JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
08
VOL.27

ISSN1006-8961
CN11-3758/TB

工业低纹理目标检测 *P2418*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第8期（总第316期）
2022年8月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

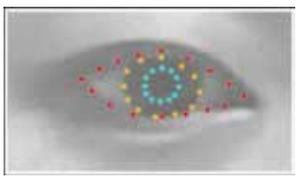
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

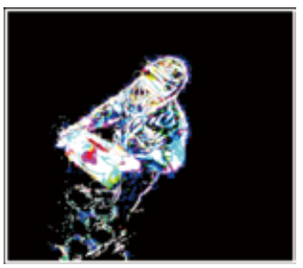
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2022年8月 第27卷 第8期（总第316期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

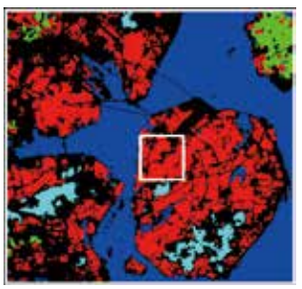
目次



自然光普通摄像头的眼部分割及特征点定位数据集ESLD(第2329页)



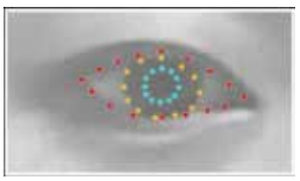
双视图三维卷积网络的工业装箱行为识别(第2368页)



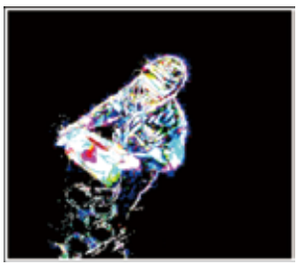
多源特征自适应融合网络的高分遥感影像语义分割(第2516页)

Volume 27, Number 8
Published August 16, 2022

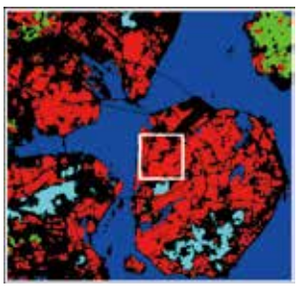
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



ESLD: eyes segment and landmark detection in the wild (P2329)



Dual-view 3D ConvNets based industrial packing action recognition(P2368)



Multi-source features adaptation fusion network for semantic segmentation in high-resolution remote sensing images(P2516)

E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP311　　**文献标识码**: A　　**文章编号**: 1006-8961(2022)08-2368-12

论文引用格式: Hu H Y, Pan J and Li Z J. 2022. Multi-view 3D ConvNets based industrial packing action recognition method. Journal of Image and Graphics,27(08):2368-2379(胡海洋,潘健,李忠金. 2022. 双视图三维卷积网络的工业装箱行为识别. 中国图象图形学报,27(08):2368-2379)
[DOI:10.11834/jig.210064]

双视图三维卷积网络的工业装箱行为识别

胡海洋*,潘健,李忠金

杭州电子科技大学计算机学院,杭州　310018

摘　要: **目的**　在自动化、智能化的现代生产制造过程中,行为识别技术扮演着越来越重要的角色,但实际生产制造环境的复杂性,使其成为一项具有挑战性的任务。目前,基于3D卷积网络结合光流的方法在行为识别方面表现出良好的性能,但还是不能很好地解决人体被遮挡的问题,而且光流的计算成本很高,无法在实时场景中应用。针对实际工业装箱场景中存在的人体被遮挡问题和光流计算成本问题,本文提出一种结合双视图3D卷积网络的装箱行为识别方法。**方法**　首先,通过使用堆叠的差分图像(residual frames,RF)作为模型的输入来更好地提取运动特征,替代实时场景中无法使用的光流。原始RGB图像和差分图像分别输入到两个并行的3D ResNeXt101中。其次,采用双视图结构来解决人体被遮挡的问题,将3D ResNeXt101优化为双视图模型,使用一个可学习权重的双视图池化层对不同角度的视图做特征融合,然后使用该双视图3D ResNeXt101模型进行行为识别。最后,为进一步提高检测结果的真负率(true negative rate,TNR),本文在模型中加入降噪自编码器和two-class支持向量机(support vector machine,SVM)。**结果**　本文在实际生产环境下装箱场景进行了实验,采用准确率和真负率两个指标进行评估,得到的装箱行为识别准确率为94.2%、真负率为98.9%。同时在公共数据集UCF(University of Central Florida)101上进行了评估,以准确率为评估指标,得到的装箱行为识别准确率为97.9%。进一步验证了本文方法的有效性和准确性。**结论**　本文提出的人体行为识别方法能够有效利用多个视图中的人体行为信息,结合传统模型和深度学习模型,显著提高了行为识别准确率和真负率。

关键词:行为识别;双视图;三维卷积神经网络;降噪自编码器;支持向量机(SVM)

Dual-view 3D ConvNets based industrial packing action recognition

Hu Haiyang*, Pan Jian, Li Zhongjin

School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: **Objective** The action recognition technology is proactive in computer vision contexts, such as intelligent video surveillance, human-computer interaction, virtual reality, and medical image analysis. It plays an important role in the automated and intelligent modern manufacturing process, but the complexity of the actual manufacturing environment has still been challenging. The research direction is attributed to the deep neural networks largely, especially the 3D convolutional networks, which mainly use 3D convolution to capture temporal information. The 3D convolutional networks can extract the spatio-temporal features of videos better with the added temporal dimension compared to 2D convolutional net-

收稿日期:2021-02-18;**修回日期**:2021-04-19;**预印本日期**:2021-04-26

*通信作者:胡海洋　huhaiyang@hdu.edu.cn

基金项目:国家自然科学基金项目(61572162,61802095);浙江省重点研发计划资助(2018C01012);浙江省自然科学基金项目(LQ17F020003)

Supported by: National Natural Science Foundation of China(61572162,61802095)

works. At present, it shows good performance in action recognition through the optical flow melting in 3D convolutional network, but it still cannot solve the problem of human body being occluded, and the computational cost of optical flow is complicated and cannot be applied in real-time scenes. The product qualification rate is required to be satisfied in the context of action recognition application in production scenes. It is necessary to rank out the unqualified products as much as possible while ensuring high accuracy and high true negative rate (TNR) of detection results. It is challenged to optimize the true negative rate among the existing action recognition methods. Our analysis facilitates a packing action recognition method based on dual-view 3D convolutional network. **Method** First, we extract motion features better through stacked residual frames as inputs, replacing optical flow that is not available in the real-time scene. The original RGB images and the residual frames are input to two parallel 3D ResNeXt101, and a concatenation layer is used to concatenate the features extracted in the last convolution layer of the two 3D ResNext101. Next, we adopts a dual-view structure to resolve the issue of human body being occluded, optimizes 3D ResNeXt101 into a dual-view model, builds up a learnable dual-view pooling layer for multifaceted feature fusion of views, and then uses this dual-view 3D ResNeXt101 model for action recognition. Finally, a noise-reducing self-encoder and two-class support vector machine (SVM) are added in our model to improve the true negative rate (TNR) of the detection results further. The dual-view pooling derived features are input to the noise-reducing self-encoder in the model, and the features are optimized and downscaled by the trained noise-reducing self-encoder, and then the two-class SVM model is used for secondary recognition. **Result** We conducted experiments in a packing scenario and evaluated using two metrics like accuracy rate and true-negative rate. The accuracy of our packing action recognition model is 94.2%, and the true negative rate is 98.9%, which optimizes current action recognition methods. Our accuracy is increased from 91.1% to 95.8% via the dual-view structure. The accuracy of the model is increased from 88.2% to 95.8% based on the residual frames module. If the residual frames module is altered by optical flow module, the accuracy rate is 96.2%, which is equivalent to the model using the residual frames module. The accuracy is only 91.5% that the unique two-class SVM structure added to the model without the denoising autoencoder. Thanks to the optimization and dimensionality reduction of the feature vectors by the denoising autoencoder, the accuracy reaches 94.2% via the combination of the denoising autoencoder and the two-class SVM both, the highest true negative rate of 98.9% obtained. After adding denoising autoencoder and two-class SVM to the model, the true negative rate of the model increased from 95.7% to 98.9%, while the accuracy rate decreased by 1.6%. Our demonstrated result is evaluated in the public dataset UCF (University of Central Florida) 101. Our single-view model obtained an accuracy of 97.1%, which achieved the second highest accuracy among all compared methods, second only to 3D ResNeXt101's 98.0%. **Conclusion** We use a dual-view 3D ResNeXt101 model for effective packing action recognition. To obtain richer features from RGB images and differential images, two parallel 3D ResNeXt101 are used to learn spatio-temporal features and a dual-view feature fusion is accomplished using a learnable view pooling layer. In addition, a stacked denoising autoencoder is trained to optimize and downscale the features extracted in terms of the dual-view 3D ResNeXt101 model. To improve the true negative rate, a two-class SVM model is used for secondary detection. Our method can recognize the boxing action of the packing workers accurately and realize the high true negative rate (TNR) of the recognition results.

Key words: action recognition; dual-view; 3D convolutional neural network; denoising autoencoder; support vector machine (SVM)

0　引　言

人体行为识别由于其广泛的应用,一直是计算机视觉中的热门研究方向,例如智能视频监控、人机交互、虚拟现实和医疗影像分析(Cao 等,2017;Wang 等,2015;Kuanar 等,2018,2019)。该领域的突破很大程度上归功于深度神经网络的出现,尤其是3D 卷积网络的提出(Ji 等,2013)。3D 卷积网络是2D 卷积网络的扩展,主要使用3D 卷积来捕获时间信息。相较于2D 卷积网络,3D 卷积网络由于增加了时间维度,可以更好地提取视频的时空特征。Hara 等人(2018)将 C3D (convolutional 3D network)卷积网络与残差网络(residual neural network, ResNet)相结合,并在 Kinetics 数据集中进行预训练,相比 C3D 卷积网络,3D ResNet 在 UCF101(University

of Central Florida-101)和 HMDB-51(Human Motion DataBase-51)中取得了很大的进步,同时运行速度也快了 2 倍。Simonyan 和 Zisserman(2014)采用 RGB 图像和光流作为两个独立的 2D 卷积网络流的输入,基于 RGB 图像的空间流从静止图像捕获空间特征以识别人的动作,而基于光流的时间流用于识别密集光流的运动,光流作为进一步表达运动信息并改善性能,是有必要的,如多种模型(Simonyan 和 Zisserman,2014;Liu 和 Hu,2019;Feichtenhofer 等,2016;Carreira 和 Zisserman,2017)都通过结合光流取得了较好的效果。但是由于提取光流的过程非常耗时,在现有硬件条件下无法在实时场景应用,例如图 1 中所展示的需要实时检测的实际生产场景。Tao 等人(2020)使用差分图像作为网络输入,用于获取额外的运动特征,并且取得了较好的结果。基于以上两个动机,本文引入差分图像,更好地获取运动特征。

与传统的行为识别研究工作相比,工厂环境下的行为识别有其复杂性和特殊性:生产制造环境中背景混乱、光线变化频繁以及人体被遮挡问题严重,给研究工作带来了挑战。现有的行为识别方法并不能很好地解决人体被遮挡的问题,如图 1 中装箱工人的右手动作被柱子和箱子遮挡。

行为识别在生产场景中应用时,由于对产品合格率的要求,需要尽可能排查出不合格的产品,同时保证检测结果的高准确率和高真负率(true negative rate,TNR)。现有的行为识别方法中,还没有较好的优化真负率的方法。

针对在工业装箱场景中存在的上述问题,本文提出了一种基于双视图 3D ResNeXt101 的行为识别方法。首先,本文引入堆叠的差分图像作为模型的输入来更好地提取运动特征,替代实时场景中无法使用的光流。原始 RGB 图像和差分图像分别输入到两个并行的 3D ResNeXt101 中,基于差分图像的 3D ResNeXt101 模块用于更好地提取运动特征,基于 3D ResNeXt101 模块用于提取运动特征和弥补差分图像中缺失的外观特征。其次,采用双视图结构来解决人体被遮挡的问题,将 3D ResNeXt101 优化为双视图模型,使用一个可学习权重的双视图池化层对不同角度的视图特征进行融合,利用该双视图 3D ResNeXt101 模型进行行为识别。最后,为进一步提高检测结果的真负率,在模型中加入降噪自编码器和 two-class SVM(support vector machine)模型。

图 1 工厂中的装箱环境

Fig. 1 Packing environment in the factory

1 相关工作

近些年,行为识别技术取得了非常大的进步。行为识别技术的主要方法可以分为两类,基于视频手工特征的方法和基于深度学习的方法。

基于时空兴趣点(space-time interest points,STIP)的方法作为手工特征的代表性方法,广泛应用于行为识别。该方法从视频中提取运动变化的关键区域来识别动作。STIP 中的时空兴趣点通常是指时空维度上变化最大的位置(Das Dawn 和 Shaikh,2016)。3D-Harris 时空兴趣点方法(Laptev,2005)是 STIP 最具代表意义的方法,主要思想是将局部特征检测技术从 2 维空间兴趣点扩展到 3 维时空兴趣点,然后计算特征描述符,统计像素直方图形成描述行为的特征向量。该方法不需要做运动物体分割,在背景复杂的场景也有较好的效果,但是在人体被遮挡和光线变化大时效果较差。基于运动轨迹提取的方法也同样受到许多研究者的关注,Wang 和 Schmid(2013)提出了改进的稠密轨迹方法(improved dense trajectories,IDT),综合 HOG(histogram of gradient)、HOF(histogram of flow)和 MBH(motion boundary histograms)的特征,对轨迹全局施加平滑约束,获得了很好的鲁棒性。此类方法的主要优点是不需要将运动物体分割,因此在光照变化和复杂背景下依然有较好的效果,但是时空兴趣点很容易受相机视图变化的影响,在视角变化和遮挡下效果不佳。使用 IDT 是基于手工特征中效果最好、应对场景最丰富的算法。但由于计算复杂度较高,该方法速度较慢。

与手动制作的动作特征不同,深度学习方法从图像中自动学习特征方面表现较好,目前许多研究人员已尝试使用深度学习方法从 RGB 图像、光流、差分图像和人体骨骼数据中提取动作特征(Simonyan 和 Zisserman,2014;Tao 等,2020;Wu 等,2018;Wu 等,2021),深度学习可以从单模数据或多模融合数据中学习人体行为特征。Simonyan 和 Zisserman(2014)采用 RGB 图像和光流作为两个独立的 2D 卷积网络流的输入,空间流从静止图像捕获空间特征以识别人的动作,而时间流用于识别密集光流的运动。该体系结构能够从每帧以及帧之间的运动中提取互补特征。与 2D 卷积相比,3D 卷积网络更能捕捉复杂的运动信息(Ji 等,2013;Hara 等,2018)。Feichtenhofer 等人(2016)提出了双流 3D 卷积网络模型来学习时空特征,与在 softmax 层进行融合的传统方式不同,他们发现在最后的卷积层上融合时空流,在不损失性能的同时还可以节省大量参数。一些研究(Crasto 等,2019;Peng 等,2016;Hong 等,2019)也是通过探索不同融合策略以显著提高行为识别的性能。另一方面,Carreira 和 Zisserman(2017)以及 Ullah 等人(2019)通过改善网络结构来提升行为识别的性能。Carreira 和 Zisserman(2017)将 Inception-V1 的网络结构从 2 维扩展到3 维,并提出了用于行为识别的双流膨胀 3D 卷积网络,该方法在公共数据集 UCF101 和 HMDB-51 中都取得了非常好的结果。

在行为识别模型中加入自编码器成为提高准确性的一种有效方法(Ullah 等,2019;Budiman 等,2014),Ullah 等人(2019)使用经过训练的降噪自编器来有效获取原始视频帧中的动作信息,将 VGG-16(Visual Geometry Group layer 16)卷积网络模型中全连接层中的高维特征转换为低维,并学习相邻帧之间的信息变化。该方法通过在卷积网络中加入自编码器,较大提升了行为识别的准确率。

关于双视图行为识别,Su 等人(2015)提出了双视图卷积神经网络(dual-view convolutional neural network,MVCNN),该网络成功实现了 3D 模型和 CNN 的结合,结合后的模型经过训练可以独立地对多个 2D 投影图像进行分类。Zeng 等人(2019)针对双视图池化方法进行改进,提出一种基于学习的多池融合方法(learning-based multiple pooling fusion,LMPF),通过学习一组最佳权重以使多个不同视图的融合效果达到最佳。

2　本文方法

2.1　行为识别框架

图 2 展示了本文装箱行为识别模型。模型使用两个不同视角的 RGB 视频作为输入,两个 RGB 视频分别为不同视角的摄像头在实际装箱场景中获取的同一时间段的视频。然后使用差分法处理输入的 RGB 视频得到差分图像(residual frames,RF),作为模型的另一个输入。

单个视图的每一批输入数据(原始 RGB 视频)的形状为 $T \times H \times W \times C$,也就是 T 帧高度为 H、宽度为 W 且通道数 C 为 3 的 RGB 图像。每个 3D 卷积层同时在 3 个维度计算输入的数据。模型的每个视图都包含一个基于差分图像的 3D ResNeXt101 模块和一个基于 RGB 图像的 3D ResNeXt101 模块,本文将两个 3D ResNeXt101 的最后一个卷积层使用一个串接层(concatenation layer)进行融合,然后由一个可学习的双视图池化层将多个视图的特征融合,传递到一个全连接层,最后输出判别结果。在此基础上,本文引入降噪自编码器和 two-class SVM 来进一步排除不存在装箱动作的装箱过程,提高真负率。将模型经过双视图池化之后的特征作为降噪自编码器的输入,在经过降噪自编码器优化之后送入 two-class SVM 进行二次判别。通过两次的判别结果得到最终装箱行为识别的结果。

2.2　RGB 和差分图像结合的网络结构

大多数行为识别方法通过对 RGB 视频进行特征提取来获得行为识别结果(Ji 等,2013;Carreira 和 Zisserman,2017;Tran 等,2015),本文在使用 RGB 图像作为输入的同时引入一个差分图像模块来优化模型性能。

通过相邻帧相减来获取差分图像,保留两帧间的不同。由于差分图像的特殊性,在单个差分图像中,运动信息存在于空间轴。将差分图像用于 2D 卷积中,已经证明是有效的(Wu 等,2018)。但是较为复杂的动作持续时间相对更长,并不是单独一帧可以表示,需要连续的多帧差分图像。这时动作信息不单单存在于空间轴,还存在于时间轴上,相邻帧之间的联系也作为运行信息的一部分。

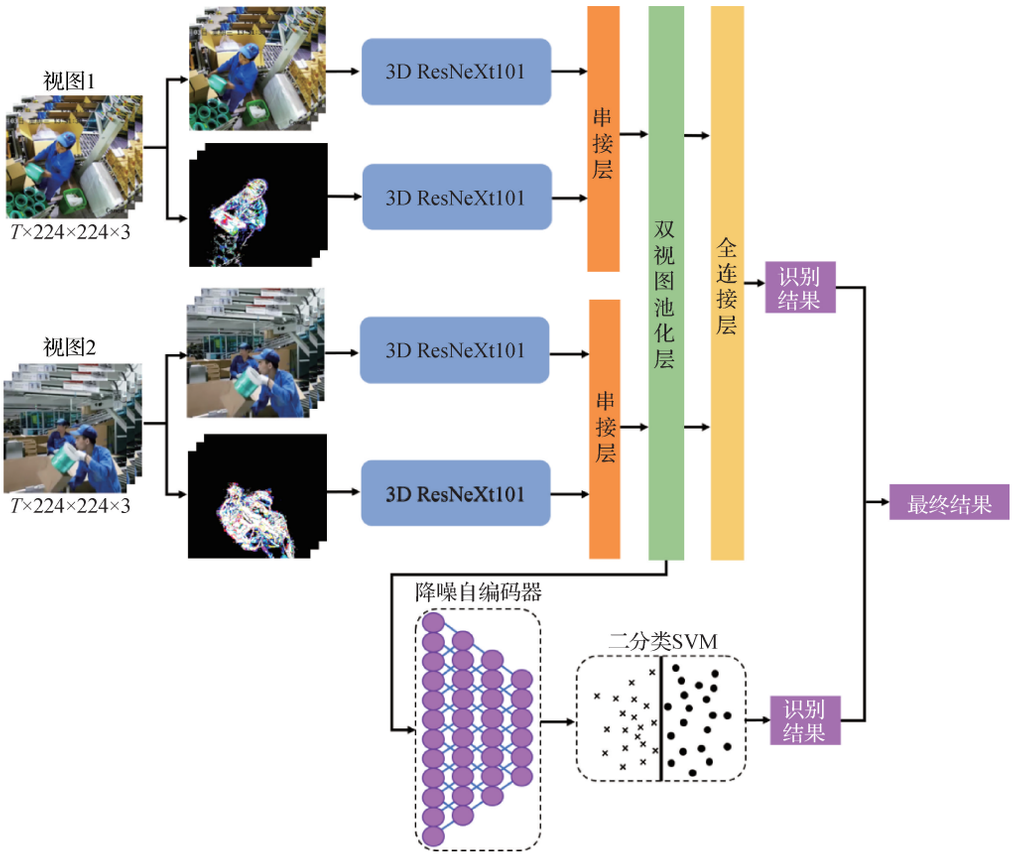


图 2　装箱行为识别模型

Fig. 2　Packing action model

同时，相较于 2D 卷积网络，3D 卷积网络由于增加了时间维度，可以更好地提取视频的时空特征，事实证明使用 3D 卷积网络具有更高的准确性（Tao 等，2020）。3D ResNet 是 3D 卷积网络与残差网络的结合，相比 C3D 具有更高的准确性和更快的运行速度。本文使用性能较好的 3D ResNeXt101 模型来处理输入的差分图像，从中获取运动特征。

本文使用 $\boldsymbol{F}_i$ 表示第 i 帧，$\boldsymbol{F}_{i-j}$ 表示第 $i-j$ 帧的连续堆叠帧，$\boldsymbol{RF}_{i-j}$ 表示第 $i-j$ 帧的连续差分图像。图 3 所示为使用差分法获取差分图像的示意图，获取差分图像的过程可以表述为

$$\boldsymbol{RF}_{i-j} = \left| \boldsymbol{F}_{i-j} - \boldsymbol{F}_{i+1-j+1} \right| \tag{1}$$

与光流的计算成本相比，差分图像的计算成本非常低，甚至可忽略不计，可以应用在实时场景中。而且差分图像中只包含运动物体的信息，处于静止状态的背景不会出现在差分图像上，可以剔除工厂复杂背景对识别效果的影响。但是差分图像只显示运动的部分，特别是在工人装箱过程中运动主要集中在双手，除工人双手以外的外观信息有时不会出现在差分图像上，只使用差分图像作为模型的输入，可以一定程度地区分出工人的投放动作，但同时外观信息的丢失也会降低结果的精度。Carreira 和 Zisserman（2017）使用 3D 卷积中在 Kinetics 数据集上的实验表明，同样的 3D 卷积模型在只使用 RGB 图像作为输入时效果相对只使用光流更好，RGB 图像仍然是 3D 卷积模型获取时空特征的主要途径。因此，本文使用另一个基于 RGB 图像的 3D ResNeXt101 模块来获取差分图像中缺失的外观特征和 RGB 图像中额外的运动特征。使用基于差分图像的 3D ResNeXt101 的目的是用来获取 RGB 图像中没有的运动特征，进而提升模型的准确性。在行为识别领域，多个网络框架的融合一般使用直接得到各个框架的得分，然后通过加权或者平均得分的方式得到最终结果的方法（Simonyan 和 Zisserman，2014），或者通过一个串接层（concatenation layer）将各个模块提取到的特征进行融合（Rastgoo 等，2020）的方法。为了更好地利用多视图之间互补的特性，便于多模块特征融合之后的双视图特

(a) 前一帧RGB图像　(b) 后一帧RGB图像　(c) 差分图像

图 3　差分图像获取过程

Fig. 3　The process of obtaining residual frame((a) front RGB;(b) latter RGB;(c) residual frame)

征融合,本文使用串接层将 RGB 和差分图像模块提取的特征融合。

本文使用两个并行的 3D ResNeXt101 模块,两个模块分别接收原始 RGB 图像和差分图像作为输入。在两个 3D ResNeXt101 模块的最后一层卷积层之后加入一个串接层将两个模块的特征串联到一起,用于后续的双视图特征融合。

2.3　多视图学习

由于实际生产环境的限制,装箱工人身体的关键运动部位易被遮挡,只使用单个视图进行检测会由于缺失关键的运动信息导致识别准确率下降。为了使本文模型在训练和测试阶段可以获取到完整的运动信息,本文使用双视图的方式,使用两个不同视角的摄像头来获取装箱过程的图像,作为模型的输入。利用 3D 卷积网络和双视图的学习能力对装箱过程中的工人进行装箱行为识别。

相比单视图,多视图数据还包含额外的一致信息和互补信息,可以在这些额外的信息中学习到有意义的输出。其中,一致信息用于平衡多视图信息,互补信息用于不同视图之间的信息互补。利用多视图学习的一致信息和互补性信息,使图像信息得到最有效的提取和表示。使用两个并行的 3D ResNeXt101 模型对不同视图进行特征提取,然后采用一个视图池化层来融合双视图信息。直接采用最大池化和平均池化是融合过程中较为常用的方法,但是最大池化方法容易出现过拟合,网络的泛化能力差;平均池化不能很好地反映池化区域的特征,较小的元素会削弱较大元素对激活值的贡献,两者都会导致部分信息丢失。针对这个问题,本文使用一种可学习权重的视图池化层来融合多个视图的特征(权重学习视图池化层)。其计算为

$$O_L(p,q) = w_1 \times O_{\max}(p,q) + w_2 \times O_{\text{mean}}(p,q) \tag{2}$$

式中,w_1 和 w_2 分别是最大池和平均池的权重,初始值分别为 1 和 0,权重优化过程中保证两个权重的总和为 1。具体来讲,权重学习视图池化层在对位置 (p,q) 做双视图特征融合时,分别计算一个使用最大池化方法做特征融合的结果 $O_{\max}(p,q)$ 和一个使用平均池化方法做特征融合的结果 $O_{\text{mean}}(p,q)$。最后使用学习到的权重 w_1 和 w_2,计算出最终位置 (p,q) 的双视图特征融合结果 $O_L(p,q)$。本文使用 BP(back propagation)算法实现整个训练阶段的最佳权重搜索,本文使用的双视图池化方法如图 4 所示。

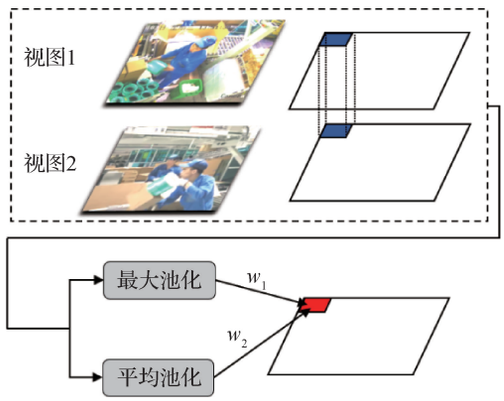


图 4　双视图池化方法

Fig. 4　Dual-view pooling fusion

本文双视图池化层是针对多个特征图的最大池化策略和平均池化策略的融合,通过端到端的训练,可以学习到一组最佳池化权重 w_1 和 w_2。使用该方法能将最大池化和平均池化有效地结合起来,从而

减少双视图池化阶段的信息丢失。

2.4 二分类 SVM 模型

使用双视图和差分图像相结合的模型,在识别装箱动作上已经有了较好的表现。但是,实验显示在一些复杂工业场景的影响下,本文模型仍会将一部分不存在装箱动作的装箱过程错误判断为存在装箱动作。为满足生产场景需求,尽可能排查出不合格装箱产品,保证产品合格率,这样的错误需要尽量避免。针对这个问题,本文加入降噪自编码器和 two-class SVM 来进一步排查没有出现装箱动作的装箱过程,提高装箱行为识别的真负率。本文将模型中经过双视图池化后的特征经过一个降噪自编码器优化、降维之后得到的特征向量作为 two-class SVM 的输入,然后对输入的特征样本进行分割,将结果分为存在装箱动作和不存在装箱动作两类。

two-class SVM 的主要方法是将训练数据通过核函数映射到一个高维的特征空间,然后在这个高维的特征空间寻找一个最佳超平面将这些特征向量分割为存在装箱动作和不存在装箱动作。具体而言,给定训练样本集 $\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots(\boldsymbol{x}_n,y_n)\}$,其中,$\boldsymbol{x}_i\in\mathbf{R}^d$ 为 d 维样本输入,$y_i\in(+1,-1)$ 为样本输出。为了确保超平面以最优的边界将样本进行分类,需要优化问题,即

$$\min_{w,b}\max_{\alpha}L(\boldsymbol{w},b,\alpha)=\frac{1}{2}\|\boldsymbol{w}\|^2-\sum_{i=1}^{n}\alpha_i(y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)-1)\tag{3}$$

式中,$\boldsymbol{x}_i$ 为训练样本,y_i 为类别标号,b 是一个偏移量,$\boldsymbol{w}\in\mathbf{R}^d$ 是需要学习的权重向量,第 i 个训练样本的拉格朗日系数 $\alpha_i>0$。其最终决策函数计算为

$$p_{w,b}(x)=\mathrm{sign}(\boldsymbol{w}\cdot\boldsymbol{x}+b)=\mathrm{sign}\left(\sum_{i=1}^{n}\alpha_iK(\boldsymbol{x}_i,\boldsymbol{x})+b\right)\tag{4}$$

式中,α_i 通过优化式(3)得到;sign( )为符号函数。当 $p_{w,b}(x)<0$ 时,表示 two-class SVM 模型将特征判定为存在装箱动作;否则 $p_{w,b}(x)>0$ 时,SVM 模型判定为不存在装箱动作。

常用的核函数有线性核函数 $K(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{x}\cdot\boldsymbol{y}$,多项式核函数 $K(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x}\cdot\boldsymbol{y}+1)^p$,sigmoid 核函数 $K(\boldsymbol{x},\boldsymbol{y})=\tanh(\gamma\cdot\boldsymbol{x}\cdot\boldsymbol{y}+1)$,径向基函数(radial basis function, RBF)核函数 $K(\boldsymbol{x},\boldsymbol{y})=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}\right)$。根据实际使用经验,在对非线性问题进行求解时,RBF 核函数的分类效果相对其他核函数较好,因此本文使用基于 RBF 核函数的 two-class SVM 对特征样本进行分类。

2.5 三层堆叠的降噪自编码器

降噪自编码器(Vincent 等,2008)是一种有效的无监督特征表达技术。它具有多个可学习的隐藏层,隐藏层的参数不是手动设置的,而是根据给定的数据自动学习的。已有工作表明(Budiman 等,2014),经过训练的降噪自编码器可以有效表现视频帧中的原始动作信息,并且学习到相邻帧之间的信息变化。同时降噪自编码器还可以利用编码、解码的特点,将高维特征压缩到低维,这解决了双视图池化之后特征维度过高的问题。本文引入降噪自编码器来优化从双视图池化层中获取到的特征,用于 two-class SVM 的输入,从而使 two-class SVM 模型更好地将有装箱动作的视频片段和无装箱动作的视频片段分隔。

自编码器包括两个阶段,即编码阶段和解码阶段,这两个阶段共享一个隐藏层。在编码阶段,数据从输出层到隐藏层,即

$$h(\boldsymbol{x})=sigm(\boldsymbol{W}\boldsymbol{x}+b)\tag{5}$$

在解码阶段,数据从隐藏层到输出层,即

$$\hat{\boldsymbol{x}}=sigm(\boldsymbol{W}(h(\boldsymbol{x}))+b)\tag{6}$$

式中,$\boldsymbol{W}$ 是权重,$\boldsymbol{x}$ 是数据,b 是偏置值,$sigm$ 表示常用的激活函数 sigmoid,$\boldsymbol{x}$ 编码之后得到 $h(\boldsymbol{x})$,$h(\boldsymbol{x})$ 解码之后得到 $\hat{\boldsymbol{x}}$。相比自编码器,降噪自编码器(denoising autoencoder, DAE)的关键就是在输入层和隐藏层之间添加噪声处理,将输入的原始数据 $\boldsymbol{x}$ 加入噪声,将其变为 $\tilde{\boldsymbol{x}}$,然后将带有噪声的 $\tilde{\boldsymbol{x}}$ 进行自编码器常规的变换工作。具体来讲,首先通过输入一个损坏的 $\tilde{\boldsymbol{x}}$,随后在隐藏层获得一个压缩后的特征 $h(\boldsymbol{x})$,然后通过 $h(\boldsymbol{x})$ 尽可能还原出加入噪声之前的 $\boldsymbol{x}$。降噪自编码器的损失函数为

$$L_{\mathrm{loss}}=\frac{1}{2N}\sum_{i=1}^{n}\|\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i\|_2^2+\lambda\|\boldsymbol{W}\|_2^2\tag{7}$$

在训练过程中,降噪自编码器通过最小化损失函数来学习参数 $(\boldsymbol{W},b)$。为了提升收敛速度和获取更高级的特征,本文使用由 3 个降噪自编码器构建而成的一个 3 层的堆叠降噪自编码器。

本文的堆叠降噪自编码器如图 5 所示，训练阶段以逐层方式进行。在测试阶段，去掉堆叠自编码器的解码部分，通过编码部分的 3 个隐藏层对输入的特征进行优化。3 个隐藏层分别将 2 048 维特征向量编码为 1 024 维、512 维、256 维。然后直接使用瓶颈隐藏层（bottleneck hidden layer）输出的 256 维的特征向量作为 two-class SVM 的输入。

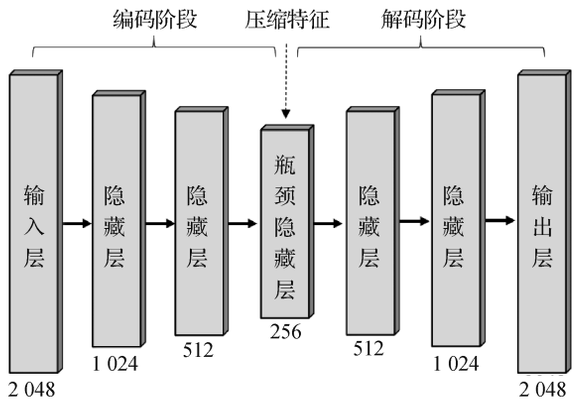


图 5　堆叠降噪自编码器

Fig. 5　Stacked denoising autoencoder

2.6　装箱行为识别

针对工业装箱场景中的装箱动作，本文提出了双视图 3 维卷积网络的工业装箱行为识别方法，通过识别是否出现装箱动作来判断装箱工人是否投放配件。主要工作包括：1）针对人体遮挡问题，设计了双视图结构，使用两个不同角度摄像头同一时间获取到的 RGB 视频作为双视图模型的输入，并使用一个可学习权重的双视图池化层融合两个视图的时空特征。2）针对光流的巨大计算成本，本文使用堆叠的差分图像作为模型的输入来更好地提取运动特征，替代实时场景中无法使用的光流。原始 RGB 图像和差分图像分别输入到两个并行的 3D ResNeXt101 中。3）为提高模型的真负率，满足装箱场景需求，在模型中加入堆叠降噪自编码器和 two-class SVM。

本文提出的装箱行为识别方法具体流程如下：将两个视图的每次装箱过程的 RGB 视频作为模型的输入，连续 RGB 图像和对 RGB 图像做差分计算得到的连续差分图像作为输入数据，分别输入到 4 个 3D ResNeXt101 中，经过模型计算后得到初步识别结果。为了进一步排查无装箱动作的装箱过程，提高真负率，将模型中经过双视图池化后的特征输入到降噪自编码器，通过训练好的降噪自编码器对特征优化和降维，然后利用 two-class SVM 模型进行二次判断，得到二次识别结果。只有两个识别结果都表明这次装箱过程中存在装箱动作，才最终判定这次装箱过程存在装箱动作，认为装箱工人在这次装箱过程中投放了配件。否则判定这次装箱过程不存在装箱动作，认为装箱工人在这次装箱过程中没有投放配件，装箱行为识别模型流程可形式化表述为图 6 所示。总结来讲，通过两次的装箱动作判别，只有两次的识别结果都判定存在装箱动作时，最终结果才为“装箱成功”，否则“装箱失败”。

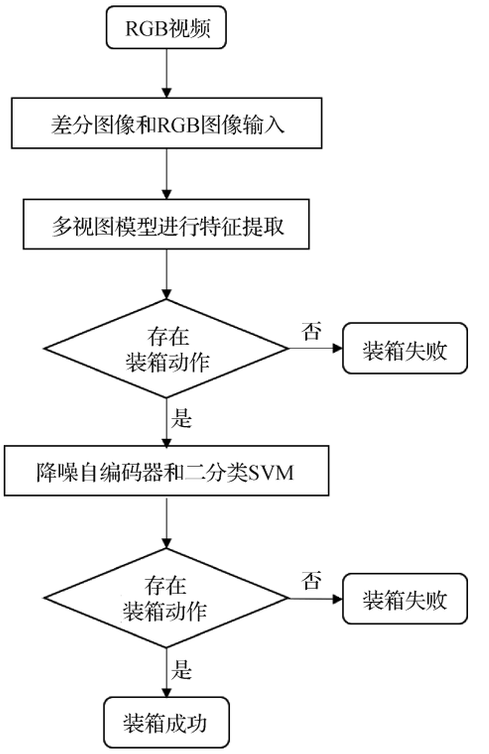


图 6　装箱行为识别模型执行流程

Fig. 6　The execution process of packing action model

该判别策略是在模型中加入了降噪自编码器和 SVM 的基础上制定的，会将没有装箱动作的装箱过程尽可能识别出来，进而提高真负率。但由于经过两次判别，部分存在装箱动作的装箱过程会被误判为不存在装箱动作，可能导致准确率下降。

3　实　验

3.1　双视图装箱动作数据集

双视图装箱动作数据集（dual-view packing action data，MPAD）由 7 个装箱工人在实际装箱过程

中执行,包含 2 个不同视角,3 个不同的装箱场景,共 2 400 个 RGB 视频。每个 RGB 视频就是装箱工人的一次装箱过程,视频帧率为 25 帧/s,视频时长约为 5 s。根据实际生产过程,数据集中包含为两个类别:有装箱动作的装箱过程(70%)、没有装箱动作的装箱过程(30%)。本文将数据集 MPAD 随机划分为训练集(80%)和测试集(20%)。数据集中的装箱动作为装箱工人投放配件的动作。在该数据集上,对单视图行为识别方法进行评估,本文将两个视图的 RGB 视频拆分成两个独立的 RGB 视频,作为单视图行为识别方法的输入。

3.2 实验细节

使用 Pytorch 作为深度学习平台,并使用 2 个 NVIDIA RTX 2080Ti 来评估本文方法。数据在输入网络之前,会先执行水平翻转,在训练阶段中应用时间抖动。批尺寸(batch size)设置为 32,使用随机梯度下降作为优化策略,动量(momentum)设置为 0.9,学习率初始值设置为 0.1,每 50 轮训练减小为初始值的 1/10。本文模型共进行 400 轮训练。为了验证模型中双视图结构的作用,加入了单视图模型进行比较。通过将双视图模型的视图池化层删除并由串接层直接与全连接层相连,完成本文的单视图模型,单视图模型的输入也从两个视图修改为一个视图。

引入降噪自编码器和 two-class SVM 来提高真负率,在不需要分析真负率,只需评估准确率的实验中,不使用降噪自编码器和 two-class SVM,剔除降噪自编码器和 two-class SVM 对模型准确率的干扰。在 UCF101 数据集中的实验中也使用去除降噪自编码器和 two-class SVM 的单视图模型来和主流行为识别方法进行对比,原因为 UCF101 是单视图数据集且由于 UCF101 包含多个动作类别无法使用二分类支持向量机。

3.3 MPAD 实验结果分析

在数据集 MPAD 上,对主流的行为识别方法和本文方法进行比较。表 1 展示了实验结果,从实验结果可以看出,本文方法获得了最高的准确率和真负率。表 2 评估了基于 RGB 图像和基于差分图像的两个 3D ResNeXt101 的效果,只使用差分图像作为输入使大量的外观信息丢失,导致较低的准确率。基于 RGB 图像的 3D ResNeXt101 相比差分图像取得较好的效果,RGB图像和差分图像结合的方法可以捕获到更丰富的特征,在数据集 MPAD 上取得了最高的准确率。

表 1 在 MPAD 中与主流行为识别方法的比较

Table 1 Comparison with mainstream action recognition methods on MPAD

方法	准确率/%	真负率/%
two-stream (Simonyan 和 Zisserman,2014)	83.2	84.1
C3D (Ji 等,2013)	78.2	78.9
3D ResNet101 (Hara 等,2018)	84.4	84.0
3D ResNeXt101 (Hara 等,2018)	86.4	86.4
two-stream I3D (Carreira 和 Zisserman,2017)	90.1	90.0
R(2+1)D (Girdhar 等,2019)	89.5	88.9
本文	**94.2**	**98.9**

注:加粗字体表示各列最优结果。

表 2 基于 RGB 和差分图像的 3D ResNeXt101 评估结果

Table 2 Evaluation results of 3D ResNeXt101 based on RGB frames and residual frames

方法	准确率/%
RGB 3D ResNeXt101	86.4
RF 3D ResNeXt101	71.2
RGB + RF 3D ResNeXt101	**91.1**

注:加粗字体表示最优结果,RF 为差分图像。

为了分析本文模型中各个模块的必要性,对不同模块组合策略进行了评估,并额外加入光流来和差分图像进行对比,表 3 展示了组合策略评估实验的结果。结果表明,使用双视图加光流的方法具有最高的准确率,但是对应的帧率只有 6.2 帧/s,在实

表 3 在 MPAD 中不同组合策略评估结果

Table 3 Evaluation results of different combination strategies on MPAD

方法	准确率/%	帧率/(帧/s)
RGB + 单视图	86.4	**116.4**
RGB + 双视图	88.2	61.0
RGB + 光流 + 单视图	91.5	11.5
RGB + 光流 + 双视图	**96.2**	6.2
RGB + RF + 单视图	91.1	82.4
RGB + RF + 双视图	95.8	49.8

注:加粗字体表示各列最优结果,RF 为差分图像。

时场景应用时会出现视频延迟的情况。同时使用双视图加差分图像的方法得到了较高准确率且帧率达到 49.8 帧/s，可以在实时场景应用。进一步说明差分图像是网络捕获运动特征的有效方法，通过使用差分图像替代光流，可以避免光流的复杂计算，解决无法在实时场景中应用的问题。另一方面，表 3 显示了本文引入的差分图像模块和双视图结构都能有效提升行为识别精度。

为了提升模型的真负率，引入了降噪自编码器和 two-class SVM，图 7 分别展示了只使用双视图模型、只使用降噪自编码器和 two-class SVM 以及两者结合的识别结果。图中的“双视图模型”是指不使用降噪自编码器和 two-class SVM 的方法；“DAE + SVM”表示 3D ResNeXt101 提取到的高层特征经双视图池化后直接由 DAE + SVM 来得到识别结果的方法。需要强调的是，图中的“双视图模型 + （DAE + SVM）”为两者结合的方法，即本文使用的装箱行为识别方法。该方法只有在双视图模型和（DAE + SVM）的识别结果都为“有装箱动作”时，才判定识别结果为“有装箱动作”，否则都判定为“没有装箱动作”，从而尽可能识别出所有没有装箱动作的装箱过程。

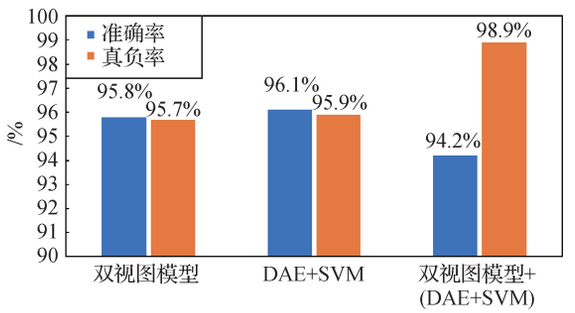


图 7　在 MPAD 中对双视图模型和 DAE + SVM 的评估结果

Fig. 7　Evaluation results of DAE + SVM and dual-view model in MPAD

图 7 中双视图模型的准确率和真负率稍低于 DAE + SVM。双视图模型和 DAE + SVM 的组合方法的准确率相较其他两种方法，下降了约 2%，但仍然保持在一个较高的水平，同时该方法将真负率提高到了 98.9%。造成准确率下降和真负率大幅提高的原因是在原双视图模型的基础上增加了 DAE 和 SVM，并在此基础上修改了识别结果的判别方式，由于经过两次判别，将没有装箱动作的装箱过程尽可能识别出来，从而提高了真负率。但同时，在数据集 MPAD 中正样本占比 70% 的情况下，部分存在装箱动作的装箱过程被误判为不存在装箱动作，会导致准确率下降。

表 4 评估了在双视图模型中加入降噪自编码器和 two-class SVM 对真负率和准确率的影响，实验结果中，只加入 two-class SVM 而没有降噪自编码器的组合准确率只有 91.5%。由于降噪自编码器对特征向量的优化和降维，同时加入降噪自编码器和 two-class SVM 的组合准确率达到 94.2%，并获得最高真负率 98.9%。使用降噪自编码器和 two-class SVM 与本文模型结合可以在有效提升真负率的同时保证较高的准确率，这样的组合更加符合实际装箱行为识别场景。

表 4　对降噪自编码器和 two-class SVM 的评估结果

Table 4　Evaluation results of denoising autoencoder and two-class SVM

方法		准确率/%	真负率/%
DAE	SVM		
×	×	**95.8**	95.7
×	√	91.5	98.1
√	√	94.2	**98.9**

注：加粗字体表示各列最优结果，√表示使用该方法，×表示未使用该方法。

3.4　UCF101 实验结果分析

UCF101 数据集（Reddy 和 Shah，2013）包含 101 种动作类别，共计 13 320 个视频片段，每种动作类别包含 25 组，每组包含 4 ～7 个视频片段，视频片段的平均长度为 7.21 s，包含相机运动、杂乱背景、不同光照条件、遮挡和低质量等不确定因素。UCF101 为单视图动作数据集，本文使用单视图模型进行评估。表 5 将本文方法与之前的方法进行比较，可以看出本文方法达到较好的检测结果。在对多个动作类别进行识别时，本文单视图模型也获得较高的准确率。

4　结　论

提出一种用于实际生产场景的装箱行为检测方法，该方法使用双视图 3D ResNeXt101 模型进行有效的装箱行为识别。使用两个并行的 3D ResNeXt101，分别从RGB图像和差分图像中学习时

表 5　在 UCF101 中与主流行为识别方法的比较

Table 5　Comparison with mainstream action recognition methods on UCF101

方法	准确率/%
two-stream	88.0
IDT	86.4
C3D	82.3
P3D(Qiu 等,2017)	88.6
3D ResNeXt-101	94.5
two-stream I3D	**98.0**
R(2+1)D	94.2
TesNet(Huang 和 Bors,2020)	95.2
本文(单视图)	97.1

注:加粗字体表示最优结果。

空特征,以获得更丰富的特征,并使用可学习的视图池化层做双视图特征融合。此外本文训练了一个堆叠的降噪自编码器对双视图 3D ResNeXt101 模型提取的特征进行优化和降维,并使用 two-class SVM 模型进行二次检测来提高真负率(TNR)。实验结果和分析表明,本文装箱行为识别方法在数据集 MPAD 中得到的准确率和真负率分别为 94.2%、98.9%,均优于其他 6 种主流行为识别方法。在人体频繁被遮挡的实际装箱场景中,本文方法可以精确识别出装箱工人的装箱动作,且同时保证识别结果的高真负率,满足实际生产场景需求。

在实验过程中发现,本文模型中的双视图结构能够有效降低人体遮挡的干扰,但由于人体遮挡导致的运动信息不足,仍会有部分样本的装箱行为识别会受到遮挡的影响。未来的研究工作将致力于解决人体遮挡问题,通过增加视图数量来获取更多不同角度的运动信息,同时加入使用双目视觉得到的 3 维人体骨架作为模型的输入,进一步降低人体遮挡的干扰。

参考文献(References)

Budiman A, Fanany M I and Basaruddin C. 2014. Stacked denoising autoencoder for feature representation learning in pose-based action recognition//Processings of the 3rd IEEE Global Conference on Consumer Electronics. Tokyo, Japan: IEEE: 684-688 [DOI: 10.1109/GCCE.2014.7031302]

Cao Z, Simon T, Wei S E and Sheikh Y. 2017. Realtime multi-person 2D pose estimation using part affinity fields//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1302-1310 [DOI: 10.1109/CVPR.2017.143]

Carreira J and Zisserman A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4724-4733 [DOI: 10.1109/CVPR.2017.502]

Crasto N, Weinzaepfel P, Alahari K and Schmid C. 2019. MARS: motion-augmented RGB stream for action recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7874-7883 [DOI: 10.1109/CVPR.2019.00807]

Das Dawn D and Shaikh S H. 2016. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. The Visual Computer, 32(3): 289-306 [DOI: 10.1007/s00371-015-1066-2]

Feichtenhofer C, Pinz A and Zisserman A. 2016. Convolutional two-stream network fusion for video action recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1933-1941 [DOI: 10.1109/CVPR.2016.213]

Girdhar R, Tran D, Torresani L and Ramanan D. 2019. Distinit: learning video representations without a single labeled video//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 852-861 [DOI: 10.1109/ICCV.2019.00094]

Hara K, Kataoka H and Satoh Y. 2018. Can spatiotemporal 3D CNNs retrace the history of 2D cnns and ImageNet?//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6546-6555 [DOI: 10.1109/CVPR.2018.00685]

Hong J, Cho B, Hong Y W and Byun H. 2019. Contextual action cues from camera sensor for multi-stream action recognition. Sensors, 19(6): #1382 [DOI: 10.3390/s19061382]

Huang G X and Bors A G. 2020. Learning spatio-temporal representations with temporal squeeze pooling//Proceedings of ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain: IEEE: 2103-2107 [DOI: 10.1109/ICASSP40776.2020.9054200]

Ji S W, Xu W, Yang M and Yu K. 2013. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1): 221-231 [DOI: 10.1109/TPAMI.2012.59]

Kuanar S, Athitsos V, Mahapatra D, Rao K R, Akhtar Z and Dasgupta D. 2019. Low dose abdominal CT image reconstruction: an unsupervised learning based approach//Proceedings of 2019 IEEE International Conference on Image Processing (ICIP). Taipei, China:

IEEE: 1351-1355 [DOI: 10.1109/ICIP.2019.8803037]

Kuanar S, Athitsos V, Pradhan N, Mishra A and Rao K R. 2018. Cognitive analysis of working memory load from eeg, by a deep recurrent neural network//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, Canada: IEEE: 2576-2580 [DOI: 10.1109/ICASSP.2018.8462243]

Laptev I. 2005. On space-time interest points. International Journal of Computer Vision, 64(2/3): 107-123 [DOI: 10.1007/s11263-005-1838-7]

Liu Z and Hu H F. 2019. Spatiotemporal relation networks for video action recognition. IEEE Access, 7: 14969-14976 [DOI: 10.1109/ACCESS.2019.2894025]

Peng X J, Wang L M, Wang X X and Qiao Y. 2016. Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. Computer Vision and Image Understanding, 150: 109-125 [DOI: 10.1016/j.cviu.2016.03.013]

Qiu Z F, Yao T and Mei T. 2017. Learning spatio-temporal representation with pseudo-3 d residual networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5534-5542 [DOI: 10.1109/ICCV.2017.590]

Rastgoo R, Kiani K and Escalera S. 2020. Hand sign language recognition using dual-view hand skeleton. Expert Systems with Applications, 150: #113336 [DOI: 10.1016/j.eswa.2020.113336]

Reddy K K and Shah M. 2013. Recognizing 50 human action categories of web videos. Machine Vision and Applications, 24(5): 971-981 [DOI: 10.1007/500138012-04504]

Simonyan K and Zisserman A. 2014. Two-stream convolutional networks for action recognition in videos//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 568-576

Su H, Maji S, Kalogerakis E and Learned-Miller E. 2015. Multi-view convolutional neural networks for 3D shape recognition//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 945-953 [DOI: 10.1109/ICCV.2015.114]

Tao L, Wang X T and Yamasaki T. 2020. Rethinking motion representation: residual frames with 3D convnets for better action recognition [EB/OL]. [2021-02-18]. https://arxiv.org/pdf/2001.05661.pdf

Tran D, Bourdev L, Fergus R, Torresani L and Paluri M. 2015. Learning spatiotemporal features with 3D convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 4489-4497 [DOI: 10.1109/ICCV.2015.510]

Ullah A, Muhammad K, Haq I U and Baik S W. 2019. Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments. Future Generation Computer Systems, 96: 386-397 [DOI: 10.1016/j.future.2019.01.029]

Vincent P, Larochelle H, Bengio Y and Manzagol P A. 2008. Extracting and composing robust features with denoising autoencoders//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM: 1096-1103 [DOI: 10.1145/1390156.1390294]

Wang H and Schmid C. 2013. Action recognition with improved trajectories//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE: 3551-3558 [DOI: 10.1109/ICCV.2013.441]

Wang Y S, Liu J Y, Zeng Q H and Liu S. 2015. Visual pose measurement based on structured light for MAVs in non-cooperative environments. Optik, 126(24): 5444-5451 [DOI: 10.1016/j.ijleo.2015.09.041]

Wu C Y, Zaheer M, Hu H X, Manmatha R, Smola A J and Krähenbühl P. 2018. Compressed video action recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: CVPR: 6026-6035 [DOI: 10.1109/CVPR.2018.00631]

Wu Q Y, Zhu A C, Cui R, Wang T, Hu F Q, Bao Y P and Snoussi H. 2021. Pose-guided inflated 3D convnet for action recognition in videos. Signal Processing: Image Communication, 91: #116098 [DOI: 10.1016/j.image.2020.116098]

Zeng H, Wang Q, Li C and Song W. 2019. Learning-based multiple pooling fusion in dual-view convolutional neural network for 3D model classification and retrieval. Journal of Information Processing Systems, 15(5): 1179-1191 [DOI: 10.3745/JIPS.02.0120]

作者简介

胡海洋，1977 年生，男，教授，主要研究方向为机器视觉、智能制造。
E-mail：huhaiyang@hdu.edu.cn

潘健，男，硕士研究生，主要研究方向为机器视觉、行为识别。
E-mail：panjian@hdu.edu.cn

李忠金，男，讲师，主要研究方向为云计算、工作流调度。
E-mail：lizhongjin@hdu.edu.cn