International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 4, August, 2023, pp. 1723-1729
Journal homepage: http://iieta.org/journals/ts

Enhanced Tool Detection in Industry 4.0 via Deep Learning-Augmented Human Intent Recognition: Introducing the Industry-RetinaNet Model

Yaqiao Zhu[1,2*], Zhiwu Shang[1,3], Jin Wu[4]

[1] School of Mechanical Engineering, Tiangong University, Tianjin 300387, China
[2] School of Aviation and Aerospace, Tianjin Sino-German University of Applied Sciences, Tianjin 300350, China
[3] Tianjin Modern Electromechanical Equipment Technology Key Laboratory, Tianjin 300387, China
[4] School of Mechanical Engineering, Tianjin University, Tianjin 300072, China

Corresponding Author Email: julin_2017@126.com

https://doi.org/10.18280/ts.400441

Received: 8 February 2023
Revised: 30 May 2023
Accepted: 16 June 2023
Available online: 31 August 2023

Keywords:
Industry 4.0, human intention recognition, tool detection, Industry-RetinaNet, deep learning, real-time processing

ABSTRACT

In the context of Industry 4.0, a transformative shift in industrial manufacturing, product enhancement, and distribution methods has been observed, emphasizing the critical need for precise recognition of human intention to ensure operational reliability, safety, and efficiency. Central to this recognition, especially in equipment manufacturing, is the accurate identification of tools manipulated by human operators. In this study, a novel object detection model, referred to as 'Industry-RetinaNet', has been proposed for advanced tool detection. Improvements upon the conventional RetinaNet are evident in the form of optimized anchor box shapes derived from advanced anchor generation techniques, an augmented number of detection boxes, and the reinforcement of an alternate backbone architecture. When validated against a test dataset, the model demonstrated notable performance metrics with an F1-score of 0.904, an mAP of 0.903, and a recall of 0.809, while preserving real-time processing capabilities. It is anticipated that the implementation of this methodology will pave the way for improved interpretation of worker intentions, potentially enhancing overall efficiency in the burgeoning arena of intelligent factories.

1. INTRODUCTION

With the emergence of Industry 4.0, transformative shifts in manufacturing and product distribution methodologies have been observed [1, 2]. These transitions have been driven by the integration of nascent technologies, including the Internet of Things, artificial intelligence, and machine learning, into manufacturing facilities and operational processes [3, 4]. Consequently, the birth of a more intelligent and automated industrial landscape has been witnessed, characterized by heightened product quality, diminished production costs, and enhanced safety standards.

In the evolving landscape of intelligent factories, the significance of recognizing and classifying human intention is paramount [5-7]. Such recognition not only ensures the reliability and safety of human operations but also gauges the holistic efficiency of production systems. This understanding is pivotal, especially when extended to the interactions between humans and robots, where the goal is to establish more safe and efficient exchanges. Outside the confines of factory settings, discerning human intention has crucial implications in domains like autonomous driving [8], pedestrian intention prediction [9], and surveillance and security [10].

It is crucial to distinguish intention prediction from action recognition and action prediction, as the former aims to fathom the underlying intentions behind a sequence of evolving actions [11-15]. A notable area of application is equipment manufacturing, where the accurate detection of tools manipulated by operators remains an integral component of human intention recognition.

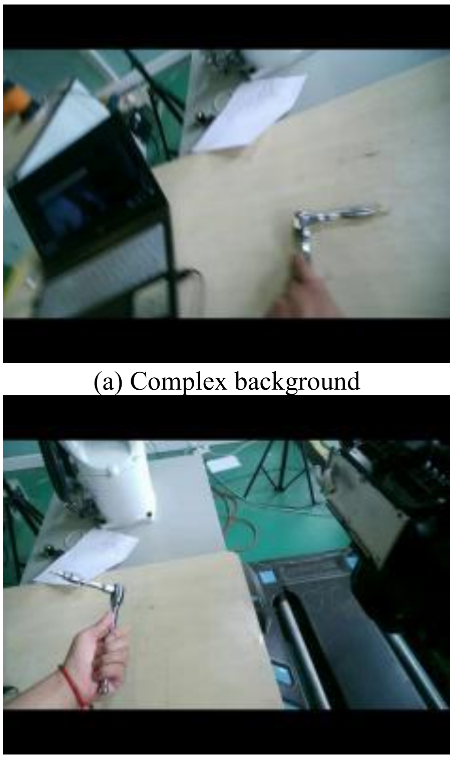
Amidst the swift advancements in deep learning, various object detection models such as RCNN, Fast RCNN, Faster RCNN, SSD, YOLO, and RetinaNet have been presented in the literature [16-21]. However, the unique challenges presented by tool detection, owing to the diversity in tool sizes, shapes, and appearances amidst intricate backgrounds, necessitate the formulation of a specialized detection mechanism (Figure 1).

To bridge this gap, an innovative object detection model named 'Industry-RetinaNet' has been proposed, specifically tailored for tool detection. This novel model incorporates a unique backbone with subnet architectural adjustments, amplifying the dilated convolution operation evident in the conventional RetinaNet. Such modifications grant the network an expanded receptive field, optimizing multi-scale performance at the same depth. For the detection of objects across various scales, adjustments in the anchor configuration and an attention-gated function have been introduced during the training phase. Additionally, post-processing mechanisms, such as Soft-NMS, have been incorporated to elevate detection performance [22].

The salient contributions of this study are delineated as:

- Compilation of a specialized tool detection dataset tailored for human intention recognition in equipment manufacturing.
- Proposition of the innovative 'Industry-RetinaNet' model for tool detection.
- Achieving noteworthy tool detection results via 'Industry-RetinaNet' whilst preserving real-time processing proficiencies.

The remainder of the article is organized as follows: Section 2 delves into existing work pertinent to object detection. The methodologies adopted in this study are elucidated in Section 3. Section 4 offers insights into the implementation specifics of the proposed model and the empirical results obtained. Concluding remarks are presented in Section 5.



(a) Complex background

(b) Blurred image

Figure 1. Example of tool. Tools have various sizes and shapes, present different appearances in complex background

2. RELATED WORKS

In object detection, the primary objective remains the localization of specific object instances. Prior to the resurgence of deep learning, feature extraction for traditional detection algorithms was primarily dependent on detectors that utilized, for instance, HOG and SIFT features. Following the introduction of DPMs and their variants, a dominance of sliding-window-based detection methods was observed, a claim further corroborated by repeated successes in the PASCAL competition [23].

With the integration of Convolutional Neural Networks (CNNs) for feature extraction, significant strides were made in the domain. CNN-based detectors are typically bifurcated into two primary components: a feature extractor and a regression component. Differentiation between these detectors was based on the methodology adopted for generating potential bounding boxes, further categorizing them into either single-stage or two-stage paradigms.

Within the two-stage paradigm, an initial stage is tasked with producing a series of candidate proposals, perceived to encapsulate objects of interest. These proposals then undergo refinement in a subsequent stage, wherein classification into foreground and background is executed. This approach was first manifested by the Selective Search network and later refined by R-CNN, wherein CNN was employed for proposal classification [24]. Following the acclaim of R-CNN, several advancements, epitomized by Fast R-CNN's use of feature maps and RoI pooling, were introduced to minimize convolution operations per image. Furthermore, the Faster R-CNN implemented a Region Proposal module to generate regions of proposals. However, a notable limitation of these two-stage detectors was discerned: they often failed to meet real-time processing demands for video sequences [25].

On the contrary, single-stage detectors, prioritizing high frames per second (fps), opted for anchor boxes as a substitute for proposals. Such boxes, characterized by predefined ratios and scales, reflected prior knowledge regarding the given task. Earlier iterations of this paradigm, as represented by SSD and YOLO, emphasized speed, albeit often at the cost of accuracy [26]. A consistent challenge faced by these networks was the class-imbalance problem, which arose due to the disparity between background and foreground images. To mitigate this, focal loss was proposed, leveraging specific coefficients to diminish the overwhelming influence of background images on classification tasks. Subsequent single-shot network designs not only maintained commendable speed but also achieved performance metrics comparable to their two-stage counterparts [27]. A majority of these advanced designs either directly incorporated focal loss or introduced specialized structures to counter the class imbalance issue.

To address the persistent class-imbalance issue, various strategies were investigated, ranging from Online Hard Example Mining (OHEM), class-specific sampling, and focal loss modifications, to class re-weighting [28]. These methodologies spanned from accentuating hard examples during training (as observed in OHEM) to altering the training sampling strategy, thereby ensuring a balanced representation between foreground and background classes. Modifications of the focal loss function aimed at amplifying the model's resilience against class imbalances were also introduced. Furthermore, in the class re-weighting strategy, differential weights were allocated to diverse classes during training, harmonizing their contributions to the overall loss function.

Such dedicated strategies, in tandem with the integration of focal loss, have been instrumental in significantly elevating the performance metrics of contemporary designs, particularly in managing the inherent class-imbalance problem in object detection tasks. The aforementioned advancements lay a strong groundwork for further exploration and development in the object detection domain.

3. METHOD

This section delineates the methodology of the proposed tool detector, offering detailed insights into the network architecture, including the Feature Pyramid Network (FPN) and the anchor configuration. Essential components aimed at enhancing detection accuracy are discussed, as illustrated in Figure 2, which presents a schematic of the proposed network architecture. In the initial processing phase, DetNet is leveraged for feature extraction, followed by the construction of the FPN atop the DetNet output and residual blocks. The pyramid network's output features are then segmented into four stages and distributed across two sub-networks.

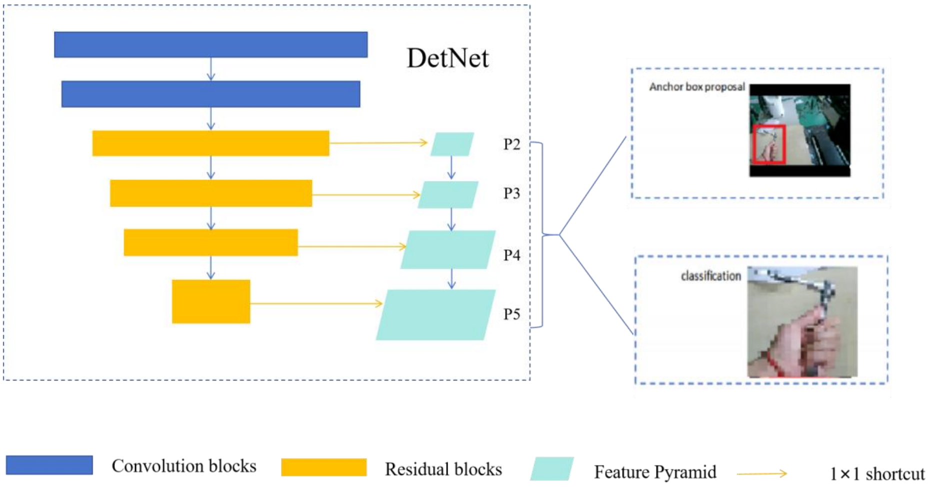


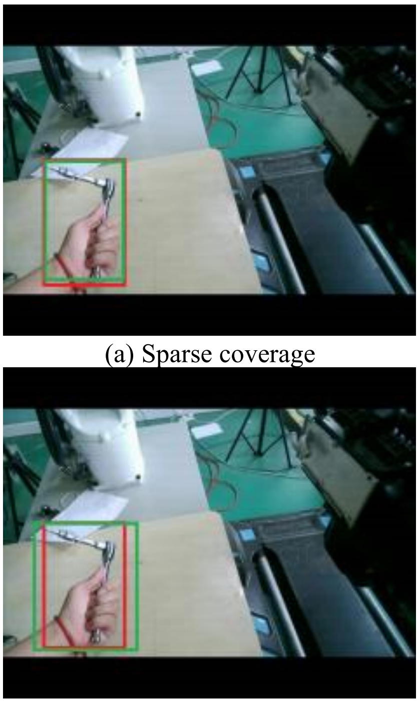
Figure 2. Proposed network architecture: DetNet feature extraction and FPN construction with segmentation into four stages distributed across dual sub-networks

3.1 Proposed network architecture

RetinaNet, as a single-stage detector, addresses the class imbalance challenge through the incorporation of focal loss. A FPN architecture, designed for constructing multi-scale outputs, is also integrated into RetinaNet. The foundational design of the proposed model mirrors that of RetinaNet, and the choice of ResNet-50 as the backbone over VGG is substantiated in the study [21]. The intrinsic FPN structure, as employed by RetinaNet, spans five stages, ranging from P3 to P7. Each of these stages is equipped to detect objects across varying scales. Notably, the mechanism labeled as the "shortcut output" is designed for deducing levels P3 to P5. This mechanism employs both lateral and top-down connections to the ResNet blocks C3 to C5. Here, the "shortcut output" represents feature maps generated by associating the output of a specific ResNet block with its corresponding FPN level through a lateral link. This design enables FPN levels to extract and utilize information spanning different phases of the backbone network. The P6 level is derived through a 2-strided convolution performed on P5, while P7 is formulated as a down-sampled variant of P6, actuated by ReLU. It is observed that the coarser levels, namely P6 and P7, are tailored to detect larger objects and are derived based on data from P5, bypassing the features originally imbibed by the ResNet backbone. Such a design strategy optimally balances computational speed with only a marginal compromise in large object detection. In the proposed architecture, an alternate method, aimed at enhancing speed, is incorporated and further elucidated in the subsequent sections. The computation of all five levels, spanning from P3 to P7, adheres to the approach detailed in the study [27].

3.2 Anchor optimization technique

Anchors of diverse sizes and aspect ratios are generated atop each pyramid level and subsequently directed to dual subnets, which are tasked with label classification and bounding box regression. Configurations for these anchors are found to be consistent with sizes of 32, 64, 128, 256, and 512. Notably, for each of these sizes, three scales - 20, 2, and 2 - and three distinct aspect ratios, namely 1:2, 1:1, and 2:1, are observed. Owing to the pivotal role that anchor quality plays in detection, conventional anchor settings have been identified to be potentially restrictive, particularly when detecting multiple tools smaller than 32 within a singular frame. To address this shortcoming, an evolutionary algorithm is introduced to deduce optimal anchors [28]. Optimization of these anchor configurations, as delineated in the study [29], is realized through the formulation of a specific objective function. The overlap between the anchor and the object's bounding box is incrementally augmented using a refined set of candidate proposals. An increased quantity of anchors per level, paired with a variety of aspect ratios, ensures a more comprehensive anchor coverage across every pyramid scale. Deviating from the methodology described in the study [29], scales in this context are optimized in relation to their inherent stride value. Through rigorous experimentation, combinations of scales and ratios have been discerned that most aptly suit the task at hand.



(a) Sparse coverage

(b) Dense coverage

Figure 3. Anchor optimization outcomes

Note: Results of the optimization process can be observed in the depicted anchor boxes. Enhanced ratios and sizes offer a superior overlay of the ground truth label. Ground truth is symbolized by the red box, while predictions are denoted by the green box.

3.3 Implementation of dilated convolutions

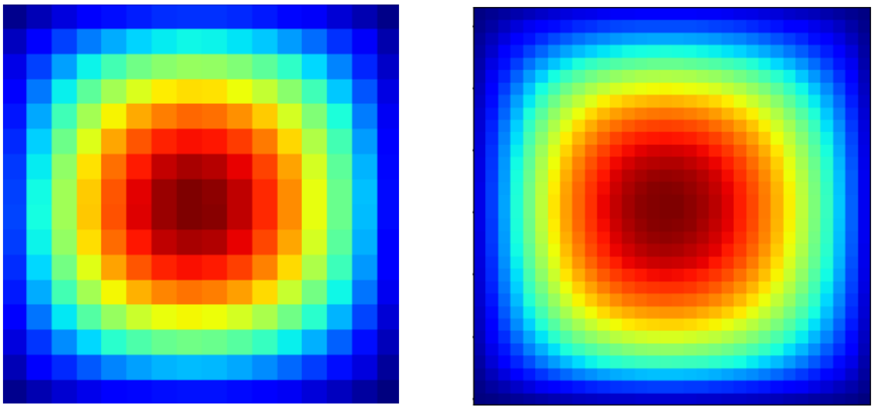
In preceding sections, optimal methodologies for anchor configurations were discussed, revealing their potential in augmenting detection performance, particularly for finer objects at the boundaries. However, it was recognized that the introduction of a higher anchor count necessitates a refined backbone structure, due to the added computational intricacy.

Classical CNNs are often observed to reduce resolution while expanding the receptive field, a trait that can result in the derived feature map being too granular for precise discernment. To address this challenge, the Dilated Residual Network (DRN) was introduced.

Dilated convolutions, occasionally referred to as atrous convolutions, are depicted by Eq. (1) [30]:

$$(F *_l k)(p) = \sum_{s+lt=p} F(s)k(t) \tag{1}$$

Within this equation, l denotes the dilation rate. It is noteworthy that this operation induces spaces within the convolutional kernel. For example, a dilation of a 3×3 convolution at a rate of 2 has been found to produce a receptive field size analogous to a 5×5 convolution. A salient advantage of this approach is the capability to maintain the same receptive field size, but with reduced parameters and at an enhanced resolution. Hence, performance improvements can be secured without additional depth or convolutional complexity.



(a) Gridded feature map representation (b) Degridded feature map depiction

Figure 4. Feature map illustrations

Note: The degridded feature map, as compared to its gridded counterpart, exhibits a smoother spread of heat values.

Nevertheless, a prominent artifact known as the gridding effect has been associated with dilated convolutions. This artifact manifests particularly when the dilation sampling rate falls below the frequency of the feature map, resulting in the generation of the aforementioned gridded pattern, as presented in Figure 4(a). Several corrective mechanisms have been proposed and evaluated. Among these, the global pooling layer's replacement with a ResNet block, along with an increase in residual blocks, was reported in studies [31, 32].

Drawing inspiration from DetNet, the backbone structure was redefined. Specifically, the native bottleneck blocks C4 and C5 of ResNet-50 were substituted with a dilated bottleneck integrated with a 1×1 convolution projection. Furthermore, a degridding mechanism was applied to this novel structure, with P5 being determined within the backbone. The resultant degridded feature map is visualized in Figure 4(b).

3.4 Implementation of the attention-gated block

The effectiveness of attention gates, especially when targeted at small and variable objects, has been previously established [29]. Such a mechanism, requiring merely a 1×1 convolution for the formation of an attention matrix, is not only lauded for improving accuracy but is also acknowledged for its efficiency and lightweight nature.

The underlying principle of the attention mechanism is its ability to discern feature saliency, which pertains to the degree of relevance certain features in the input possess. It has been observed that, through the generation of an attention map, models can be guided to predominantly focus on these salient regions. Concurrently, less pertinent regions are either suppressed or altogether overlooked. Within the scope of this discussion, 'non-task information' encompasses elements of the input that are deemed extraneous to the primary task. Such elements may comprise background nuances, noise, or any distractions unrelated to object detection.

To proficiently segregate this non-task information, integration of the global feature vector 'G', derived from a coarser spatial level, with the attention gate mechanism is advocated. It is postulated that this global feature vector equips the model with the capability to differentiate between pertinent features and the non-task counterparts. Consequently, the attention gate is facilitated in its suppression of the non-task information, simultaneously emphasizing the salient features, thus promising heightened performance metrics.

The operational principles of the attention gate modules are encapsulated in Eq. (2).

$$\begin{gathered} q^l_{att,i} = \psi^T(\sigma_1(W_x^{\mathrm{T}} x_i^l + W_g^T g + b_{xg})) + b_\psi \\ a^l = \sigma_2(q^l_{att}(x^l, g; \Theta_{att})) \end{gathered} \tag{2}$$

For pragmatic implementations, it is noted that the output of each pyramid has been utilized as the gate signal g, and concurrently, the output of the skip connection is employed as the input feature vector X.

4. EXPERIMENTATION AND RESULTS

4.1 Configuration of the experimental environment and dataset description

A novel tool detection was incorporated into the study. The experiments were conducted using an Nvidia GTX1080 Ti for both the training and testing phases. The open-source deep learning framework, TensorFlow, was employed to implement the model.

A dataset was curated to discern human intentions in equipment manufacturing scenarios. Consisting of 3,440 pre-processed images, the dataset exhibits varied pixel resolutions. Four categories were identified: screwdriver, spanner, screwdriver in hand, and spanner in hand. Expert personnel carefully annotated each image using the Label-me software. To gauge detection performance, a series of experiments were devised and executed.

4.2 Process of model training

Initially, images were resized to a 300×300 scale, followed by the application of data augmentation techniques. The proposed network was trained using Adam as the optimizer, incorporating a weight decay of 0.0001 and a momentum of 0.9. Furthermore, focal loss and L1 loss were utilized during network training.

4.3 Detection results for tools

For this dataset, partitioning was undertaken to create training, validation, and testing subsets, adhering to a 5:3:2 distribution. Given the absence of pre-trained weights suitable for the proposed architecture, networks were trained from their foundational parameters. To appraise the detection method, Precision, Recall, F1-score, Average Precision (AP), and Frame rate were selected as evaluative indicators. The performance in tool detection was quantified using metrics presented in Eq. (3).

$$\begin{aligned} &\text{Precision}=\frac{\text{TP}}{\text{TP+FP}},\ \text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \\ &\text{F1}=\frac{2\times\text{Prec}\times\text{Rec}}{\text{Prec+Rec}},\ \text{AP}=\int_0^1 p(r)dr \end{aligned} \tag{3}$$

Table 1. Tool detection outcomes on the test set

Methods	Params Size	F1-Score	mAP	Recall	Frame Per Second
RetinaNet [33]	35.33M	0.822	0.87	0.815	0.0291s
RetinaNet + AO	35.24M	0.828	0.891	0.834	0.0422s
RetinaNet + AO + AG	35.39M	0.827	0.894	0.851	0.0422s
RetinaNet + AO + AG + DilatedConv	29.62M	0.894	0.863	0.806	0.0198s
RetinaNet + AO + AG + DilatedConv + Degrid	33.73M	0.904	0.903	0.809	0.0229s

The test set outcomes for tool detection are depicted in Table 1, which encompasses various components. In the case of RetinaNet, ResNet-50 served as the backbone. For DilatedConv, the backbone was substituted with DetNet-59. It is discernible from the results that the inclusion of anchor optimization resulted in a 3% ascent in mAP, albeit at the cost of diminished processing speed. As a subsequent measure, the backbone architecture transitioned to DetNet, demanding fewer parameters without inducing a pronounced decrement in performance. It was observed that potential gridding effects could emanate from the backbone architecture. Evidently, the attention module mitigated these effects, bolstering detection performance without considerable hindrance to detection velocity.

5. CONCLUSION

In this research, Industry-RetinaNet, a distinct object detector tailored for tool detection, has been introduced. Factors contributing to the observed non-competitiveness in RetinaNet's performance were systematically investigated. Through this examination, it was determined that the incorporation of a task-specific anchor optimization strategy markedly augmented the original RetinaNet furnished with a ResNet-50 backbone.

Furthermore, an attention gate module was seamlessly integrated with the degridding backbone of DetNet, ensuring the preservation of real-time detection speed. Subsequent testing on a novel dataset revealed that the refined model yielded remarkable outcomes in tool detection, thereby highlighting its efficacy and broad applicability.

During the course of this research, a new dataset was curated. The proposed method, as evidenced by its superior performance and real-time inference capabilities, reaffirms its potential practical application, particularly within the ambit of Industry 4.0. Coupled with the model's resilience in diverse scenarios, a promising avenue for its deployment in future industry-centric applications emerges.

The broader ramifications of this study underscore the profound impact of targeted modifications to object detection models on enhancing their proficiency within niche industrial domains. Such findings not only pave the way for advancements in machine vision but also delineate a prospective route for bolstering operational efficiency in industrial settings. As a natural progression, it is recommended that subsequent studies endeavor to extrapolate this approach to alternative industrial contexts and datasets, aiming to further establish the adaptability and scalability of the solution presented.

ACKNOWLEDGMENT

This work was supported by 2021 Tianjin Applied Basic Research Multiple Investment Fund (Grant No.: 21JCQNJC00830).

REFERENCES

[1] Chen, S.L., Chen, Y.Y., Hsu, C. (2014). A new approach to integrate Internet-of-Things and software-as-a-service

model for logistic systems: A case study. Sensors, 14(4): 6144-6164. https://doi.org/10.3390/s140406144

[2] Lee, J., Bagheri, B., Kao, H.A. (2015). A cyber-physical systems architecture for Industry 4.0-based manufacturing systems. Manufacturing Letters, 3: 18-23. https://doi.org/10.1016/j.mfglet.2014.12.001

[3] Xu, X. (2012). From cloud computing to cloud manufacturing. Robotics and Computer-Integrated Manufacturing, 28(1): 75-86. https://doi.org/10.1016/j.rcim.2011.07.002

[4] Ooi, K.B., Lee, V.H., Tan, G.W.H., Hew, T.S., Hew, J.J. (2018). Cloud computing in manufacturing: The next industrial revolution in Malaysia. Expert Systems with Applications, 93: 376-394. https://doi.org/10.1016/j.eswa.2017.10.009

[5] Wang, Z., Boularias, A., Mülling, K., Schölkopf, B., Peters, J. (2017). Anticipatory action selection for human–robot table tennis. Artificial Intelligence, 247: 399-414. https://doi.org/10.1016/j.artint.2014.11.007

[6] Koppula, H.S., Jain, A., Saxena, A. (2016). Anticipatory planning for human-robot teams. In Experimental Robotics: The 14th International Symposium on Experimental Robotics, pp. 453-470. https://doi.org/10.1007/978-3-319-23778-7_30

[7] Townsend, E.C., Mielke, E.A., Wingate, D., Killpack, M.D. (2017). Estimating human intent for physical human-robot co-manipulation. arXiv Preprint, arXiv:1705.10851. https://doi.org/10.48550/arXiv.1705.10851

[8] Kim, I.H., Bong, J.H., Park, J., Park, S. (2017). Prediction of driver’s intention of lane change by augmenting sensor information using machine learning techniques. Sensors, 17(6): 1350. https://doi.org/10.3390/s17061350

[9] Kwak, J.Y., Ko, B.C., Nam, J.Y. (2017). Pedestrian intention prediction based on dynamic fuzzy automata for vehicle driving at nighttime. Infrared Physics & Technology, 81: 41-51. https://doi.org/10.1016/j.infrared.2016.12.014

[10] Phule, S.S., Sawant, S.D. (2017). Abnormal activities detection for security purpose unattainded bag and crowding detection by using image processing. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 1069-1073. https://doi.org/10.1109/ICCONS.2017.8250631

[11] Feichtenhofer, C., Pinz, A., Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 1933-1941. https://doi.org/10.1109/CVPR.2016.213

[12] Ma, S., Sigal, L., Sclaroff, S. (2016). Learning activity progression in LSTMs for activity detection and early detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 1942-1950. https://doi.org/10.1109/CVPR.2016.214

[13] Ryoo, M.S. (2011). Human activity prediction: Early recognition of ongoing activities from streaming videos. In 2011 International Conference on Computer Vision, Barcelona, Spain, 1036-1043. https://doi.org/10.1109/ICCV.2011.6126349

[14] Xu, Z., Qing, L., Miao, J. (2015). Activity auto-completion: Predicting human activities from partial videos. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 3191-3199. https://doi.org/10.1109/ICCV.2015.365

[15] Li, S., Zhang, L., Diao, X. (2018). Improving human intention prediction using data augmentation. In 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Nanjing, China, 559-564. https://doi.org/10.1109/ROMAN.2018.8525781

[16] Reddy, S.P.K., Harikiran, J. (2022). Cast Shadow Angle Detection in morphological aerial images using faster R-CNN. Traitement du Signal, 39(4): 1313-1321. https://doi.org/10.18280/ts.390424

[17] Mohammed, H., Tannouche, A., Ounejjar, Y. (2022). Weed detection in pea cultivation with the faster RCNN ResNet 50 convolutional neural network. Revue d'Intelligence Artificielle, 36(1): 13-18. https://doi.org/10.18280/ria.360102

[18] Tao, M.J., Lou, J.S., Wang, L. (2022). MRI liver image assisted diagnosis based on improved faster R-CNN. Traitement du Signal, 39(4): 1347-1355. https://doi.org/10.18280/ts.390428

[19] Yildiz, E.N., Cengil, E., Yildirim, M., Bingol, H. (2023). Diagnosis of chronic kidney disease based on CNN and LSTM. Acadlore Transactions on AI and Machine Learning, 2(2): 66-74. https://doi.org/10.56578/ataiml020202

[20] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788. https://doi.org/10.1109/cvpr.2016.91

[21] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2980-2988. https://doi.org/10.1109/iccv.2017.324

[22] Bodla, N., Singh, B., Chellappa, R., Davis, L.S. (2017). Soft-NMS--improving object detection with one line of code. In Proceedings of the IEEE International Conference on Computer Vision, pp. 5561-5569. https://doi.org/10.1109/iccv.2017.593

[23] Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W. (2013). Selective search for object recognition. International Journal of Computer Vision, 104: 154-171. https://doi.org/10.1007/s11263-013-0620-5

[24] Zhong, Q., Li, C., Zhang, Y., Xie, D., Yang, S., Pu, S. (2020). Cascade region proposal and global context for deep object detection. Neurocomputing, 395: 170-177. https://doi.org/10.1016/j.neucom.2017.12.070

[25] Shrivastava, A., Gupta, A., Girshick, R. (2016). Training region-based object detectors with online hard example mining. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 761-769. https://doi.org/10.1109/cvpr.2016.89

[26] Shrivastava, A., Sukthankar, R., Malik, J., Gupta, A. (2016). Beyond skip connections: Top-down modulation for object detection. arXiv Preprint, arXiv:1612.06851. https://doi.org/10.48550/arXiv.1612.06851

[27] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI,

USA, pp. 2117-2125. https://doi.org/10.1109/CVPR.2017.106

[28] Storn, R., Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization, 11: 341-359. https://doi.org/10.1023/a:1008202821328

[29] Zlocha, M., Dou, Q., Glocker, B. (2019). Improving RetinaNet for CT lesion detection with dense masks from weak RECIST labels. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, pp. 402-410. https://doi.org/10.1007/978-3-030-32226-7_45

[30] Yu, F., Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv Preprint, arXiv:1511.07122. https://doi.org/10.48550/arXiv.1511.07122

[31] Yu, F., Koltun, V., Funkhouser, T. (2017). Dilated residual networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 472-480. https://doi.org/10.1109/CVPR.2017.75

[32] Triggs, B. (2001). Empirical filter estimation for subpixel interpolation and matching. In Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vancouver, BC, Canada, 2: 550-557. https://doi.org/10.1109/iccv.2001.937674

[33] Chen, M., Du, P., Zhang, D. (2018). Massive colonoscopy images oriented polyp detection. In Proceedings of the 2018 5th International Conference on Biomedical and Bioinformatics Engineering, pp. 95-99. https://doi.org/10.1145/3301879.3301903