DOI:10.13718/j.cnki.xsxb.2021.07.020

基于 Spark 的动作识别特征提取①

荆于勤[1],　夏书银[2]

1. 重庆工商职业学院 电子信息工程学院,重庆 401520;2. 重庆邮电大学 计算机学院,重庆 400065

摘要:针对大规模动作识别时间长、识别精度低等问题,本文提出基于 Spark 框架的特征提取并行解决方法,利用 Spark 的内存计算能力,将视频数据分割成视频或帧,并将其放置到弹性分布式数据集(Resilient Distributed Datasets Sets,RDDS)中进行后续处理,针对主流的深度学习特征提取方法:轨迹池深度卷积描述符(Trajectory-Pooled Deep-Convolutional Descriptors,TDD)、潜在概念描述符(Latent Concept Descriptor,LCD)和改进密集轨迹(Improved Dense Trajectories,IDT),给出分布式并行算法,并设计了局部特征聚合描述符(Vector Of Locally Aggregated Descriptors,VLAD)分布式编码算法,将提取的特征聚合到全局表示中,然后输入到深度学习模型分类器中识别视频中的动作. 实验结果表明:本文方法提高了人类动作识别的实时性,且 LCD 在识别精度和处理时间之间的权衡优于其他几种方法.

关　键　词:Spark;弹性分布式数据集;特征提取;深度学习;动作识别

中图分类号:TP311　　**文献标志码**:A　　**文章编号**:1000-5471(2021)07-0135-05

相比其他技术,如环境传感器和可穿戴传感器,视频图像动作识别技术具有更高的效率和更低的成本,然而由于人类姿势和图像质量的大量变化,人类行为的可靠检测对于研究者来说仍然是一项极具挑战性的工作[1-3]. 人类行为识别(Human Action Recognition,HAR)是将人类行为转化为数字行为的过程,具有复杂的动作理解能力,在智能监控、网络视频搜索和检索、病人监护、运动分析、人机交互等多媒体应用中起着重要的作用[4].

在人类行为识别领域,许多研究者提出了不同的方法来促进该方面的进步. Jalal 等[5]实现了用于姿势估计的身体部位估计与检测,Uddin 等[6]使用深度递归神经网络对翻译和尺度不变特征进行活动识别. 现有典型的两种特征类型为:人造局部特征和深度学习特征提取. 吴亮等[7]提出了基于时空兴趣点和概率潜动态条件随机场模型的在线行为识别方法,应用时空兴趣点(STIP)对行为特征进行提取,Nguyen 等[8]提出了用于动态纹理识别的密集轨迹(DT)定向光束方法,这类方法在识别上有局限性[9-10]. 对于深度学习特征方法,Xiao 等[11]提出了分层动态贝叶斯网络的动作识别方法,Yang 等[12]提出的非对称 3D 卷积神经网络的动作识别方法,打破了识别上的局限性[13].

对于大规模的特征识别,传统深度方法具有局限性,需要考虑并行化处理方法. 文献[14]实现了 MapReduce框架下的深度神经网络特征提取,但是局限性是 MapReduce 不适合迭代算法. Apache Spark 通过使用弹性分布式数据集(RDDS)高效地执行分布式应用程序,更适合分布式视觉算法的开发.

在现有动作识别特征提取算法的基础上,本文提出基于 Spark 框架的特征提取并行解决方法,实现分

① 收稿日期:2020-07-13
基金项目:重庆市教育委员会科学技术研究计划项目(KJQN201904007);重庆市教育委员会科学技术研究计划项目(KJQN202004002).
作者简介:荆于勤,硕士,讲师,主要从事大数据及机器学习研究.

布式环境中视频序列提取局部特征. 该方法基于 Spark 框架，针对现有轨迹池深度卷积描述符(TDD)特征、改进密集轨迹(IDT)和潜在概念描述符(LCD)特征，设计特征提取并行算法，最后设计局部特征聚合描述符(VLAD)并行实现，将提取的局部特征聚合到全局表示中，识别视频中的动作.

1 特征提取方法

IDT 框架与 DT 的基本框架一致，不同之处是对光流图像的优化、特征正则化方式的改进[8]. ①估计相机运动来消除背景上的光流以及轨迹；②特征正则化方式由 L1 范数取代原理的 L2 范数正则化，能够提升分类准确率.

TDD 特征具有人造设计特征和深度学习特征的优点，有区分的卷积特征映射通过深度结构来学习，然后使用轨迹控制的 pooling 方法融合卷积特征. 首先设计深度的 ConvNet 提取卷积特征映射，选择具有较好性能的双流 ConvNet，该双流 ConvNet 包含两个单独的 ConvNet，即空间网和时间网. 空间网旨在捕获静态外观线索，这些线索在单帧图像上训练，而时间网旨在描述动态运动信息，其输入是堆叠的光流场体积.

双流 ConvNets 训练完成后，将其视为通用特征提取器，以获取视频的卷积特征映射. 对于每帧或每卷，将其作为空间网络或时间网络的输入. 对时空网络进行两种修改，①删除目标图层之后的图层进行特征提取；②在每个卷积或池化层之前，对层输入进行零填充，通过这种填充可以很容易地将视频中轨迹点的位置映射到卷积特征映射的坐标上. 空间网络和时间网络的输出是卷积特征映射，该卷积特征映射将在下一部分中用于提取 TDD.

TDD 的提取包括两个步骤：特征映射正则化和轨迹合并. 时空正则化方法可确保每个卷积特征通道在相同间隔内变化，从而对最终 TDD 识别性能做出同等贡献. 在特征正则化之后，基于轨迹和正则化的卷积特征映射，使用轨迹池提取 TDD.

对于卷积神经网络(CNN)潜在概念描述符(LCD)特征，本文 CNN 架构采用的是 2014 年 ImageNet 大规模视觉识别挑战赛牛津大学视觉几何组卷积神经网络分类任务获胜解决方案中具有 16 个权重层的配置，前 13 个权重层是卷积层，其中 5 个紧随其后的是最大合并层，最后 3 个权重层是全连接层.

2 基于 Spark 的分布式特征提取及编码表示

本节设计了在 Spark 环境中并行 LCD 提取方法，然后给出了 TDD 的并行实现方法，给出 IDT 并行描述，最后设计了 VLAD 编码的并行实现.

2.1 LCD 的分布式表示

在 Spark 上提取潜在概念描述符：利用 CNN 特征映射提取深层的局部特征，给定一帧 I_t，$t=1$，…，T，T 为视频持续时间，将 CNN 中间层的过滤器作为特征提取器，将 CNN 特征映射 M_t 的像素变成帧 I_t 中相应补丁的局部特征. 其中，$M_t \in R^{H\times W\times C}$ 是帧 I_t 的特征映射，H 是高度，W 是宽度，C 是通道数.

局部特征称为潜在概念描述符(LCD)，为了使群集内存受益，原始视频数据将从分布式文件系统(HDFS)加载到 Spark RDDS. 最初，flatMap()函数将视频输入文件作为输入，读取所有帧并将其放入帧 RDD 中，flatMap()函数由 Spark 执行，并应用于每个视频以获取所有 RGB 帧.

进行 flatMap()转换后，使用 BigDL 加载预训练的卷积神经网络(VGG19)模型并将其传递给 Map()函数，该函数利用 Conv5 层将所有 RGB 帧转换为 CNN 特征映射. 最后，将帧 I_t 的 CNN 特征映射传递给 flatMap()函数以获取 LCD 特征$\{LCD_{t_j}\}$，该特征将存储在 HDFS 中.

2.2 TDD 的分布式表示

在 Spark 上提取轨迹合并的深度卷积描述符：首先采用 CNN 的中间层来计算视频序列中每个帧的特征映射，通过使用改进轨迹的方法来检测一组轨迹，然后遵循轨迹约束的采样和合并策略，获得深度卷积

描述符. 通过在以轨迹点为中心的时空网络上合并局部 CNN 响应，将卷积特征映射与改进轨迹组合在一起，并将多个标度上的采样点作为 IDT 的原始实现进行跟踪.

2.3　VLAD 编码的分布式表示

在特征提取阶段之后，对局部特征进行编码生成全局表示，该全局表示将在随后的分类阶段中用于训练和测试.

3　实验结果与分析

为了评估基于 Spark 框架的特征提取分布式算法的性能，在 9 个节点(包括 1 个 Master 节点和 8 个 Slave 节点)的计算机集群上进行人类动作识别实验，每个节点具有相同的配置：Win10 操作系统，I7 处理器、8 GB 运行内存，使用 Hadoop 版本为 2.7，Spark 版本为 2.3.3，所有数据都放在同一 HDFS 群集上. 实验数据集为动作识别数据集 UCF101，该数据集是最大的动作数据集之一，在实际场景中从 YouTube 收集了 13 320个具有 101 个动作类的视频剪辑，每类动作由 25 个人做动作，分辨率为 320×240，共 6.5 G.

图 1 给出了分布式特征(IDT,LCD 和 TDD)的运行时间，由图 1 可以看出本文所提的 Spark 分布式特征方法的可行性和可扩展性. 另外，也可以看出当将节点数目从 1 增加到 8 时，特征提取过程明显加快. 即随着运行节点数量的增加，分布式特征提取的时间几乎呈线性下降. 这种良好的可伸缩性性能是因为 Spark 的内存计算能力，可以最大程度地减少 I/O 和网络通信的时间.

此外，LCD 的提取过程比 IDT 的提取过程快，这是因为从 LCD 的 RGB 帧中提取 CNN 特征映射，而 IDT 中的光流计算和特征跟踪操作需要更多的时间来执行. 当在大量节点上运行时，IDT 会花费额外的时间在本机程序和 Spark workers 之间进行通信. 同时，LCD 的所有特征提取过程都完全在 Spark workers 上运行，而无需本机库与 Spark 之间的通信. TDD 是 3 种特征提取方法中最耗时的，其执行时间几乎等于 LCD 和 IDT 提取时间之和. 这是因为从给定的原始视频中提取 TDD 的过程，是计算 CNN 特征映射和提取顺序执行轨迹的组合.

图 2 给出了不同特征提取方式执行分布式特征编码时执行时间的比较结果.

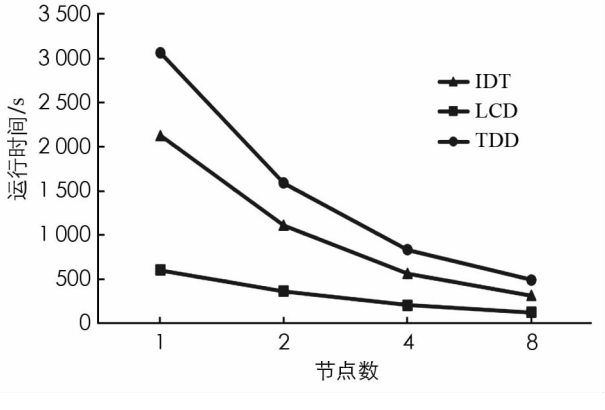

图 1　分布式特征提取执行时间

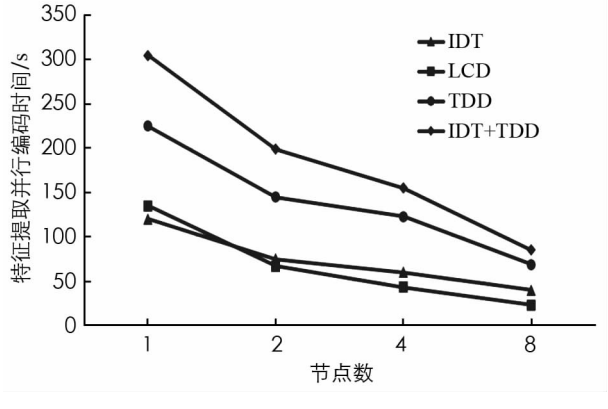

图 2　特征提取分布式特征编码时执行时间

从图 2 中可以看出，随着运行节点数量的增加，编码过程显著加快. 从图 2 中还可以看到 IDT 和 LCD 的编码时间没有太大差异，原因是编码时间受特征数量和特征维数影响，并且对于这两个特征，这些因子的值相近. IDT 特征的数量相当于 TDD 特征的数量，这是因为它们都采用了同样的轨迹约束采样策略. 本文进一步比较了在描述符级别将 IDT 和 TDD 组合在一起的结果，由于特征尺寸较大，组合特征的编码时间比其余 3 个特征更长.

然后使用目标检测、动作识别问题中最常用的度量标准——平均精度均值(Mean Average Precision，MAP)作为指标，来验证本文所提的分布式特征提取方法的有效性，实验结果如图 3 所示.

从图3中可以看出，使用VLAD编码，深度学习特征MAP要优于手工制作特征MAP. 尽管LCD是TDD的简化版本，具有更简单的采样策略，但LCD的精度更高. 另外，TDD拥有外观信息，而IDT捕获运动信息，由于IDT和TDD彼此具有很强的互补性，因此组合特征可提高MAP.

由图1—图3中数据可以得出，在4个特征中，LCD在精度和处理时间之间的权衡要比其他特征更好，组合特征方法可以提高准确性，但同时又会牺牲时间.

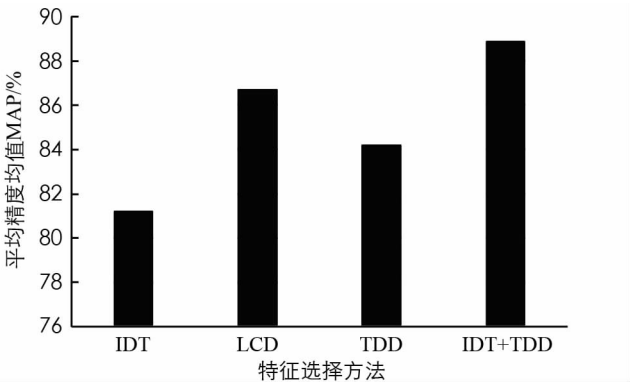

图3 分布式特征选择的平均精度均值对比

4 结 语

利用Spark提供的内存计算和容错功能来解决大规模的人类动作识别问题，本文提出了基于Spark的分布式动作识别特征提取方法. 设计了用于人类动作识别的几个特征提取的分布式解决方案，包括IDT，LCD,TDD以及VLAD编码的分布式实现. 在数据集UCF101上的实验可以得出，本文方法提高了人类动作识别的实时性能，并具有令人满意的可扩展性，其中LCD在精度和处理时间之间的权衡要比其他特征更好.

参考文献：

[1] 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.

[2] K S，S S M. Human Detection and Tracking Using HOG for Action Recognition [J]. Procedia Computer Science，2018，132：1317-1326.

[3] JALAL A，KAMAL S，AZURDIA-MEZA C A. Depth Maps-Based Human Segmentation and Action Recognition Using Full-Body Plus Body Color Cues via Recognizer Engine [J]. Journal of Electrical Engineering & Technology，2019，14(1)：455-461.

[4] KIM J，CHI S. Action Recognition of Earthmoving Excavators Based on Sequential Pattern Analysis of Visual Features and Operation Cycles [J]. Automation in Construction，2019，104：255-264.

[5] JALAL A，NADEEM A，BOBASU S. Human Body Parts Estimation and Detection for Physical Sports Movements [C]//2019 2nd International Conference on Communication，Computing and Digital systems (C-CODE). Islamabad：IEEE，2019.

[6] UDDIN M Z，KHAKSAR W，TORRESEN J. Activity Recognition Using Deep Recurrent Neural Network on Translation and Scale-Invariant Features [C]//2018 25th IEEE International Conference on Image Processing (ICIP). Athens：IEEE，2018.

[7] 吴 亮，何 毅，梅 雪，等. 基于时空兴趣点和概率潜动态条件随机场模型的在线行为识别方法[J]. 计算机应用，2018，38(6)：1760-1764.

[8] NGUYEN T T，NGUYEN T P，BOUCHARA F，et al. Directional Beams of Dense Trajectories for Dynamic Texture Recognition[M]//Advanced Concepts for Intelligent Vision Systems. Cham：Springer International Publishing，2018.

[9] YI Y，WANG H L. Motion Keypoint Trajectory and Covariance Descriptor for Human Action Recognition [J]. The Visual Computer，2018，34(3)：391-403.

[10] YU T Z，WANG L F，DA C，et al. Weakly Semantic Guided Action Recognition [J]. IEEE Transactions on Multimedia，2019，21(10)：2504-2517.

[11] XIAO Q K，SONG R. Action Recognition Based on Hierarchical Dynamic Bayesian Network [J]. Multimedia Tools and Applications，2018，77(6)：6955-6968.

[12] YANG H，YUAN C F，LI B，et al. Asymmetric 3D Convolutional Neural Networks for Action Recognition [J]. Pattern

Recognition，2019，85：1-12.

[13] ZHU J G，ZHU Z，ZOU W. End-to-End Video-Level Representation Learning for Action Recognition [C]//2018 24th International Conference on Pattern Recognition (ICPR). Beijing：IEEE，2018.

[14] SINTHONG P，MAHADIK K，SARKHEL S，et al. Scaling DNN-Based Video Analysis by Coarse-Grained and Fine-Grained Parallelism [C]//2020 IEEE International Conference on Multimedia and Expo (ICME). London：IEEE，2020.

Feature Extraction of Action Recognition Based on Spark

JING Yu-qin[1]，　XIA Shu-yin[2]

1. *College of Electronic Information Engineering，Chongqing Technology and Business Institute，Chongqing 401520，China*；

2. *College of Computer Science，Chongqing University of Posts and Telecommunications，Chongqing 400065，China*

Abstract：Aiming at the problems of large-scale motion recognition time and low recognition accuracy，a parallel solution method for feature extraction based on the Spark framework has been proposed. Using the memory computing advantage of Spark，the video data is divided into videos or frames and placed into an elastic distribution. the subsequent processing in the RDDS，for the mainstream deep learning feature extraction methods：trajectory-pooled deep-convolutional descriptors，latent concept descriptor and improved dense trajectory，distributed parallel steps are given and designed the vector of locally aggregated descriptors VLAD distributed encoding algorithm aggregates the extracted features into a global representation and then inputs them into the deep learning model classifier to identify the actions in the video. Experimental results show that the method in this paper improves the real-time performance of human action recognition，and the trade-off between recognition accuracy and processing time of LCD is better than other methods.

Key words：spark；resilient distributed dataset；feature extraction；deep learning；action recognition

责任编辑　夏　娟