基于深度学习的视频行为分类方法综述*

杨 戈[1,2],邹武星[1,2]
(1.北京师范大学珠海分校 智能多媒体技术重点实验室,广东 珠海 519087;
2.北京师范大学自然科学高等研究院,广东 珠海 519087)

摘 要:过去几年,视频行为分类从手工选择特征方式逐步向采用深度学习端到端网络模型方式转变。讨论了传统手工选择特征的行为分类方法以及基于深度学习的行为分类方法,着重对包括基于卷积神经网络、长短期记忆网络和时空融合网络等不同的深度学习方法进行了论述,并对常用视频行为分类数据集做了概述,对视频行为分类方法的发展进行总结和展望。

关键词:视频行为分类;数据集;深度学习
中图分类号:TP391 **文献标识码**:A DOI:10.16157/j.issn.0258-7998.212388

中文引用格式:杨戈,邹武星. 基于深度学习的视频行为分类方法综述[J].电子技术应用,2022,48(7):1-7,12.
英文引用格式:Yang Ge,Zou Wuxing. A survey on video action classification methods based on deep learning[J]. Application of Electronic Technique,2022,48(7):1-7,12.

A survey on video action classification methods based on deep learning

Yang Ge[1,2],Zou Wuxing[1,2]
(1.Key Laboratory of Intelligent Multimedia Technology,Beijing Normal University,Zhuhai 519087,China;
2.Advanced Institute of Natural Sciences,Beijing Normal University,Zhuhai 519087,China)

Abstract:In the past few years, video action classification has gradually changed from manual feature selection to deep learning end-to-end model. This article discusses the traditional action classification method of manually selecting features and the action classification method based on deep learning, focusing on different deep learning methods including convolutional neural networks, recurrent neural network, dual-stream network, long and short-term memory network, etc., and it summarizes the commonly used video action classification data sets, summarizes and prospects the development of video action classification methods.

Key words:video action classification;data set;deep learning

0 引言

视频行为分类的目的是根据视频内容将视频行为归类为预设类别。随着数字摄像机、智能手持终端等各种视频拍摄设备的普及,网络上视频产生数量出现飞速增长。截至2019年6月,中国网络视频用户规模近7.59亿,中国短视频用户规模为6.27亿[1],最新兴起的短视频业务用户规模以及用户日均短视频移动应用(Application,APP)停留时长均出现爆发式增长。图像本身就包含大量信息,而视频是图像在时间维度的扩展,每秒往往包含24帧左右的图像,所占存储空间较之图像可以说是呈数量级倍数关系。存储、分析这些视频内容需要花费巨大的财力和人力,在计算机自动分析视频数据得到广泛应用前,视频内容的行为分类一般依靠人工实现,不仅效率低而且误判、漏判率高。自动化视频内容分析技术推广的现实意义广泛而深远。

计算机视频行为分析技术不仅可以同时自动监控多路信号,且不会产生疲劳,降低误判的可能性;在视频内容检索领域的应用更是将极大减轻公共安全从业人员的视频检索工作量,提高他们的检索效率,降低漏检率。自动化视频内容分析技术推广的现实意义广泛而深远,深度学习在视频分类的应用主要有以下方面。

(1)智慧校园

校园安全问题目前是国家非常重视的一个问题,通过行为识别技术在校园里的应用,能自动对校园监控到的视频进行行为分析,对可能出现危险的包括斗殴、坠楼、禁区侵入等情况及时通知管理人员,排除险情;另一

*基金项目:广东高校省级重大科研项目(2019KZDXM015,2020ZDZX3058);广东省学科建设专项资金(2013WYXM0122);校级智能多媒体技术重点实验室(201762005);校级教学团队(202012);校级课程思政(201932);2020年广东省教改项目(655)

方面，可以通过行为识别技术分析课堂教学情况，对学生听课状态进行数字量化，辅助教师管理班级，提高家长对孩子的知情度。

(2)智能安防

安防是行为识别非常好的应用领域，在目前的安防监控条件下，一位监控人员仅能观察几路视频，且长时间重复劳动容易产生疲劳、走神以及主观判断偏差等问题，导致危险的发生。行为识别技术的引入可以自动分析所有监控信号，减少人力消耗。同时，机器的工作效率不会随着运行时长的增加而降低，提高了高危行为报警的可靠性。

(3)视频内容搜索与检测

目前相似图片搜索和图片内容搜索已经在各大搜索引擎网站得到普及，为有图像搜索需求的用户带来极大的便利。但视频内容搜索还没有得到很好的应用。一旦行为识别技术能成熟地应用于视频内容搜索与检测领域，一方面，将为有视频内容搜索需求的用户带来革命性的使用体验，可以方便进行诸如人或物品的视频查找，视频事件检索与回溯；另一方面，在短视频审核领域的应用可以大大降低抖音、快手、微视等短视频 APP 后台审核人员的工作量，带来巨大的经济效益。

目前视频行为分类技术主要分为两大方向：采用传统选择特征的方式和使用深度学习建立端到端预测网络模型的方式[2]。传统的视频行为分类方法先手工选择并提取相关视觉特征，然后对特征进行编码，最后采用统计机器学习中相关分类技术得到预测分类结果。视频本身具有大量时间空间信息，基于传统手工提取特征的算法往往针对特定类型的视频有较好的分类效果，但提取特征过于依赖人工选择，泛化性能较差。深度学习的方法在 2012 年的图像分类等计算机视觉任务中取得了良好的效果[3]，深度学习的方法越来越受到研究人员的关注。

1 基于传统手工提取特征的分类

前面提到行为分类技术分为两个大类，传统手工提取特征的视频行为分类典型流程主要有两个阶段，即手工提取特征和特征编码。需要注意的是这里说的手工是指研究人员依据视频数据的特点手工设计相关特征或者滤波器，以此来描述、编码这些视频数据。

1.1 视觉特征提取

1.1.1 方向梯度直方图

方向梯度直方图(Histogram of Oriented Gradient，HOG)[4]是图像方向梯度的统计直方图。由于 HOG 能较好描述图像空间信息，因此在行人检测(Pedestrian Detection)领域有很好的表现[5]。对于 HOG 的计算，首先是计算梯度值(Gradient)。这个梯度值是灰度图像的梯度，使用 x 和 y 两个方向上的滤波核 $f_x=[-1\ 0\ 1]$ 与 $f_y=[-1\ 0\ 1)]^T$ 对灰度化后的图像进行滤波。对于灰度图像 P，得到灰度图像在 x 和 y 方向上的卷积滤波特征图 $P_x=P\cdot f_x$ 和 $P_y=P\cdot f_y$。继而计算到梯度的大小 $|g|$ 和方向 θ，根据梯度的大小和方向进行加权投票，最后做局部归一化(Local Response Normalization，LRN)。这些来自所有块的归一化直方图分量的向量就是 HOG 特征。

1.1.2 光流梯度直方图

光流梯度直方图(Histogram of Flow，HOF)[6]像 HOG 一样，也是一种视觉特征的统计直方图，不过这个特征不再是梯度，而是光流。通俗地说，光流是视觉感觉到的色彩动态的运动，是动态画面中像素点的瞬时速度[7]，常常用于分析图像的时间特征，如行为分类。获得 HOF 特征首先要提取出图像的光流，对视频的每帧画面计算其光流向量矩阵；再像 HOG 一样统计直方图，通过水平轴与光流矢量夹角建立相应的直方图的横坐标，用对应方向光流的大小加权作为纵坐标。同样为了使算法能有更好的健壮性，需要做局部归一化。

1.1.3 运动边界直方图

运动边界直方图(Motion Boundary Histograms，MBH)[8]也是为获得图像时序信息所经常使用的一种视觉特征。HOG 统计的指标是图像灰度后的梯度，HOF 是光流，而 MBH 统计的指标量区别于 HOG 和 HOF，是图片水平 x 方向和垂直 y 方向光流场灰度的梯度。具体说来，是先把水平和垂直方向光流场灰度化，再对这两个灰度的光流场像 HOG 一样计算梯度。

本节介绍的三种特征中 HOG 属图像空间(Spatial)特征，HOF、MBH 能体现视频的动态性，是图像或视频的时间(Temporal)特征。

1.2 特征编码方式

提取到视频的相关局部、全局特征后，需要进行编码，可采用像视觉词包编码(Bag Of Visual Words，BOVW)[9]、Fisher Vector[10]等方案，最后应用支持向量机(Support Vector Machine，SVM)或 softmax 得到行为分类的结果。

1.2.1 视觉词袋模型

词包(Bag-of-Words，BoW)[11]最早出现在信息检索和自然语言处理领域，核心思想是忽略句子本身的语法和语序，而用一系列词语(Word)的集合来表征一段文字或一段话。首先对句子进行分词建立词典，对词典中的词语进行唯一索引，并统计词频。索引的顺序与词汇在原文中出现的顺序没有关联。这样，使用词典对原文编码成向量。向量的各个维度按索引序对应词典中的词汇，大小就是这个词汇在原文中出现的词频。由此可见 BoW 本质也是一种统计直方图，只不过统计的不再是每个图像单元的梯度、光流或是其他视觉特征。在自然语言分析或文本内容检索的任务场景，BoW 可以描述不同句子或文字(Text)的相似度。由此带来启示，BoW 被应用到计算机视觉任务中[12]，文献[13]提出视觉词包(Bag of Visual Words，BOVW)，用图像块的特征如色彩直方图、尺度不变特征变换(Scale-Invariant Feature Transform，SIFT)[14]、局

部二值模式[15]等代替词汇，这样 BOVW 编码就成了由所有图像块的特征得到的直方图。

1.2.2 Fisher Vector

Fisher Vector[16](FV)和 BOVW 都是编码算法，都可以表示图像特征，也都可以对这些特征矩阵归一化。不同的是 FV 不仅存储特征在图像中出现的频率，还统计全局特征频率与局部特征的差异。FV 采用混合高斯模型(Gaussian Mixed Model，GMM)构建词典或 SVM，前者可以衡量同类型特征之间的相似度，后者属于判别异构数据间的不同。BOVW 得到的是一个极其稀疏的向量，只关注了关键词的数量信息，仅含有 0 阶信息。较之 BOVW，FV 并不稀疏，还包含了 1 阶信息(期望)、2 阶信息(方差)，在表征图像时信息更加丰富。

2 基于卷积神经网络

文献 [3] 采用深度学习卷积神经网络(Convolutional Neural Networks，CNN)在 2012 ILSVRC 竞赛(图片分类竞赛)中对 ImageNet 图片分类任务的预测准确率领先亚军超过 10%，引起学界强烈关注。紧跟其后，文献[16]率先将深度学习的方法从图像分类任务领域引入视频行为分类领域。

CNN 的运作方式与标准的神经网络相似，不过，一个关键的区别是 CNN 层中的每个单元都是一个二维(或更高维)的过滤器，它与该层输入(可能是图像，也可能是卷积后的特征图)进行卷积。CNN 过滤器通过使用类似(但更小)的空间形状作为输入合并空间上下文，并使用参数共享来显著减少需要学习参数的数量。CNN 通过检测大量抽象的特征来自适应将给定数据(图像或视频)映射到它相应的类别。

支持向量机(Support Vector Machine，SVM)是一种基于统计的机器学习模型，它在所有样本在当前空间都可实现线性可分的假设前提下，通过对所有样品进行拟合，最终获得一个利用支持向量(Support Vector)描述的并在样本之间具有最大间隔的一个超平面(Hyperplane)，也即使用这个超平面实现对不同类别的样本进行区分。对于在二维的情形，SVM 对样本集 A 和样本集 B 进行分类的本质是找到一个超平面对这些样本进行划分。

2.1 2D-CNN

CNN 在图像分类任务中有很好应用后，文献[16]研究了 2D-CNN(2 Dimensions-CNN)在视频时间维度上的扩展方式，提出多分辨率融合的模型，与单帧的 2D-CNN 相比性能有了显著提升。文献[16]提出了几种 2D-CNN 在时间维度的扩展方案。Late Fusion 是将两个有一定时间间隔的 RGB(Red-Green-Blue)帧分别输入到两个独立的单帧网络中，各自通过对应网络中的卷积、池化层，在第一个全连接层处合并两个网络的特征图。Single Frame 网络不能提取视频的时间信息，但 Late Fusion 的全连接层融合了两帧网络的输出以获得全局时间信息。Early Fusion 是把网络的第一层卷积层在时间范围做了扩展以对多帧进行卷积，是在原始图像帧级别上对时间信息进行融合。这种对像素底层的直接融合使网络较好地获取视频局部运动信息。Slow Fusion 可以认为是 Late Fusion 和 Early Fusion 的混合，首先由多个 CNN 对多段有重叠的视频帧进行融合，并且随着网络层数的提高，逐步融合更多的时间、空间信息，在全连接层前最后一组卷积、池化层进行全局融合。

文献[16]将这三种融合方式与单帧模型(不融合)对比，实验发现 Slow Fusion 优于 Late Fusion 和 Early Fusion，在 UCF101 数据集上达到 60.9%的预测准确率。但作者认为单帧模型已经显示出一定的性能，在 UCF101 数据集上预测准确率为 59.3%。

2.2 C3D

从文献[16]结果中可以看到 2D-CNN 的几种不同融合方式相比单帧网络并没有特别大的效果提升。这可能是由于 2D-CNN 没有很好提取视频的时间动态信息造成的。文献[17]提出的 3D 卷积是通过将 3D 内核卷积应用到视频中来实现的，是一种对 2D-CNN 的扩展，称作 C3D(3 Dimensions-CNN)。C3D 之前 2D-CNN 在行为分类领域探索的效果并不好，2D-CNN 不能很好地捕获视频的时间特征，两者的区别如下。

通常的 CNN 是 2D 的，它对视频的单帧画面进行处理；2D-CNN 应用在连续图像帧，得到的多帧特征图经融合后输出，即多帧的信息丢失了很多，被压缩了；3D-CNN，其卷积核得到的特征图也是 3D 的，并没有融合。给定视频数据是 RGB 三通道，分辨率为 $h\times w$，C3D 的卷积核深度为 d，即规格为 $k\times k\times d$。比起 2D-CNN，C3D 显然更适合学习带有时序信息的图像数据(视频)。

C3D 的结构并不复杂，如图 1 所示，包括 8 个卷积层，5 个池化层，2 个全连接层，最后经过 softmax 输出。网络上有学者通过实验得到了 3×3×3 的最佳 3D 卷积核规格。

特别需要说明的是池化层的规格，除最左侧 2×2×1，其他是 2×2×2。说明第一次池化操作是在单帧上进行

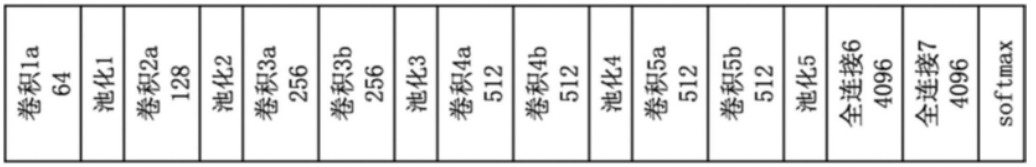

图 1 C3D 网络结构

的，其他是在相邻帧完成的，单帧池化对开始阶段保留时序特征有利，不会一开始就抽象化。比起2D-CNN，C3D显然更适合学习带有时序的图像数据。C3D在UCF101数据集上分类准确率达到了85.2%，较2D-CNN来说有了很大的提升。

3 基于循环神经网络

循环神经网络(Recurrent Neural Network，RNN)是一种反馈神经网络，之所以称为循环，是因为具有时间特征的数据往往存在上下文(Context)关联性。RNN的隐藏层的输出还会反馈(Feed-back)到该层的输入，即隐藏层在t时处理的信息不单是从输入层传递过来的I_t，还有本层前一次计算的输出M_{t-1}，中间层节点之间有连接，实现了“记忆”功能。RNN这种本身能记忆早期输入数据的有状态特性，适合处理上下文信息。而视频是一种包含时序上下文信息的图像数据，有研究人员尝试将RNN的各种变种如长短期记忆网络(Long Short-Term Memory，LSTM)、GRU(Gated Recurrent Unit)引入到视频行为分类中来。

3.1 CNN+LSTM

文献[18]将CNN与LSTM结合，提出了LRCN(Long-term Recurrent Convolutional Networks)。LRCN在浅层使用CNN获得视频图像帧的空间特征，再将这些信息以视频时间作为LSTM的输入，依靠LSTM网络学习视频时间信息。LRCN不限于固定长度输入帧，因此，可以学习分类更复杂的动作视频。在LRCN中，单个视频帧首先通过具有共享参数的CNN，然后连接到单层LSTM网络。确切地说，LRCN模型结合了一个深层视觉特征抽取器(CNN)和一个LSTM，后者可以以端到端方式学习时序信息的变化。这种方法通常通过在视频帧上应用2D-CNN，然后将CNN激活输出传递给LSTM以便表征视频的时间变化。

视频有序帧画面首先进入网络，首先应用2D-CNN获取帧图像的特征图，之后把这些特征图按时序输入后面的LSTM并得到一组时序向量。这种网络有很好的适应性，因为可以看到该结构是对CNN和LSTM的结合，输入既可以是单帧画面，也可是时序视频，与之对应得到的输出是图像预测或序列预测。

LRCN的灵活性使得只需简单调整就可以应用在多种视觉任务场景中，除行为分类、视频描述(Video Caption)外，在视频预测(Video Prediction)任务中有着很好的效果。LRCN在三种不同视觉任务中，网络结构的差异可参考文献[18]。

LRCN在UCF101数据集上达到了82.9%的分类准确率。

3.2 CNN嵌入RNN

文献[19]认为视频小的区域的动态性一般局限于邻近帧，所以获取全局动态信息的LRCN不太适合提取时间信息的细节。为此引入了RCN(Recurrent Convolution Networks)结构，在2D-CNN中嵌入LSTM的变种GRU。这样卷积层激活或卷积特征图保留了输入视频更精细的空间细节，用于提取局部时空特征[19]。

GRU也是一种RNN，是LSTM的一种变种，将输入门和遗忘门合并成更新门，最终的结构比LSTM简单。将RNN直接应用于CNN特征图，不可避免生产由于表征卷积特征图输入到RNN隐藏变换的大量参数。此外，CNN特征图保留了帧空间拓扑。文献[19]提出通过在RNN单元中引入稀疏性和局部性，减少内存需求来利用此拓扑，扩展了文献[20]提出的GRU-RNN模型，并采用卷积替换全连接的RNN线性乘积运算。GRU网络允许每个循环单元自适应捕获不同时间尺度的依赖关系。基于RCN的方法在UCF101数据集上达到80.7%的准确率，实验表明，利用多种分辨率的感知来建模时间变化，可以提高网络模型的性能。

4 基于时空融合网络

4.1 Two Stream

文献[21]提出双流网络(Two-Stream Network)模型，使用两个独立的时空CNN，通过后期融合将两个网络输出合并。空间网络从单视频RGB帧进行行为分类，而时间网络则从密集的光流中进行行为分类。双流网络模型背后的思想与以下事实有关：大脑视觉皮层包含用于对象和运动分类两条路径[21]。

(1)空间流卷积网络

双流网络中空间CNN结构类似文献中的单帧结构，即给定一个行为视频，每个视频帧将分别通过空间网络，并为每个帧分配一个行为标签。需要注意的是，对于同一个行为的视频的所有帧，给定的行为标签是一样。

(2)时间流卷积网络

双流网络中时间CNN则对几个连续帧之间的光流位移场进行分析，以学习时间特征。下面说明基于光流输入可选的三种输入变型：光流堆叠、轨迹堆叠和双向光流堆叠。

光流堆叠是通过堆叠L个连续的图像密集光流形成CNN的输入。第t帧中点(u,v)的光流是2D位移(水平和垂直位移分量)，它将该点移至下一帧$t+1$中的对应点。因此，L个连续帧堆叠的密集光学流形成$2L$个通道的输入图像，将其输入到时间CNN中。

轨迹堆叠与光流堆叠算法在L个连续帧中的相同位置采样位移矢量不同，轨迹堆叠算法通过沿运动轨迹采样L个2D点来表示$2L$通道输入图像中的运动。光流堆叠的输入通俗说就是某个固定位置的像素点到该点下一帧位置的位移矢量，而轨迹堆叠是某个锚点在各相邻帧之间的位置矢量。

光流堆叠和轨迹堆叠算法均在前向光流上运行。双向光流堆叠算法通过计算向前和向后光流位移场来扩

展这些算法。更准确地说，通过在帧 t 和 $t+\frac{L}{2}$ 之间堆叠前向光流和在帧 $t-\frac{L}{2}$ 和 t 之间堆叠后向光流，将运动信息编码在 $2L$ 通道的输入图像中。

(3)双流融合

双流网络模型分为两个网络，一个获取时间特征，另一个获取空间特征。时间网络中删去空间网络中第二个归一化层以减少内存消耗，其他部分空间网络和时间网络的结构相似。在两个网络的输出端将预测分数融合来组合空间和时间网络的 softmax 最终分数。不同的任务可以用于不同的融合方式。文献[21]实验得出在堆叠的 L2 归一化 softmax 分数上训练线性 SVM 分类器的效果优于作简单的平均融合。基于双流的方法在 UCF101 数据集上达到 88.0%的准确率。

4.2 Two Stream Fusion

文献[22]认为文献[21]的 Two Stream 网络模型依然不能充分利用时间维度的信息，无法学习到时间特征和空间特征像素级的关系。换言之，将空间特征和时间特征结合起来考虑，能为动作分类提供更多线索，也就有希望提升网络模型的性能。而且 Two Stream 网络模型对时间维度的利用很有限，空间网络只用了 1 帧，时间网络只用了 10 帧。文献[22]同样讨论了用于处理视频中空间和时间信息的两个网络。不同于文献 [21]，Two-Stream Fusion 重点在时间信息的处理上，网络模型结构参见文献[22]。

可以看到，网络中有两个数据流(two-stream)，分别用来提取时间和空间特征，最后每个支路都会输出一个 softmax 层，然后把两个 softmax 层的输出进行融合。在原文中仅提供了两种融合方法：空间融合是在隐藏层中间对两个网络进行融合；时间融合是用 3D 卷积(Conv 3D)和 3D 池化(Pool 3D)提取时间维度的特征，在时间维度上进行融合。

文献[22]在这个基础上继续讨论了多种其他融合方法。通过实验，得出三个结论：(1)在中间的卷积层进行融合比在最后的 softmax 层融合更能够提升性能。(2)在最后一层卷积层融合的效果最好，且在全连接层再次融合能进一步提高准确率。(3)融合之后进行 3D 池化能进一步提高网络模型性能。基于双流融合的方法在 UCF101 数据集上获得了 92.5%的准确率。

4.3 TSN

文献[23]提出的 TSN(Temporal Segment Network)也属于时空网络融合的方式。Two Stream 的弊端是只能对连续的几帧提取时间上下文信息，不能进行长时(long-range)的分析。文献[23]认为 CNN 在基于视频的行为分类任务方面难以展示好的效果，原因是长时结构在理解视频行为上起着重要作用，但主流的神经网络结构通常只关注空间信息和短期运动。另外，在实际中，训练深度卷积神经网络需要较大的训练样本以使性能最佳，但是这方面的数据资源有限。TSN 采用稀疏时间采样策略，利用整个视频长时信息支持有效地学习。

文献[23]提出了将视频分成 N 个部分，然后从每个部分中随机选出一个短的片段，对这个片段应用上述的 two-stream 方法，最后对于多个片段上提取到的特征做一个融合。TSN 网络结构能够在一段长的视频序列中通过稀疏采样的方法提取短片断(Short snippets)，这些样本在时间维度上服从均匀分布，从采样得到的片段中挖掘信息。

文献[23]还分别尝试了 4 种 two-stream 卷积神经网络的输入：RGB 图像、堆叠 RGB 差异、堆叠光流场、堆叠翘曲(warped)光流场。TSN 在 UCF101 数据集上达到了 94.2%的准确率。

5 数据集

随着计算机视觉算法的兴起，如何对比不同算法或模型的性能成为需要考虑的问题。为了在视频领域进行研究，一些机构花费了大量的人力和资源来收集、标记视频数据集。标准数据集的建立为对不同网络模型进行比较提供了平台。表 1 是常用视频行为数据集。早期发布的数据集样本数量少，种类也少，场景单一，而近年发布的数据集的视频样本增多，类别也多，场景复杂。其中，Weizmann、KTH 和 Hollywood 发布时间早，规模较小，行为数在 20 以下，但标签明确。UCF101(University of Central Florida 101)、Thumbos′14 和 HMDB51(a large Human Motion Database 51)是中型数据集，行为数量在 20 至 101 之间。大型数据集行为数通常在 200 以上，例如 YouTube-8M、Sports-1M、ActivityNet、Kinetics 等，其中 YouTube-8M 行为数量达到 4800 类，视频样本数量达 800 多万个，对网络模型提出了挑战。其中 HMDB51 和 UCF101 数据集由于视频样本数量适中，为近年大多数行为分类研究采用以比较网络模型的性能。

表 1 常用视频行为数据集

名称	行为数	视频数量	发布年
Kinetics[24]	400	306 245	2017
YouTube-8M[25]	4800	8 264 650	2016
ActivityNet[26]	203	27 901	2015
Thumos′14[27]	101	18 394	2014
Sports-1M[16]	487	1 133 158	2014
HMDB51[28]	51	6 474	2013
UCF50[29]	50	6 676	2012
UCF101[30]	101	13 320	2012
Hollywood2[31]	12	3 669	2009
Weizmann[32]	9	81	2005
KTH[33]	6	2 361	2004

5.1 UCF 系列

University of Central Florida(UCF)自 2007 年以来发

布的一系列行为分类数据集[30](UCF50,UCF101)引起广泛关注。这些数据集的原始视频大多收集自电视台以及互联网视频网站。其中UCF101数据集由于视频数和行为类别数多,成了近年视频行为分类领域研究普遍使用的数据集。

UCF101由从互联网视频网站YouTube整理出的时长在几秒到十几秒不等、每秒25帧、分辨率为320×240的13 320个视频样本组成。这些视频样本被分类并打上了101种行为标签。UCF101所包含的视频样本有很大的多样性,行为按属性分类情况如表2所示。UCF101包括不同类型的摄像机运动;多样的物体外观和人体姿势;有差异的物体比例、视角;杂乱的背景、亮度条件等。数据集中的101种行为被分作25个组,每个组含有4到7个行为类型。同组的行为有一些共性,如相似的视角,相似的背景等。UCF101中的行为有:人和物的互动、人的肢体行为、人之间的交互、音乐演奏、体育运动。

表2 UCF101数据集行为类型

	行为类别	行为数量	行为类别	行为数量
肢体运动	翻转	13	跳上跳下	10
	走	14	向前跳跃	13
	跑步	14	跳过障碍	3
	骑行	8	旋转	11
	下来	9	爬	2
	拉	6	水平	39
	提升	4	垂直向上	22
	推	6	垂直向下	26
	潜水	6	弯曲	26
人数	单人	89	双人	19
	多人	7		
场景	户外	65	室内	78
姿势	坐着	33	躺下	10
	坐在桌前	2	倒立	7
	站	77		

5.2 HMDB51

HMDB51数据集由Brown University SERRE实验室建立[28],视频样本大多源自YouTube网站和电影。HMDB51数据集大体可归成一般脸部动作、与外界互动的脸部动作、一般肢体动作、与外界互动的肢体动作和人间互动肢体动作5个类别,共计51种行为标签,6 766段视频。每个行为标签包含多于102段分辨率为320×240的视频。

除了行为类别的标签之外,每个视频还带有一些标签以及描述剪辑属性的元标签。因为HMDB51视频序列是从商业电影以及YouTube中提取的,光线条件、情况和周围环境多种多样;使用了不同的相机类型和录制技术捕获行为的出现;有全方位的覆盖,可区分运动正面、侧面(左右)和后方视角。另外,包括两个不同的类别,即"不运动"和"相机运动",后者是变焦、便携式旅行镜头拍摄和相机抖动等的结果。视频质量分3级,被评为"好"的视频样本,质量足以在运动过程中识别出单个手指。如果行为过程中身体部位或四肢消失或模糊化,则被评为"中级"或"不良"。数据集划分为训练集和测试集,训练集占70%用作训练,另外30%的测试集用作测试。

5.3 评估标准

准确率(precision)和查全率(recall)不能很好地衡量多分类任务的性能,视频行为分类常用的评估标准是平均准确率均值(mean Average Precision,mAP)。mAP的计算公式如式(1)所示,mAP值越大,分类的准确率越高。

$$\mathrm{mAP}=\sum_{m=1}^{M}\mathrm{Ave}P_m/M \tag{1}$$

其中P_m为m类的分类准确率。为了保证在同一数据集比较的公平性,一般制定了训练集和测试集。以UCF101数据集为例,划分约70%作为训练集,剩余30%作为测试集。划分三次,分别测试取平均值。

6 结论

本文首先介绍了视频行为分类领域的相关数据集,介绍了基于CNN、RNN、时空融合网络的深度学习的行为分类方法。表3是目前相关深度学习网络模型在UCF101数据集上的分类准确率。

表3 UCF101数据集分类准确率

网络模型	mAP/%
ConvNets-Temporal Fusion[16]	63.9
C3D[17]	85.2
Two-stream VGG[21]	88.0
LRCN[19]	82.9
Convolutional RNNs	80.7
Snippets Fusion[34]	88.2
Two-stream Fusion[22]	92.5
Temporal Segment Network[23]	94.2

这些基于深度学习的视频行为分类网络模型都取得了一些成果,在常用数据集上相比传统方法都有了一定幅度的提升。但随着更丰富、更大规模数据集的诞生,现有技术面临更多的挑战。更大规模的数据集对于网络模型需要更高的硬件性能,更长的训练时间,如何利用有限的训练时间或计算性能获得更好的分类准确率是一个研究方向。视频行为分类的实际场景中的画面较数据集可能更为复杂,如何让分类网络模型在实际应用领域中落地,特别是对异常行为,细微差别动作的准确分类也还需要大量工作;时空的人体骨架、动作信息也可以用于行为分类;随着深度摄像机技术的发展,深度对行为分类的作用同样需要进一步进行研究。

参考文献

[1] 艾媒大文娱产业研究中心.2019-2020中国文娱行业运行监测与头部企业布局研究报告[EB/OL].(2020-03-10)

[2021-11-29].https://www.iimedia.cn/c400/69705.html.
[2] SINGH T,VISHWAKARMA D K.Human activity recognition in video benchmarks: A survey[C]//Advances in Signal Processing and Communication.Springer,2019:247-259.
[3] KRIZHEVSKY A,SUTSKEVER I,HINTON G E.Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems,2012:1097-1105.
[4] JAIN A,SINGH D.A review on histogram of oriented gradient[J].IITM Journal of Management and IT,2019,10(1):34-36.
[5] CHEE K W,TEOH S S.Pedestrian detection in visual images using combination of HOG and HOM features[C]//10th International Conference on Robotics,Vision,Signal Processing and Power Applications,2019:591-597.
[6] RAGUPATHY P,VIVEKANANDAN P.A modified fuzzy histogram of optical flow for emotion classification[J].Journal of Ambient Intelligence and Humanized Computing,2019:1-8.
[7] CARMONA J M,CLIMENT J.Human action recognition by means of subtensor projections and dense trajectories[J].Pattern Recognition,2018,1:81-85.
[8] FAN M,HAN Q,ZHANG X,et al.Human action recognition based on dense sampling of motion boundary and histogram of motion gradient[C]//2018 IEEE 7th Data Driven Control and Learning Systems Conference(DDCLS).IEEE,2018:1033-1038.
[9] NAZIR S,YOUSAF M H,VELASTIN S A.Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition[J].Computers & Electrical Engineering,2018,72:660-669.
[10] AI S,LU T,XIONG Y.Improved dense trajectories for action recognition based on random projection and Fisher vectors[C]//MIPPR 2017:Pattern Recognition and Computer Vision.International Society for Optics and Photonics,2018:10609-10615.
[11] SILVA F B,WERNECK R O,GOLDENSTEIN S,et al.Graph-based bag-of-words for classification[J].Pattern Recognition,2018,74:266-285.
[12] NAZIR S,YOUSAF M H,VELASTIN S A.Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition[J].Computers & Electrical Engineering,2018,72:660-669.
[13] FEI-FEI L,PERONA P,A bayesian hierarchical model for learning natural scene categories[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR'05).IEEE,2005,2:524-531.
[14] LI R,LIU Z,TAN J.Reassessing hierarchical representation for action recognition in still images[J].IEEE Access,2018,6:61386-61400.
[15] SINGHAL S,TRIPATHI V.Action recognition framework based on normalized local binary pattern[C]//Progress in Advanced Computing and Intelligent Engineering,2019:247-255.
[16] KARPATHY A,TODERICI G,SHETTY S,et al.Large-scale video classification with convolutional neural networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition,2014:1725-1732.
[17] TRAN D,BOURDEV L,FERGUS R,et al.Learning spatiotemporal features with 3d convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision,2015:4489-4497.
[18] DONAHUE J,ANNE HENDRICKS L,GUADARRAMA S,et al.Long-term recurrent convolutional networks for visual recognition and description[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015:2625-2634.
[19] BALLAS N,YAO L,PAL C,et al.Delving deeper into convolutional networks for learning video representations[C]//International Conference on Learning Representations,2016:1320-1331.
[20] CHO K,VAN MERRI?NBOER B,GULCEHRE C,et al.Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Conference on Empirical Methods in Natural Language Processing,2014:1406-1418.
[21] SIMONYAN K,ZISSERMAN A.Two-stream convolutional networks for action recognition in videos[C]//Advances in Neural Information Processing Systems,2014:568-576.
[22] FEICHTENHOFER C,PINZ A,ZISSERMAN A.Convolutional two-stream network fusion for video action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016:1933-1941.
[23] WANG L,XIONG Y,WANG Z,et al.Temporal segment networks:towards good practices for deep action recognition[C]//European Conference on Computer Vision,2016:20-36.
[24] KAY W,CARREIRA J,SIMONYAN K,et al.The kinetics human action video dataset[EB/OL].(2017-05-19)[2020-03-29].https://arxiv.org/abs/1705.06950.
[25] ABU-EL-HAIJA S,KOTHARI N,LEE J,et al.Youtube-8m:a large-scale video classification benchmark[EB/OL].(2016-09-27)[2020-03-29].https://arxiv.org/abs/1609.08675.
[26] CABA HEILBRON F,ESCORCIA V,GHANEM B,et al.Activitynet:a large-scale video benchmark for human activity understanding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015:961-970.
[27] IDREES H,ZAMIR A R,JIANG Y G,et al.The THUMOS

(下转第 12 页)

将成为下一步研究和关注的焦点。

本文结合物联网发展趋势，概述了国际国内物联网安全标准化组织，从技术和应用视角对物联网安全标准进行了分类，尝试给出了物联网安全标准体系框架，并针对性提出了物联网安全标准化推进建议。下一步将结合物联网行业应用和新技术发展，深化物联网安全标准研究，加大安全标准应用推进。

参考文献

[1] 周开宇.ITU-T 物联网标准化综述[J].电信技术，2016(5)：13-15.
[2] 中国互联网发展报告[R].第二十届中国互联网大会，2021.
[3] 张玉清，周威，彭安妮.物联网安全综述[J].计算机研究与发展，2017，54(10)：2130-2143.
[4] 肖益珊，张尼，刘廉如，等.物联网安全标准及防护模型研究概述[J].信息技术与网络安全，2020，39(11)：1-7.
[5] 赵佩，陶鹏，李翀，等.物联网信息安全技术标准研究与解读[J].河北电力技术，2019，38(5)：1-3.
[6] 付敏，蒲小英，秦伟强.基于网络安全等级保护 2.0 标准的物联网安全体系架构[C]//2019 中国网络安全等级保护和关键信息基础设施保护大会论文集，2019：4.
[7] 刘巍，王冬鸽.物联网安全体系结构研究[J].物联网技术，2016，6(4)：61-63.
[8] 何艾玲，汤学军，陈卫平，等.基于三维结构模型的医疗健康物联网信息标准体系表编制方法设计与研究[J].中国标准化，2016(12)：127-131.
[9] 杨林.农业物联网标准体系框架研究[J].标准科学，2014(2)：13-16.
[10] 全国信息安全标准化技术委员会通信安全标准工作组.物联网安全标准化白皮书[R].2019-10.
[11] GB/T 33745-2017，物联网术语[S].2017.
[12] 吴晗.基于 AMQP 的消息中间件的设计和实现[D].南京：东南大学，2019.
[13] 陶伟，潘丰，崔恩隆，等.MT7628 与 OpenWrt 的 MQTT 异构协议设计[J].单片机与嵌入式系统应用，2021，21(9)：14-17，22.
[14] 雷煜卿，仝杰，张树华，等.能源互联网感知层技术标准体系研究[J].供用电，2021，38(7)：14-20，33.
[15] 宋金圣.大容量物联网网关技术研究与实现[D].成都：电子科技大学，2021.
[16] 卢周正.基于 USB 3.0 总线标准的疲劳试验机控制器设计与实现[D].杭州：浙江大学，2017.
[17] 曲至诚.天地融合低轨卫星物联网体系架构与关键技术[D].南京：南京邮电大学，2020.
[18] 中共中央、国务院.《国家标准化发展纲要》[Z].2021.
[19] 周丽莎，孔勇平，陆钢.物联网安全政策解读及技术标准综述[J].广东通信技术，2017，37(12)：39-41，45.
[20] 布轩.《物联网新型基础设施建设三年行动计划(2021-2023 年)》解读[N].人民邮电，2021-09-30(008).

(收稿日期：2021-10-27)

作者简介：

李峰(1982-)，男，博士，高级工程师，主要研究方向：物联网、智能物流。

陈亮(1991-)，男，硕士，工程师，主要研究方向：物联网。

李凯(1992-)，男，博士，工程师，主要研究方向：物联网。

扫码下载电子文档

(上接第 7 页)

challenge on action recognition for videos in the wild[J]. Computer Vision and Image Understanding，2017，155：1-23.
[28] KUEHNE H，JHUANG H，STIEFELHAGEN R，et al.Hmdb51：A large video database for human motion recognition[C]// High Performance Computing in Science and Engineering'12，2013：571-582.
[29] REDDY K K，SHAH M.Recognizing 50 human action categories of web videos[J].Machine Vision & Applications，24(5):971-981.
[30] SOOMRO K，ZAMIR A R，SHAH M.UCF101：a dataset of 101 human actions classes from videos in the wild[J].Computer Science，2012：231-243.
[31] LAPTEV I，MARSZALEK M，SCHMID C，et al.Hollywood2：human actions and scenes dataset[Z].2008：12-16.
[32] ARUNNEHRU J，CHAMUNDEESWARI G，BHARATHI S P. Human action recognition using 3D convolutional neural networks with 3D motion cuboids in surveillance videos[J]. Procedia Computer Science，2018，133：471-477.
[33] SHORE T，ANDROULAKAKI T，SKANTZE G.KTH tangrams：a dataset for research on alignment and conceptual pacts in task-oriented dialogue[C]//11th International Conference on Language Resources and Evaluation，LREC 2018，2019：768-775.
[34] YUE-HEI NG J，HAUSKNECHT M，VIJAYANARASIMHAN S，et al.Beyond short snippets：Deep networks for video classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2015：4694-4702.

(收稿日期：2021-11-29)

作者简介：

杨戈(1974-)，男，博士，副教授，主要研究方向：人工智能技术、计算机视觉技术、网络智能化技术。

邹武星(1990-)，男，硕士研究生，主要研究方向：人工智能技术、计算机视觉技术。

扫码下载电子文档

版权声明