Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Civil Engineering—Article

基于人体姿态估计信息的工人施工活动分析

周绪红[a], 李帅[a,*], 刘界鹏[a], 伍洲[b,*], Yohchia Frank Chen[c]

[a] School of Civil Engineering, Chongqing University, Chongqing 400044, China
[b] School of Automation, Chongqing University, Chongqing 400044, China
[c] Department of Civil Engineering, The Pennsylvania State University, Middletown, PA 17057, USA

ARTICLE INFO

Article history:
Received 11 April 2022
Revised 27 September 2023
Accepted 25 October 2023
Available online 25 November 2023

关键词
姿态估计
活动分析
目标跟踪
建筑工人
自动化系统

摘要

识别工人的施工活动或行为可以帮助管理者更好地监控劳动效率和施工进度。然而，目前的施工工人活动分析依赖于人工观察和记录，这消耗了大量的时间和人力成本。研究人员已经致力于监测工地上工人的施工行为，但当多名工人一起工作时，现有的研究不能准确和自动地识别施工活动。本研究提出了一个深度学习框架，用于自动分析多名工人的施工活动。在这个框架中，设计并使用了多个深度神经网络模型来完成工人关键点提取、工人跟踪和工人施工活动分析。设计的框架在实际建筑工地进行了测试，对多名工人的活动进行了识别，表明该框架用于自动监控工作效率的可行性。

© 2023 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. 引言

在建筑行业中，工人是最重要的资源之一，他们的行为直接影响到项目的进度和成本。建筑行业员工的效率低下最终会导致生产力下降，浪费时间和资源。因此，需要对工人进行现场监控，跟踪他们的动作行为是衡量他们生产率的有效方法之一。建筑行业的管理者可以获取关于工人状态的最新信息作为参考，并据此调整策略。

传统的监控系统依赖于现场工头进行监控，他们负责监督工人的状态。一些工人在没有监控的情况下，可能无法有效地履行职责，导致生产力下降。这种监控方式有很大的局限性，需要许多工头来监控工人，这将增加人工成本，结果的可靠性也无法保证，存在主观性。因此，需要一个自动分析建筑工人行为的系统，以保证工人高效地工作。

研究人员开发了各种技术进行自动提取和分析现场信息，并设计了相应的系统来监控施工现场。目前，有两种类型的方法：非视觉和视觉。非视觉方法包括全球定位系统（GPS）、惯性测量单元（IMU）系统、射频识别（RFID）系统和超宽带（UWB）系统[1–4]。这些方法基于电子传感器，持续收集信息（如对象的速度、加速度和方向）以对工人的活动进行分类。然而，工人需要佩戴传感器设备进行数据采集，带来一些不便。此外，使用非视觉方法难以对工人详细的活动进行分类。例如，GPS用于记录位置信息的变化，但它无法识别细微的活动。

视觉方法相对于非视觉方法没有上述的劣势。这些系

* Corresponding authors.
E-mail addresses: li_shuai@cqu.edu.cn (S. Li), wuzhsky@gmail.com (Z. Wu).

2095-8099/© 2023 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
英文原文：Engineering 2024, 33(2): 225–236
引用本文：Xuhong Zhou, Shuai Li, Jiepeng Liu, Zhou Wu, Yohchia Frank Chen. Construction Activity Analysis of Workers Based on Human Posture Estimation Information. *Engineering*, https://doi.org/10.1016/j.eng.2023.10.004

统无需直接接触工人，可以利用图片和视频数据分析建筑工人的行为。目前，监控系统收集了大量建筑工地的视频数据[5]。许多研究使用非接触式的普通红-绿-蓝（RGB）摄像机或红-蓝-绿-深度（RGB-D）摄像机利用算法来识别工人的活动[6–7]。RGB-D摄像机（如微软的Kinect摄像机）是一种传感系统，能够捕捉RGB图像以及每个像素的深度信息[8]。此外，深度信息被用来提取人体的轮廓图，并获取工人的关键点数据。在相关研究中，提取的关键点数据被用来实时识别不安全的工人行为[9]。与原始图像相比，关键点数据具有更小的数据量，可以更快地进行处理。然而，由于RGB-D深度摄像机的拍摄距离有限以及光线的影响，它通常只适用于室内场景，在室外的工地场景则难以获取深度信息或关键点数据。

很少有研究将活动与特定工人相关联，因为大多数方法只评估静态视频帧图像，并没有在前后帧之间建立同一工人的关联，从而进一步限制了对每名工人劳动力的评估[10]。这是基于计算机视觉的非接触方法的一个缺点，而一些基于非视觉的系统在执行这项任务方面更有效。此外，大多数工人行为识别系统依赖于原始图像，与基于关键点数据的方法相比，计算量大，且难以保证实时性。建筑工地通常有几个摄像机，处理的视频数据量很大。与基于关键点数据的方法相比，现有的基于视觉的系统依赖于原始图像来完成工人行为检测，这在计算上要求高，难以实时执行。

本文介绍了一种新的基于视觉的活动识别框架，使用普通摄像机录制的视频作为输入，自动获取每名工人的行为。在这个框架中，使用轻量级姿态估计网络从普通摄像机拍摄的视频中获取人体关键点信息。然后，采用多人跟踪算法，并利用关键点提取工人的边界帧，以完成运动和外观信息的提取，从而有效地跟踪工人和关键点。最后，设计了多层全连接（FC）神经网络和堆叠长短期记忆（LSTM）神经网络，对每名工人的关键点信息进行分类，完成行为识别和施工效率分析。利用从实际建筑工地收集的视频，测试并验证了所提出框架的可行性。本文的贡献如下：

（1）使用人体姿态估计算法从普通摄像机视频中提取工人关键点，并结合神经网络模型分析工人活动；

（2）考虑时间和空间组合对工人活动分析准确性及实际工地监控视频的影响；

（3）使用外观和运动信息在视频的前后帧之间建立工人的联系，完成多工人动作识别和效率统计。

2. 相关工作

本节介绍了现有的非视觉和视觉施工活动识别方法及相关工作的分析。

2.1. 非视觉系统

非视觉系统的研究工作主要基于物联网（IoT）。基于IoT的系统以佩戴电子传感器的目标对象作为分析对象，依靠电子传感器通过连续收集传感器信息，包括速度、加速度和方向，分析对象的活动或工作状态。基于传感器的接触式监测研究侧重于远程定位和跟踪技术。Kelm等[2]设计了一个移动RFID门户网站来检查工人的安全防护装备是否符合相应规范。Pradhananga等[11]进一步整合了GPS的结果，完成对土方作业的生产率测量。Montaser等[12]开发了一种基于RFID技术的材料位置识别和跟踪方法，可以用来获取近实时决策所需的信息。Akhavian等[13]使用移动传感器（加速度计、陀螺仪和GPS）和机器学习分类器，通过分析设备的数据来识别土方设备的活动。Cheng等[3]通过UWB获取了工人的运动和工地分布信息，完成建筑工人活动的识别。Zhao等[14]和Sanhudo等[15]使用深度神经网络（DNN）模型从可穿戴IMU传感器捕获的运动数据中识别建筑工人的活动。一些研究侧重基于可穿戴无线IMU测量的运动数据，研究了可能导致与施工相关的工作相关肌肉骨骼疾患（WMSDs）的人体运动[16–18]。

传感器的主要优点之一是可以有效地提供穿戴者身份识别信息（ID），并分析不同工人佩戴的传感器的活动[19]。然而，IoT系统的应用要求每个建筑工人都佩戴IoT传感器，而传感器的佩戴可能会影响正常的施工活动，通常不被工人所接受[20]。

2.2. 视觉系统

研究人员使用非接触式普通或深度摄像机结合计算机算法来识别工人活动。近年来，这一研究课题引起了越来越多的关注。随着深度摄像机的出现，研究人员开始使用它们获取关键点，完成工人活动识别。例如，使用Kinect深度摄像机提取人体轮廓，使用快速骨架化技术获取人体姿态描述符。针对建筑工人经常患有各种肌肉骨骼疾病的问题，使用深度摄像机进行姿态采集，检测不正确的活动[21]。Khosrowpour等[6]使用了微软的Kinect传感器来估计人体姿态，并利用获得的关键点信息来分析工人的活动。相关研究证实，姿态估计数据足以用于理解和分类人类动作行为[22]，但目前基于关键点的活动识别依赖于深

度摄像机提取，成本高且测量距离有限[23]。

在基于普通摄像机的智能监控领域，随着计算机视觉技术的显著进步，计算机视觉和模式识别已显示出深度学习方法优于传统机器学习方法，包括图像分类[24]、物体检测[25]和动作识别任务[26]。Fang等[27]和Kim等[28]基于区域检测卷积神经网络（CNN）从RGB数据中检测建筑工人和设备。Fang等[29]使用基于区域提名的更快CNN（Faster-RCNN）来判断工人是否佩戴安全帽。Zhang等[30]提出了一种基于深度学习的实时裂缝检测方法，用于检测桥面的裂缝。上述研究主要基于图像来完成不安全活动或物体的识别。这些方法仅将视频中每一帧单独作为图片处理，没有建立前后帧对象之间的联系，因此前后帧之间的处理结果独立存在，导致无法在存在多名工人的场景中完成检测。考虑到活动的连续性，研究人员基于连续的视频帧数据进行处理。Yang等[5]使用一种图像特征提取的方法研究基于视觉的工人行为识别。Ding等[31]开发了一种结合CNN和LSTM [32]的混合深度学习模型，自动识别不安全的工人行为。Luo等[10,33]使用图像和光流图像获取空间和时间信息，并使用时间段网络（TSN）[34]方法，通过双流数据识别建筑工地上工人的活动。Chen等[35]使用三维（3D）CNN准确识别挖掘机活动，并将其分类为详细类型（如挖掘、装载和摆动）。然而，在杂乱的建筑场景中，除工人外还有许多其他物体会对分类结果产生影响。上述方法在前后帧之间建立了联系，并将前后帧数据结合进行动作识别。但当前后帧中存在许多工人时，难以区分每名工人，同时基于双流或3D卷积，计算量大，难以保证实时性能。深度学习方法的缺点是，由于视频帧的整个区域都提供了学习表征，因此学习到的表征可能不专门针对人类行为。此外，很少有方法可以自动跟踪工人或在无人工干预的情况下识别和跟踪进入视频的新的工人[10]。

与普通摄像机相比，深度摄像机可以以更快的数据处理速度获取工人关键点信息，且关键点信息消除了多余背景干扰信息[36]。现有研究中可以用基于视觉的方法来获取工人的关键点信息[37]。Roberts等[38]使用关键点提取算法alphapose进行姿态估计，同时使用3D卷积处理双流和关键点信息。由于alphapose是一种自上而下的估计方法，因此当工人数量增加时，计算量会大大增加[39]，对于双流数据的3D卷积处理无法保证实时性能。此外，提取的边界框中ORB特征可能并非工人本身的特征，算法在背景和工人之间寻找特征点，这使得难以通过保证前后帧特征的关联来实现跟踪。关键点提取方法已被证明适用于活动识别任务[22,40–41]。骨架特征提供了关于人体关节和骨骼的量化信息。与RGB流相比，骨架特征在复杂背景的动态情况下可以提供更精简和有用的信息。

3. 方法框架

多工人施工动作和工作效率识别框架如图1所示，包括四个主要模块：建筑工人关键点提取、工人位置提取与跟踪、行动识别和工作效率分析。首先，姿态估计检测器完成视频分析处理，获取工人关键点数据。其次，使用关键点数据获得每名工人当前帧的位置信息，并使用基于CNN的跟踪算法来识别每名工人的位置和移动轨迹。跟踪结果返回每名工人的ID编号。通过工人的ID号，将检测的关键点信息在时间帧上映射到对应的工人：即连续获取每名工人对应的关键点信息。然后，设计了一个时空分析关键点数据的神经网络模型，用以识别工人的行为。最后，对行为识别结果进行汇总，计算每名工人在一段时间内的工作效率和全体工人的工作效率。

3.1. 关键点提取

关键点数据量相对较小，且仅包含工人的信息，因此可消除复杂背景的影响。该模块从普通摄像机的视频中提取人体关键点信息，不依赖深度摄像机进行关键点提取。为了快速连续提取建筑工地上多名工人的关键点信息，本文提出了一种轻量级姿态估计方法。在框架中使用改进的轻量化网络结构的OpenPose模型[40]来提取工人关键点数据，并结合几何关系分析安全帽的不规范使用情况[42]。

OpenPose模型是一种高效的多人姿态估计方法，在多个公开基准数据集上测试效果出色。该模型将宽度为w、高度为h的图像作为输入。经过处理后，模型为每个人输出二维（2D）关键点。在原始OpenPose网络中，图像被一个CNN处理，其前十层由视觉几何小组的19层网络结构（VGG-19）[43]组成，生成一组特征图$\boldsymbol{F}$。为了进一步提高估计速度，本研究采用了轻量级架构，使用MobileNetV2 [44] 代替VGG-19作为特征提取网络。特征提取网络结构如图2所示，主要由瓶颈残差块（BRB）结构组成。该结构首先将低维压缩表示扩展到高维，并通过轻量级深度卷积进行计算[44]。随后，通过线性卷积将特征投影回低维表示。其余结构包括标准卷积（SC）、深度可分离卷积（DSC）[45]和上采样层（UPS-L），通过网络分析原始图像生成一组特征图$\boldsymbol{F}$。

然后，网络由多阶段卷积网络（图2）组成，并分为多个相似的阶段，每个阶段包括两个分支：一个用于获得身体部位位置的置信度图，如图2中橙色虚线框所示；另一个用于预测部分关联字段（PAF），以编码身体部分之

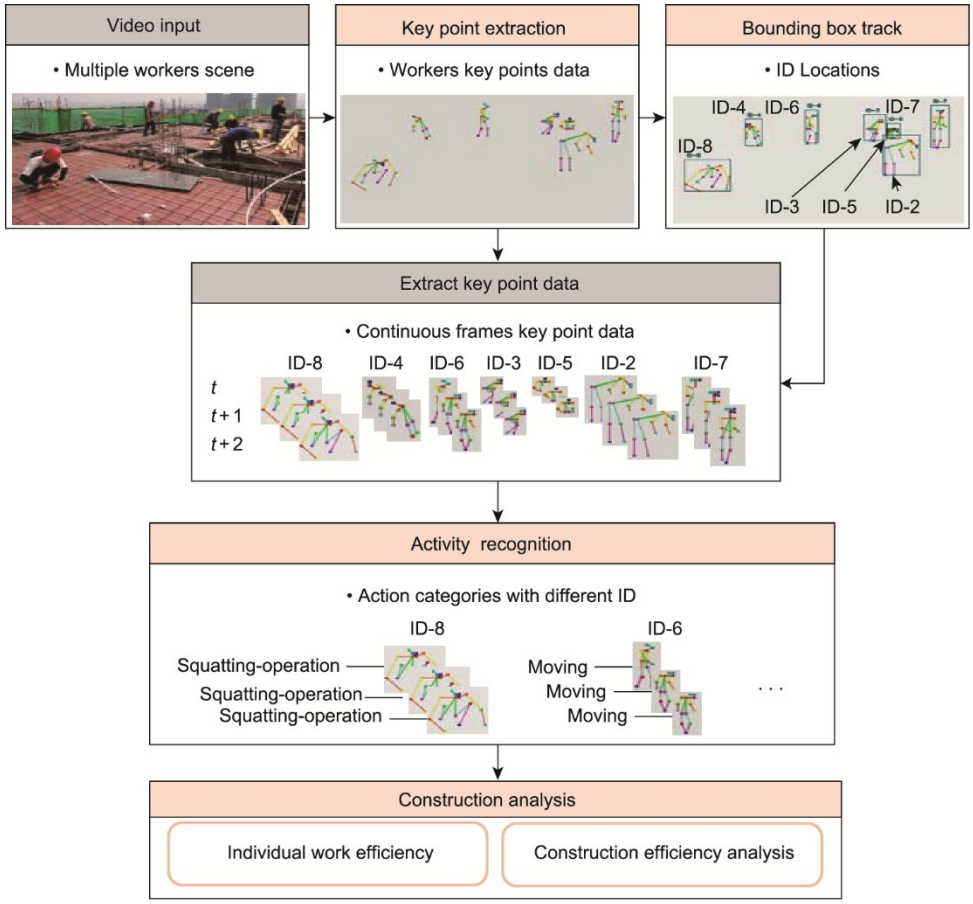

图1. 多工人施工行为和工作效率识别框架。*t*: 阶段。

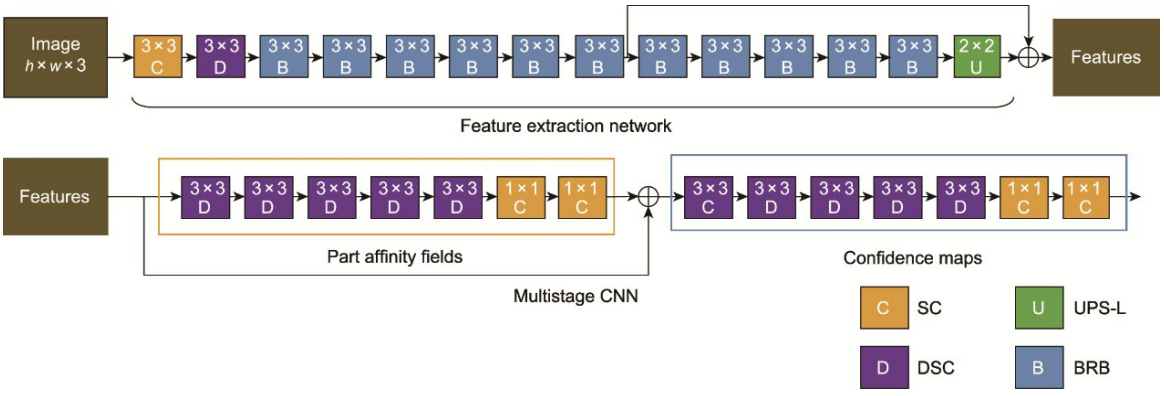

图2. 姿态估计网络结构。SC：标准卷积；DSC：深度可分离卷积；UPS-L：上采样层；BRB：瓶颈残差块。

间的关联度，如图2中蓝色虚线框所示。最后，通过贪婪算法推理解析置信度图和PAF，输出所有工人的二维关键点。

第一阶段将特征图$\boldsymbol{F}$作为输入，并生成一组PAF $\boldsymbol{L}^1$。

$$\boldsymbol{L}^1=\varphi^1(\boldsymbol{F}) \tag{1}$$

式中，φ^1是第一阶段橙色虚线框内的多阶段卷积网络，用于第一阶段的推理。

在每个后续阶段，将特征图$\boldsymbol{F}$和上一阶段的PAF $\boldsymbol{L}^{t-1}$预测结果连接作为多阶段CNN的输入，其中t表示阶段数，$\boldsymbol{L}^t$表示第t阶段的PAF预测矢量场，$\boldsymbol{L}^t$中的每个元素代表属于肢体的像素的方向和大小。

$$\boldsymbol{L}^t=\varphi^t\left(\boldsymbol{F},\boldsymbol{L}^{t-1}\right),\forall 2\le t\le T_{\mathrm{P}} \tag{2}$$

式中，φ^t是第t阶段橙色虚线框内的多阶段CNN，T_{P}是总PAF阶段的数量。

经过T_{P}次迭代后，从最近更新的PAF预测开始，置信图检测重复与PAF阶段类似的过程：

$$\boldsymbol{S}^{T_{\mathrm{P}}}=\rho^t\left(\boldsymbol{F},\boldsymbol{L}^{T_{\mathrm{P}}}\right),\forall t=T_{\mathrm{P}} \tag{3}$$

$$\boldsymbol{S}^t=\rho^t\left(\boldsymbol{F},\boldsymbol{L}^{T_P},\boldsymbol{S}^{t-1}\right),\forall T_P<t\le T_P+T_C \tag{4}$$

式中，ρ^t是第t阶段蓝色虚线框内的多阶段CNN；T_C是置信度图总阶段数；$\boldsymbol{S}$是对应阶段的置信度图；$\boldsymbol{S}^t$是第t阶段置信度图的输出，$\boldsymbol{S}^t$中的每个元素代表像素属于关键点的概率；$\boldsymbol{S}^{t-1}$是第$t-1$阶段置信度图的输出，其形状和含义与$\boldsymbol{S}^t$相同；$\boldsymbol{L}^{T_P}$表示第T_P阶段的PAF预测矢量场，$\boldsymbol{S}^{T_P}$是第T_P阶段置信度图的输入。

在每个阶段的末尾，应用损失函数，使网络在第一分支迭代检测PAF，在第二分支迭代检测置信度图。第t_i阶段PAF分支的损失函数和第t_k阶段置信度图分支的损失函数分别由公式（5）和（6）表示：

$$f_L^t=\sum_{c=1}^{C}\sum_{p}\boldsymbol{W}(p)\left\|\boldsymbol{L}_c^t(p)-\boldsymbol{L}_c^*(p)\right\|_2^2 \tag{5}$$

$$f_S^t=\sum_{j=1}^{J}\sum_{p}\boldsymbol{W}(p)\left\|\boldsymbol{S}_j^t(p)-\boldsymbol{S}_j^*(p)\right\|_2^2 \tag{6}$$

式中，C和c都是表示人体部位之间连接的索引，C是连接的总数，c是特定连接；J和j都是表示人体部位的索引，J是部位的总数，j是特定部位；$\boldsymbol{L}_c^*$是第c个身体部位到部位关联度的真实向量场；$\boldsymbol{S}_j^*$是第j个身体部位位置的真实置信度图；$\boldsymbol{W}$是一个二进制掩码。对于矩阵$\boldsymbol{W}$，当图像像素p缺少注释时，$\boldsymbol{W}$中对应于p的元素为0。整体损失函数为：

$$f=\sum_{t=1}^{T_P}f_L^t+\sum_{t=T_P+1}^{T_P+T_C}f_S^t \tag{7}$$

在本研究中，多阶段网络中的卷积被替换为可分离卷积。通过使用DSC结构作为替换，改进了姿态估计网络结构，如图2底部所示。对于多阶段网络，使用可分离卷积进行替换，减少了网络计算量。如果卷积核的大小为动态卷积核（DK）× DK，IN是输入通道的数量，ON是输出通道的数量。SC的参数量可以表示为SC = DK × DK × IN × ON。深度可分离体积积分执行相同的处理操作，表示为DSC = DK × DK × IN + IN × ON。

关键点提取如图3所示，图3（a）显示输入的原始图像，图3（b）显示提取的工人关键点信息。每名工人由18个关键点组成，如图3（c）所示。每个关键点的对应含义或身体部位见表1。

3.2. 工人位置的提取和跟踪

这一模块的目的是处理关键点信息，以获得工人的边界框位置信息，然后使用跟踪算法来跟踪和分析边界框和边界框区域图像，获得工人ID号。对于只有一名工人的情况，可以使用第3.1节中描述的方法从视频中提取关键点信息。设$P=\{P^0,P^1,\cdots,P^K\}$表示连续帧号K的关键点，

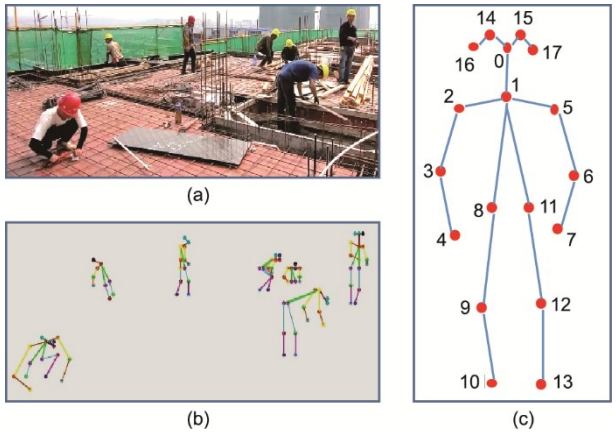

图3. 关键点提取。（a）原始图像；（b）关键点提取结果；（c）关键点组成。

表1 关键点位置

Number	Key point	Number	Key point
0	Nose	9	Left knee
1	Neck	10	Left foot
2	Left shoulder	11	Right hip
3	Left elbow	12	Right knee
4	Left wrist	13	Right foot
5	Right shoulder	14	Left eye
6	Right elbow	15	Right eye
7	Right wrist	16	Left ear
8	Left hip	17	Right ear

其中$P^k=\{(x^0,y^0),(x^1,y^1),\cdots,(x^{17},y^{17})\}$表示第$k$帧中工人的关键点。当视频中有多名工人时，需要获取每一帧中所有工人的关键点信息。关键点检测方法可以同时提取每一帧中所有工人的关键点信息，但对于每名工人无法将连续帧的数据关联起来。设N为获得帧中检测到的个体数量，不同帧中工人的个体数量不同。对于每一帧中每名工人的ID号n是不变的，其中工人ID = n的第k帧的关键点表示为$P_n^k=\left\{\left(x_n^0,x_n^0\right),\left(x_n^1,x_n^1\right),\ldots,\left(x_n^{17},x_n^{17}\right)\right\}$。图4是跟踪结果的示意图，显示了连续获取工人关键点的过程。

考虑到新工人的出现、算法的漏检、移动到视野范围外以及工人的移动和遮挡，保持每名工人的编号不变并获得连续帧工人关键点数据是困难的。为了解决上述问题，本文使用多目标跟踪算法实现工人和关键点数据之间的连续帧关联。在跟踪模块中，应用了深度在线实时（SORT）跟踪器，将前一步检测到的相同工人在视频的所有帧中关联起来[46]。该方法已被用于跟踪预制墙，并从监控视频中收集预制墙的位置和时间信息[47]。

x_{max}和y_{max}表示18个关键点坐标的最大值，x_{min}和y_{min}表示最小值，(w,h,x,y)表示边界框的坐标，其中(x,y)是

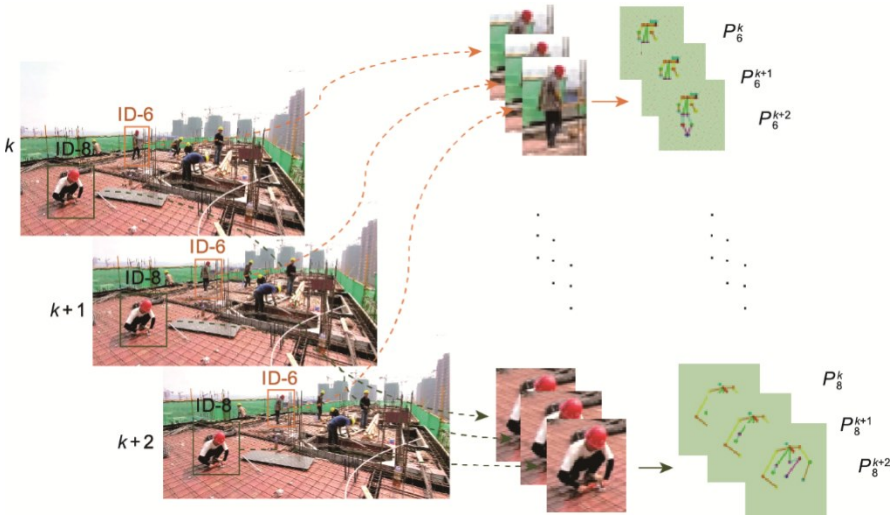

图4. 跟踪结果的示意图。

边界框中心点的坐标，w, h 分别是边界框的宽度和高度。边界框的坐标可以通过使用最大值和最小值的简单计算得到。图5显示使用关键点信息完成工人边界框检测。

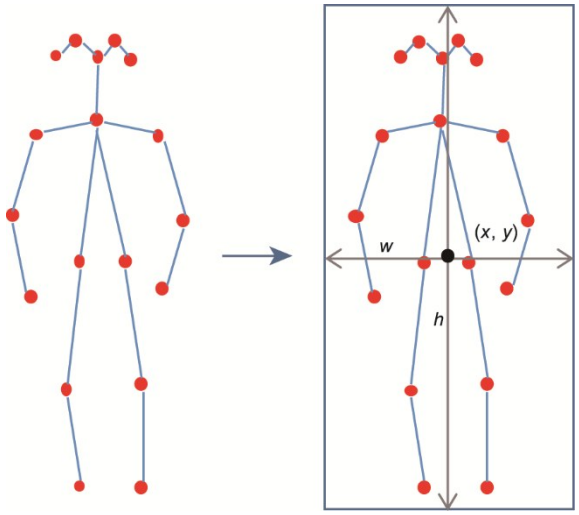

图5. 提取边界框。

在第 k 帧提取了工人的边界框 $b_n^k = [x_n^k, y_n^k, w_n^k, h_n^k]$，其中 (x_n^k, y_n^k) 是第 k 帧中 ID = n 工人边界框的中心位置，w_n^k 和 h_n^k 分别是其像素宽度和高度。在本研究中，选择深度SORT算法作为目标跟踪方法，这是一种实时性能优异的跟踪器。该方法将跟踪场景定义为一个八维向量 $\boldsymbol{d}^k = [x, y, \gamma, h, \dot{x}, \dot{y}, \dot{\gamma}, \dot{h}]^{\mathrm{T}}$，其中包含边界框的中心位置 (x, y) 和高度 h、纵横比 $\gamma = w/h$，以及图像坐标中的各自速度。

使用标准卡尔曼滤波器和线性观测模型，预测更新的轨迹，该滤波器具有恒定速度运动[44]。然后，根据预测值匹配当前检测值，基于第 $k+1$ 帧状态的 $\boldsymbol{d}^k$ 预测，可以获得 $\boldsymbol{d}^{k+1}$。

深度SORT算法使用运动和外观信息来实现有效的跟踪。使用外观信息（128维特征）和运动信息（由卡尔曼滤波器预测的跟踪位置）来确定跟踪和检测同一人的可能性。对于不同的场景，可以调整两个距离关联度的权重。运动信息关联速度相对较快，适用于短期内没有遮挡的情况。外观描述符与提取工人的特征与先前存储的特征进行比较，可应用于工人遮挡的场景中。通过跟踪算法，将ID编号 n 和检测的关键点信息在时间上与对应的工人映射。

3.3. 动作识别

上述两个模块将普通摄像机拍摄的视频信息转换为连续帧中每名工人对应的关键点序列，代表不同工人的施工动作信息。本研究选择了多层FC神经网络来处理一帧中各个关键点之间的相对距离和角度等空间特征。堆叠LSTM网络可以处理连续帧的关键点信息提取的时间特征。通过Softmax函数提供最后九个施工动作的得分，完成行为分类任务。

上述网络输入数据的形状为 $10 \times 18 \times 2$，表示18个关键点的 x 和 y 坐标，10表示连续10帧的数据（图6）。空间特征提取网络的设计主要依赖于四个FC层，具体网络参数列在表2中。网络中每层的dropout层设置为0.1，以防止过拟合。$s^0, s^1, s^2, \ldots, s^{31}$ 形成特征 $\boldsymbol{s}$，空间特征提取网络输出的特征 $\boldsymbol{s}$ 用作网络下一阶段的输入。

通过提取网络的空间网络特征，可以获得空间特征，并将相应结果输入到时间特征提取网络。时间特征提取网络采用LSTM对时间序列进行处理，不限于固定长度的输入或输出。因此，它不仅可以分析单帧特征信息，还可以结合连续帧进行特征分析，这更有利于解决动作分类问题。本研究使用两层堆叠LSTM网络，具体网络参数列在

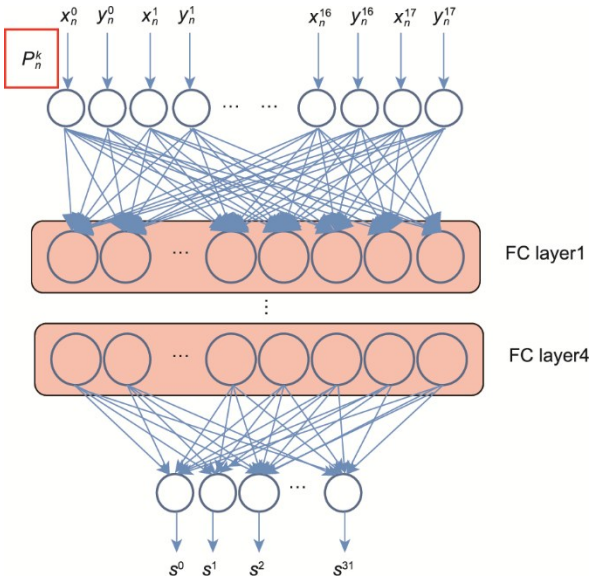

图6. 空间特征提取网络。$s^0, s^1, s^2, ..., s^{31}$ 构成特征 $\boldsymbol{s}$，空间特征提取网络输出的特征 $\boldsymbol{s}$ 用作网络下一阶段的输入。

表2中。维度在括号内标注为（批量大小、时间步长、输入维度）。当LSTM1的输入维度为（32, 10, 32）时，批量大小中每个输入到LSTM模型的尺寸为时间步长 × 输入维度，执行次数为30。上层LSTM的输出序列信息用作下层LSTM的输入信息（图7）。经过空间网络提取的特征经过LSTM层处理后，输出到Softmax激活函数和九个神经元的FC层。这九个输出中的每一个都以交叉熵的形式提供相应动作的概率，其中最大的概率作为当前动作的类别。动作分类网络共有七层，共训练了36 169个参数。与其他神经网络模型相比，计算量较小。

4. 实验和结果

在本节中，对所提出的框架进行测试。使用轻量级姿态检测网络处理视频，并完成工人关键点动作数据集的制作。使用上述数据集训练动作识别网络，以获得动作分类模型。

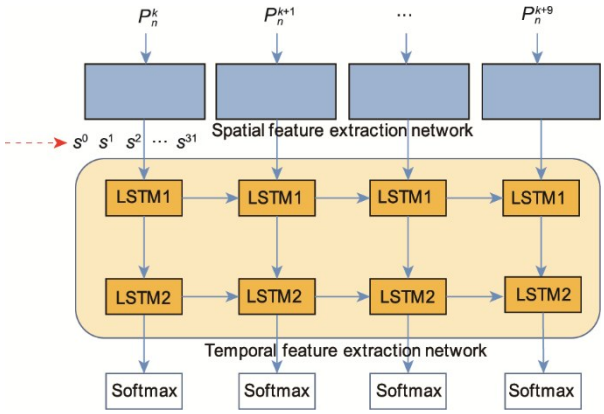

图7. 时间特征提取网络。

4.1. 数据集制作

视频由移动摄像机在多个建筑工地拍摄，每秒30帧，分辨率为1920 × 1280。摄像机不是固定安装的，可以移动到不同的角度记录和观察多工人施工行为，视频中包含多名工人，场景如图8所示。

图8. 施工场景的视频帧。

4.1.1. 包含工人施工行为的视频数据集的制作

在关键点数据集中，为了确保每名工人在每帧对应关键点的动作分类正确，需要创建一个动作视频数据集，使

表2 动作识别网络的参数

Name	Input size	Output size	Number of neurons	Number of parameters	Activation function
FC layer 1	(32, None, 36)	(32, None, 128)	128	4736	ReLU
FC layer 2	(32, None, 128)	(32, None, 64)	64	8256	ReLU
FC layer 3	(32, None, 64)	(32, None, 64)	64	4160	ReLU
FC layer 4	(32, None, 64)	(32, None, 32)	32	2080	ReLU
LSTM1	(32, 10, 32)	(32, 10, 32)	32	8320	Tanh
LSTM2	(32, 10, 32)	(32, 10, 32)	32	8320	Tanh
FC layer 5	(32, None, 32)	(32, None, 9)	9	297	Softmax

The dimensions are denoted as (batch size, time step, input dimension) inside the parentheses in the input and output size columns. ReLU: rectified linear unit; Tanh: hyperbolic tangent function.

每个短视频只包含一名工人和一个动作类别。

为了满足上述视频数据集的要求，需要在空间和时间上对收集的视频进行剪辑。在空间上，逐帧剪辑每个工人边界框对应的图像。在时间上，对连续视频帧中每名工人的相同动作时间帧进行剪辑。剪辑九个动作的短视频，每个视频包含一名工人的动作类别。在本实验中，有九个动作类别，这个数字可以根据不同的实际需求进行调整。不同的类别被划分为施工状态（工人状态=1）和非施工状态（工人状态=0）。具体类别列在表3中。“其他”类别表示与建筑活动无关的类别。

4.1.2. 关键点数据集的制作

在创建短视频数据集之后，使用改进的姿态提取算法将所有短视频转换为关键点数据。关键点信息包含了每张图片中工人的具体位置信息。采样单位为录制的视频帧率，视频中的关键点逐帧提取。通过数据处理，得到了每个短视频片段的连续关键点序列。短视频片段的动作类别即为生成的关键点序列的类别。相同类别的关键点序列放在一起，形成一个长序列。在网络训练期间，使用设置的批量大小来划分整个数据集序列。每个视频片段的持续时间不同，视频数据集中的动作持续时间长短不同。因此，每个关键点序列的长度不固定，长度大约为90。图9显示了关键点数据集，数据经过归一化。通过公式（8）归一化，数据集包含36个坐标点和动作类别编号。原始关键点数据集中的身体关节位置是视频中的像素位置，因此相对于视频的大小进行了归一化。本文中的原始数据根据视频的长度和宽度进行了归一化，因此$x_{\min}$和$y_{\min}$均为0，$x_{\max}$是视频帧宽度w，$y_{\max}$是视频帧高度h。视频尺寸为1920 × 1080，因此$x_{\max}$为1920，$y_{\max}$为1080。

表3 工人动作类别和帧数

Action name	Category number	Worker state	Number of frames
Shouldering materials	0	1	2160
Carrying materials	1	1	2160
Sitting-rest	2	0	2430
Standing-rest	3	0	2520
Squatting-operation	4	1	2430
Squatting-rest	5	0	2160
Standing-operation	6	1	2520
Moving	7	0	2520
Other	8	0	2400

$$\begin{cases} \hat{x}_i = \dfrac{x_i - x_{\min}}{x_{\max} - x_{\min}} \\ \hat{y}_i = \dfrac{y_i - y_{\min}}{y_{\max} - y_{\min}} \end{cases} \tag{8}$$

式中，i是视频中关键点的索引，范围从0到17。

4.2. 动作识别模型的训练

使用关键点数据集训练动作分类网络。本文中循环网

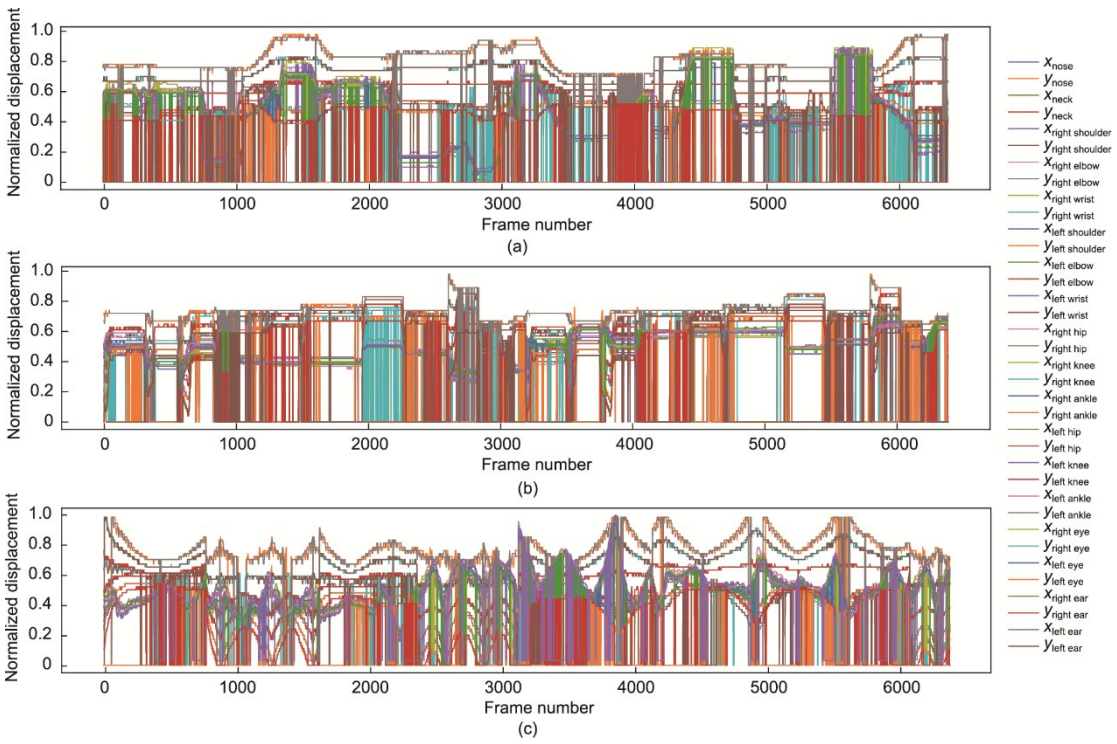

图9. 关键点数据集。（a）站立操作关键点数据；（b）蹲着休息关键点数据；（c）肩扛物料关键点数据。

络的时间步长为10，每个输入到网络的样本序列由10个连续帧的关键点组成，每帧的关键点数据维度为36。实验的关键点总数为63 900，即样本总数为63 900，其中训练集、测试集和验证集的比例为8∶1∶1。本实验的样本采集是连续的，即如果选取时刻k和时刻k+9之间的关键点作为一个样本，下一个样本则由时刻k+1和时刻k+10之间的关键点数据组成。

动作分类网络训练环境是Python 3.6，深度学习框架为Keras 2.2.0，处理器为Intel i7 7700K，显卡为Nvidia GTX 1050ti图形处理器，系统为Windows 10 64位。训练参数设置如下：批量大小 = 32，迭代次数 = 200，初始学习率 = 0.0001。使用Adam优化器控制学习率。当训练达到200次迭代（图10）时，训练损失函数变得稳定，耗时690 s。由于网络参数数量相对较少，与原始图像数据相比，训练时间非常短。训练期间的损失曲线如图10（a）所示，训练损失稳定。

4.3. 动作分类模型的测试结果

在本研究中，使用准确率和混淆矩阵作为性能指标。

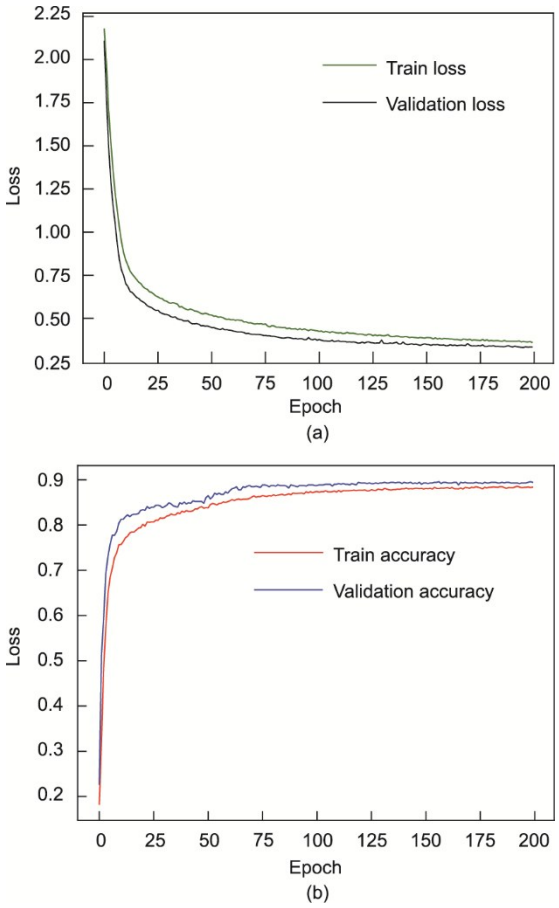

图10. 动作识别模型的训练和验证结果。（a）损失曲线；（b）准确率曲线。

准确率给出正确预测的概率，一直是活动识别中广泛使用的评价方法。混淆矩阵不仅可以完整地显示正确和错误预测的数量，还可以完整地显示相应活动预测错误的结果的具体预测类别。此外，还可以计算每个类别预测结果的准确率。

准确率也可以衡量分类网络做出正确预测的概率。它是动作识别领域中公认的评估指标。图10（b）显示了训练的准确率，约为89%。

混淆矩阵分析了测试数据集上活动识别的准确性，并且可以用来详细分析每个类别的预测情况。在测试数据集上得到的混淆矩阵如图11所示，它显示了每个类别的正确预测数量和其他类别的具体识别情况。测试数据集的总体准确率约为90%。图11中“蹲着操作”类别识别和“其他”类别识别的准确率较低。“蹲着操作”主要被误识别为“搬运材料”和“蹲着休息”类别。这可能的原因是数据本身的相似性和“搬运材料”时下半身遮挡，导致腿部关键点的丢失。由于数据本身存在较大差异，“其他”类别中的动作部分被识别为“蹲着操作”，识别准确率最低。其余类别的准确率高于90%，具有较高的识别准确率。

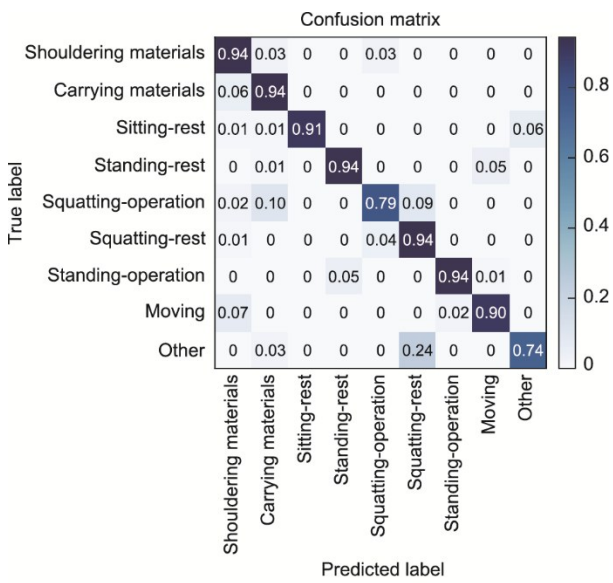

图11. 动作识别混淆矩阵图。

4.4. 实际场景中的活动识别和效率分析结果

在本节中，将所提出的框架在实际场景中进行了验证，包括活动识别和效率分析。首先，跟踪工人的位置是将关键点数据与个体工人编号相关联的必要条件。如第3.2节所述，采用深度SORT算法进行跟踪，实现了关键点与各自工人编号的关联。因此，获得了每名工人在一段时间内的连续关键点数据，用于后续动作识别。如图12

所示，记录了两个场景中几名工人的运动轨迹，用不同的颜色标记了每名工人的运动路径。实验结果证实了前述跟踪方法可以准确地确定每名工人的位置，同时在整个场景中持续而精确地追踪他们的运动轨迹。

在施工过程中，工人可能会执行多种动作。在本文中，施工类的动作被归为状态1（表3）。对于工人来说，实际的工作时间（状态1的帧数之和）可以用来衡量工人的表现。将实际工作时间与总时间（总帧数）相比，可以确定一段时间内工人的工作效率。类似地，可以通过每帧中的建筑工人数量（每帧中状态1的人数）和每一帧中所有工人的工作效率，来计算某一施工阶段监控范围内所有工人的施工进度统计。上述信息可以随时间记录，以分析每名工人的工作效率和同期所有工人的工作效率。

本节使用了一段施工工人监控视频，来测试所提出的框架，解决多工人施工活动识别和工作效率分析统计的问题。为了验证施工活动识别，识别视频中所有工人的行为类别，施工视频中有七名工人。图13显示每位工人活动识别结果，每位工人都由所提出的框架自动分配一个特定的ID号，并显示工人当前帧的活动类别。此外，系统统计了人员活动的类别数量，用于工作效率的统计。

为了获得详细的工人工作效率信息，对图14（a）中的三工人场景进行了进一步统计。这段视频的长度是

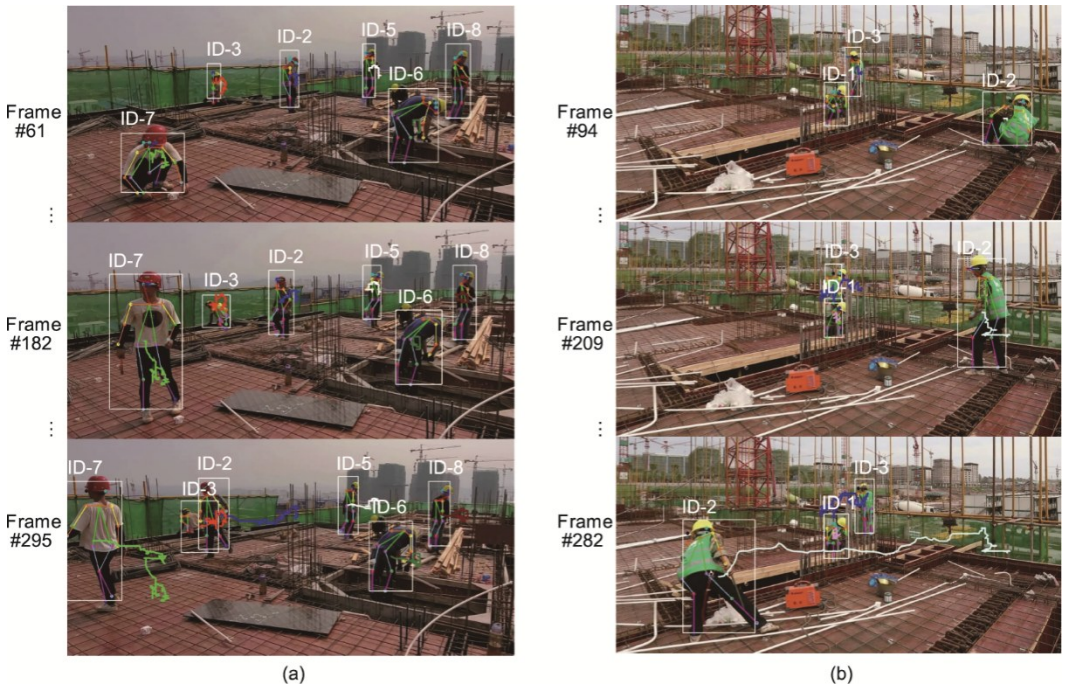

图12. 工人跟踪结果。（a）场景1中跟踪六名工人；（b）场景2中跟踪三名工人。

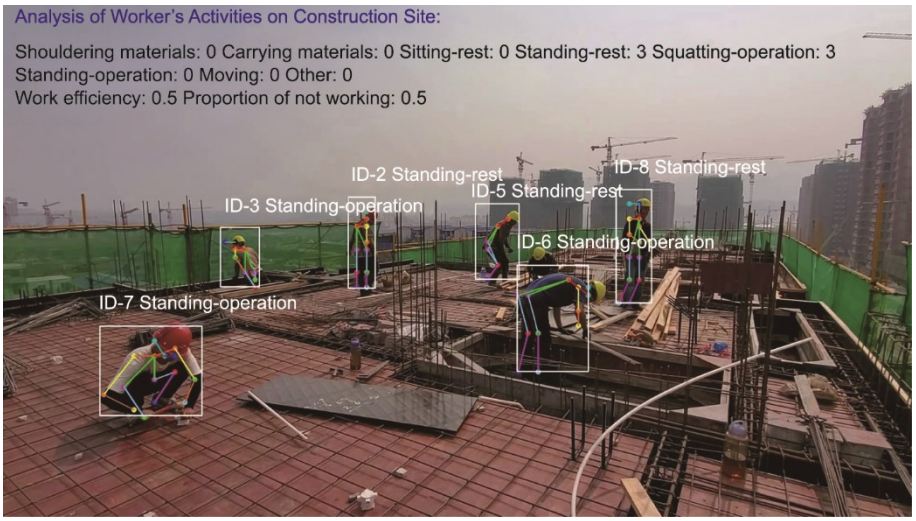

图13. 活动识别结果。视频帧中整体施工效率显示。

20 s，每秒30帧，视频逐帧输入到网络中进行分析。图14（b）中的纵轴表示行为的类别，介于0和8之间，横轴表示帧数。三个数字分别分配给了三名工人，每个数字记录了工人的动作类别。图中的虚线表示由于工人被遮挡而没有检测到关键点。利用每名工人的动作信息，得到了工人效率的统计。工人效率的统计结果如图15所示，其中绿色表示工人的关键点丢失，橙色表示该时间段工人的效率，蓝色表示工人没有工作的时间。工人ID-1和ID-3大部分时间都在休息。在表3中，可以将行为3和行为5分别分类为“站着休息”和“蹲着休息”。工人ID-2多数时间在“搬运材料”和“蹲着操作”，实际工作效率更高。

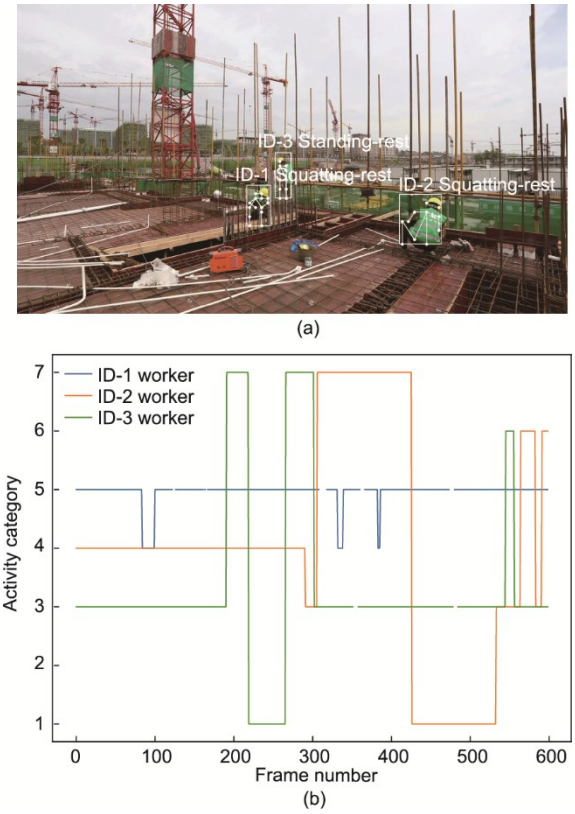

图14. 工人动作识别结果。（a）三工人场景的测试；（b）不同帧中每名工人的动作类别分布图。

4.5. 讨论

与使用目标检测进行不同工人活动识别和安全行为检测相比，本研究不仅将空间信息作为关键点坐标信息，还分析了连续帧的关键点信息，从而更符合活动随时间连续变化的特点。本研究中用于活动识别的方法可以完成端到端训练，便于后期适应数据集数量的增加或活动类别的增加。基于深度摄像机识别人体关键点，易受光线影响且成本高。本研究中，只使用普通摄像机从视频流中获取关键点信息；使用深度排序跟踪算法跟踪每名工人；初始帧不需要手动标记每名工人的检测框。在动作识别过程中，关键点无法识别，导致关键点数据的丢失。过多的数据丢失会导致无法确定工人的工作状态，影响每名工人工作效率的统计。

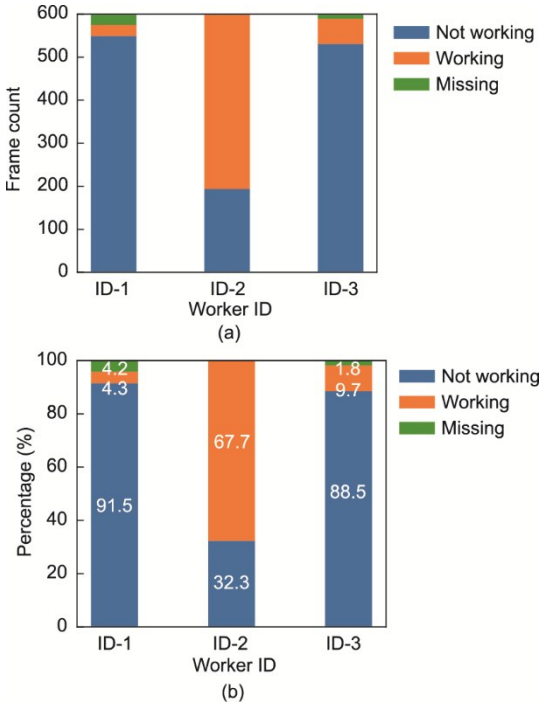

图15. 工人效率的统计结果。（a）不同工人的工作、非工作和缺失的帧数；（b）不同工人的工作效率。

本研究仍有一定局限。首先，建筑工地上有各种类型的动作，这意味着并非所有活动类别都被考虑到，尽管考虑了与建筑活动无关的类别（“其他”标签）。此外，当工人长时间离开摄像机视频范围导致对象跟踪丢失时，所提出的框架将为活动识别的工人分配新的ID编号。

5. 结论

本文提出了一个基于深度学习的活动分析框架，用于处理涉及多名工人的施工信息。该框架集成了关键点提取、跟踪、活动识别和效率分析模块。姿态估计检测器处理从RGB视频流中获得的2D姿态信息。还提出了施工现场的关键点动作数据集，以完成活动分类网络的训练。少量的关键点数据可以代表人类的动作和位置。所提出的框架可以用于视频中的多工人活动的识别和工作效率的统计。通过在从实际施工现场收集的施工监控视频上验证所提出的方法，证明了所提出的框架在监控建筑工人活动方面的有效性。实验结果表明了2D姿态信息对于建筑工人活动分析和效率分析的有效性。

后续研究将集中于通过添加工具检测，更准确、更细致地识别不同工种的工人行为类别。工人行为类别可以更加细化，更有利于施工进度的统计。此外，由于施工现场的场景通常很大，而一个摄像机的范围有限，因此还将探索跨摄像机的工人跟踪和活动识别。行为识别的结果也可以通过引入更多包含不安全行为的行为类别（如违规攀爬、翻越等），及时进行安全分析。后续的研究还将考虑对工人行为的预测；预测危险行为的发生，如工人失去平衡时的跌落预测；对视频数据进行实时分析；以及及时生成预警，提醒工人和管理人员。

致谢

本研究得到了国家自然科学基金（52130801、U20A20312、52178271、52077213）和国家重点研发计划（2021YFF0500903）的支持。

Compliance with ethics guidelines

Xuhong Zhou, Shuai Li, Jiepeng Liu, Zhou Wu, and Yohchia Frank Chen declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Slaton T, Hernandez C, Akhavian R. Construction activity recognition with convolutional recurrent networks. Autom Constr 2020;113:103138.
[2] Kelm A, Laußat L, Meins-Becker A, Platz D, Khazaee MJ, Costin AM, et al. Mobie passive radio frequency identification (RFID) portal for automated and rapid control of personal protective equipment (PPE) on construction sites. Autom Constr 2013;36:38–52.
[3] Cheng T, Teizer J, Migliaccio GC, Gatti UC. Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data. Autom Constr 2013;29:24–39.
[4] Kim J, Ham Y, Chung Y, Chi S. Systematic camera placement framework for operation-level visual monitoring on construction jobsites. J Constr Eng Manage 2019;145(4):04019019.
[5] Yang J, Shi Z, Wu Z. Vision-based action recognition of construction workers using dense trajectories. Adv Eng Inf 2016;30(3):327–36.
[6] Khosrowpour A, Niebles JC, Golparvar-Fard M. Vision-based workface assessment using depth images for activity analysis of interior construction operations. Autom Constr 2014;48:74–87.
[7] Han SU, Lee SH. A vision-based motion capture and recognition framework for behavior-based safety management. Autom Constr 2013;35:131–41.
[8] Yu N, Wang S. Enhanced autonomous exploration and mapping of an unknown environment with the fusion of dual RGB-D sensors. Engineering 2019; 5(1): 164–72.
[9] Luo X, Li H, Cao D, Dai F, Seo JO, Lee SH. Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks. J Comput Civ Eng 2018; 32(3): 04018012.
[10] Luo X, Li H, Cao D, Yu Y, Yang X, Huang T. Towards efficient and objective work sampling: recognizing workers' activities in site surveillance videos with two-stream convolutional networks. Autom Constr 2018;94:360–70.
[11] Pradhananga N, Teizer J. Cell-based construction site simulation model for earthmoving operations using real-time equipment location data. Visualization in Eng 2015;3(1):12.
[12] Montaser A, Moselhi O. RFID indoor location identification for construction projects. Autom Constr 2014;39:167–79.
[13] Akhavian R, Behzadan AH. Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers. Adv Eng Inf 2015;29(4):867–77.
[14] Zhao J, Obonyo E. Convolutional long short-term memory model for recognizing construction workers' postures from wearable inertial measurement units. Adv Eng Inf 2020;46:101177.
[15] Sanhudo L, Calvetti D, Martins JP, Ramos NMM, Mêda P, Gonçalves MC, et al. Activity classification using accelerometers and machine learning for complex construction worker activities. J Build Eng 2021;35:102001.
[16] Valero E, Sivanathan A, Bosché F, Abdel-Wahab M. Musculoskeletal disorders in construction: a review and a novel system for activity tracking with body area network. Appl Ergon 2016;54:120–30.
[17] Yan X, Li H, Li AR, Zhang H. Wearable IMU-based real-time motion warning system for construction workers' musculoskeletal disorders prevention. Autom Constr 2017;74:2–11.
[18] Golabchi A, Han SH, Seo JO, Han SU, Lee SH, Al-Hussein M. An automated biomechanical simulation approach to ergonomic job analysis for workplace design. J Constr Eng Manage 2015;141(8):04015020.
[19] Kanan R, Elhassan O, Bensalem R. An IoT-based autonomous system for workers' safety in construction sites with real-time alarming, monitoring, and positioning strategies. Autom Constr 2018;88:73–86.
[20] Chi S, Caldas CH. Automated object identification using optical video cameras on construction sites. Comput Aided Civ Infrastruct Eng 2011;26(5):368–80.
[21] Seo JO, Lee SH, Seo J. Simulation-based assessment of workers' muscle fatigue and its impact on construction operations. J Constr Eng Manage 2016; 142(11):04016063.
[22] Gatt T, Seychell D, Dingli A. Detecting human abnormal behaviour through a video generated model. In: 2019 11th International Symposium on Image and Signal Processing and Analysis; 2019 Sep 23–25; Dubrovnik, Croatia. Piscataway: IEEE; 2019. p. 264–70.
[23] Wang D, Li W, Liu X, Li N, Zhang C. UAV environmental perception and autonomous obstacle avoidance: a deep learning and depth camera combined solution. Comput Electron Agric 2020;175:105523.
[24] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2012;25:1097–105.
[25] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;39(6):1137–49.
[26] Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Darrell T, et al. Long-term recurrent convolutional networks for visual recognition and description. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7–12; Boston, MA, USA. Piscataway: IEEE; 2015. p. 2625–34.
[27] Fang W, Ding L, Zhong B, Love PED, Luo H. Automated detection of workers and heavy equipment on construction sites: a convolutional neural network approach. Adv Eng Inf 2018;37:139–49.
[28] Kim H, Bang S, Jeong H, Ham Y, Kim H. Analyzing context and productivity of tunnel earthmoving processes using imaging and simulation. Autom Constr 2018;92:188–98.
[29] Fang Q, Li H, Luo X, Ding L, Luo H, Rose TM, et al. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. Autom Constr 2018;85:1–9.
[30] Zhang Q, Barri K, Babanajad SK, Alavi AH. Real-time detection of cracks on concrete bridge decks using deep learning in the frequency domain. Engineering 2021;7(12):1786–96.
[31] Ding L, Fang W, Luo H, Love PED, Zhong B, Ouyang X. A deep hybrid learning model to detect unsafe behavior: integrating convolution neural networks and long short-term memory. Autom Constr 2018;86:118–24.
[32] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997; 9(8):1735–80.
[33] Luo X, Li H, Yang X, Yu Y, Cao D. Capturing and understanding workers' activities in far-field surveillance videos with deep action recognition and Bayesian nonparametric learning. Comput-Aided Civ Infrastruct Eng 2019; 34(4):333–51.
[34] Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, et al. Temporal segment networks for action recognition in videos. IEEE Trans Pattern Anal Mach Intell 2018;41(11):2740–55.

[35] Chen C, Zhu Z, Hammad A. Automated excavators activity recognition and productivity analysis from construction site surveillance videos. Autom Constr 2020;110:103045.
[36] Silva V, Soares F, Leão CP, Esteves JS, Vercelli G. Skeleton driven action recognition using an image-based spatial-temporal representation and convolution neural network. Sensors 2021;21(13): 4342.
[37] Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks. In: Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. Piscataway: IEEE; 2014. p. 1653–60.
[38] Roberts D, Calderon WT, Tang S, Golparvar-Fard M. Vision-based construction worker activity analysis informed by body posture. J Comput Civ Eng 2020; 34(4):04020017.
[39] Cao Z, Simon T, Wei SE, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017 July 21 – 26; Honolulu, HI, USA. Piscataway: IEEE; 2017. p. 1302–10.
[40] Chen B, Hua C, Li D, He Y, Han J. Intelligent human–UAV interaction system with joint cross-validation over action–gesture recognition and scene understanding. Appl Sci 2019;9(16):3277.
[41] Okumura T, Urabe S, Inoue K, Yoshioka M. Cooking activities recognition in egocentric videos using hand shape feature with OpenPose. In: CEA/MADiMa'18: Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management; 2018 Jul 15; Stockholm, Sweden. New York: New York; 2018. p. 42–5.
[42] Chen S, Demachi K. Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph. Autom Constr 2021;125:103619.
[43] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv:1409.1556.
[44] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. Piscataway: IEEE; 2018. p. 4510–20.
[45] Chollet F. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21 – 26; Honolulu, HI, USA. Piscataway: IEEE; 2017. p. 1800–7.
[46] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric. In: Proceedings of 2017 IEEE International Conference on Image Processing; 2017 Sep 17–20; Beijing, China. Piscataway: IEEE; 2017. p. 3645–9.
[47] Wang Z, Zhang Q, Yang B, Wu T, Lei K, Zhang B, et al. Vision-based framework for automatic progress monitoring of precast walls by using surveillance videos during the construction phase. J Comput Civ Eng 2021; 35(1):04020056.