Multi-person Detection and Identification Method in Complex Industrial Sites under Intelligent Manufacturing

Jihong Yan*
School of Mechatronics Engineering
Harbin Institute of Technology
Harbin 150001, China
jyan@hit.edu.cn

Chao Chen
School of Mechatronics Engineering
Harbin Institute of Technology
Harbin 150001, China
chenchao97@foxmail.com

Abstract

Under the trend of intelligent manufacturing and flexible manufacturing, human-machine collaboration has been the mainstream in production. Therefore, as a key factor affecting the efficiency of human-machine collaboration, it is very important to identify and analyze human actions in production line. However, cameras are often unable to get close to the operators considering the complex environment of industrial sites, while the identification accuracy reduces greatly because multi-person wearing the same uniform in industrial sites at the same time. Therefore, this paper proposes an automatic multi-person detection and pose estimation method based on object detection and pose estimation framework, which not only effectively distinguishes the identity information of the multiple operators, but also provides accurate body joint features for identifying the actions of multiple target operators.

Keywords
Intelligent manufacturing, Multi-person detection, Pose estimation, Action analysis

1. Introduction

With the concept of Industry 4.0 put forward, the manufacturing industry is gradually transforming from mass production to personalized and customized production. The production process becomes more complex due to the requirement of customization, which poses a great challenge to the flexibility and intelligence of the production process. Compared with automation equipment such as robots, human beings have stronger flexibility and adaptability. Therefore, human still play an irreplaceable role of core productivity which determines the production efficiency and quality as well as flexibility (Wang et al. 2019). In the context of intelligent manufacturing, the detection and identification of human is of great importance. The emergence and development of action recognition and pose estimation technology provides an effective method to solve this problem. Human action recognition technology monitors and analyzes actions in real time to find the existing problems, thus eliminating time waste, improving quality and efficiency.

In the era of Industry 4.0, large factories have embraced intelligent manufacturing one after another. At present, surveillance cameras in manufacturing factories with a certain scale can produce TB-level effective video data every day (Bonci et al. 2016), and most of these videos are only used to monitor workers' production, and the rich action information contained in them has not been fully utilized. With the application of mass customization production, various production lines often need to be reorganized according to personalized needs. However, workers are limited by knowledge and ability, and it is difficult to adapt to this changeable production mode, which often leads to the omission of key inspection procedures, the wrong use of assembly tools and substandard assembly operations. Action

© IEOM Society International

recognition and monitoring technology is to enable computers to make judgments on unknown actions through learning action information and then assist human to make more complex decisions. In human-centered manufacturing tasks, online action monitoring and identification technology provides information about errors or difficulties that workers may encounter, making it possible to solve these problems in a real-time manner.

Traditional action analysis methods, such as stopwatch method, usually identify and record actions through manual operation. This method has a large workload and low efficiency. The commonly used industrial analysis software, such as OTRS and ECRS, can realize the automatic analysis of video stream (Bortolini et al. 2020), but there are also many problems: (1) This kind of system only supports offline data analysis, unable to support real-time monitoring and abnormal correction; (2) It does not have the ability of automatic recognition of actions and requires manual interval division to complete data analysis. And this kind of system is difficult to transmit with the robot in real time in the actual industrial sites, so it is very important to study the online and automatic human detection and identification method.

Munaro applied HOG-like Descriptors to extract features from the whole detection window and trained the Support Vector Machine classifier to detect nearby workers on the factory floor (Munaro et al. 2016). Hao Zhang proposed people detection techniques that estimate the ground plane equation and exploit a depth-based clustering followed by color- based classification (Zhang et al. 2013). But in the actual acquisition process, cameras are often unable to get close to the operators considering the complex environment of industrial sites, while the identification accuracy reduces greatly because multi-person wearing the same uniform in industrial sites at the same time. Therefore, this paper proposes an automatic multi-person detection and pose estimation method based on object detection and pose estimation framework, which not only effectively distinguishes the identity information of the multiple operators, but also provides accurate body joint features for identifying the actions of multiple target operators.

The rest of the paper is structured as follows: Section 2 discusses the algorithm and experiment of human detection using YOLOv3. Section 3 discusses the algorithm and experiment of human detection using OpenPose. Section 4 proposes the multi-person detection and identification method. Finally, Section 5 concludes the paper and discusses our future work.

2. Human detection in complex industrial sites based on YOLOv3

The traditional target recognition algorithm is based on R-CNN. The YOLOv3 algorithm is chosen to meet the needs of complex environments in the industrial sites. In the experiment, the YOLOv3 algorithm is applied to the real-time video stream of a vehicle inspection production line. Finally, great recognition results and frame rate are obtained.

2.1 Comparison of detection algorithm

When convolutional neural network is used to recognize a single image, the speed can meet the real-time requirements. However, in the detection process, it is not only to recognize the target category but also to detect the target position information, so it is difficult to meet the real-time requirements in the actual use by relying solely on traditional detection methods such as R-CNN. In the detection process, two stage algorithms such as R-CNN need to traverses the target, extract features and return frames in the candidate box body of the image (Girshick et al. 2014). In the whole process, each step is separated from each other, which is not only tedious and time-consuming, but also has a high demand on the computing power of the equipment. In addition, the application object of this study is the industrial production line, where the environment is noisy and the illumination is uneven, which will put forward higher requirements for the detection algorithm (Dewantoro et al. 2020). Therefore, we need a detection algorithm that can not only ensure target recognition and positioning, but also stable and accurate target classification and position regression. Considering the above reasons and requirements, the most successful recognition and positioning algorithm YOLOv3 is adopted.

2.2 Network architecture of YOLOv3

YOLO stands for You Only Look Once. It's an object detector that uses features learned by a deep convolutional neural network to detect an object. YOLO v3 uses the new network Darknet-53, which structure is shown in Figure. 1 (416×416 for example). Darknet-53 uses YOLO v2, Darknet-19 and Resnet. This model uses a lot of well-behaved 3×3 and 1×1 convolution layer, and some shortcut connection structures. Eventually it has 53 convolutional layers, so named Darknet-53. First YOLO v3 extracts features from the input image through the network, which have a certain size of feature map, such as 13×13 (Redmon et al. 2018). Then the input image is divided into 13×13 grid cell. If the center coordinates of an object in the ground truth is in the grid cell, the object will be predicted. Each grid cell predicts 3 bounding boxes which are not the same size (13×13, 26×26, 52×52). The object prediction predicts the IOU (Intersection over Union) of the ground truth and the proposed box. And the class predictions predict the probability of that class given that there is an object. There are two dimensions (width and height) in the output feature map such as 13×13.

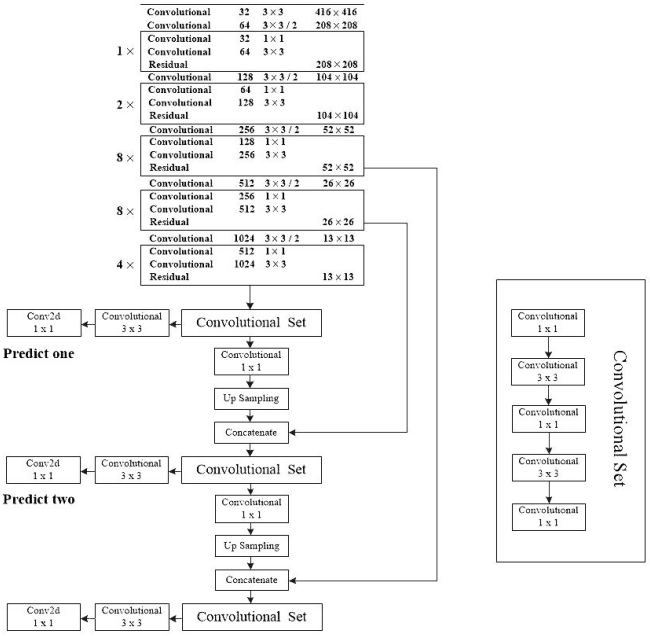

Figure 1. The network architecture of YOLOv3

2.3 Detection results on video streams on an actual industrial site

In this research, a human detection simulation on a vehicle inspection production line is conducted to test the ability of the YOLOv3 to detect human targets in complex industrial sites. This test is carried out under various conditions that may occur when the human targets detection system is working. Before conducting the test, we will conduct the import of model trained on COCO dataset. The goal is that the YOLO algorithm is able to recognize all the human targets captured in videos by cameras.

The computer specifications and environment configuration for training and testing is shown in Table 1.

Table 1. The computer specifications and environment configuration

Software and hardware experiment platform	The specific models
Operating System	Windows 10
CPU	Intel i5-9400
RAM	DDR4 2x8GB 2666MHz
GPU	Nvidia GTX 1660ti
Machine Learning Framework	Tensorflow 1.6.0
Software	Anaconda

After parameter configuration of YOLO framework, the model is applied to detect targets in the video stream collected from complex industrial sites, the frame rate for video streams is about 8FPS, and results indicated by the box is as shown in Figure. 2.

Figure 2. Experiment results of the movement process detection

3. Bottom-up pose estimation based on OpenPose

Although the YOLOV3 framework has achieved considerable recognition capability, it cannot accurately locate the key points of the human body. Therefore, OpenPose is used for human pose estimation to further improve the joint recognition failure of dedicated depth sensors such as Kinect in complex industrial environments (Tsai et al. 2020). OpenPose algorithm is an open source human body detection project based on Caffe. It adopts a bottom-up method and can detect facial key points, human body key points and hand key points. It is suitable for single and multi-person detection and has strong robustness.

3.1 Network architecture of OpenPose

OpenPose human skeleton extraction method is an open source library based on convolutional neural network and supervised learning proposed by Carnegie Mellon University in the United States, which calculates posture estimation of single or multiple human movements, facial expressions and finger movements in 2D images (Cao et al. 2019). The real-time performance of OpenPose is excellent, and reliable key point information can be obtained by using a monocular camera, without the need for a dedicated depth camera. Applications based on OpenPose involve many fields, such as sports and fitness, action acquisition, 3D fitting, public monitoring and so on. In this paper, the human pose estimation method of OpenPose was used to extract the coordinates of 25 key body points in the 2D videos of human skeleton.

The core of this method is a bottom-up body pose estimation algorithm using Part Affinity Fields (PAFS), that is, the key points are detected first and then the skeleton is obtained. The image is analyzed by a CNN (initialized by the first 10 layers of VGG-19 (Simonyan et al. 2014) and fine-tuned), generating a set of feature maps F that is input to the first stage. At this stage, the network produces a set of part affinity fields $L^1 = \phi^1(F)$, where ϕ^1 refers to the CNNs for inference at Stage 1 (Cao et al. 2019). In each subsequent stage, the predictions from the previous stage and the original image features F are concatenated and used to produce refined predictions. The process is repeated for the confidence maps detection such as S^1, S^2.

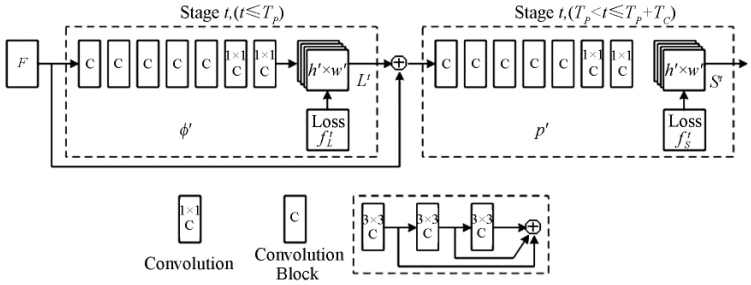

Figure 3. Structure of OpenPose network

3.2 Estimation results of key body points on video streams on an actual industrial site

The computer specifications and environment configuration for estimating key body points is shown in Table 2.

Table 2. The computer specifications and environment configuration

Software and hardware experiment platform	The specific models
Operating System	Windows 10
CPU	Intel i5-9400
RAM	DDR4 2x8GB 2666MHz
GPU	Nvidia GTX 1660ti
Machine Learning Framework	Caffe
Software	Anaconda

Through the experiment on the video collected on the industrial site, the positions of the following 25 points are obtained, as shown in Figure 4 and Table 3. The probability of the existence of the key points is obtained, and the average value of the confidence of the key points is calculated for ranking.

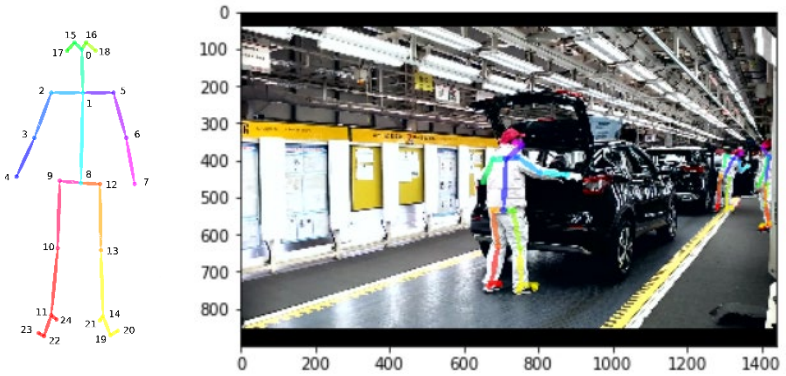

Figure 4. Experiment results on industrial sites video stream

Table 3. Identification results - coordinates and confidence

Keypoints	x-axis	y-axis	Confidence
{0, "Nose"},	1346.364	383.0324	0.065107666
{1, "Neck"},	1321.912	397.6459	0.97220474
{2, "RShoulder"},	1338.96	402.4251	0.9630301
{3, "RElbow"},	1343.845	429.5043	0.9188754
{4, "RWrist"},	1363.472	417.1754	0.7863996
{5, "LShoulder"},	1304.799	387.8806	0.89068025
{6, "LElbow"},	1299.671	417.1752	0.13964765
{7, "LWrist"},	0	0	0
{8, "MidHip"},	1302.284	444.2387	0.93086481
{9, "RHip"},	1314.433	449.129	0.83507699
{10, "RKnee"},	1307.17	490.7625	0.85685718
{11, "RAnkle"},	1307.212	527.5472	0.84256667
{12, "LHip"},	1289.974	441.7733	0.87360764
{13, "LKnee"},	1285.076	481.115	0.92809576
{14, "LAnkle"},	1282.552	522.6704	0.90771586
{15, "REye"},	1344.041	383.0076	0.097028174
{16, "LEye"},	0	0	0
{17, "REar"},	1343.793	383.0371	0.8270126
{18, "LEar"},	1324.328	380.419	0.26927939
{19, "LBigToe"},	1285.067	525.0721	0.64952445
{20, "LSmallToe"},	1280.163	522.699	0.65160811
{21, "LHeel"},	1280.194	532.429	0.83334446
{22, "RBigToe"},	1319.3	525.1654	0.60373986
{23, "RSmallToe"},	1321.774	532.4262	0.65875554
{24, "RHeel"},	1304.75	534.839	0.88451415

4. Multi-person Detection and Identification Method

The camera model “HYUNDAI HYS-015” is used to shoot the person to be estimated. It is a photographic equipment that transmits data using USB2.0, with 1920×1080 resolution and 100° field of view.

Figure 5. The vison sensor used to collect data

As can be seen from the key point diagram, since Neck and MidHip connect all joints and are the core nodes, Neck and MidHip are regarded as the main key points. In a video clip, the detection threshold of two key points, Neck and MidHip, was set to 0.9. Considering occlusion, if one of the key points above exists, the location of the person with this identity in the video can be determined.

By combining the advantages of YOLO and OpenPose frameworks, we propose to carry out parallel processing of video streams with two frames, and finally obtain high-precision identification effect by summarizing information flows as is shown in Figure 6. Through testing the OpenPose framework, the recognition frame of video stream is 8.2FPS. Finally, the image is divided into eight regions, and movement pattern of the target between regions is compared with the templates to identify the person on sites.

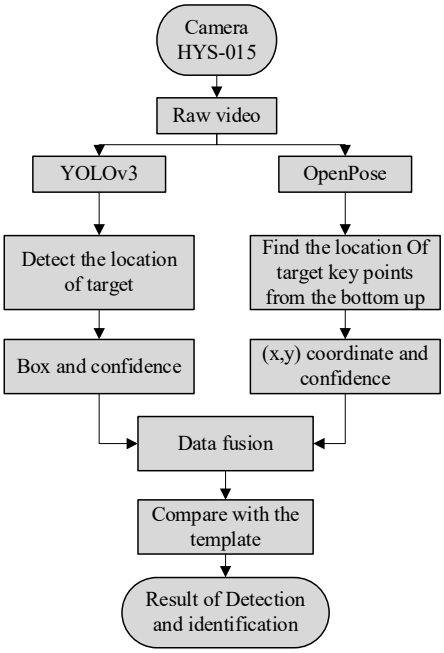

Figure 6. Process of detection and identification

5. Conclusions and future work

By integrating both frameworks of YOLOv3 and OpenPose, desirable detection and identification outcomes of people on complex industrial sites are achieved. This method has obvious significance for analyzing worker's action and improving production efficiency. However, this method is faced with the problems of huge computation and requirement of high performance hardware.

In the future, we will further integrate the advantages of the two identification methods and use data enhancement methods to further improve their application capability on industrial sites. At the same time, the model size is to be reduced to carry out more efficient detection and identification for human targets.

Acknowledgements

This work was supported in part by the JCKY Project of China (#JCKY2019603C016).

References

Wang, Z., Qin, R., Yan, J., & Guo, C. (2019). Vision Sensor Based Action Recognition for Improving Efficiency and Quality Under the Environment of Industry 4.0. *Procedia CIRP*, 80, 711-716.

Bonci, A., Pirani, M., & Longhi, S. (2016). A database-centric approach for the modeling, simulation and control of cyber-physical systems in the factory of the future. *IFAC-PapersOnLine*, 49(12), 249-254.

Bortolini, M., Faccio, M., Gamberi, M., & Pilati, F. (2020). Motion Analysis System (MAS) for production and ergonomics assessment in the manufacturing processes. *Computers & Industrial Engineering*, 139, 105485.

Munaro, M., Lewis, C., Chambers, D., Hvass, P., & Menegatti, E. (2016). RGB-D human detection and tracking for industrial environments. *In Intelligent Autonomous Systems 13* (pp. 1655-1668). Springer, Cham.

Zhang, H., Reardon, C., & Parker, L. E. (2013). Real-time multiple human perception with color-depth cameras on a mobile robot. *IEEE Transactions on Cybernetics*, 43(5), 1429-1441.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).

Dewantoro, N., Fernando, P. N., & Tan, S. (2020, February). YOLO Algorithm Accuracy Analysis in Detecting Amount of Vehicles at the Intersection. *In IOP Conference Series: Earth and Environmental Science* (Vol. 426, No. 1, p. 012164). IOP Publishing.

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv*:1804.02767.

Tsai, Y. S., Hsu, L. H., Hsieh, Y. Z., & Lin, S. S. (2020). The real-time depth estimation for an occluded person based on a single image and OpenPose method. *Mathematics*, 8(8), 1333.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556.

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2019). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1), 172-186.