 LSTM 中（这可以减少帧间的噪声影响），最后将 LSTM 模型学到的特征与视频的其他 2 种特征（STP 和 IDT）经过线性 SVM 得到的分数融合获取视频分类的最终结果，在 UCF101 上获得了 89.4%的识别准确度，比传统的 LSTM 的识别准确度高了 2.4%。

4　行为识别算法分析评价

本节主要介绍历年来较有代表意义的检验行为识别算法性能的公用数据集，并对前述比较典型的行为识别算法进行了分析、总结和比较。

4.1　基准数据集介绍

判断一个行为识别算法的优劣需要在同一个环境中和其他的同类算法进行比较，这就促使了一些公开数据库的诞生。表 1[58-67]列出了行为识别研究发展历程中常用的一些数据库的信息，包括每个数据库的发布年份、动作类、简介以及近 3 年被引用次数。由表 1 中的 2015-2017 年引用次数可以看出，随着深度学习的流行，在选择测试评估的数据集时逐渐倾向选取 UCF101、HMDB51 这种大型的、与现实环境一致的数据集。且深度学习算法需要用到大量的数据进行训练，而小型的数据库不能满足此类需求。随着行为识别在智能看护、人机交互等现实场景应用的普及，人们对于行为识别算法的准确度、适应性、实时性等要求越来越高，固定条件或场景录制的视频已很难满足人们的实际需求。其中，HMDB51 的识别难度较高，因为它的视频片段均来源于真实世界，背景杂乱，视角变化、类内差异较大。

4.2　行为识别算法分析与比较

本节在 3 个具有代表性的数据集 KTH、HMDB51 和 UCF101 上分析比较了一些有代表性的基于手动提取特征的方法以及基于深度学习的方法，分别如表 2~表 4[68-89]所示。由表 2~表 4 中 2 类方法近几年的识别准确度来看，基于手动提取特征的方法逐渐走向一个研究瓶颈，很难再开发出比改进稠密轨迹效果更好的描述子，大多数的研究都是围绕改进稠密轨迹展开的，通过不同的编码方法获取顽健性更强的独立表示，但效果并不显著。相反，基于深度网络学习特征的方法虽然最初的识别准确度并不高，但经过几年的发展，准确度有了很大提升，逐渐超越了基于改进稠密轨迹的方法。Peng 等[68]通过对改进稠密轨迹特征进行叠加费舍尔编码，在 HMDB51 数据集上获得了 66.79%的识别准确度；Duta 等[69]通过在改进稠密轨迹特征中融入位置信息进行编码，在 UCF101 上获得了 91.5%的识别准确度；四流深度卷积网络模型[70]在 UCF101 和 HMDB51 数据集上取得了目前最高的识别准确度，分别为 96%和 74.2%。

4.3　行为识别中亟待解决的问题及未来发展趋势

目前，行为识别的研究虽然取得了一定的进展，但还是面临很多的挑战，还有许多亟待解决的问题。首先，目前大部分的研究方法需要足够多的标签样本进行训练，才能达到比较好的识别预测效果。但现实中许多情况下提供不了足够多的样本，那么如何依靠现有的少量监督样本达到较高的识别准确度是目前亟待解决的一个问题。其次，相比动作幅度大的人体行为（如踢足球、跳舞等人体行为），比较细微的人体行为识别的难度很大，现存方法的效果非常不理想，例如，根据眼皮的下沉情况判断正在驾驶车辆的司机是否有打瞌睡的迹象，或根据犯罪审问中罪犯的微表情、微动作判断罪犯

表 1　　**历年来常用数据库简介**

数据库名	发表年份	动作数	数据库简介	2015-2017 年被引用次数
MU MoBo[58]	2001	4	该数据库包含 4 类不同的行为，分别是慢走、快走、斜走以及带球走，以上动作由 25 个演员在 3D CMU 房间的跑步机上演示	62
KTH[59]	2004	6	该数据库包含 6 类动作，共计 2 391 个视频样本，由 25 个演员在 4 个不同场景下完成。数据库中的视频样本中包含了尺度变化、衣着变化和光照变化，但其背景比较单一，相机角度也是固定的	492
Weizmann[33]	2005	10	该数据库包含 10 类动作，每类动作有 9 个不同的样本。相机视角是固定的，背景相对简单，每一帧中只有一个人在做动作。数据集包含类别标签、剪影和背景序列	230
IXMAS[60]	2006	14	该数据库为多视角数据库，包含 14 类动作，由 11 个演员完成，每个动作重复 3 次。相机分布在 5 个位置，分别是室内 4 个角落和头顶位置	4
UCF-Sports[61]	2008	10	该数据库的视频来源于电视频道 ESPN 和 BBC，包含 10 个运动动作类	220
Hollywood1[36]	2008	8	该数据库包含 8 类动作，这些动作从 32 部电影当中收集	674
Hollywood2[62]	2009	12	该数据库包含 12 类动作，共 3 669 个视频，所有视频都是从 69 部 Hollywood 电影中抽取出来的。视频样本中行为人的表情、姿态、穿着以及相机运动、光照变化、遮挡、背景等变化很大，接近于真实场景下的情况，因而对于行为的分析识别极具挑战性	272
HumanEva[63]	2009	6	该数据库中的视频采用 3 个色彩相机、4 个灰度相机拍摄而成，由 4 个演员演示了 6 个动作类	81
UCF-YouTube[64]	2009	11	该数据库包含 11 个动作类，其中的视频存在抖动、视觉变化、光照变化和背景遮挡等问题	167
MSR Action3D[65]	2010	20	该数据库包含 20 类动作，总计 557 个深度图视频序列	309
HMDB51[42]	2011	51	该数据库包含 51 类动作，总计 6 849 个视频，视频多数来源于电影以及 YouTube 等网络视频库，每个动作类至少包含有 101 段样本	390
UCF101[11]	2012	101	该数据库是目前公开数据库中最大的数据库之一，它的视频来源 YouTube，包含 101 个动作类	459
UCF-50[66]	2013	50	该数据库视频来源 YouTube，它依据视频的标签被划分为 50 个动作类，共有 6 618 个视频序列	161
UCF Kinect[67]	2013	16	该数据库中的骨架序列是使用单个 Kinect 和 OpenNI 框架采集获取的，一共有 16 个行为，都是为游戏场景所设计的	70

表 2　　**KTH 数据集上行为识别方法分析比较**

方法类型	文献	出版源	发表年份	准确度
基于手动提取特征表示方法分析比较	文献[71]	IEEE Conference on Computer Vision and Pattern Recognition	2011	94.5%
	文献[72]	IEEE Conference on Computer Vision and Pattern Recognition	2011	91.59%
	文献[17]	IEEE Conference on Computer Vision and Pattern Recognition	2011	95%
	文献[73]	IEEE Conference on Computer Vision and Pattern Recognition	2012	98.2%
	文献[23]	British Machine Vision Conference	2013	95.6%
	文献[74]	Multimedia Tools & Applications	2015	97.41%
深度网络学习特征表示方法	文献[75]	IEEE Transactions on Pattern Analysis & Machine Intelligence	2012	93.5%
	文献[76]	IEEE Transactions on Pattern Analysis & Machine Intelligence	2013	90.2%
	文献[77]	European Conference on Computer Vision	2014	96.6%
	文献[78]	IEEE Conference on Computer Vision and Pattern Recognition	2014	93.1%

是否撒谎从而辅助警察办案。

未来行为识别的研究发展将更加贴近实际应用，朝着更少样本、更快速度以及更精细动作识别的研究方向发展。

5　结束语

人体行为识别在现实生活中有非常大的应用需求，受到越来越多的计算机视觉研究者的关注。为了

表 3　　HMDB51 数据集上行为识别方法分析比较

方法类型	文献	出版源	发表年份	准确度
基于手动提取特征表示方法分析比较	文献[17]	IEEE Conference on Computer Vision and Pattern Recognition	2011	46.6%
	文献[79]	European Conference on Computer Vision	2012	40.7%
	文献[73]	IEEE Conference on Computer Vision and Pattern Recognition	2012	26.9%
	文献[20]	IEEE International Conference on Computer Vision	2013	57.2%
	文献[80]	IEEE Conference on Computer Vision and Pattern Recognition	2013	33.7%
	文献[68]	European Conference on Computer Vision	2014	66.79%
	文献[21]	IEEE Conference on Computer Vision and Pattern Recognition	2015	63.7%
深度网络学习特征表示方法	文献[40]	Neural Information Processing Systems	2014	59.4%
	文献[81]	IEEE International Conference on Computer Vision	2015	59.1%
	文献[82]	IEEE Conference on Computer Vision and Pattern Recognition	2015	65.9%
	文献[83]	IEEE Winter Conference on Applications of Computer Vision	2016	54.9%
	文献[84]	European Conference on Computer Vision	2016	70.4%
	文献[85]	Multimedia Tools & Applications	2016	74.7%
	文献[70]	IEEE Transactions on Pattern Analysis & Machine Intelligence	2016	74.9%

表 4　　UCF101 数据集上行为识别方法分析比较

方法类型	文献	出版源	发表年份	准确度
基于手动提取特征表示方法分析比较	文献[79]	ICCV 2013 workshop of THUMOS'13 Action Recognition Challenge	2013	87.46%
	文献[86]	IEEE International Conference on Computer Vision	2014	73.1%
	文献[87]	European Conference on Computer Vision	2014	87.7%
	文献[88]	IEEE Conference on Computer Vision and Pattern Recognition	2015	89.1%
	文献[37]	Computer Vision & Image Understanding	2016	87.9%
	文献[69]	International Conference on MultiMedia Modeling	2017	91.5%
深度网络学习特征表示方法	文献[40]	Neural Information Processing Systems	2014	88%
	文献[56]	IEEE Conference on Computer Vision and Pattern Recognition	2015	88.6%
	文献[82]	IEEE Conference on Computer Vision and Pattern Recognition	2015	91.5%
	文献[83]	IEEE Winter Conference on Applications of Computer Vision	2016	89.1%
	文献[85]	Multimedia Tools & Applications	2016	91.6%
	文献[89]	IEEE Conference on Computer Vision and Pattern Recognition	2017	94.6%
	文献[70]	IEEE Transactions on Pattern Analysis & Machine Intelligence	2016	96%

帮助初学者快速掌握行为识别的流程，把握研究热点，本文在前人的研究基础上，综述了基于手动提取特征的行为识别方法以及典型的多流卷积神经网络模型。介绍了行为识别研究常用的公开数据集，在此基础上分析比较了传统手工提取特征方法和深度学习方法的性能。基于改进稠密轨迹特征的行为识别方法是传统方法中效果较好的，因为改进稠密轨迹依据光流进行稠密采样，获取到的特征信息较为丰富，表征能力较强，缺点是计算量较大。近年来，基于复杂深度模型的行为识别研究取得了相较于传统方法更好的效果。未来的行为识别研究可能朝着更实用、更精细、需要更少训练数据的方向发展。

参考文献：

[1] MOESLUND T B, HILTON A, KRUGER V. A survey of advances in vision-based human motion capture and analysis[J]. Computer Vision & Image Understanding, 2006, 104(2):90-126.

[2] CHENG G C, WAN Y F, SAUDAGAR A N, et al. Advances in human action recognition: a survey[J]. Computer Science, 2015,2015(1):1-30.

[3] JI X, LIU H. Advances in view-invariant human motion analysis: a review[J]. IEEE Transactions on Systems Man & Cybernetics Part C, 2009, 40(1):13-24.

[4] DHAMSANIA C J, RATANPARA T V. A survey on human action recognition from videos[C]//Online International Conference on Green Engineering and Technologies. 2017:1-5.

[5] CANDAMO J, SHREVE M, GOLDGOF D B, et al. Understanding transit scenes: a survey on human behavior recognition algorithms[J].

IEEE Transactions on Intelligent Transportation Systems, 2010, 11(1):206-224.

[6] POPPE R. A survey on vision-based human action recognition[J]. Image & Vision Computing, 2010, 28(6):976-990.

[7] WEINLAND D, RONFARD R, BOYER E. A survey of vision-based methods for action representation, segmentation and recognition[J]. Computer Vision & Image Understanding, 2011, 115(2):224-241.

[8] CHAUDHARY A, RAHEJA J L, DAS K, et al. A survey on hand gesture recognition in context of soft computing[C]//International Conference on Computer Science and Information Technology. 2011:46-55.

[9] LAPTEV I. On space-time interest points[J]. International Journal of Computer Vision, 2005, 64(2-3):107-123.

[10] HARRIS C J. A combined corner and edge detector[J]. Proc Alvey Vision Conf, 1988, 1988(3):147-151.

[11] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild[J]. Computer Science, 2012.

[12] OIKONOMOPOULOS A, PATRAS I, PANTIC M. Spatiotemporal salient points for visual recognition of human actions[J]. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society, 2006, 36(3):710-719.

[13] DOLLAR P, RABAUD V, COTTRELL G, et al. Behavior recognition via sparse spatio-temporal features[C]//IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. 2006:65-72.

[14] RAPANTZIKOS K, AVRITHIS Y, KOLLIAS S. Spatiotemporal saliency for event detection and representation in the 3d wavelet domain: potential in human action recognition[C]//ACM International Conference on Image and Video Retrieval. 2007:294-301.

[15] RAPANTZIKOS K, AVRITHIS Y, KOLLIAS S. Dense saliency-based spatiotemporal feature points for action recognition[C]//Computer Vision and Pattern Recognition. 2009:1454-1461.

[16] WILLEMS G, TUYTELAARS T, GOOL L. An efficient dense and scale-invariant spatio-temporal interest point detector[C]//European Conference on Computer Vision. 2008:650-663.

[17] WANG H, KLASER A, SCHMID C, et al. Action recognition by dense trajectories[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2011:3169-3176.

[18] MURTHY O V R, GOECKE R. Ordered trajectories for human action recognition with large number of classes[J]. Image & Vision Computing, 2015, 42(C):22-34.

[19] CHO J, LEE M, CHANG H J, et al. Robust action recognition using local motion and group sparsity[J]. Pattern Recognition, 2014, 47(5):1813-1825.

[20] WANG H, SCHMID C. Action recognition with improved trajectories[C]//IEEE International Conference on Computer Vision. 2014:3551-3558.

[21] FERNANDO B, GAVVES E, ORAMAS M J, et al. Modeling video evolution for action recognition[C]//IEEE Conference Computer Vision and Pattern Recognition. 2015:5378-5387.

[22] JHUANG H, SERRE T, WOLF L, et al. A biologically inspired system for action recognition[C]//International Conference on Computer Vision. 2007: 1-8.

[23] PENG X, QIAO Y, PENG Q, et al. Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition[C]//British Machine Vision Conference. 2013.

[24] ALI S, BASHARAT A, SHAH M. Chaotic invariants for human action recognition[C]//International Conference on Computer Vision. 2007:1-8.

[25] YILMA A, SHAH M. Recognizing human actions in videos acquired by uncalibrated moving cameras[C]//Tenth IEEE International Conference on Computer Vision. 2005:150-157.

[26] JHUANG H, GALL J, ZUFFI S, et al. Towards understanding action recognition[C]//IEEE International Conference on Computer Vision. 2014:3192-3199.

[27] SINGH V K, NEVATIA R. Action recognition in cluttered dynamic scenes using pose-specific part models[C]//International Conference on Computer Vision. 2011:113-120.

[28] DU Y, WANG W, WANG L. Hierarchical recurrent neural network for skeleton based action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1110-1118.

[29] WU D, SHAO L. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2014: 724-731.

[30] WANG C, WANG Y, YUILLE A L. An approach to pose-based action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2013:915-922.

[31] JIANG Z, LIN Z, DAVIS L S. Recognizing human actions by learning and matching shape-motion prototype trees[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(3):533-547.

[32] HUANG M, SU S Z, CAI G R, et al. Meta-action descriptor for action recognition in RGBD video[J]. IET Computer Vision, 2017, 11(4):301-308.

[33] GORELICK L, BLANK M, SHECHTMAN E, et al. Actions as space-time shapes[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(12):2247-2253.

[34] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Computer Vision and Pattern Recognition. 2005:886-893.

[35] DALAL N, TRIGGS B, SCHMID C. Human detection using oriented histograms of flow and appearance[C]//European Conference on Computer Vision. 2006: 428-441.

[36] LAPTEV I, MARSZALEK M, SCHMID C, et al. Learning realistic human actions from movies[C]//Computer Vision and Pattern Recognition. 2008:1-8.

[37] PENG X, WANG L, WANG X, et al. Bag of visual words and fusion methods for action recognition: comprehensive study and good practice[J]. Computer Vision & Image Understanding, 2016, 150(C):109-125.

[38] PERRONNIN F, MENSINK T. Improving the fisher kernel for large-scale image classification[C]//European Conference on Computer Vision. 2010:143-156.

[39] JEGOU H, DOUZE M, SCHMID C, et al. Aggregating local descriptors into a compact image representation[C]//Computer Vision and Pattern Recognition. 2010:3304-3311.

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

[41] WANG L, GE L, LI R, et al. Three-stream CNNs for action recogni-

tion[J]. Pattern Recognition Letters, 2017, 92(C):33-40.

[42] KUEHNE H, JHUANG H, STIEFELHAGEN R, et al. HMDB51: a large video database for human motion recognition[C]//IEEE International Conference on Computer Vision. 2011:2556-2563.

[43] GKIOXARI G, GIRSHICK R, MALIK J. Contextual action recognition with R*CNN[J]. CoRR, 2016, 40(1):1080-1088.

[44] GKIOXARI G, GIRSHICK R, MALIK J. Actions and attributes from wholes and parts[C]//International Conference on Computer Vision. 2015: 2470-2478.

[45] HOAI M. Regularized max pooling for image categorization[C]//British Machine Vision Conference. 2014:94-100.

[46] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014.

[47] OQUAB M, BOTTOU L, LAPTEV I, et al. Learning and transferring mid-level image representations using convolutional neural networks[C]//Conference on Computer Vision and Pattern Recognition. 2014:1717-1724.

[48] CHERON G, LAPTEV I, SCHMID C. P-CNN: pose-based CNN features for action recognition[C]//International Conference on Computer Vision. 2015:3218-3226.

[49] ROHRBACH M, AMIN S, ANDRILUKA M, et al. A database for fine grained activity detection of cooking activities[C]//Conference on Computer Vision and Pattern Recognition. 2012:1194-1201.

[50] ZHOU Y, NI B, HONG R, et al. Interaction part mining: a mid-level approach for fine-grained action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2015:3323-3331.

[51] ZHOU Y, NI B, YAN S, et al. Pipelining localized semantic features for fine-grained action recognition[C]//European Conference on Computer Vision. 2014:481-496.

[52] GRAVES A, MOHAMED A, HINTON G. Speech recognition with deep recurrent neural networks[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. 2013:6645-6649.

[53] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[54] NIEBLES J C, WANG H, LI F F. Unsupervised learning of human action categories using spatial-temporal words[J]. International Journal of Computer Vision, 2008, 79(3):299-318.

[55] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description [C]//IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2625-2634.

[56] NG Y H, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: deep networks for video classification[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2015:4694-4702.

[57] YU S, CHENG Y, XIE L, et al. A novel recurrent hybrid network for feature fusion in action recognition[J]. Journal of Visual Communication & Image Representation, 2017, 49:192-203.

[58] GROSS R, SHI J. The CMU motion of body (MoBo) database[J]. Monumenta Nipponica, 2001, 45(4).

[59] SCHULDT C, LAPTEV I, CAPUTO B. Recognizing human actions: a local SVM approach[C]//International Conference on Pattern Recognition. 2004:32-36.

[60] WEINLAND D, RONFARD R, BOYER E. Free viewpoint action recognition using motion history volumes[J]. Computer Vision & Image Understanding, 2011, 104(2):249-257.

[61] RODRIGUEZ M D, AHMED J, SHAH M. Action MACH a spatio-temporal maximum average correlation height filter for action recognition[C]//Conference on Computer Vision and Pattern Recognition. 2008:1-8.

[62] MARSZALEK M, LAPTEV I, SCHMID C. Actions in context[C]//Conference on Computer Vision and Pattern Recognition. 2009:2929-2936.

[63] SIGAL L, BALAN A O, BLACK M J. HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion[J]. International Journal of Computer Vision, 2006, 87(1-2):4-27.

[64] LIU J,LUO J,SHAH M. Recognizing realistic actions from videos in the wild[C]//Computer Vision and Pattern Recognition. 2009:1996-2003.

[65] LI W, ZHANG Z, LIU Z. Action recognition based on a bag of 3D points[C]//Conference on Computer Vision and Pattern Recognition. 2010:9-14.

[66] REDDY K K, SHAH M. Recognizing 50 human action categories of web videos[J]. Machine Vision & Applications, 2013, 24(5):971-981.

[67] ELLIS C, MASOOD S Z, TAPPEN M F, et al. Exploring the trade-off between accuracy and observational latency in action recognition[J]. International Journal of Computer Vision, 2013, 101(3):420-436.

[68] PENG X, ZOU C, QIAO Y, et al. Action recognition with stacked fisher vectors[C]//European Conference on Computer Vision. 2014:581-595.

[69] DUTA I C, LONESCU B, AIZAWA K, et al. Spatio-temporal VLAD encoding for human action recognition in videos[C]//International Conference on Multimedia Modeling. 2017:365-378.

[70] BILEN H, FERNANDO B, GAVVES E, et al. Action recognition with dynamic image networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99):1.

[71] WU X, XU D, DUAN L, et al. Action recognition using context and appearance distribution features[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2011:489-496.

[72] LIU J, KUIPERS B, SAVARESE S. Recognizing human actions by attributes[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2011:3337-3344.

[73] CORSO J J. Action bank: a high-level representation of activity in video[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2012:1234-1241.

[74] CHEN M, GONG L, WANG T, et al. Action recognition using lie Algebrized Gaussians over dense local spatio-temporal features[J]. Multimedia Tools & Applications, 2015, 74(6):2127-2142.

[75] ZHANG Z, TAO D. Slow feature analysis for human action recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(3):436-450.

[76] 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.

[77] HASAN M, ROY-CHOWDHURY A K. Continuous learning of human activity models using deep nets[C]//European Conference on Computer Vision. 2014:705-720.

[78] SUN L, JIA K, CHAN T H, et al. DL-SFA: deeply-learned slow feature analysis for action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2014:2625-2632.

[79] JIANG Y G, DAI Q, XUE X, et al. Trajectory-based modeling of

human actions with motion reference points[C]//European Conference on Computer Vision. 2012:425-438.

[80] WANG L M, QIAO Y, TANG X. Motionlets: mid-level 3d parts for human motion recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2013: 2674-2681.

[81] SUN L, JIA K, YEUNG D Y, et al. Human action recognition using factorized spatio-temporal convolutional networks[C]//IEEE International Conference on Computer Vision. 2015: 4597-4605.

[82] WANG L, QIAO Y, TANG X. Action recognition with trajectory-pooled deep-convolutional descriptors[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4305-4314.

[83] PARK E, HAN X, BERG T L, et al. Combining multiple sources of knowledge in deep CNNs for action recognition[C]//IEEE Winter Conference on Applications of Computer Vision. 2016:1-8.

[84] SOUZA C R D, GAIDON A, VIG E, et al. Sympathy for the details: dense trajectories and hybrid classification architectures for action recognition[C]//European Conference on Computer Vision. 2016:697-716.

[85] YU S, CHENG Y, SU S, et al. Stratified pooling based deep convolutional neural networks for human action recognition[J]. Multimedia Tools & Applications, 2017, 76(11):13367-13382.

[86] MURTHY O V R, GOECKE R. Ordered trajectories for large scale human action recognition[C]//IEEE International Conference on Computer Vision. 2014:412-419.

[87] PENG X, WANG L, QIAO Y, et al. Boosting VLAD with supervised dictionary learning and high-order statistics[C]//European Conference on Computer Vision. 2014:660-674.

[88] LAN Z, LIN M, LI X, et al. Beyond gaussian pyramid: multi-skip feature stacking for action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2015:204-212.

[89] FEICHTENHOFER C, PINZ A, WILDES R P. Spatiotemporal multiplier networks for video action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2017:7445-7454.

[作者简介]

罗会兰（1974-），女，江西上高人，博士，江西理工大学教授、硕士生导师，主要研究方向为机器学习、模式识别。

王婵娟（1992-），女，江西鄱阳人，江西理工大学硕士生，主要研究方向为计算机视觉、行为识别。

卢飞（1994-），男，江西赣州人，江西理工大学硕士生，主要研究方向为图像处理、机器视觉。