·多媒体技术及应用· 文章编号：1000-3428(2019)12-0274-07 文献标志码：A 中图分类号：TP391

基于运动区域差分与卷积神经网络的动作识别

陈晓春[1]，林博溢[2,3]，孙 乾[2]，张坤华[3]

（1. 深圳清华大学研究院 电子设计自动化实验室，广东 深圳 518057；2. 鹏城实验室，广东 深圳 518082；
3. 深圳大学 电子与信息工程学院，广东 深圳 518060）

摘 要：针对视频动作识别中数据处理效率不高的问题，建立一种基于视频帧间差分序列的动作识别模型。利用帧间差分检测视频帧中的运动区域，以该区域为中心进行相应的图像剪切和增强处理。整个识别模型采用双流架构，在数据样本制作时通过适当的隔帧差分来扩大样本的时间跨度。采用分阶段逐步增加训练样本量的方法，以提升模型识别性能并解决训练过程中易出现的过拟合问题。实验结果表明，该模型可以在 CPU 级配置的电脑中完成快速动作识别，且在 UCF11 和 UCF25 数据集中的识别准确率均高于 85%。

关键词：帧间差分；动作识别；双流架构；卷积神经网络；运动区域

开放科学（资源服务）标志码（OSID）：

中文引用格式：陈晓春，林博溢，孙乾，等. 基于运动区域差分与卷积神经网络的动作识别[J]. 计算机工程，2019，45(12)：274-280，293.

英文引用格式：CHEN Xiaochun，LIN Boyi，SUN Qian，et al. Action recognition based on motion region difference and convolutional neural network[J]. Computer Engineering，2019，45(12)：274-280，293.

Action Recognition Based on Motion Region Difference and Convolutional Neural Network

CHEN Xiaochun[1]，LIN Boyi[2,3]，SUN Qian[2]，ZHANG Kunhua[3]

（1. Key Laboratory of Electronic Design Automation，Research Institute of Tsinghua University in Shenzhen，Shenzhen，Guangdong 518057，China；2. Peng Cheng Laboratory，Shenzhen，Guangdong 518082，China；
3. College of Electronics and Information Engineering，Shenzhen University，Shenzhen，Guangdong 518060，China）

【Abstract】Aiming at the low efficiency of data processing in video action recognition，this paper proposes an action recognition model based on the difference sequences between video frames. First，this paper uses inter frame difference to detect the motion region in the video frame，and this region is taken as the center where corresponding image clipping and enhancement are carried out. Then，the dual-stream architecture is applied to the recognition model and the time span of the samples is extended by the appropriate frame difference when data samples are made. Finally，the number of training samples is gradually increased，so as to improve the performance of model recognition and tackle the over fitting problem in the training process. Experimental results show that the proposed model can complete fast action recognition in CPU level computers，and its recognition accuracy in UCF11 and UCF25 datasets is higher than 85%.

【Key words】inter frame difference；action recognition；dual-stream architecture；Convolutional Neural Network（CNN）；motion region

DOI：10.19678/j.issn.1000-3428.0053623

0 概述

视频动作识别有助于人机交互中机器对人类行为动作的理解与判断。动作视频包括描述场景和运动目标的空域部分以及运动轨迹对应的时域部分。在实际应用中，摄像效果通常受到场地和光照的干扰，且人类在进行同一类动作时存在明显的类内或类间区别，使得同一动作在不同视频中存在不同程度的差异，这给视频动作的特征提取带来了困难。

基金项目：广东省科技计划项目（2016B010126003）；深圳市基础研究项目（JCYJ20170816151958999）。

作者简介：陈晓春（1972—），男，博士，主研方向为机器学习、多媒体信息处理；林博溢、孙 乾，硕士研究生；张坤华，副教授、博士。

收稿日期：2019-01-09 **修回日期：**2019-03-01 **E-mail：**chenxc@tsinghua-sz.org

光流场(Optical Flow Field,OFF)是用来描述视频图像序列中像素点灰度值变化趋势的有效手段。基于稠密光流(DT)[1]的动作描述方案自提出后经过不断改进,目前已成为主流的动作识别技术。早期的光流动作检测方法是采集相邻视频帧的稠密光流,再通过光流图提取定向光流直方图(HOF)、方向梯度直方图(HOG)和运动边界直方图(MBH)等特征作为动作描述子。由于动作特征数据规模庞大,为便于分类器学习并克服冗余信息干扰,研究者们对提取的光流进行 FisherVector 或 BoW 编码[2],以提高视频中动作特征的质量。基于光流的改进动作识别方法有 iDT[3]、双流[4]和 LSTM[5]等,其他的动作检测方法还有以 C3D[6]为代表的 3D 卷积神经网络(Convolutional Neural Network, CNN)处理方案[7-8]。双流方法使用 2 个 CNN 分别处理时域光流信息和空域图像信息,最后将 2 个网络的检测结果进行融合以得到动作识别结果。在该方法中,时域卷积网络的输入为栈化限长连续视频帧,空域卷积网络的输入为单帧静态图片,模型的检测速度和精度受光流图与时域长度的影响较大。在双流方法的基础上,研究者们提出了较多改进方法[9-11],如 TSN 采用稀疏采样方式将动作视频分为数段,在每一段中随机抽取一帧进行处理并作为双流卷积网络的输入,最后将所有输出的预测结果进行整合得到判断结果。TSN 方法通过采用模型预训练、正则化或数据增强等策略而获得了较高的识别准确率。文献[11]采用批归一化变换与 GoogLeNet 相结合的网络结构,通过融合时空网络来识别动作,其准确率较高。

然而,现有的动作识别大多处于动作视频分类阶段。在光流的基础上发展起来的双流识别方法虽具有较高的识别准确率,但是稠密光流的提取需要大量的 CPU/GPU 计算资源和时间消耗,基于稠密光流的网络结构也较复杂,难以实现端到端的动作检测。采用 3D 卷积网络[8,12-13]虽然可以实现端到端的训练和检测,但是 3D 卷积计算需要消耗较大的处理资源。文献[13]将 3D 卷积与双流网络相结合,利用视频中的空时信息来识别人物动作,其首先利用 2 个 CNN 分别抽取视频动作片段的空间和时间特征,然后融合这 2 个 CNN 并提取中层时空特征,最后将提取的中层特征输入到 3D 卷积神经网络中,以完成视频中的人物动作识别。但是,目前识别率较高的 3D 卷积网络模型参数规模庞大,在识别过程中都需要在 GPU 的协助下才能完成动作分类。

视频差分运算常被用于定位目标运动区域,如混合高斯模型[14]和帧间差分[15]等。当视频中出现异常运动目标时,差分运算往往能准确捕获图像中的运动区域以及位移轮廓。本文研究动作视频的差分图像提取、运动区域检测以及基于 CNN 的特征提取与动作识别问题,在此基础上,提出基于帧间差分的运动目标表征与动作识别方法。

1　图像运动区域与栈化处理

目前常用的运动目标提取方法有光流法、背景差分法及帧间差分法等。考虑到实际场景中背景复杂多变会对目标提取造成较大干扰,以及人机交互对实时性检测的需求,本文采用像素级别的帧间差分法。由于视频序列中的动作运动幅度不一,而动作背景多数情况下处于相对静止的状态,通过帧间差分计算可以得到 2 帧之间的像素差分图。根据这些像素差值和设定好的阈值,可以实现运动变化像素提取。

已知一个视频由多帧组成,表示为 $X=[x_1,x_2,\cdots,x_t]$,其中,x_t 表示视频中的第 t 帧,t 为整数。帧间差分计算方法具体如下:

$$D\_x_{(k,k+d)}=\begin{cases}|x_{k+d}-x_k|, & |x_{k+d}-x_k|\geqslant T\\ 0, & |x_{k+d}-x_k|<T\end{cases} \tag{1}$$

其中,x_k 表示视频中的第 k 帧图像,x_{k+d} 表示第 $(k+d)$ 帧图像,x_k 和 x_{k+d} 之间有 $(d-1)$ 帧图像,一般情况下取 $d=1$,$D\_x_{(k,k+d)}$ 表示 x_k 和 x_{k+d} 在差分运算后的目标图像,T 为阈值。差分图像中小于 T 的像素值判定为静止,即认为是动作发生时的背景,反之,认为该像素属于运动目标。

考虑视频动作的时域特性,可以在视频中选择一系列连续的帧间差分图像来表征动作的连续形态变化,本文称之为栈化帧间差分流。设栈化帧间差分流包含的图像数目为 K,则可以在视频中连续取 $(K+1)$ 帧图像,使用式(1)计算得到栈长度为 K 的栈化帧间差分流,每幅图都包含了运动显著区域。单个栈化帧间差分流可表示为 $S=[s_1,s_2,\cdots,s_K]$。在有监督的分类中,可为动作视频数据集建立如下训练集:

$$X=\{(S_i,y_i)\}_{i=1}^{N}$$

其中,S_i 为训练集中的一个栈化帧间差分流样本,y_i 为该样本对应的视频动作标签,N 为训练样本数。

由于光照强度、运动幅度和图像背景等因素的影响,经过差分算法得到的差分图像的像素值往往很小,目标轮廓不是特别明显,导致后续生成的栈化帧间差分图不能很好地表示视频中动作连续变化的情形,因此,有必要对差分图像进行增强处理,本文将式(1)的结果乘以一个大于 1.0 的系数,然后进行归一化操作。帧间差分可以检测到视频中的运动区域,对帧间差分图的进一步预处理包括:1)获取运动的显著性区域;2)以此区域为中心进行图像随机剪

切;3)图像作翻转和归一化操作。上述过程可以生成训练样本和测试样本。

2 CNN 的特征提取与分类

2.1 CNN 特征提取

本文参考双流方法[4]构建 CNN 模型,如图 1 所示,模型分空域流、时域流和融合评估 3 个部分。其中,空域流和时域流网络采用了相同的结构。CNN 模型由输入层、卷积层、最大池化层、全连接层和融合层构成[16],采用大卷积核和小卷积核相结合的方法。前两层 CONV_1 和 CONV_2 分别采用(7×7)和(5×5)卷积核,后面的CONV_3a 和CONV_3b 组成一组、CONV_4a 和 CONV_4b 组成一组,都采用 3×3 卷积核,对应有 4 个池化层。由于运动视频中相邻图像差分计算后的图像像素跳跃性较大,因此本文分别在 CONV_1、CONV_3a 和 CONV_4a 之后增加了批量归一化(Batch Normalization,BN)处理[17],主要用于中间特征的归一化,以避免数据训练过程中出现漂移现象。全连接层 FC_5 中加入了正则化器 L2,主要用于神经网络在训练过程中对各层的参数或激活情况进行惩罚。

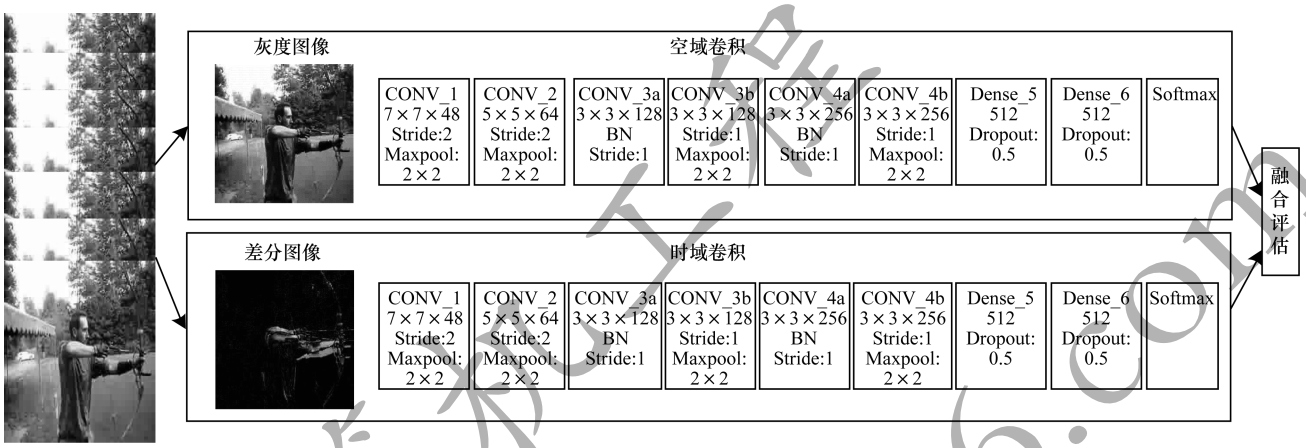


图 1 CNN 模型结构

在图 1 中,空域卷积网络的输入为单通道灰度图像,时域卷积网络的输入为栈化帧间差分图像,2 个网络的输出特征在融合评估模块中被合并。时域卷积网络输入的栈化差分图像包括一系列连续的帧间差分图像,栈长度对应栈中差分图像的数目。本文的时域和空域卷积网络模型分别进行训练,输入的图像尺度为(112×112),训练过程的最优化方法采用 SGD(Stochastic Gradient Descent),初始学习率设为 0.001。

2.2 特征表征与动作分类

本文方法对卷积特征进行可视化表征,此处主要针对时域卷积网络,采用 t-SNE[18]实现栈化帧间差分流特征的可视化。t-SNE 是对 SNE 技术的改进,其可以根据高维空间数据的内在结构将其在低维空间中进行显示。t-SNE 解决了数据在降维过程中的拥挤问题,在低维空间中,同类样本的聚集性更好,不同类样本具有分离性。

本文方法的动作分类采用图 1 所示的模型结构,在双流模型中,空域卷积网络和时域卷积网络的后面部分各包括 2 个全连接层和 1 个 Softmax 层,Softmax 层的评估结果输入融合评估模块,然后基于平均法得到最终输出。

3 实验结果与分析

3.1 数据集与测试方法

本文研究人机交互场景下的动作快速识别方法,因此,在选择数据集时考虑了非 GPU 场景下的机器处理能力。选取如下 3 个数据集对模型性能进行评估:

1)KTH 数据集[19]:包括“走”“慢跑”“跑”“拳击”“挥手”“拍掌”等常见的动作类别,共有 600 段视频剪辑,每个视频分成 4 个片段,拆分后得到2 400 个视频片段。

2)UCFsport 数据集[20]:包括“跳水”“打高尔夫”“踢”“吊装”“骑马”“跑步”“滑板”“荡秋千”“摆动双侧”“步行”等 13 个常见动作,但仅包含150 段视频剪辑。

3)UCF-101 数据集[21]:包括 13 320 个视频剪辑和 101 个动作类别,该数据集是从 YouTube 采集的真实场景。本文在 101 个动作类别中随机选择11 类和 25 类的全部动作视频用于研究,分别简称为 UCF11 和 UCF25 数据集。

在实验过程中,KTH 和 UCFsport 2 个数据集的训练和识别在 CPU 级别电脑下进行,采用的机器配置为:Intel i7 CPU,内存 8 GB,操作系统 Windows10。UCF11 和 UCF25 数据集采用的机器配置为:处理器 Intel i7,16 GB 内存,单显卡 GTX1080Ti,操作系统 Ubuntu 16.04。

3.2 时域卷积网络的测试结果

3.2.1 时域卷积网络训练

时域卷积网络训练结果如图 2、图 3 所示,每个数据集的训练都至少包括 2 个阶段,即先在较少数据样本条件下训练得到初始模型,然后增加样本量并调整模型训练参数继续进行训练,直到损失函数不再下降。

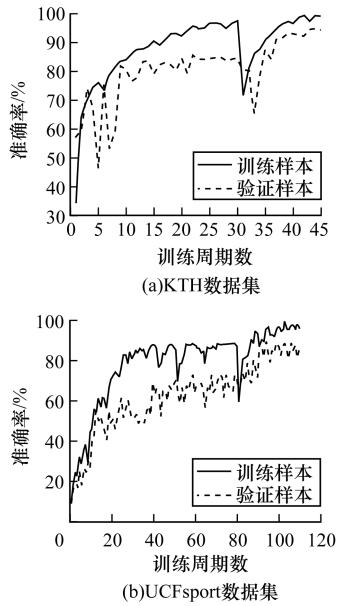


图 2　KTH 和 UCFsport 数据集的时域卷积网络训练结果

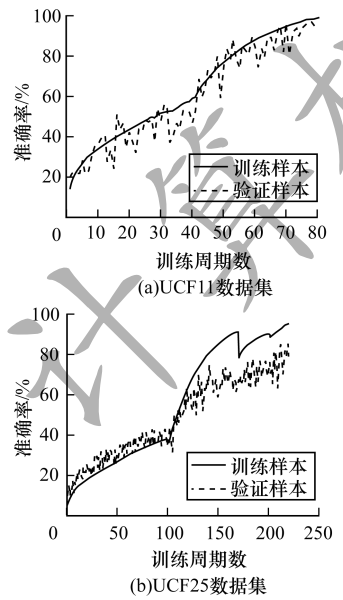


图 3　UCF11 和 UCF25 数据集的时域卷积网络训练结果

图 2 中所采用的 KTH 和 UCFsport 数据集较小,视频图像背景较简单,因此,直接在 Intel i7 CPU 环境下训练。KTH 每个视频包括 4 个片段,每个片段至少可提取一个栈化帧间差分流,图像转换成灰度图像后缩放到(60×80)大小,然后输入到 CNN 模型中进行训练,最后识别准确率约为 95%。UCFsport 数据集由于视频数目较少,为便于机器学习,本文在每个动作中提取了多组栈化帧间差分流样本以用于机器学习,训练结果如图 2(b)所示。可以看出,由于样本数目较少,训练曲线震荡较明显,总共训练次数为 110,分 2 轮进行,大约在训练 80 次后增加了训练数据,最后平均识别准确率约为 85%。此外,由于 KTH 和 UCFsport 数据集的动作类别、样本数量和图像复杂度较低,本文采用了更加简化的 CNN 模型,如 KTH 的卷积层只有 3 层,对应的卷积核数目分别为 16、32 和 64,即 16-32-64 结构;UCFsport 采用了 32-64-128 结构,均为(3×3)卷积核。

图 3 是在 Ubuntu 16.04 + Inel i7 + GPU 环境下的训练曲线,其中,图 3(a)的 UCF11 数据集包括 11 个动作类别,训练周期为 80 次,前 40 次训练结束后验证数据收敛速度降低,准确率在 43% 左右。随后,为降低过拟合现象,增加训练和验证数据样本量,将栈化差分流数据从每视频 2 组提高到 3 组以上,最终训练模型的平均准确率达到 92%。UCF25 数据集包括 25 个动作类别,其训练难度明显增加。在训练周期达 100 次时,增大了视频中栈化帧间差分流的样本数目,此时验证数据准确率达到 65% 左右,但训练数据准确率和验证数据准确率开始出现背离现象。继续增大训练次数和测试样本量,在训练达 220 次后,测试准确率稳定在 80%~86% 之间。

3.2.2　CNN 特征的可视化表征

为进一步了解训练后模型的分类效果,采用 t-SNE 对模型提取的卷积特征进行可视化表征。t-SNE 可以将高维数据降维、聚类并投影到二维空间进行显示[17,22]。以每个数据集的验证数据为对象,首先用上述训练后的模型来提取所有验证数据样本的卷积特征,然后用 t-SNE 对这些特征进行降维和可视化,其中,不同特征用不同颜色标识。

由于栈化差分流中包括连续的多幅图像,为便于 t-SNE 降维和聚类处理,需将它们转化成一维数据。对于原始栈化差分流,如 KTH 的 CNN 输入图像为(60×80)的灰度差分图像,在栈长度为 8 的情况下则转化为(1×38 400)的一维向量,维度非常高,这时数据分类的难度也非常大。未经时域卷积网络处理的 UCF11 的栈化差分流样本,用 t-SNE 处理后的形状大致如图 4(a)所示,其中,每个点代表一个样本。可以看出,本文所采用的 4 个数据集的栈化差分流在时域卷积网络处理前的 t-SNE 图像和图 4(a)基本类似。对于 UCF11 和 UCF25 数据集,采用图 1 所示的时域卷积网络进行处理。栈化差分流图像包括多个尺度为(112×112)的灰度差分图像序列,经多层卷积-池化处理后得到长度为 512 的向量表示(1×512)。经 t-SNE 处理后得到的特征分布如图 4(b)、图 4(c)所示。可以看出,UCF11 与 UCF25 数据集的特征分类效果明显改善,说明视频帧间差分图像可以用于动作分类。

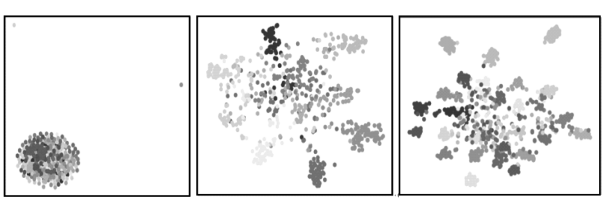
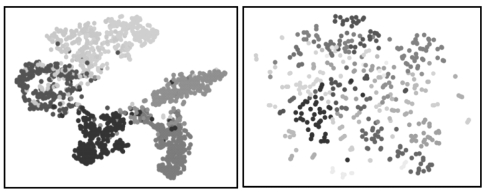


图 4　UCF11 与 UCF25 卷积特征的 t-SNE 可视化效果

类似地,利用训练后的 CNN 模型对 KTH 和 UCFsport 的验证样本集进行 t-SNE 聚类和可视化表征,结果如图 5所示,其中,UCFsport 的可视化分类效果差于 KTH,这是由于 UCFsport 样本数较少,模型训练得不够充分,但其准确率仍超过了 80% 。可以看出,对于小样本数据集,在 CPU 级条件下进行训练,也能达到较好的动作分类效果。



(a)KTH卷积特征 (b)UCFsport卷积特征

图 5 KTH 与 UCFsport 卷积特征的 t-SNE 可视化效果

3.3 帧间差分分析

3.3.1 图像预处理

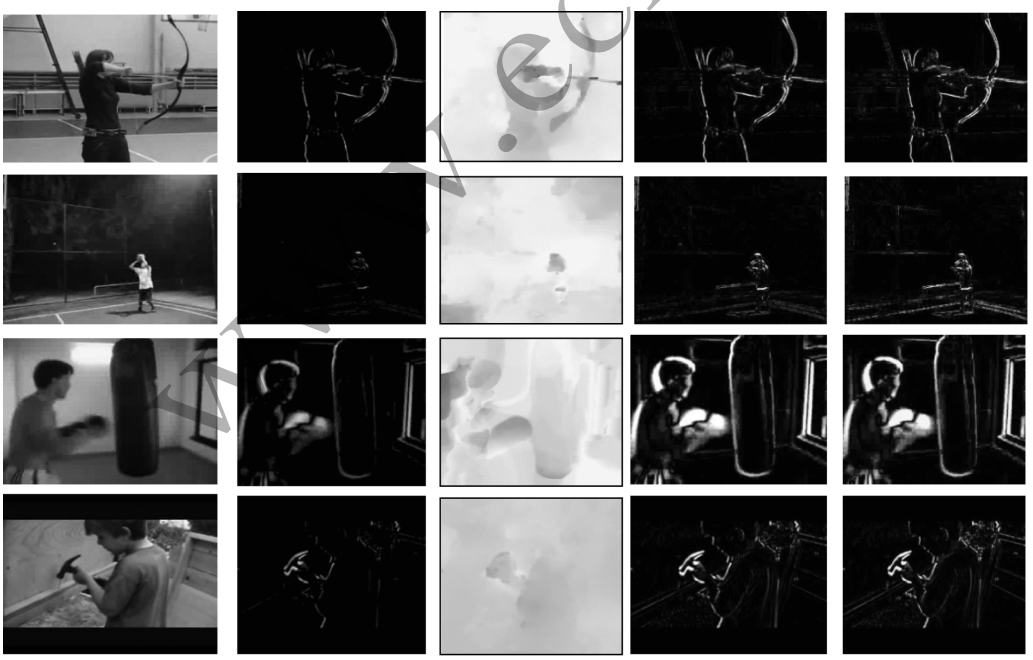
通过对前后 2 幅相邻图逐像素计算可得稠密光流图,其包含了目标运动信息。如图 6 所示,对相邻图像同一位置的像素值进行差分可得帧间差分图,当动作缓慢时,相邻 2 幅图之间变化不大,对应的差分图很可能接近空白,如图 6(b)所示。因此,为较好地获得运动区域特征,可采用相隔一帧或多帧的方式进行差分运算。由于传统的连续帧间差分图像存在颜色较浅、轮廓不连续和目标内部空洞等问题,因此本文在提取差分图时采用隔帧的差分方式,以获得具有更多运动区域信息的差分图,克服由于视频内一些缓慢动作造成的运动目标信息提取失败或运动目标信息提取不足的问题。图6(c)所示为隔一帧的2幅图像在帧间差分运算后的差分图,可以看出,其运动目标信息有明显改善,但是这种隔帧区间不宜太长,原因是会导致图 6(d)所示的“重影”现象,影响后续特征提取效果。

(a)原视频帧 (b)相邻帧差分图

(c)隔1帧差分图 (d)隔2帧差分图

图 6 相邻差分图和隔帧差分图的比较

为强化差分图中的运动目标信息,本文对差分图的亮度进行了像素增强处理,并将它们归一化到 0~255 之间,最后得到的部分灰度差分图、RGB 差分图以及相应的光流图如图 7 所示。其中,第 1 列为原始视频帧,第 2 列为 2 帧直接差分后的效果图,第 3 列为光流图,第 4 列为经过增强处理的灰度帧间差分图,第 5 列为经过增强处理的 RGB 帧间差分图。本文主要关注灰度差分图在视频动作识别中的应用,对比图 7 的第 2 列和第 4 列可以看出,经过增强处理后,灰度差分图可呈现清晰的运动区域轮廓。



(a)原始视频帧 (b)灰度帧差分图 (c)光流图 (d)灰度差分图增强 (e)RGB帧间差分图增强

图 7 部分运动图像对应的光流图和帧间差分图

3.3.2　差分图对动作识别的影响

图 7 包括部分运动图像的灰度和 RGB 帧间差分图。相比灰度差分图,RGB 差分图能更全面地反映像素亮度的变化情况,在 CNN 模型训练中,一般能得到更高的准确率。但是灰度图像也能较好地反映动作的轮廓变换,这种动作轮廓对于不同颜色的同类动作视频可能更具有共性,此外,由于灰度图只有一个通道,能够使 CNN 模型的训练过程更加快捷,因此,本文选择灰度差分图像进行实验。

比较图 7 中的帧间差分图和稠密光流图可以看出,对于运动视频中相邻的两帧而言,它们的稠密光流图和帧间差分图中运动区域和形状基本相同。帧间差分图像往往只包括运动目标的轮廓信息,内部可能空洞;稠密光流图内部和边缘一致,其包括了运动目标边缘和内部的运动信息,这种图像更加稳定。此外,帧间差分图比光流图更简洁,计算复杂度和时间复杂度较低,但其不足之处在于像素值较低,图像往往不清晰,描述运动目标之间的相似性时可能不稳定。

帧间差分图像去除了大部分背景和目标内部信息,仅包含运动目标边缘以及小位移信息,这种图像尽管和动作的相关性很大,但缺乏上下文信息辅助,将它们应用于复杂动作识别是一个具有挑战性的任务。文献[9]将差分图像作为光流和空间流模式的补充,但其未涉及样本生成和模型训练的细节。此外,光流通常被认为代表着视频的运动或时序信息,包含水平和垂直方向的运动信息,图 7 中的深浅颜色表征了不同的运动速度。

从图 2、图 3 可以看出,利用栈化帧间差分流训练时域卷积网络模型到一定程度后,验证数据的准确率不再提高,出现了过拟合现象,这是由训练样本数量不足导致的,在动作训练中较常出现,可以采用深度残差网或基于大型 Activity 数据集的预训练模型来解决该问题[8]。本文由于数据集较小,动作类别最多为 25 个,在帧间差分样本的训练过程中同样出现了过拟合问题,可以通过调整学习率或增加训练样本的采集密度,然后继续训练模型以解决该问题。

3.3.3　实时性分析

稠密光流需计算图像上所有点的偏移量,从而形成一个稠密的光流场,但是该过程中的计算量非常大。在动作训练中,通常要预先准备好视频帧对应的光流图数据集,然后才能实施后续的机器学习过程,这给数据量庞大的动作识别带来了不便。帧间差分计算在这方面有明显优势,本文计算两视频帧之间的光流或帧差图像,在 CPU 级电脑上的处理时间如表 1 所示。其中,光流采用 TV-L1[23] 和 Farneback[24] 算法进行实现。由表 1 可以看出,相邻两帧光流图的计算时间远大于帧间差分计算时间。

表 1　帧获取时间对比　　ms

项目	计算时间
灰度图帧间差分	1.7
RGB 图帧间差分	3.4
光流计算(TV-L1)	3 735
光流计算(Farneback)	3 262

由于帧间差分计算较快,因此可以直接从视频中生成训练样本并进行 CNN 模型训练,如本文中采用的 KTH 和 UCFsport 数据集,从而实现端到端的机器学习和分类预测。

3.4　融合输出

在训练好空域卷积模型和时域卷积模型后,可以对输入的视频进行卷积特征提取,然后实现相应的视频动作识别。视频帧数据首先被分成两路:一路输入到空域卷积网络,经卷积和全连接层后向融合评估模块输出 Softmax 评估结果;另一路经过帧间差分后生成栈化差分流,再被输入到时域卷积网络,经卷积和全连接层后向融合评估模块输出 Softmax 评估结果。最后,2 个评估结果相结合以完成动作识别。图 8 所示为融合评估模块结构,该模块接收时域和空域两路输入,包括时间片分配、数据扁平化和结果合并等步骤,其中,结果合并时采用平均求解的方式。

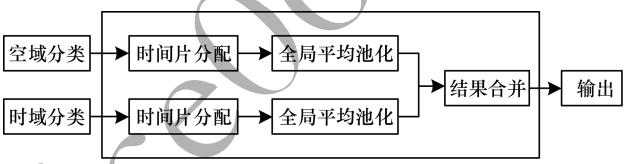


图 8　融合评估模块结构

在采用时域单网和 UCF101 数据集的环境下,文献[2]中 iDT + Fisher Vector 方法的准确率为 0.879,文献[4]方法的准确率达到 0.837。目前基于光流栈的多模检测方法在 UCF101 数据集上的动作识别准确率已超过 0.900[9]。如表 2 所示,本文时域卷积网络在 UCF11 数据集上的准确率达到了 0.920,在 UCF25 数据集上的准确率为 0.830,融合评估结果的准确率也在 0.850 以上。本文方法目前检测的动作类别不超过 25 个,但考虑到帧间差分计算的高效率特征,在室内人机交互环境下将该方法应用于数量有限的动作实时检测时,具有较高的可行性。

表 2　本文模型准确率结果

数据集	空域模型准确率	时域模型准确率	融合评估结果准确率
UCF11	0.850	0.920	0.930
UCF25	0.820	0.830	0.853

统计本文模型在视频分类过程中的计算速度,方法为:在 CPU 环境下,从测试集中选取一段视频,提取多个视频帧和多组连续视频帧,连续视频帧随后被转换为栈化差分流,统计上述过程的处理时间

并求出平均值,测试结果如表 3 所示,其中选取了 3 个动作的多个视频,每个视频随机取 4 个静态帧和 4 个帧间差分栈,得出单个视频经过本文算法处理时的平均耗时。从表 3 可以看出,本文模型基本达到了实时分类的效果。

表 3 本文模型对各视频的分类时间统计

视频名称	视频数	静态帧数	差分栈数	平均耗时/s
Archery	38	4	4	0.259
BabyCrawling	32	4	4	0.258
BalanceBeam	33	4	4	0.232

4 结束语

与光流图相比,传统的帧间差分图像存在颜色较浅、轮廓不连续和目标内部空洞等问题,在 CNN 训练中易出现过拟合现象。本文将帧间差分序列应用于视频动作识别中,利用帧间差分检测视频帧中的运动目标,并以该目标为中心进行相应的图像剪切、增强和标准化处理。在数据样本的制作过程中通过适当隔帧差分来扩大样本的时域跨度,从而使差分图包含更完整的动作特征信息。本文识别模型采用改进的双流架构,在融合评估时将 Softmax 层和特征层相结合,并进行相应的动作类别评估。实验结果表明,帧间差分计算速度较快、对硬件配置要求较低,在动作类别较少的情况下可以在 CPU 级电脑中完成快速动作识别,在进行帧间差分后,本文时域卷积模型在 KTH 和 UCFsport 数据集上分别取得了 95% 和 85% 的准确率,结合帧间差分的融合评估方案在 UCF11、UCF25 数据集上的识别准确率分别达 93.0% 、85.3% 。

由于本文实验只采用单通道灰度图形,KTH 的图像尺度仅为(60 ×80),UCF11 和 UCF25 的图像尺度仅为(112 ×112),因此下一步将采用 RGB 三通道和更高的图像尺度,结合帧间差分图与快速光流计算,以提高动作识别的准确率。

参考文献

[1] WANG Heng, ALEXANDER K, SCHMID C, et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International Journal of Computer Vision,2013,103(1):60-79.

[2] PENG Xiaojiang,WANG Limin,WANG Xingxing,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.

[3] WANG Heng,SCHMID C. Action recognition with improved trajectories[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press,2013:3551-3558.

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

[5] 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,2014,39(4):677-691.

[6] DU T,BOURDEV L,FERGUS R,et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Computer Society,2015: 4489-4497.

[7] QIU Zhaofan, YAO Ting, MEI Tao. Learning spatio-temporal representation with pseudo-3D residual networks[C]// Proceedings of 2017 IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017:5534-5542.

[8] KENSHO H, HIROKATSU K, YUTAKA S. Towards good practice for action recognition with spatiotemporal 3D convolutions[C]//Proceedings of the 24th International Conference on Pattern Recognition. Washington D. C., USA: IEEE Press,2018:2516-2521.

[9] WANG Limin, XIONG Yuanjun, WANG Zhe, et al. Temporal segment networks: towards good practices for deep action recognition[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2016:20-36.

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

[11] WANG Jingzhong, HU Kai. Human angle fitting based on BP neural network[J]. Computer Systems and Applications,2019,28(8):235-240. (in Chinese)
王景中,胡凯. 基于 BP 回归神经网络的人体角度拟合研究[J]. 计算机系统应用,2019,28(8):235-240.

[12] ZHANG Rui, LI Qishen, CHU Jun. Human action recognition algorithm based on 3D convolution neural network[J]. Computer Engineering,2019,45(1):259-263. (in Chinese)
张瑞,李其申,储珺. 基于 3D 卷积神经网络的人体动作识别算法[J]. 计算机工程,2019,45(1):259-263.

[13] LI Wei. Analysis of character motion based on single role video[D]. Jinan: Shandong University,2018. (in Chinese)
李伟. 基于单角色视频的人物运动分析[D]. 济南:山东大学,2018.

[14] ZIVKOVIC Z. Improved adaptive Gaussian mixture model for background subtraction[C]//Proceedings of International Conference on Pattern Recognition. Washington D. C., USA: IEEE Press,2004:28-31.

[15] QU Jingjing, XIN Yunhong. Combined continuous frame difference with background difference method for moving object detection[J]. Acta Photonica Sinica, 2014,43(7):219-226. (in Chinese)
屈晶晶,辛云宏. 连续帧间差分与背景差分相融合的运动目标检测方法[J]. 光子学报,2014,43(7):219-226.

(下转第 293 页)

(上接第 280 页)

[16] ZHENG Changyan, MEI Wei, WANG Gang. Deep convolutional neural networks for the image recognition of "S-Maneuver" target[J]. Fire Control and Command Control, 2017, 42(4): 66-70. (in Chinese)
郑昌艳,梅卫,王刚. 基于深度卷积神经网络的蛇形机动航迹图像识别[J]. 火力与指挥控制,2017,42(4):66-70.

[17] SERGEY I, CHRISTIAN S. Batch normalization: accelerating deep network training by reducing internal covariate shift[EB/OL]. [2018-12-20]. https://arxiv.org/pdf/1502.031 67.pdf.

[18] MAATEN L V D, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(3): 2579-2605.

[19] SCHULDT C, LAPTEV I, CAPUTO B. Recognizing human actions: a local SVM approach[C]//Proceedings of International Conference on Pattern Recognition. Washington D. C., USA: IEEE Press, 2004: 32-36.

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

[21] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild[EB/OL]. [2018-12-20]. http://export.arxiv.org/pdf/1212.0402.

[22] ZHANG Yahong, LI Yujian. Fisher information metric based on stochastic neighbor embedding[J]. Journal of Beijing University of Technology, 2016, 42(6): 862-869. (in Chinese)
张亚红,李玉鑑. 基于费希尔信息度量的随机近邻嵌入算法[J]. 北京工业大学学报,2016,42(6):862-869.

[23] ZHANG Congxuan, CHEN Zhen, WANG Mingrun, et al. Non-local TV-L1 optical flow estimation using the weighted neighboring triangle filtering[J]. Journal of Image and Graphics, 2017, 22(8): 1056-1067. (in Chinese)
张聪炫,陈震,汪明润,等. 非局部加权邻域三角滤波 TV-L1 光流估计[J]. 中国图象图形学报,2017,22(8):1056-1067.

[24] GUNNAR F. Two-frame motion estimation based on polynomial expansion[C]//Proceedings of the 13th Scandinavian Conference on Image Analysis. Berlin, Germany: Springer, 2003: 363-370.

编辑　吴云芳