视频技术大作业

题目：*IntelliFlow Insighter*—流水线操作智能分析框架

姓 名	张瑞程
学 号	22354189
院 系	智能工程学院
专 业	智能科学与技术
指导老师	李熙莹

2024 年 6 月

摘要

流水线操作作为现代工业社会的核心生产组织方式，其效率和准确性对企业竞争力至关重要。人工智能和数字化管理的融合，为提升流水线生产效率、降低人为错误提供了强大动力。本项目从实际生产环境出发，开发了一款名为 IntelliFlow Insighter（IFI）的智能分析框架。IFI 主要由检测模块和动作分析模块两部分构成。IFI 能够实时捕捉人工流水线手部操作的视频流，运用先进的目标检测算法，精确检测关键操作对象。随后，结合检测结果与视频内容，IFI 的动作评估模块能够预测和分析操作的类型、流程及执行的标准化程度。为深化模型对人类手部动作的理解，项目还集成了手部关键点检测技术，利用手部位姿信息以实现更精细的手部操作分析。IFI 项目填补了人工流水线操作智能分析领域的空白，并在准确率和响应时间等关键性能指标上取得了优秀的成绩。

关键词：人工流水线操作，视频理解，动作分析，目标检测，手部关键点检测

引言

1.1 研究背景

1.1.1 流水线操作

自工业革命以来，流水线作业以其高效的生产模式，一直是大规模制造的基石。特别是人工操作的流水线，它结合了人的灵活性和机器的重复精度，广泛应用于各类工业生产中。尽管自动化技术为流水线操作带来了革命性的变化，但在某些领域，如服装制造、精密组装等，人工操作因其灵活性和适应性仍然不可替代[1]。然而，随着全球市场竞争的加剧和消费者需求的多样化，传统的人工操作流水线面临着越来越多的挑战。

人工操作流水线依赖于工人的技能和经验，同时受人为因素、环境因素影响较大，导致生产效率和产品质量存在较大波动。世界范围内，每年由于生产者的操作不规范而导致的流水线生产损失高达数万亿元[2]。同时，由于缺乏有效的人工操作和件数监控手段，人工操作的生产/装配线上生产状态和工件比率通常是事后统计的，这对生产线高效智能化管理提出了挑战。因此，如何实现对人工流水线操作的智能监管与评估，成为了制造业亟待解决的问题。

1.1.2 机器视觉技术

数据采集、存储和通信技术的进步促进了视频图像在各种应用中的使用。视频内容分析能够自动分析视频并检测和确定某些事件。在制造环境中，机器视觉技术主要用于质量控制[3]、监控自动化生产线以及测量工人的人体工程学负荷[4]。

机器视觉和深度学习技术，特别是视频行为预测和目标检测算法，为监控和分析工人在生产/装配线上的操作行为提供了全新的解决思路[5]。视频行为分类技术能够识别和分析流水线上工人的行为模式，从而实现对生产流程的智能检测和评估。目标检测技术则能够实时监测流水线上的各个操作对象，确保操作的准确性和安全性。

1.2 相关工作

1.2.1 人工流水线行为分析

现有研究多聚焦于流水线的自动化和机械化改进，而人工生产/装配线上的行为检测与分析尚存在广阔的研究空间。对人工操作

的流水线监测与分析的方法大致可以分为两类：基于结果的分析和基于过程的分析。本文的方法属于基于过程的分析策略。

基于结果的分析：指通过对生产结果的测量、计数、检测等操作间接分析流水线工人的工作状态。P. Pierleoni 等人 [6] 基于 BLOB（二进制大型对象）分析开发了一种机器视觉系统，能够识别、计数操作员传递给下一个操作员的物料。Ping Lou 等人[5]提出了一种基于机器视觉的非接触式监控框架，以实现实时的生产/装配过程监控和重复计数。其优点为技术难度相对较低，但忽视了生产操作的过程，且对于产品的精细化检测通常是事后统计，无法做到实时分析。

基于过程的分析：与只关注生产结果的策略不同，基于过程的分析聚焦生产操作过程，通过捕获动作、声音等过程量实现对流水线生产的实时分析。Dongbao Ma 等人 [7] 提出了一种端到端的基于的物联网的自动化生产在线监控方法，使用生产视频作为输入，直接输出产品数量、工作效率等分析结果，但不涉及对工人个体行为的检测和分析。Tao 等人 [8] 使用惯性测量单元和表面肌电图信号与 CNN（卷积神经网络）来识别工人活动，以及量化和评估工人绩效，但其使用的接触式设备过于复杂。与上述只关注视觉信息的方法相反，Zhang 等人 [9]提出了一种将视频与音频信息相结合的方法，该方法在视觉条件较差的环境中显示出了独特的优势。基于过程的分析方法为模型提供了更为丰富的信息数据，为模型完成复杂任务提供了可能。

1.2.2 基于深度学习的目标检测

目标检测的主流方法大致可分为 one-stage 和 two-stage 两类。对于 two-stage 检测，例如 Fast R-CNN[10]，通过区域建议网络（RPN）生成区域建议，从而实现目标识别的高精度。对于 one-stage 检测，如 YOLO 系列[11]，直接进行对象分类和边界框回归，不使用预先生成的区域建议，因此推理速度比两阶段检测更快。而单次多盒检测（SSD）[12]结合了上述两种方法的优点，使用多个预测分支来检测不同尺度的物体，平衡精度和速度。在我们的应用场景中，对模型推理速度要求较高，因此选择使用 one-stage 的 YOLO 算法。

现有工作在一些特定领域中取得了较好的效果，但存在以下主要问题：

- 模型智能化程度不足，大多数工作仅限于重复计数、过程计时等简单任务。
- 人工流水线上，生产/装配操作主要通过人手完成，而现有模型忽视了对手部关键信息的有效利用。
- 未能使用最前沿的机器学习算法，模型性能存在较大提升空间。

技术方法

2.1 方法介绍

本文提出了一个基于过程的双阶段人工流水线操作分析评估模型 IntelliFlow Insighter（IFI）。IFI 接收由 MPV 软件预处理后的视频流，在第一阶段并行化输入目标检测网络和手部关键点检测网络，得到该帧图像中的关键物体信息和人手位姿信息，随后二者连同视频流一起输入第二阶段的行为分类网络，最终得出视频中操作的种类、流程以及标准度。

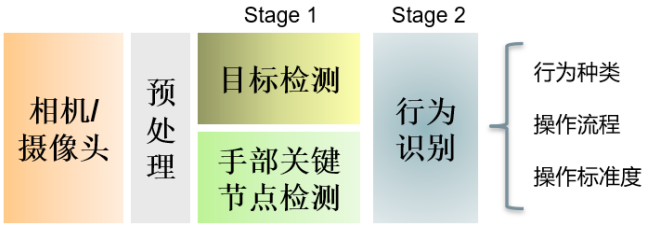

图 1：工作流图

2.2 创新与贡献

本文的主要贡献如下：

（1） 提出了一种新颖的基于机器视觉的非接触式在线实时感知模型 IFI，用于监控、分析生产/装配线工人的手工操作，并取得优秀效果。

（2） 提出了一种两阶段式的分析方法，通过复杂任务解耦，分阶段完成子任务，在简化任务的复杂性的同时，为分析提供了更为丰富的信息。

（3） IFI 使用模块化设计，每个子任务对应一个独立的模块，易于移植和更换。

（4） 据我所知，本项目为第一个使用机器视觉技术，对人工流水线手部操作进行智能监测和分析的工作。

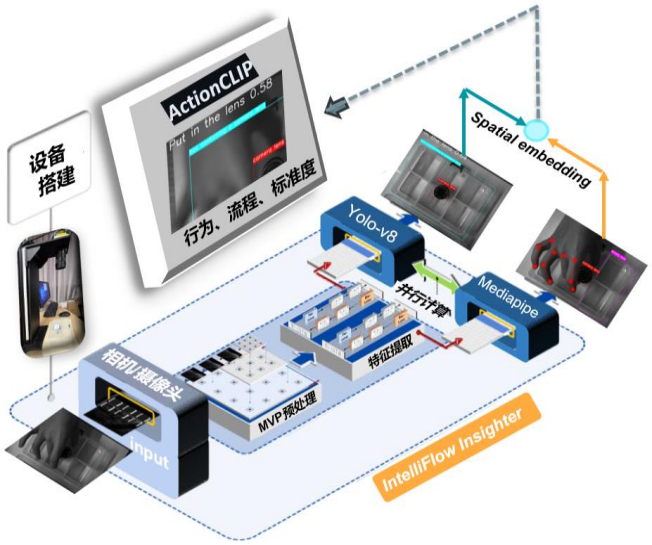

图 2：模型架构：IntelliFlow Insighter 接收实时视频数据，经 MPV 软件的图像预处理后输入特征提取网络，随后并行化进行关键物体检测和手部关键点检测。检测结果通过 Spatial Embedding 的方法与原始视频信息融合，共同输入动作分析模块。

方案细节

3.1 第一阶段方案细节

在第一阶段工作中，IFI 将完成对流水线操作中关键物品的检测和人手位姿检测。为提高运算效率，两任务采用并行化计算方法，视频流同时输入物品检测和手部关键点检测网络（如图 2），完成第一阶段任务。

3.1.1 关键物品检测

在流水线作业中，操作对象的齐全是正确完成流水作业的基本保障，因此，对关键物品的检测和识别是必要的。在 IFI 模型中，评估工人操作是否专业规范的最低标准为：在操作流程中使用了全部的必要物品。此设计旨在防止生产中出现遗漏操作环节、偷工减料等现象。

YOLOv8[13]基于深度学习和计算机视觉领域的尖端技术，在检测速度和准确性方面具有无与伦比的性能。IFI 使用 YOLOv8 的 backbone 作为公共特征提取模块，并在随后的物品检测分支中接入 YOLOv8 的 Decoupled Head，输出检测框和物品类别。

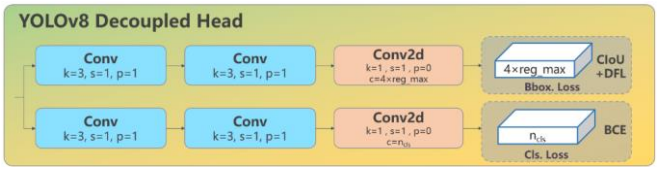

图 3：YOLOv8 Decoupled Head

3.1.2 手部关键点检测

IFI 的手部关键点检测使用 MediaPipe 工具包[14]实现。MediaPipe 是一个由 Google 开发的开源框架，专门用于构建应用机器学习模型的多媒体处理管道。它支持多模式数据，例如视频、音频和任何时间序列数据。在手势识别方面，MediaPipe 提供了一个专门的解决方案，名为 mediapipe.hands，它能够实时检测和跟踪手部在三维空间中的位置和运动。IFI 的手部关键点检测模块中定义了一个 mediapipe.hands 模型类，输入视频流，以字典的形式输出手部位置坐标。

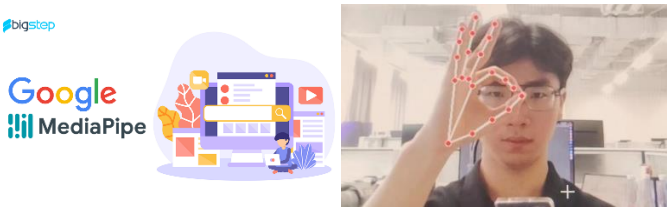

图 4：MediaPipe 展示（作者本人）

3.2 第二阶段方案细节

在第二阶段中，IFI 将完成人工流水线作业中的操作动作识别，和操作标准度评估。

3.2.1 操作动作识别

人工流水线作业中，一个任务往往可以划分成多个具有一定逻辑顺序的子环节，每一个子环节的最小单元为一项基本行为动作，保证子环节的完整无缺和操作规范对完成整体任务至关重要。因此在 IFI 模型中，评估工人操作是否专业规范的二级标准为：子环节数量完整、置信度高。对于一些对操作顺序敏感的流水线作业，IFI 还可以在评估时考虑子环节的顺序是否正确。

视频动作识别的传统方法要求神经模型执行经典的 1/N 多数投票任务。它们经过训练可以预测一组固定的预定义类别，但这限制了它们在存在未曾见过的动作的新数据集上的可转移能力。ActionCLIP[15]通过重视标签文本的语义信息而不是简单地将它们映射到数字，为动作识别提供了新的视角。具体来说，ActionCLIP 将此任务建模为多模态学习框架中的“视频-文本”匹配问题，通过语义语言监督加强了视频表示，并使模型能够胜任零镜头动作识别。

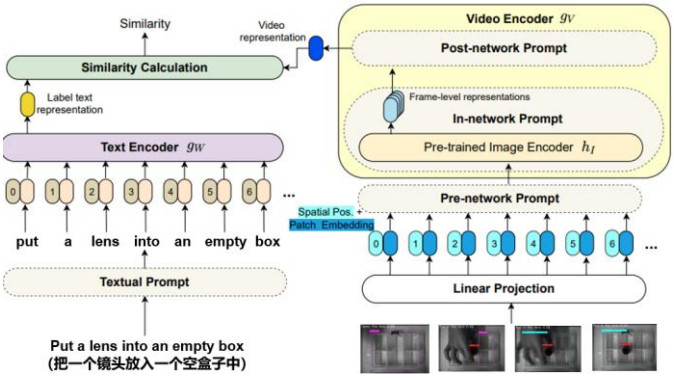

图 5：模型原理：ActionCLIP 由文字文字提示(左)和视频分析(右)两部分组成，通过对文字和视觉双模态信息的深刻理解，实现对视频内容的准确预测，上图展示了本文实验场景下模型工作原理。

IFI 的操作动作识别模块使用 ActionCLIP 实现，输入包括原始视频流和在第一阶段得到的物品检测信息和手部关键点信息。物品检测信息和手部关键点检测信息以字典形式存放（图 6），首先需要对其进行编码操作。IFI 在此处使用空间位置编码（Spatial embedding）[16]，将编码信息与图像 Patch 进行可学习性相加后输入 VIT -B/16 的骨干网络。在此模块中，本文经过尝试后设定滑动窗口数量为 4，也就是说对于每帧视频的行为识别，模型将参考前后 4 帧（包括自己）图像后给出最终预测结果。

```
classification {
  index: 0
  score: 0.6538159847259521
  label: "Left"
}
```

```
hand_landmarks: landmark {
  x: 0.27405625581741333
  y: 0.4661455750465393
  z: 3.558965389061086e-08
}
```

图 6：第一阶段检测结果的数据结构

3.2.2 操作标准度评估

在模型前向推理阶段，物品识别模块和动作识别模块会分别将检测结果存入“物品列表”和“动作列表”中，在推理结束阶段对全流程进行标准度评估。

本项目设计了简单有效的操作准确度评估算法，具体流程如下图所示。从大方面来看，IFI 通过判断流程中关键物品和关键动作的完整性将操作划分为“合格”和“不合格”两大方面，在合格的基础上，会进一步对操作标准度进行评估。标准度估计使用 ActionCLIP 输出的预测置信度进行，跟据实验得出的经验数据，IFI 将置信度 0.6 对应为标准度满分（100），置信度大于 0.6 的情况按 0.6 计算，计算公式为：

$$标准度 = （置信度 + 0.4）\times 100$$

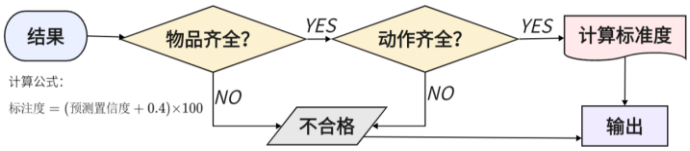

图 7：标准度评估算法流图

实验与结果

4.1 实验设备

本项目使用大华工业相机平台完成流水线模拟和实时数据采集。相机型号为 Dahua AX7B96MG051，焦距 60 毫米，最大光圈 F3.5，镜头直径 46 毫米，镜头距底座约 35 厘米。在实验开始前，首先按照如图 8(c)方式安装实验设备并完成设备间接线。然后进入 MV Viewer 软件，在左侧“设备列表”栏修改相机 IP 并完成主机与相机间的连接（注意※需要将工业相机设置成和 MV Viewer 运行电脑相同的网段），连接成功后，右侧窗口将会实时显示相机画面。在断开后 MV Viewer 与相机的连接后，进入 MVP 软件，添加已连接的镜头，通过逻辑模块对视频流进行预处理操作（彩色转灰度、对比度增强、平滑滤波），随后使用 OpenCV 库将处理后的视频流导入 python 程序，进行检测和分析。

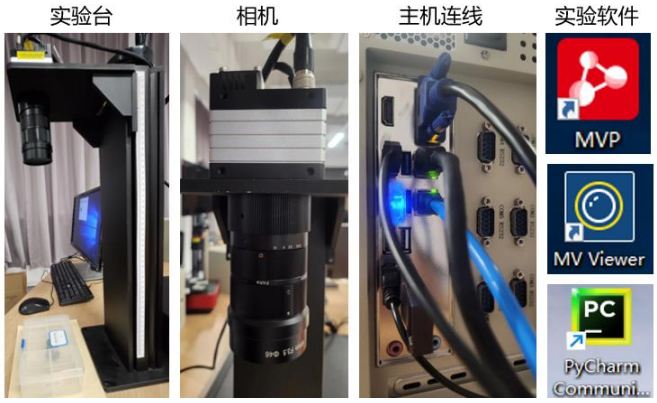

图 8：实验设备与软件

4.2 数据准备

由于模型面向全新的应用场景，本项目需要自己构建模型训练、评估的数据集。囿于时间、资源限制，项目无法得到真实流水线场景下的大量数据，而是通过实验室中的操作任务模拟了人工流水线作业。本人使用上述实验平台采集了 261 张流水线作业模拟图片，使用 Makesense.ai 网站[17]对图片进行标注，即可得到含有检测框金标准的 txt 文件（与 YOLOv8 的数据标注格式完全一致，为后续训练提供了极大的便利），随后按照 4：1 比例划分为训练集和测试集。

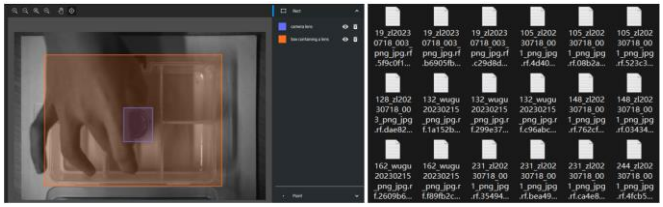

图 9：Makesense.ai 是一个在线的图像标注工具，它允许用户通过网页版界面进行目标检测和图像识别的标签制作。这个工具支持多种标签导出格式，如 YOLO、VOC XML、VGG JSON 和 CSV。

4.3 模型训练

为了解决训练数据不足问题并利用海量网络数据，IFI 使用“预训练—微调”的训练范式。具体来说，YOLOv8 首先使用公开数据集 Coco 2017 中的海量数据进行预训练（未使用全部 Label），在获得较好检测性能后，使用本项目自己构建的目标数据集进行微调；类似的，ActionCLIP 首先加载原作者提供的 Kinetics-400[18]数据集上的预训练模型参数，然后使用目标数据集进行微调，从而实现平滑的任务迁移。而手部关键点检测模型已完全集成于 python 的 Mediapipe 工具包中，由于人手形态具有高度一致性，因此在本项目中无需重新训练，直接调用即可（封装在 Mediapipe.solutions.hands 类中）。

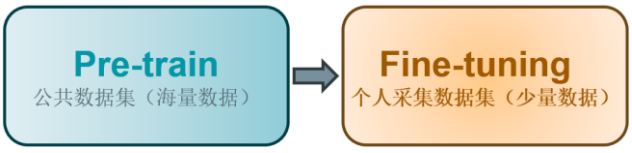

图 9：训练策略。考虑到实验中收集到的数据量较小，IFI 的物品检测和动作分析模块采用“公开数据集预训练+个人数据集微调”的训练策略

IFI 模型使用一张 NVIDIA RTX 3090 GPU 训练，在微调阶段，YOLOv8 和 ActionCLIP 分别训练 1000 轮和 250 轮。

4.4 实验结果

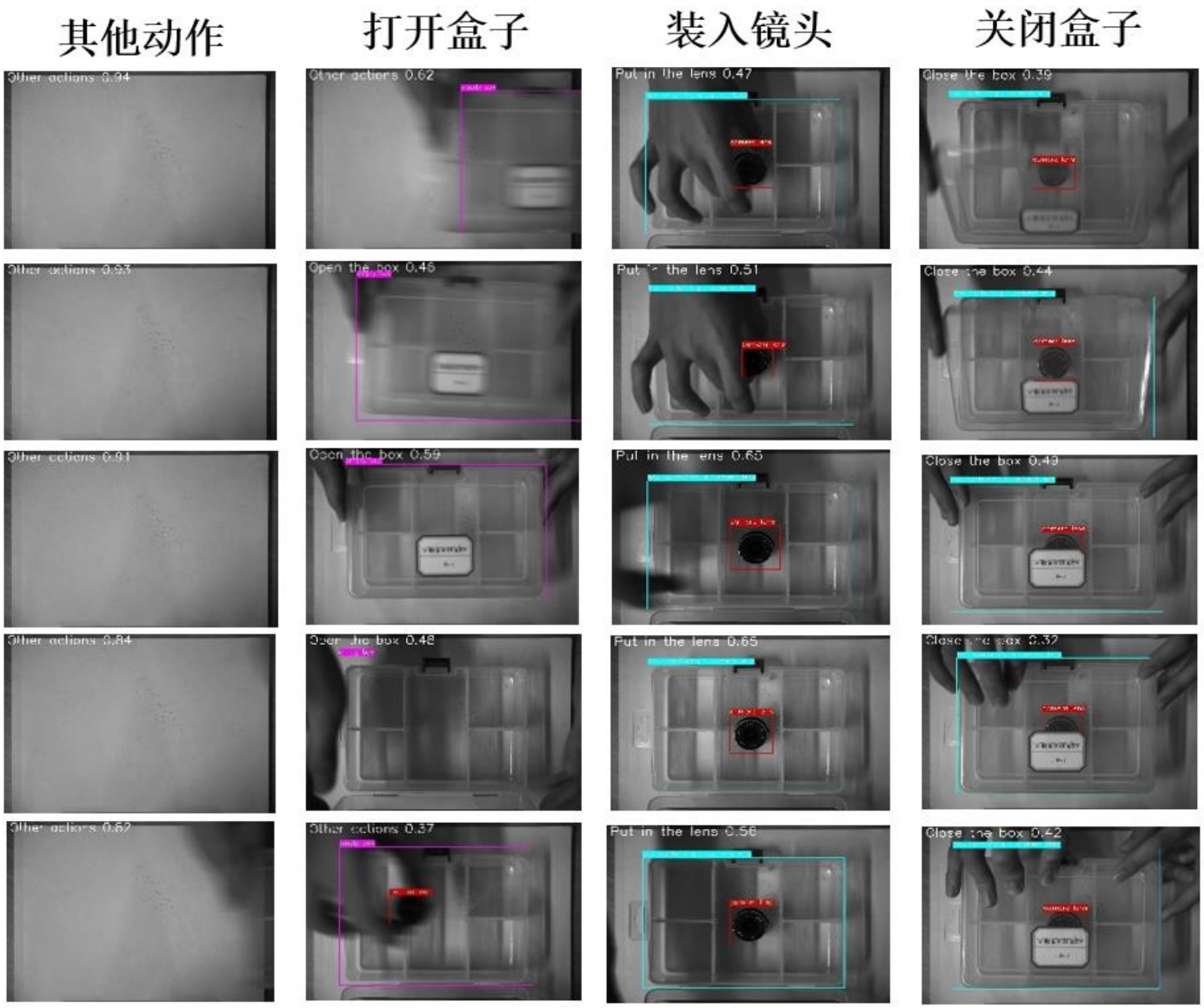

图 10：模拟流水线分析结果展示。图片右上方输出该段视频动作检测结果以及置信度，图像内包含关键物品检测框

4.4.1 动作流程分析

本项目使用了一个长达 20s 的流水线模拟操作视频进行评估，结果如图 10 所示。视频模拟了在精密仪器装配线上，流水线工人将相机镜头装箱的操作。该模拟视频中关键物品有三个：相机镜头（camera lens）、空盒子（empty box）、装有相机的盒子（box containing a camera lens），操作流程可划分为 4 个子动作：打开盒子（Open the box）、装入镜头（Put in the lens）、关闭盒子（Close the box）、其他动作（Other actions）。在正确的操作流程中，需要保证在视频中检测到全部三个关键物品和 3 个子动作（除去 Other actions），且 3 个子动作的顺序保持逻辑正确。IFI 的输出结果由处理后的视频流（图 10）和命令行输出（图 11、图 12）两部分组成。由上述实验结果可见，IFI 模型准确的检测出了各物品和动作，且动作顺序正确，返回结果如下图所示。（Other action 不加入动作列表）

```
结果{
    物品列表： ['empty box', 'camera lens', 'box containing a camera lens']
    动作列表： ['Open the box', 'Put in the lens', 'Close the box']
    平均动作置信度： 0.49
    操作标准度： 81.67
}
```

图 11：输出结果 1

为检测 IFI 对错误操作的监测能力，实验将上述视频强行截去了前 10s，即“打开盒子”动作消失。将截断后的视频输入 IFI 模型，返回结果如下图所示，可以发现，物品列表中缺少了“empty box”，动作列表中缺少了“Open the box”，并触发了 warnning 警告。

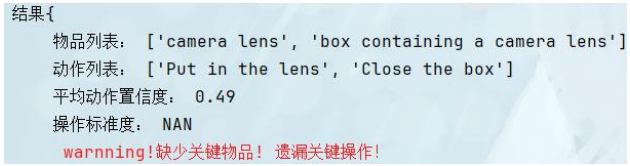

图 12：输出结果 2

4.4.2 检测性能

得益于 YOLOv8 强大性能和预训练-微调策略，关键物品检测模块取得了较好的性能，在测试集上的检测准确率为 92.34%，检测框平均 IoU(交并比)为 0.69。但手部关键节点检测效果略有逊色，主要表现在检测置信度低、准确率不足（如下图所示），这可能是因为 Mediapipe 工具包的手部关键点识别基于 RGB 彩色图片训练，而本项目中视频采用了灰度处理导致手部失去原有颜色信息，对 Mediapipe 的识别造成了干扰。归功于 IFI 的模块化的设计，后续仅需要替换使用更为强大的手部识别模型并进行针对性训练，即可解决此问题。

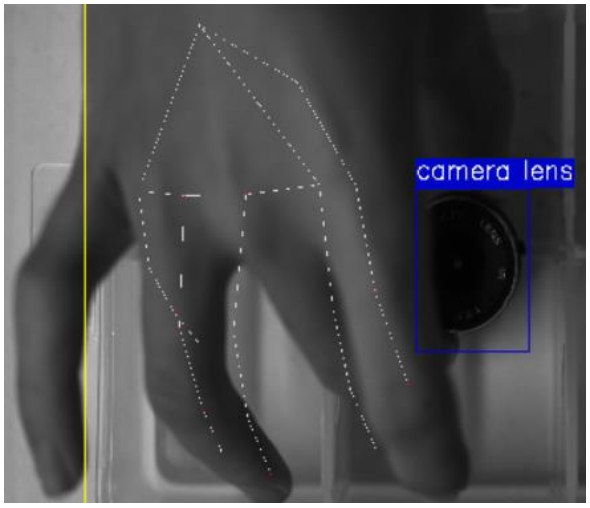

图 13：手部关键点检测异常结果展示

平均准确率	平均 IoU	平均速度
92.34%	0.69	0.68s

表 1：检测性能指标

4.5 今后工作

本文提出了一种基于视觉的人工流水线操作自动分析系统，在初步实验中取得了令人满意的效果。

进一步的研究将集中在加速视频分析上，以发展成为近乎实时的分析工具。此外，手部关键点检测的效果存在较大提升空间，使用更为先进的检测模块对模型性能的进一步提升具有重要意义。最后，本文中使用的数据集是在实验室中创建的，并且相当有限。为了展示所开发工具的真正价值，应使用真实工业案例进行研究。

结论

本文提出了一种利用机器视觉监控生产/装配线手动操作的全新框架 IFI，并设计了一种两阶段检测分析方法。在两阶段方法中，首先利用 YOLOv8 对每个采集的视频帧进行关键物品检测，随后检测结果连同视频流一起输入动作分析网络实现操作分析和标准度评估。同时 IFI 尝试使用人手关节位姿信息以加强模型对手部精细化操作的理解。IFI 项目填补了人工流水线操作智能分析领域的空白，并在准确率和响应时间等关键性能指标上取得了优秀的成绩。

引用

[1] Wenjin Tao, Ming C. Leu, Zhaozheng Yin, Multi - modal recognition of worker activity for human-centered intelligent manufacturing, Engineering Applications of Artificial Intelligence, Volume 95, 2020, 103868, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2020.103868.

[2] https://www.36kr.com/p/2067615522831489.

[3] N. Sato, Y. Murata, Quality control schemes for industrial production by workers' motion capture, Proceedings of the 22nd International Conference on Advanced Information Networking and Applications - Workshops (aina workshops 2008), Okinawa (2008).

[4] D. Mavrikios, M. Pappas, M. Kotsonis,V. Karabatsou, G. Chryssolouris, Digital humans for virtual assembly evaluation, V.V. Duffy (Ed.), Digital Human Modeling, Springer-Verlag (2007), pp. 939-948.

[5] Ping Lou, Ji Li, YuHang Zeng, Bing Chen, Xiaomei Zhang, Real-time monitoring for manual operations with machine vision in smart manufacturing,Journal of Manufacturing Systems,Volume 65, 2022, Pages 709-719, ISSN 0278-6125, https://doi.org/10.1016/j.jmsy.2022.10.015 .(https://www.sciencedirect.com/science/article/pii/S0278612522001868).

[6] P. Pierleoni, A. Belli, L. Palma, M. Palmucci and L. Sabbatini, "A Machine Vision System for Manual Assembly Line Monitoring," *2020 International Con-ference on Intelligent Engineering and Management (ICIEM)*, London, UK, 2020, pp. 33-38, doi: 10.1109/ICIEM-48762.2020.9160011.

[7] D. -b. Ma and M. -f. Qu, "Research on on-line Monitoring Method of Automatic Production Line based on Industrial Internet of Things," *2020 IEEE International Conference on Industrial Application of Artificial Intelligence (IAAI)*, Harbin, China, 2020, pp. 492-496, doi: 10.1109/IAAI51705.2020.9332863.

[8] Tao W., Lai Z.-H., Leu M.C., Yin Z. Worker activity recognition in smart manufacturing using IMU and semg signals with convolutional neural networks Procedia Manuf, 2351-9789, 26 (2018), pp. 11591166, 10.1016/-j.promfg.2018.07.152.

[9] Zhang Y, Shao L, Snoek C G M. Repetitive activity counting by sight and sound[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 14070-14079.

[10] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[11] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[12] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[13] https://github.com/ultralytics/ultralytics

[14] https://github.com/google-ai-edge/mediapipe

[15] M. Wang, J. Xing, J. Mei, Y. Liu and Y. Jiang, "ActionCLIP: Adapting Language-Image Pretrained Models for Video Action Recognition," in *IEEE Transactions on Neural Networks and Learning Systems*, doi: 10.1109/TNNLS.2023.3331841.

[16] Alberto Belussi, Sara Migliorini, and Ahmed Eldawy. 2022. Spatial embedding: a generic machine learning model for spatial query optimization. In Proceedings of the 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22). Association for Computing Machinery, New York, NY, USA, Article 26, 1–4.https://doi.org/10.1145/3557915.3560960.

[17] http://www.makesense.ai

[18] Carreira J, Noland E, Banki-Horvath A, et al. A short note about kinetics-600[J]. arXiv preprint arXiv:1808.-01340, 2018.