文章编号:1671-4598(2022)06-0182-08 DOI:10.16526/j.cnki.11-4762/tp.2022.06.028 中图分类号:TP391.4 文献标识码:A

基于关键姿态的快递场景人一物交互行为识别方法

王苡蓉[1,2], 吴静静[1,2]
(1. 江南大学 机械工程学院, 江苏 无锡 214122;
2. 江苏省食品先进制造装备技术重点实验室, 江苏 无锡 214122)

摘要: 开箱验视是邮局快递场景中的一个重要环节, 为了防止包裹内存在易燃易爆等危险品, 快递打包前工作人员需按照行业要求实施危险品开箱验视; 在人体行为识别框架中引入目标检测和关键姿态估计算法, 提出了基于深度学习的快递场景人一物交互行为识别算法; 首先, 通过改进高斯混合 (GMM, gaussian mixture model) 背景建模方法检测运动目标, 提取行为关键帧, 采用 OpenPose 算法进行姿态估计, 识别初始行为类别; 其次, 针对常规行为识别方法丢失物品语义信息的问题, 使用 YOLOv5 算法检测感兴趣物体类别和位置, 提出基于拍卖算法 (Auction) 的多人一多物最优分配算法, 构建人一物交互关系特征描述子; 最后, 将初始行为标签和人一物交互关系描述子进行决策融合得到最终识别结果; 以实际快递场景数据对所提方法进行验证分析, 实验结果表明, 该方法可以对相似目标和多人干扰的复杂环境中的开箱验视行为进行准确识别。

关键词: 快递验视; 人物交互行为识别; 目标检测; 姿态估计; 深度学习

Human-object Interaction Behavior Recognition Method in Express Scene Based on Key Gesture

WANG Congrong[1,2], WU Jingjing[1,2]
(1. School of Mechanical Engineering, Jiangnan University, Wuxi 214122, China;
2. Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology, Wuxi 214122, China)

Abstract: Unpacking inspection is an important part in the express scene. In order to prevent the presence of flammable and explosive dangerous goods in the package, staff must carry out the unpacking inspection of dangerous goods according to the industry requirements before packaging. Target detection and key pose estimation algorithms are introduced into the human action recognition framework, and a deep learning-based human-object interaction behavior recognition algorithm is proposed. Firstly, the background method of improving the Gaussian mixture model (GMM) is used to detect moving objects, and extract the key frames of behavior, by using the OpenPose algorithm for pose estimation, the initial behavior categories are identified; Secondly, the algorithm for behavior recognition misses the problem of item semantic information, using the YOLOv5 algorithm detects the type and location of the interest object, a multi-person-multi-object optimal allocation algorithm is proposed based on the auction algorithm, which constructs the feature descriptor of the human-object interaction relationship; Finally, the initial behavior label and the human-object interaction relationship are fused. the final recognition result is obtained. Taking the actual express scene as an example to verify and analyze the proposed method, the experimental results show that the method can accurately identify the unpacking inspection behavior in the complex environment with similar targets and multi-person interference.

Keywords: express inspection; human-object interaction behavior recognition; target detection; attitude estimation; deep learning

0 引言

近年来, 随着电商时代的到来, 快递行业也随之蓬勃发展, 我国快递点数量剧增, 快递揽收和运输过程中的安全问题受到日益关注。快递行业要求快递揽件时工作人员必须在快递箱封闭前进行一次开箱查验, 以确保货物能够安全送达目的地。与传统人工视频监控相比, 智能视频监控技术可以高效识别异常或者危险行为[1-2], 对快递工作人员的行为进行识别和预警, 极大提高了监控效率和监督的有效性[3]。因此, 快递场景下异常行为识别方法的研究对于实现智能安全快递的目标具有重要意义。

近年来, 人体行为识别被广泛应用于智能视频监控等日常生活场景中, 目前国内外现有的行为识别研究更多聚焦于单人和多人行为[4-7]。在危险行为识别方面, Guan[8]利用 3D-CNN 结合 LTSM 进行异常行为识别; Xu[9]等人通过提取视频的底层特征, 实现了对视频中暴力行为的检测; 吴蓬勃[10]等人基于 TensorFlow 深度学习框架, 使用

收稿日期:2022-03-21; **修回日期**:2022-03-23。
基金项目:国家自然科学基金项目(62072416);国家自然科学基金项目(61873246)。
作者简介:王苡蓉(1997-),女,浙江杭州人,硕士研究生,主要从事机器视觉与图像处理等方面的研究。
通讯作者:吴静静(1982-),女,安徽滁州人,博士,副教授,硕士生导师,主要从事图像处理与模式识别等方面的研究。
引用格式:王苡蓉,吴静静.基于关键姿态的快递场景人一物交互行为识别方法[J].计算机测量与控制,2022,30(6):182-189.

PoseNet 模型采集数据，通过 LSTM 实现了快递暴力分拣行为的识别。在动物行为识别方面，Wang[11] 等人采用 YOLOv3 模型，基于深度图像分析技术研究了一种针对蛋鸡行为的自动识别方法；Yang[12] 等人利用深度学习实现了猪行为的识别。以上方法在行为识别应用中效果较好，然而对于快递场景下开箱验视异常行为分析问题，仅使用人体运动信息描述行为往往会引起较大的识别错误率[13]，易受到复杂背景、光照变化以及寄件人行为等干扰。在快递场景中，开箱验视属于人－物交互行为，仅仅依靠人手部的骨骼和关节信息难以区分相似动作，如寄件人人手干扰、其他快递员取物品等，丢失了必要的物体和语义信息。

针对以上问题，本文分析快递场景特点和异常行为特征，提出了一种基于关键姿态的人－物交互行为识别方法。针对场景内的背景干扰和信息冗余，提出一种基于 GMM 的关键帧提取算法；针对基于 OpenPose 骨骼点的人体行为识别方法丢失上下文场景和语义信息的问题，引入目标检测方法确定感兴趣的目标物，获得目标位置和类别；针对多目标行为识别问题，提出基于 Auction 的人－物最优分配方法，确定人－物关系描述子和关键姿态向量；最后，将行为识别和人－物关系进行融合决策得到最终识别结果。

1　本文方法概述

本文提出的基于关键姿态的快递场景异常行为识别方法流程如图 1 和图 2 所示。对于网络摄像头采集并传回的视频流，首先用改进的高斯混合背景建模方法建立监控场景背景模型，检测运动目标，根据运动目标的面积阈值判断是否是关键帧；对关键帧使用 OpenPose 计算获得骨骼点和肢体特征向量，输入深度学习行为分类网络得到人体最初行为模式和位置；使用目标检测算法对关键帧中的物体进行检测和分类，获得物体类别和位置，然后提出最优分配算法获得人－物关系特征描述子和关键姿态；最后将人－物交互关键姿态特征和最初的特征识别进行融合决策得到最终行为识别结果。

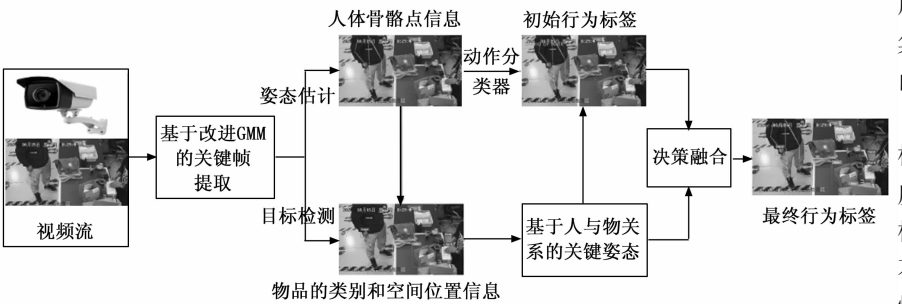


图 1　基于关键姿态的快递场景人－物交互行为识别方法

2　基于关键姿态的人－物行为识别

2.1　基于改进高斯混合模型的关键帧提取

在视频序列中由关键姿态描述的行为状态对于分析识别人的行为更有意义[14]，同时为了减少数据冗余和计算负载，本文提出基于改进高斯混合模型的关键帧提取方法。

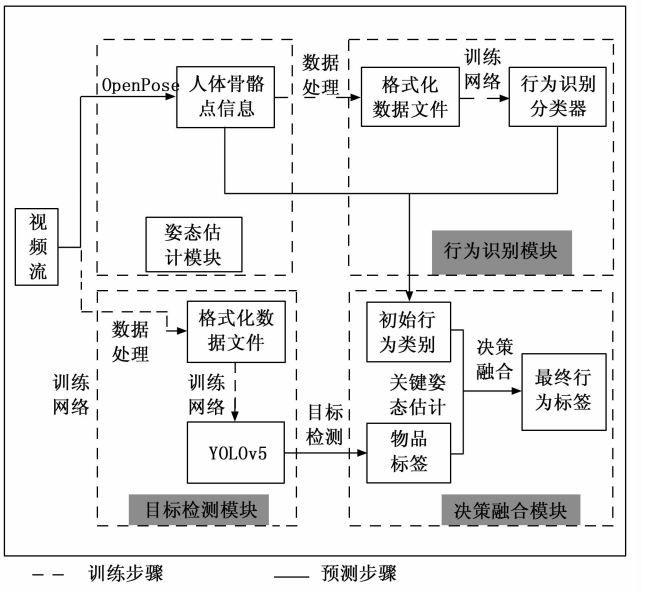


图 2　基于关键姿态的人－物交互行为识别流程

在传统高斯混合模型中，在学习的过程中学习率是固定不变的，因此在一定时间后运动目标对应的高斯分布权值会上升，逐渐更新为背景分布，这样会导致运动目标出现孔洞然后消失，尤其是运动速度较慢的目标。因此本文将运动目标的速度 v 与像素点的学习率 $\alpha_{x,y,t}$ 相关联，作为动态变量对其进行动态调整。本文定义的运动速度 $\upsilon_{x,y,t}$ 的数学表达式如式（1）所示。使用该方法进行运动目标检测，有效地提高了运行速度，加强了动态环境地自适应性。

$$\begin{cases}\upsilon_{x,y,t}=O_{x,y,t}D/\Delta t\\ \sqrt{(y_t-y_{t-1})^2+(x_t-x_{t-1})^2}\end{cases} \tag{1}$$

$$O_{x,y,t}=\begin{cases}1,\text{当前点为前景点}\\0,\text{当前点为背景点}\end{cases} \tag{2}$$

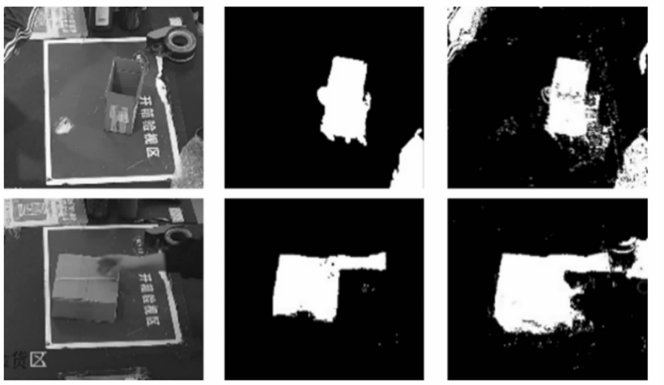
式中，Δt 代表时间间隔，为固定值。x_t 和 y_t 指代的是 t 帧图像中运动目标点集的最小外接矩形的中心像素点的行列序号。对于组成运动目标的前景像素点集合来说，其中每一个点的速度 $\upsilon_{x,y,t}$ 均由相同的速度值来表示。

为了防止固定的更新速率将低速目标识别为背景，学习率 $\alpha_{x,y,t}$ 需要随着速度 $\upsilon_{x,y,t}$ 的变化动态调整[15]。对于高速目标来说，它不会停留在固定区域，也就不存在前景分布逐步转换为背景分布的情况，所以像素点需要保持稳定、较高的学习率；而低速目标则完全相反。定义学习率 $\alpha_{x,y,t}$ 的计算公式如式（3）：

$$\alpha_{x,y,t}=\begin{cases}\dfrac{\upsilon_{x,y,t}}{\upsilon_{x,y,t}+\gamma},\text{当前为前景点且}\ \upsilon_{x,y,t}<\upsilon_0\\ \alpha_{x,y,t-1},\text{其它情况}\end{cases} \tag{3}$$

式中，υ_0 表示速度临界阈值，用于区分高速与低速的运动目标。当一个像素点满足以下条件时，它的学习率会初始

化为初值 $\alpha_{x,y,0}$：1）$t-1$ 和 t 时刻所匹配的分布模型发生了变化；2）连续 5～10 帧速度均为 0。



(a)原图 (b)改进GMM提取的前景图 (c)经典GMM提取的前景图

图 3 改进 GMM 效果对比图

如图 3 所示，图（b）为经过改进高斯混合模型之后获得的前景二值图，该模型相较于经典 GMM 提取的前景图对于消除孔洞等干扰问题有明显的优势。当相机视野范围内无运动目标进入时，检测系统处于待机模式，仅进行视频流与图像帧的获取。$m_{(k)}(x,y)$ 表示经过改进的高斯混合模型之后得出的运动目标前景二值图。当检测区域内出现运动目标时，$m_{(k)}(x,y)$ 中出现大量白色像素点，通过计算白色像素点个数与图片总像素的比例 s，设定特定阈值 V，筛选出大于阈值的帧组成关键帧，然后再进行后续的行为识别。

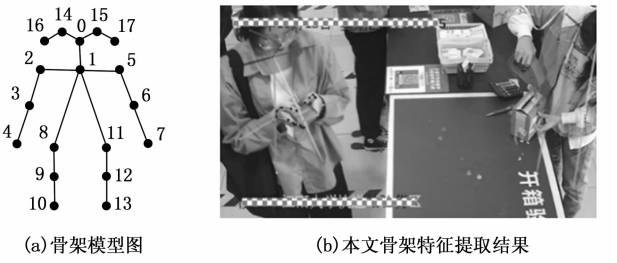
2.2 交互行为人一物状态向量提取

在完成关键帧的检测后，需要提取关键帧内的人体状态向量与物体状态向量，以便后续进一步对人一物状态向量进行匹配融合，进而实现人一物交互行为语义的描述。其中，人体状态向量可以由人体关键点的位置信息、肢体角度和行为标签来表达，物体状态向量可以由物体的类别标签、位置和尺寸来表达。

2.2.1 基于骨架建模的人体状态向量提取

OpenPose 模型[16]是由美国卡耐基大学（CMU）以 Caffe 深度学习框架开发的人体姿态估计项目，可以实现人体关节点的提取与骨架结构的聚合，从而描述人体姿态特征。作为一种自下而上的关节点检测算法，在具有较高检测精度的同时，检测速度具有优越性。如图 4 所示，该模型可以输出 18 个人体关节点信息，包括关节的坐标向量以及置信度信息。

OpenPose 模型以 RGB 图像数据作为模型输入，以 VGG19 模型[17]的前 10 层做基础特征提取，对于提取得到的特征图 F，通过关节位置回归支路（PCM）回归人体关节的位置向量集合 $\boldsymbol{S}=(S_1,S_2,\cdots,S_j,\cdots,S_n)$，$\boldsymbol{S}_j$ 表示第 j 个关节位置的坐标向量，通过关节亲和力大小预测支路（PAF）预测关节之间的亲和力场集合 $\boldsymbol{L}=(L_1,L_2,\cdots,L_c,\cdots,L_n)$，$\boldsymbol{L}_c$ 表示第 c 组关节对之间的亲和力大小分布。两条支路的输出可以表达为：



(a)骨架模型图 (b)本文骨架特征提取结果

图 4 基于 OpenPose 的骨架建模的模型图和结果图

$$\begin{cases}\boldsymbol{S}^t=\rho^t(F,S^{t-1},L^{t-1}) & t\geqslant 2\\ \boldsymbol{L}^t=\varphi^t(F,S^{t-1},L^{t-1}) & t\geqslant 2\end{cases} \tag{4}$$

式中，F 是基于图像数据提取的特征图，ρ^t 和 φ^t 分别表示在阶段 t 的 PCM 支路输出和 PAF 支路输出。对于本文来说，只需要利用 OpenPose 模型 PCM 支路输出的关节坐标向量集合 $\boldsymbol{X}=(X_1,X_2,\cdots,X_{18})$，其中包含了代表人手的关节点的位置向量 $\boldsymbol{X}_k(x_t,y_t)$。

在完成人体关节坐标向量的提取后，需要进一步实现人体状态向量的描述，即人员行为状态的预测。本文采用神经网络模型处理关节向量信息，进而实现行为类别的预测，网络结构如图 5 所示。首先将输入的 18 个关节坐标向量扁平化处理成一维向量，网络整体由 4 个结构相似的全连接层 Block 组成，对于每一个 Block 输出都进行层标准化处理约束模型的参数分布，避免模型误差反向传播时出现梯度爆炸的问题。

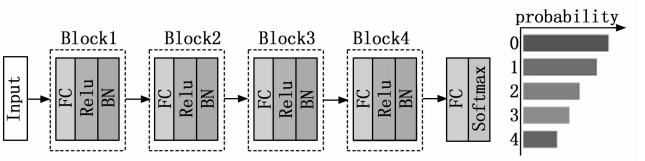


图 5 基于骨骼点的人体行为识别网络

针对行走、开箱、使用手机、包装快递、写快递单等 5 种人体状态描述，对应标签值为："0-Walking""1-Open-Box""2-UsingPhone""3-Packing""4-Writing"。据此为神经网络模型定义 5 个输出向量，并通过 Softmax 函数完成对 5 种状态预测的置信度信息做归一化处理，取最大置信度对应的行为类别为当前人体状态向量描述。

2.2.2 基于 YOLO 的物体状态向量提取

上一节内容中对快递工作人员进行了骨架建模和行为识别，但由于快递员在快递开箱验视的过程中容易产生很多相似的干扰行为，如包装快递、使用手机、写快递单以及使用工具等等，这些行为存在一定的相似性，单单凭借人体行为无法有效区分识别。因此需要对场景中的目标物进行分类与定位，确定物体的状态向量，以便于对后续关键姿态估计以及决策融合提供物品信息。

YOLO 作为一种经典的目标检测算法[18]，将基于图像的目标对象检测问题定义为了一个回归问题，即利用整张图作为网络的输入，直接在输出层回归待检测物体所在区域 Bounding Box 的位置信息以及所属类别信息。作为端到

端的模型结构，YOLO 在检测速度上具有显著优势。利用该模型实现目标检测的流程如图 6 所示。

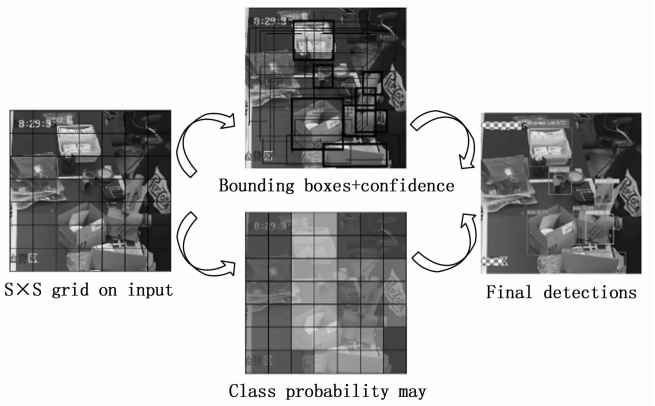


图 6　YOLO 模型目标检测流程示意图

YOLO 模型同样以 RGB 数据作为模型输入，首先将图像划分为 7×7 的网络区域，然后执行两个 Branch 分支。第一个分支进行目标位置框区域估计，即基于每个网格 Cell 给出两个指定宽高比的预测框，输出 Bounding Box 的 4 个顶点坐标与置信度信息，后续基于置信度非极大抑制以及框选位置尺寸矫正实现候选框筛选与位置优化。第二个分支负责预测每个网格 Cell 的所属目标类别，结合第一个分支 Bounding Box 位置的估计结果，实现对待检测目标所在区域的 ROI 位置以及类别预测。

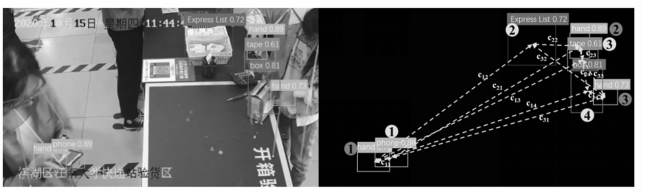
综合考虑各种因素，针对箱子、手机、胶带、快递单和小刀 6 种物体状态向量描述，对应标签值为："1-Box""2-Phone""3-Tape""4-Express List""5-Knife"（初始化类别标签为"0-Nothing"），目标检测的实验结果如图 7 所示。从图中可以看出，该网络可以准确地检测出快递站场景中的目标物品，同时返回被测物品的位置和尺寸。最终得到物品的位置和类别状态向量集 $\boldsymbol{Y}_k$（L_t，x_t，y_t，w_t，h_t），其中，L_t 表示物品标签，x_t 和 y_t 代表 Bounding Box 的中心点坐标，w_t 和 h_t 分别表示 Bounding Box 的宽和高。

2.3　基于 Auction 的关键姿态估计

与单人动作相比，多人－多物交互行为在快递场景中更为常见。如图 8（a）所示，一个常见的快递场景中往往有多个工作人员（揽件员），在工作人员周围还存在多个寄件人，桌面上除了包括多个包装纸箱，还有手机、快递单收纳篮、计算器、胶带卷等，多人和各种类型的桌面物品极大地影响了开箱验视行为识别的准确率。为了提高多人－多物复杂环境下的开箱验视行为识别性能，本文提出一种基于 Auction 的关键姿态估计方法，根据多人和多物的位置和类别状态向量，设计全局最优分配代价函数，推断出开箱验视人－物交互关系候选对集合[19]。

在快递场景中，假设 YOLO 算法检测到物品位置和类别状态向量集为 $\boldsymbol{X}_k$；OpenPose 输出的人手位置状态向量集 $\boldsymbol{Y}_k$。在本文提出的人物交互行为识别算法中，把关键姿态估计问题转化为分配问题，即将当前 k 时刻的物品状态估计 $\boldsymbol{Y}_k$ 分配给人手状态 $\boldsymbol{X}_k$。而分配问题的关键在于设计一个代价函数来衡量物品状态和人手状态的相关性，两个状态估计的相关性越大则人－物的匹配可能性越高。由于目标状态指示的是一个包含位置、标签、附属关系等的向量，要确立两个状态向量间的匹配程度或相似性，需要借助向量特征来进行相似度的度量。本节设计的代价函数的原理为采用人手状态和物品状态的巴氏距离[20]来衡量二者的相似性，相似越大则两个目标状态的距离越小，进而匹配的代价越小。基于 Auction 的关键姿态估计原理如图 9 所示，具体算法可总结如下：

图 7　目标检测识别结果图



（a）行为识别和目标检测结果图　　（b）关键姿态估计逻辑图

图 8　基于 Auction 的关键姿态估计算法

Step1：计算关联矩阵，即提取当前帧任意人手状态和任意物品状态，计算任一确认人手状态和物品状态相匹配的代价函数，并生成关联矩阵。假定 k 时刻 YOLO 输出的物品位置和类别状态估计集合 $\boldsymbol{X}_k$ 包含 $\boldsymbol{N}$（k）个状态估计，而人手状态估计集合 $\boldsymbol{Y}_k$ 有 $L_{(k)}$ 个元素，则对任一当前帧的估计状态 $x_k^i=(p_k^i,\ v_k^i,\ w_k^i,\ h_k^i)^T\in\boldsymbol{X}_k$ 和任一前帧确认状态 $y_k^j=(p_k^j,\ v_k^j)^T\in\boldsymbol{Y}_k$ 计算巴氏距离。如果 $\boldsymbol{N}(k)\geqslant\boldsymbol{L}(k)$，则任一当前人手状态 $y_k^jy_k^j$ 分配给物品状态估计 x_{k+1}^i 的代价函数可定义为：

$$c_{i,j} = exp\left(-\sqrt{1-\rho\left[\dot{q}(p_k^i),\dot{q}(p_k^j)\right]}\right) \tag{5}$$

其中：$\rho\left[\dot{q}\left(p_k^i\right),\ \dot{q}\left(p_k^j\right)\right] = \sum_{u=1}^{m}\dot{q}_u\left(p_k^i\right)\dot{q}_u\left(p_k^j\right)$ 为巴氏系数。遍历计算所有物品状态估计与所有人手状态的关联代价，可以得到一个 $\boldsymbol{N}(k)\times\boldsymbol{L}(k)$ 的二维关联矩阵。当 $\boldsymbol{N}(k)<\boldsymbol{L}(k)$ 时，可以将代价矩阵的计算看成将所有物品状态分配给当前帧人手状态估计的关联代价，则可以得到一个 $\boldsymbol{N}(k)\times\boldsymbol{L}(k)$ 的关联矩阵。实际上，$\boldsymbol{N}(k)<\boldsymbol{L}(k)$ 时的关联矩阵计算可以作为 $\boldsymbol{N}(k)\geqslant\boldsymbol{L}(k)$ 时的逆问题来看待，且二者得到的关联矩阵互为转置矩阵。

Step2：初始化 $\boldsymbol{X}_k$ 中所有未分配成功的状态估计并将关联代价（Price）设定为 0；

Step3：根据公式（6）对任一未分配成功的物品状态 x_k^m 找到与其对应的“最佳”人手 y_k^n。如果所有状态分配成功，则结束计算转到 Step7；

$$c_{mn}-\boldsymbol{P}_n = \max_{j=1,\cdots,L(k)}\left(c_{mj}-\boldsymbol{P}_j\right) \tag{6}$$

Step4：解除上一个循环计算后分配给人手 y_k^n 的状态 x_k^m 而将分配给 y_k^n；

Step5：按照下式更新人手 y_k^n 的价格：

$$\boldsymbol{P}_n = \boldsymbol{P}_n + d_n + \varepsilon \tag{7}$$

其中：d_n 为状态估计 x_k^m 的“最佳”和“次最佳”分配代价的差值，而 ε 为一个设定常数或函数用以防止死循环。

Step6：返回至 Step3；

Step7：输出带有标签的多人－多物配对集 $\boldsymbol{P}_k=\{p_1,\ p_2,\ p_3,\ \cdots,\ p_k\}$，其中，$\boldsymbol{P}_k$ 为包含第 k 个人手状态与所有物品状态价格的向量。

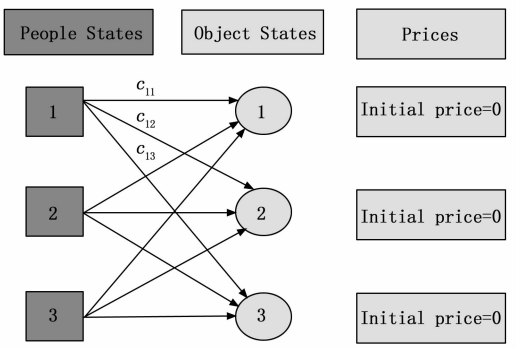


图 9 关键姿态估计原理示意图

2.4 融合决策

在得到带标签的多人－多物配对集 $\boldsymbol{P}_k=\{p_1,\ p_2,\ p_3,\ \cdots,\ p_k\}$ 之后，选择与人手状态关联价格最高的物品状态进行直接配对，得到配对完成后的潜力人－物对，直接与初始行为进行融合决策。融合决策是根据一定的准则，将经过人体行为识别网络输出的行为类别结果、经过 YOLO 检测的物品信息以及通过 Auction 关键姿态估计得到的人物相关性配对集进行融合判断，最终获得人体行为的识别结果，初始行为类别和物体类别的数字标签对应表如表 1 所示。根据初始行为类别和物体类别以及对于几种行为的综合判断，具体的融合决策策略如下：

Case1：当初始行为或者相匹配的物体类别的编号为 0 时，行为类别即为“Irrelevant Behavior”；

Case2：当初始行为与相匹配的物体类别的编号相同时，表示行为类别无需修正；

Case3：当初始行为为 1，若相匹配的物体类别为 2，则将初始行为修正为“Using Phone”；若相匹配的物体类别为 3，则将初始行为修正为“Packing”；若相匹配的物体类别为 4，则将初始行为修正为“Writing”；若相匹配的物体类别为 5，表示行为类别无需修正；

Case4：当初始行为为 2，若相匹配的物体类别为 1，表示行为类别无需修正；若相匹配的物体类别为 3，则将初始行为修正为“Packing”；若相匹配的物体类别为 4，则将初始行为修正为“Writing”；若相匹配的物体类别为 5，则将初始行为修正为“OpenBox”；

Case5：当初始行为为 3，若相匹配的物体类别为 1，表示行为类别无需修正；若相匹配的物体类别为 2，则将初始行为修正为“UsingPhone”。若相匹配的物体类别为 4，则将初始行为修正为“Writing”；若相匹配的物体类别为 5，则将初始行为修正为“OpenBox”；

Case6：当初始行为为 4，若相匹配的物体类别为 1，表示行为类别无需修正；若相匹配的物体类别为 2，则将初始行为修正为“Using Phone”；若相匹配的物体类别为 3，则将初始行为修正为“Packing”；若相匹配的物体类别为 5，则将初始行为修正为“OpenBox”。

表 1 初始行为类别和物体类别的数字标签对应表

ActionResult	ActionNumber	ObjectResult	ObjectNumber
Walking	0	Nothing	0
OpenBox	1	Box	1
Using Phone	2	Phone	2
Packing	3	Tape	3
Writing	4	Express List	4
/	/	Knife	5

在实际场景下，对于快递场景人物交互行为识别的应用意义就是判断工作人员是否进行了开箱验视。因此行为类别为“Open Box”是本文需要重点关注的行为类别，而“Packing”“Using Phone”“Writing”以及“Irrelevant Behavior”这 4 种行为属于快递站常见的其它行为，通过对这几种行为进行识别可以更好地对工作人员的行为进行规范，设计更合理的工作流程，提高快递开箱验视工作效率。

3 实验分析与设计

3.1 实验数据集介绍

本文研究的快递站场景人物交互行为识别属于具体场景应用，通用行为识别数据集不适合用来验证本文所提方法。因此，本文在真实快递站环境下采集了工作人员和顾客行为的视频片段，包含了以下 4 种行为类别，共 200 组视频，截取图像共 10 000 帧，其中行为“Opened Box”有 2

353 帧，行为“Packing”有 3 382 帧，行为“Using Phone”有 2 645 帧，行为“Writing”有 1 079 帧，行为“Irrelevant Behavior”有 541 帧，本文数据集示例如图 10 所示。

图 10　数据集图像示例

3.2　实验步骤

根据第二章所阐述的行为识别和目标检测的方法，本文设计的实验步骤如下：

Step1：数据集划分。将 200 组邮局实验样本按照 4：1 的比例划分为训练集和测试集；

Step2：将训练集输入至 OpenPose 进行骨架建模，得到人体上半身的骨骼点数据；

Step3：将骨骼点数据作为输入，制作行为识别的数据集，对如图 5 所示的人体行为识别网络进行训练，学习率设置为 0. 0001，Epoch 设置为 100，BatchSize 设置为 32。

Step4：制作目标检测的数据集，训练 YOLOv5 网络，得到网络的训练模型并测试结果；

Step5：得到两种网络模型的测试结果后，根据基于 Auction 的关键姿态估计方法得到人－物配对集；

Step6：通过融合决策得出实验结果。

3.3　实验结果分析

本文采用精确率（Precision）和召回率（Recall）作为评价指标[21]，用于评价该识别方法的优劣，具体计算方法如公式（8）和（9）所示：

$$\text{Precision}(c_i)=\frac{TP(c_i)}{TP(c_i)+FP(c_i)} \tag{8}$$

$$\text{Recall}(c_i)=\frac{TP(c_i)}{TP(c_i)+FN(c_i)} \tag{9}$$

其中：TP（ci）表示属于 c_i 类且被正确分为 c_i 类的样本数；FP（c_i）表示不属于 c_i 类但被分为 c_i 类的样本数；FN（c_i）表示属于 c_i 类但没有被正确分为 c_i 类的样本数[22-23]。

如上一小节所述，按照 4：1 的比例进行训练集和数据集的划分，因此测试集一共 2 000 帧，行为“Opened Box”有 470 帧，行为“Packing”有 680 帧，行为“Using Phone”有 530 帧，行为“Writing”有 210 帧，行为“Irrelevant Behavior”有 110 帧。将加入目标检测模块和融合决策之后的识别算法与加入前的识别算法进行对比实验，图 11 中的（a）和（b）分别为加入模块前和加入模块后识别结果的混淆矩阵，用测试集分别测试得出的准确率和召回率如表 2 和表 3 所示。

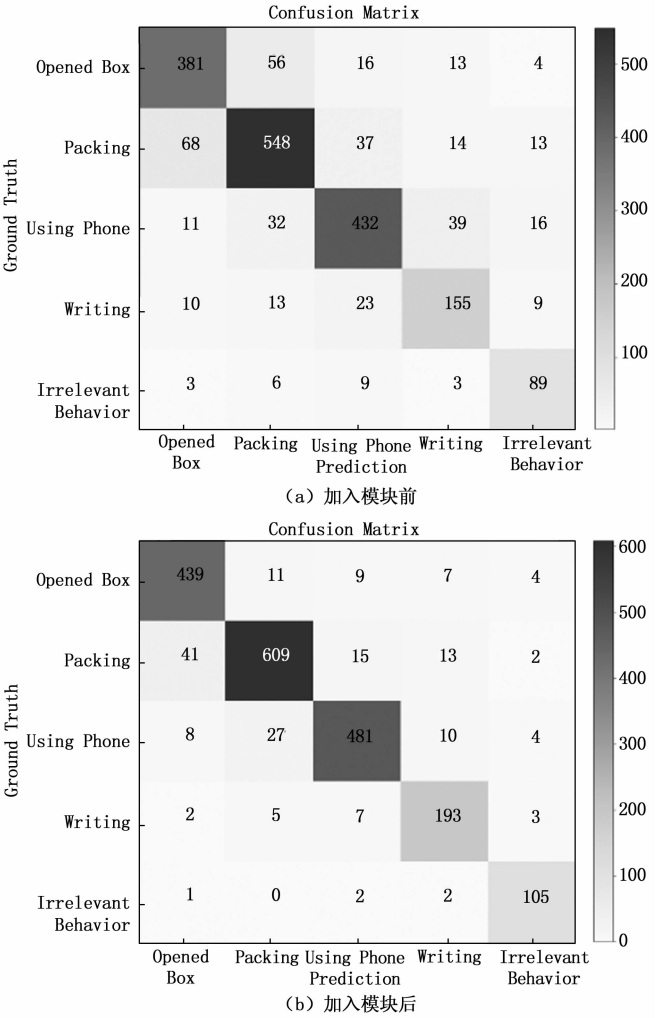


图 11　混淆矩阵

表 2　识别结果(加入模块前)　　%

	Precision	Recall
OpenedBox	80.5	81.1
Packing	83.7	80.6
Using Phone	83.6	81.5
Writing	69.2	73.8
Irrelevant Behavior	67.9	80.9

表 3　识别结果(加入模块后)　　%

	Precision	Recall
OpenedBox	89.4	93.4
Packing	93.4	89.6
UsingPhone	93.6	90.8
Writing	85.8	91.9
Irrelevant Behavior	89.0	95.5

实验结果表明，在加入目标检测模块和融合决策模块

之后，该系统的精确率和召回率有了显著的提高，为了更直观地表明本文方法的有效性，将实验结果进行可视化。快递站实际场景的实验结果如图 12 所示，分别展示了 3 种不同时间段的图像帧序列，如图所示，（1，2，3，4）－b 的右侧人员的真实行为类别为“Packing”，但是在未加入目标检测模块之前被误识别为“OpenBox”，加入目标检测模块并经过决策融合之后，识别结果得以修正；（1，2，3，4）-c 右侧人员的真实行为类别为“OpenBox”，初始误识别为“Packing”，最终识别结果被修正准确。（1，2，3，4）-a 由于该行为初始类别准确，因此最终识别结果并未发生改变。除此之外，图中仅显示了与人员行为相关的物品信息，这是通过关键姿态估计进行了人－物最优分配，去除了与人员行为无关的物品干扰。综上所述，本文算法具有良好的准确率和召回率。

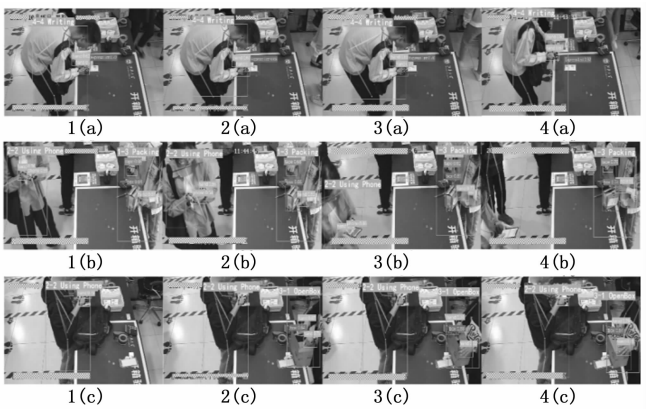


图 12 实际快递场景实验结果

4 结束语

本文综合分析快递场景的特点和异常行为特征，将基于骨架建模的人体行为识别与目标检测相结合，提出了一种基于关键姿态的快递场景人物交互行为识别方法。首先用改进的高斯混合模型进行关键帧的提取，然后用 OpenPose 进行骨架建模，继而利用基于骨骼点的人体行为识别方法获取人体的初始行为类别；使用 YOLOv5 算法获得场景内常见物品的类别和位置信息，解决了传统行为识别方法丢失上下文场景和语义信息的问题；通过提出基于 Auction 的多人－多物最优分配方法来进行关键姿态估计，最后将行为识别和人物关系进行融合决策，提高了人－物交互行为的识别精度。实验证明，本文方法的识别精度优于传统行为识别方法，解决了开箱验视过程中复杂环境干扰和相似行为难以区分这两个问题，实现了对快递场景人员开箱验视、使用手机、包装快递等行为的精确识别。但是，若目标检测算法未能检测出手机、胶带以及工具刀等小物体时，识别结果会受到影响。因此如何提高目标检测算法对于小目标的检测能力，如何将其与行为识别方法进行深度融合，将作为下一步的研究方向。

参考文献：

[1] 王生云，赵吉龙，虎晓敏，等. 利用深度学习的施工人员安全隐患行为诊断控制方法 [J]. 计算机测量与控制，2022，30 (2)：72－78.

[2] 王 婷，刘光辉，张钰敏，等. 多模态特征融合的长视频行为识别方法 [J]. 计算机测量与控制，2021，29 (11)：165－170，175.

[3] HUANG K，TAN T. Vs-star：A visual interpretation system for visual surveillance [J]. Pattern Recognition Letters，2010，31 (14)：2265－2285.

[4] AL-FARIS M，CHIVERTON J，NDZI D，et al. A review on computer vision-based methods for human action recognition [J]. Journal of Imaging，2020，6 (6)：46.

[5] ZHANG H B，ZHANG Y X，ZHONG B，et al. A comprehensive survey of vision-based human action recognition methods [J]. Sensors，2019，19 (5)：1005.

[6] REZAEI F，YAZDI M. Real-time crowd behavior recognition in surveillance videos based on deep learning methods [J]. Journal of Real-Time Image Processing，2021，18 (5)：1669－1679.

[7] TSAI J K，HSU C C，WANG W Y，et al. Deep learning-based real-time multiple-person action recognition system [J]. Sensors，2020，20 (17)：4758.

[8] GUAN Y，HU W，HU X. Abnormal behavior recognition using 3D-CNN combined with LSTM [J]. Multimedia Tools and Applications，2021，80 (12)：18787－18801.

[9] XU L，GONG C，YANG J，et al. Violent video detection based on MoSIFT feature and sparse coding [C] //2014 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP)，IEEE，2014：3538－3542.

[10] 吴蓬勃，张金燕，王 帆，等. 快递暴力分拣行为视觉识别系统 [J]. 包装工程，2021，42 (15)：245－252.

[11] WANG J，WANG N，LI L，et al. Real-time behavior detection and judgment of egg breeders based on YOLO v3 [J]. Neural Computing and Applications，2020，32 (10)：5471－5481.

[12] YANG Q，XIAO D. A review of video-based pig behavior recognition [J]. Applied Animal Behaviour Science，2020，233：105146.

[13] IJJINA E P，CHALAVADI K M. Human action recognition in RGB-D videos using motion sequence information and deep learning [J]. Pattern Recognition，2017，72：504－516.

[14] NUSSIPBEKOV A K，AMIRGALIYEY Y N，HAHN M. Improvement of human key posture recognition [J]. Biosciences Biotechnology Research Asia，2015，12 (2)：1139－1144.

[15] PIPERAGKAS G S，MARIOLISI，IOANNIDIS D，et al. Key-frame extraction with semantic graphs in assembly processes [J]. IEEE Robotics and Automation Letters，2017，2 (3)：1264－1271.

[16] CAO Z，SIMON T，WEI S E，et al. Realtime multi-person 2d pose estimation using part affinity fields [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recog-

nition，2017：7291－7299.

[17] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv，1409. 1556，2014.

[18] 段中兴，王　剑，丁青辉，等. 基于深度学习的盲道障碍物检测算法研究 [J]. 计算机测量与控制，2021，29 (12)：27－32.

[19] WU J，HU S，WANG Y. Particle probability hypothesis density filtering for multitarget visual tracking with robust state extraction [J]. Optical Engineering，2011，50 (9)：090502.

[20] SHAH M H，DANG X. Novel feature selection method using Bhattacharyya distance for neural networks based automatic modulation classification [J]. IEEE Signal Processing Letters，2019，27：106－110.

[21] OKSUZ K，CAM B C，KALKAN S，et al. One metric to measure them all：localisation recall precision (LRP) for evaluating visual detection tasks [J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2021 [2022-01-06]. DOI:10.1109/TPAMI.2021.3130188.

[22] KOMANG M G A，SURYA M N，RATNA A N. Human activity recognition using skeleton data and support vector machine [J]. Journal of Physics：Conference Series，2019，1192 (1)：012044.

[23] 宋真东，杨国超，马玉鹏，等. 基于注意力机制的多模态人体行为识别算法 [J]. 计算机测量与控制，2022，30 (2)：276－283.

(上接第 181 页)

光测距仪的空间非合作目标相对位姿紧耦合估计方法。该算法分为初始化和连续位姿估计两部分，在初始化部分恢复了所有参数的真实尺度，构建了真实尺度下的各个坐标系；在连续位姿估计部分，以紧耦合的形式融合相机数据与激光测距仪数据来优化相对位姿，即激光测距仪的信息不仅用于实时优化特征点三维坐标，还与图像信息一起构建目标函数来优化全局位姿，解决估计漂移问题。最后使用 Blender 软件仿真了任务航天器以 25 m 的半径相对空间非合作目标做绕飞观测的图像，以此对算法进行验证。结果显示初始化时相对位置和姿态估计的均方根误差分别不超过 0.02 m 和 0.03°，连续位姿估计时相对位置和姿态估计的均方根误差分别不超过 0.12 m 和 0.29°。初始化平均用时为 0.07 s，连续位姿估计时平均每帧图像用时为 0.12 s。表明该算法估计精度与实时性较高，理论上可以为自主交会、空间态势感知、空间碎片清理等空间任务提供所需的空间非合作目标位姿信息。

参考文献：

[1] 田林琳. 基于深度学习及 GPU 计算的航天器故障检测技术 [J]. 计算机测量与控制，2020，28 (5)：1－4，9.

[2] 孙永军，王　钤，刘伊威，等. 空间非合作目标捕获方法综述 [J]. 国防科技大学学报，2020，42 (3)：74－90.

[3] 束　安，裴浩东，周姗姗，等. 非合作航天器的立体视觉位姿测量 [J]. 光学精密工程，2021，29 (3)：493－502.

[4] 田九玲，杨永菊. 基于 Rodrigues 参数交互模型航天器相对位姿测量技术 [J]. 计算机测量与控制，2020，28 (9)：19－22，33.

[5] Committee on the assessment of options for extending the life of the hubble space telescope，National research council. Assessment of options for extending the life of the hubble spacetelescope：final report [M]. Washington：National Academies Press，2005.

[6] 张世杰，曹喜滨，陈　闽. 非合作航天器间相对位姿的单目视觉确定算法 [J]. 南京理工大学学报 (自然科学版)，2006 (5)：564－568.

[7] KAWANO I. Automated rendezvous docking system of engineering test satellite VII [C] //International Space Conference of Pacific-Basin Societies. Advances in the Astronautical Science，1997：96.

[8] OHKAMI Y，KAWANO I. Autonomous rendezvous and docking by engineering test satellite VII：a challenge of Japan in guidance，navigation and control—Breakwell memorial lecture [J]. Acta Astronautica，2003，53 (1)：1－8.

[9] WILSON J R. Satellite hopes ride on orbital express [J]. Aerospace America，2007，45 (2)：30－35.

[10] 周建平. 载人航天交会对接技术 [J]. 载人航天，2011，17 (2)：1－8.

[11] CAPUANO V，KIM K，HARVARD A，et al. Monocular-based pose determination of uncooperative space objects [J]. Acta Astronautica，2020，166：493－506.

[12] AUGENSTEIN S，ROCK S M. Improved frame-to-frame pose tracking during vision-only SLAM/SFM with a tumbling target [C] //IEEE International Conference on Robotics and Automation (ICRA)，2011：3131－3138.

[13] 郝刚涛，杜小平，宋建军. 空间翻滚非合作目标相对位姿估计的视觉 SLAM 方法 [J]. 宇航学报，2015，36 (6)：706－714.

[14] 高学海，徐科军，张　瀚，等. 基于单目视觉和激光测距仪的位姿测量算法 [J]. 仪器仪表学报，2007 (8)：1479－1485.

[15] 翟　光，赵　琪，张景瑞. 空间碎片在轨识别与精确定位方法 [J]. 红外与激光工程，2016，45 (S1)：176－183.

[16] BAY H，TUYTELAARS T，VAN GOOL L. Surf：Speeded up robust features [C] //European Conference on Computer vision，Springer，Berlin，Heidelberg，2006：404－417.

[17] MUJA M，LOWE D G. Fast approximate nearest neighbors with automatic algorithm configuration [J]. VISAPP (1)，2009，2：331－340.

[18] GHOBADI S E. Real time object recognition and tracking using 2D/3D images [D]. Siegen University，2010：3－25.

[19] LEPETIT V，MORENO-NOGUER F，FUA P. Epnp：An accurate o (n) solution to the pnp problem [J]. International Journal of Computer Vision，2009，81 (2)：155.

[20] 高　翔. 视觉 SLAM 十四讲：从理论到实践 [M]. 北京：电子工业出版社，2017.