



Comparative Study of Human Skin Detection Using Object Detection Based on Transfer Learning

Ping Li, Hongliu Yu, Sujiao Li & Peng Xu

To cite this article: Ping Li, Hongliu Yu, Sujiao Li & Peng Xu (2021) Comparative Study of Human Skin Detection Using Object Detection Based on Transfer Learning, Applied Artificial Intelligence, 35:15, 2370-2388, DOI: [10.1080/08839514.2021.1997215](https://doi.org/10.1080/08839514.2021.1997215)

To link to this article: <https://doi.org/10.1080/08839514.2021.1997215>



Published online: 03 Mar 2022.



Submit your article to this journal [↗](#)



Article views: 1196



View related articles [↗](#)




View Crossmark data [↗](#)



Citing articles: 2 View citing articles [↗](#)



Comparative Study of Human Skin Detection Using Object Detection Based on Transfer Learning

Ping Li ^{a,b}, Hongliu Yu^a, Sujiao Li^a, and Peng Xu^a

^aInstitute of Rehabilitation Engineering and Technology, University of Shanghai for Science and Technology, Shanghai, China; ^bDepartment of Biomedical Engineering, Chang Zhi Medical College, Changzhi Shanxi, China

ABSTRACT

With the increasing aging of the population, the design of automatic bath robot has the forward-looking significance. The robot needs to detect the skin position, so as to perform the bathing task. The perception of skin is the key technology to achieve the bathing task. In this paper, object detection is used to identify the skin, which provides reference information for the pose of the robot. According to the classification of the object detection algorithms, this paper selects four typical object detection algorithms, namely, Faster R-CNN, YOLOv3, YOLOv4 and CenterNet. Due to the limitation of the self-built data set, this paper adopts the transfer learning to promote the completion of new tasks, which takes the pre-trained model as the starting point. The experimental results show that the detection results of YOLOv4 is the best, with mAP of 78%. This paper proves the feasibility and effectiveness of object detection completing the human skin detection in the bathing task.

ARTICLE HISTORY

Received 9 April 2021
Revised 17 October 2021
Accepted 20 October 2021

Introduction

With the aged tendency of population which imposes enormous economic burdens on families and insurance systems (Zlatintsi et al. 2020), there will be greater demand for special nursing, especially in activities of daily living (ADL), such as toileting and bathing (Dunlop, Hughes, and Manheim 1997). Considering that bathing activity is the first lost ADL that need help (Werle and Hauer 2016), the research on an automatic bathing robot is of great significance in maintaining the independence and quality of life of the elderly. The perception of the skin is one of the key technologies to realize the bathing task. As shown in Figure 1, RGB information is used for skin detection in bathing scene. Combining detection results with transformation matrix from world coordinate system to pixel coordinate system, we can gain the pose of end-effector and then the rotation angle by robot inverse kinematics, guiding the

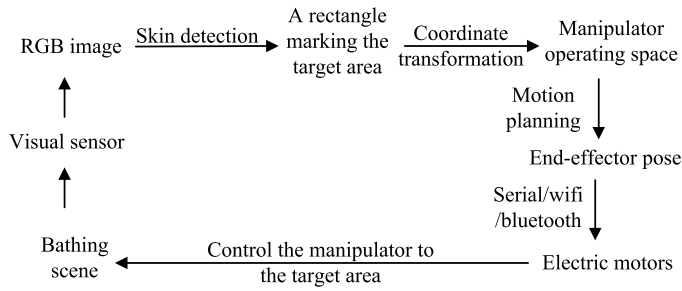


Figure 1. The key flow chart of the bathing system.

robot to the target area performing the bathing operation (scrubbing, washing, etc.). The skin detection accuracy directly affects the motion planning.

Skin Detection

According to the detection principle, skin detection can be divided into three methods: pixel-based (Nadian and Talebpour 2011), region-based (Zhu and Cai 2011) and hybrid methods (Tan et al. 2012). The pixel-based method classifies each pixel in the picture into skin or non-skin families. The region-based method is to identify regions with similar features in the picture. Based on the features used, skin detection can be roughly divided into the color-based, texture-based (Fotouhi, Rohban, and Kasaei 2009), statistical-feature-based (Fang, Kwok, and Dissanayake 2013), and other feature-based (Kawulok, Kawulok, and Nalepa 2014) methods. From the perspective of whether to explicitly establish a skin model, skin detection can be divided into machine learning-based methods and traditional methods. Machine learning methods construct a skin detector, generally using supervised methods. Traditional methods generally establish a skin model explicitly. Most of the existing skin detection works are in the medical field, such as the diagnosis of melanoma, skin cancer and other diseases (Khan et al. (2021c) and Khan et al. (2021d)).

A good skin detection algorithm should be robust to various variations. Without the deep learning method, the existing skin features or self-constructed features can be used for detection explicitly. And yet the deep learning learns the characteristics of the target from the data set automatically, which is more robust. By establishing the deep neural network and using massive data as learning samples, deep learning has analysis ability and feature representation ability. In recent years, deep learning has made rapid progress and achieved perfect results in the field of computer vision. Therefore, this paper uses deep learning to execute the skin detection.

When combined with the robot based on the vision, skin detection in the whole machine vision system is as an image processing part. In our mission, the purpose is to locate the skin and provide information for the robot control, rather than to distinguish whether each detailed pixel belongs to the skin or non-skin. The skin detection combined with the robot for the bathing task is relatively less. The existing research (Zlatintsi et al. 2020) tends to propose semantic segmentation, which assigns categories to each pixel in the image and thus is more time consuming than object detection under the same configuration of the model. Chandra, Tsogkas, and Kokkinos (Chandra, Tsogkas, and Kokkinos 2015) combine RGB and HHA-encoded depth information for semantic segmentation. However, the skin detection utilizing semantic segmentation is detrimental to real-time control of the robot. Moreover, the data set annotating required by semantic segmentation is rigorous and labor-intensive. Actually, the rectangular positioning obtained by object detection can meet the requirements. Based on the rough positioning, the robot moves toward the target area with the help of the control system. Simultaneously, the pose of the manipulator will be adjusted based on the compliance control algorithm, combined with the tactile sensor and the end-effector with variable stiffness. There is no need to determine the situation of each pixel in our mission. Adopting the object detection algorithm not only reduces the workload of data set labeling but also benefits the real-time control of the manipulator.

Deep Learning Methods for Object Detection

With the boosting growth of big data and computing power, the object detection algorithm based on deep CNN has become the mainstream algorithm in recent years, whose performance is far superior to the traditional algorithms (Hussain et al. 2020; Rashid et al. 2020, 2019). The object detection algorithm can be divided into a two-stage/one-stage algorithm and an anchor-based/anchor-free algorithm according to two classification standards. Two-stage algorithms are based on the region proposal (R-CNN (Girshick et al. 2014), Fast R-CNN (Girshick 2015), etc.), including two stages: extraction of regions of interest (ROI) and classification and regression. The features of ROI derived from the first phase will be used for classification and regression in the second phase. The bounding boxes are fine-tuned twice, with high accuracy but low speed. One-stage algorithms based on regression directly perform classification and regression and export categories and bounding boxes (YOLO (Redmon et al. 2016), SSD (Liu et al. 2016), etc.). The bounding boxes are fine-tuned only once, with low precision but high speed. Anchor-based algorithms use a large number of anchors with fixed size, which are laid on the image or feature maps as the prior of the bounding boxes. Anchors will be gradually adjusted to the bounding

Table 1. The typical algorithms under the two classification standards.

	Two-stage	One-stage
Anchor-based	Faster R-CNN , R-FCN	SSD, YOLOv3 , YOLOv4 , RetinaNet
Anchor-free	R-CNN, SPPNet, Fast R-CNN	Cornernet, CenterNet , ExtremeNet

boxes with the prediction results. Anchor-based algorithms have been dominant in the field of object detection (Wu, Sahoo, and Hoi 2020). In recent years, the emerging anchor-free algorithms, which have attracted wide attention of the academic community, abandon the anchor mechanism and re-encode the bounding box so as to obtain the bounding boxes through special processing (Zhang et al. 2020). Table 1 lists the typical algorithms under the two classification standards.

Based on the above classification, this paper selects four algorithms to identify the skin, as shown in the bold part of Table 1, which are typical algorithms under two classification standards. The introduction of an innovative anchor mechanism significantly improves the performance of two-stage algorithms, so only the anchor-based algorithm is selected for experimental research among the two-stage algorithms. Faster R-CNN (Ren et al. 2015) makes full use of the anchor mechanism and truly realizes the end-to-end training, in which the unified CNN network can realize feature extraction, ROI extraction, classification and bounding box regression. YOLOv3 (Redmon and Farhadi 2018) and YOLOv4 (Bochkovskiy, Wang, and Liao 2020) are the anchor-based and one-stage algorithms. YOLOv3 is widely used in industry. YOLOv4 achieves the best balance between accuracy and speed. CenterNet (Duan et al. 2019) is one of the representative anchor-free algorithms.

Methodology

DataSet Establishment: Collection, Screening and Annotation

We build a huge data set as the learning sample so that the neural network can learn the characteristics of the objects. Collecting images containing part or all of the skin enhances the diversity by taking the differences into account in pose, illumination, age, ethnical groups, skin color, resolution, gender, etc., which ensures that the trained model has certain reliability and robustness. A total of 1500 images were collected. Considering the image quality, 1000 images are selected to form the final data set. Furthermore, LabelImg, an image annotation tool, is used to generate a corresponding XML file in PASCAL VOC format for each image. We label the images with seven classes representing different parts of the human body, namely, (1) “Face_skin,” (2) “Trunk_skin,” (3) “Upperlimb_skin,” (4) “Lowerlimb_skin,” (5) “Hank_skin,” (6) “Foot_skin,” and (7) “Background.” Drawing the rectangle

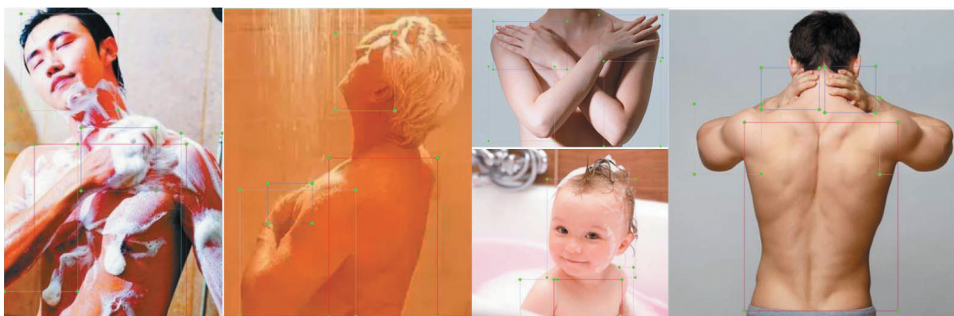


Figure 2. The labeled images in the data set.

manually specifies the label and boundary of an object to represent its category and exact location. The XML file contains information about the file name of the image, the coordinates of the ground true box (GT box), the label of the object, etc. [Figure 2](#) shows some labeled images in the data set.

Introduction to Algorithms Used

Faster R-CNN proposes RPN network, a fully convolutional structure, instead of SS (Selective Search) method ([Uijlings et al. 2013](#)) to generate proposals. Faster R-CNN is the integration of Fast R-CNN and RPN, which uses a 3×3 sliding window to slide on the feature map and generates a one-dimensional feature vector at each sliding position, the element number of which equals the number of channels in the feature map. After traversing all the positions, the obtained matrix is sent to the classifier network and the regression network in parallel.

YOLO is a pioneering work of one-stage object detection, which greatly improves the real-time performance of the network. YOLOv3 and YOLOv4 follow the idea of YOLO dividing cells for detection. YOLOv3 adopts darknet53 based on residual network ([He et al. 2016](#)) to extract features and absorbs the idea of FPN ([Lin et al. 2017](#)). Darknet53 consists of five residual blocks, which can avoid gradient vanishing during back-propagation. Each residual block is composed of multiple residual units which strengthens the feature extraction ability. The ideology of FPN is imbibed to realize multi-scale detection using the 8, 16 and 32 times down-sampling feature maps to improve the prediction accuracy. YOLOv3, without pooling layers and fully connected layers, uses batch normalization ([Ioffe and Szegedy 2015](#)) and leaky ReLU activation function ([He et al. 2015a](#)).

YOLOv4 is the model to balance the accuracy and speed, executing the prediction at three scales as YOLOv3 does. The implemented improvements on the basis of YOLOv3 are as follows:

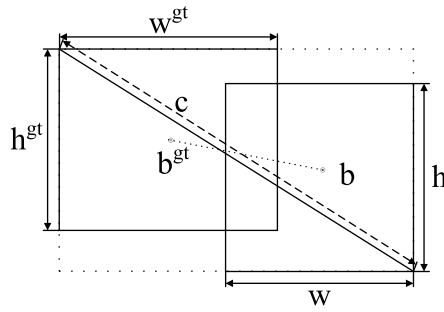


Figure 3. Schematic diagram of parameters in formula (1).

The backbone is upgraded from Darknet53 to CSPDarknet53 (Wang et al. 2020), including five basic residual blocks which contain CSP structure. With the width and height halved, the number of channels doubled;

Adding SPP (He et al. 2015b) and PANet (Liu et al. 2018) structure, which accomplish the effective feature fusion by continuous up-sampling, down-sampling and channel stacking;

Using class label smoothing, CutMix and Mosaic data augmentation;

Using CIOU loss (Zheng et al. 2020) as regression loss, solving the problem of slow convergent speed and low localization accuracy of IOU loss and GIOU loss (Rezatofghi et al. 2019). CIOU loss, fully taking overlapping area, distance between centers and aspect ratio between the prediction box and the GT box into account, is defined as:

where b and b^{gt} denote the center of the predicted box and the GT box, $\rho(\cdot)$ denotes the Euclidean distance, c denotes the diagonal length of the smallest enclosing box covering the two boxes, α is the trade-off parameter, and v measures the consistency of aspect ratio. Parameters are indicated in Figure 3. α and v are defined as follows:

$$L_{CIOU} = 1 - IoU + \rho^2(b, b^{gt})/c^2 + \alpha v \quad (1)$$

$$\alpha = v/(v+1-IoU) \quad (2)$$

$$v = 4/(\pi^2) * (\arctan(w^{gt}/h^{gt}) - \arctan(w/h))^2 \quad (3)$$

(5) Mish activation function (Misra 2019) is used as follows:

$$Mish = x * \tanh(\ln(1 + e^x)) \quad (4)$$

CenterNet, which represents the object with a central point and a pair of corner points, uses Hourglass (Newell, Yang, and Deng 2016) as the backbone and additionally adds a central point detection branch on the basis of CornerNet (Law and Deng 2018). In this way, the limitation of CornerNet, i.e., partial error detection caused by failure on reference object global

Table 2. Detailed comparison of the models used in this paper.

Models	Applicable scenarios	Advantages	Disadvantages
Faster R-CNN	Object detection	End-to-end optimization	Non-real-time, poor detection of small objects
YOLOv3	Multi-scale and real-time object detection	Clear structure, higher precision than YOLOv2, improved small object detection, real-time, and multi-scale detection	Slower than YOLOv2
YOLOv4	High-precision, multi-scale and real-time object detection	The best trade-off of speed and accuracy	The practical application is more difficult with the use of a variety of tuning techniques
CenterNet	High-precision object detection	Add center point position parameter	Non-real-time, complex feature extraction network

information, is better solved. CenterNet designs the cascade corner pooling module and the central point pooling module to enrich and integrate information. The cascade corner pooling is used to predict the location of the corner points of the object, which are mapped to the corresponding position of the input image to determine which two corners belong to the same object so as to form the final detection box. At the same time, the central point position of the object is predicted by the central point pooling and then is corrected by offset operation. The error detection boxes are eliminated by judging whether there is a predicted central point in the central area of each detection box. By means of this re-encoding manner, the detection accuracy is improved.

The mentioned models are compared in terms of applicable scenarios, advantages and disadvantages, as shown in [Table 2](#).

Transfer Learning

The model is trained on the data set which is collected for our shower mission. The input samples are imported into the network and then are transmitted to the output layers through the hidden layers. When the output values are different from the expected values, the errors are back propagated (LeCun, Bengio, and Hinton 2015) so as to correct the weight of each neural unit. The above process is repeatedly performed. Finally, the error between the actual output and the expected output reaches the minimum. However, training a network from scratch requires massive annotated data (Pattnaik, Shrivastava, and Parvathi 2020), which is mainly annotated by hand with the low efficiency and high error rate. Labeling a small data set carries a significant risk of model overfitting. Hence, transfer learning is indispensable (Khan et al. 2021a; Khan, et al.,2020). Combining the small data set with transfer learning can quickly train an acceptable model (Khan, Zhang, and Sharif 2021; Khan et al. 2021b), which can achieve the equivalent performance of training the network from scratch. The general transfer learning is what using the model weights pretrained on the ImageNet data set as the initial weights, which has learned common features such as textures and lines.

Prevent overfitting and Improve Generalization Ability

We use the following method to avoid overfitting, enhance the robustness and improve generalization performance:

Early stopping. When the loss value on the validation set is no longer decreasing while the value on the training set is decreasing, stop training to prevent overfitting caused by overtraining.

BN operation. BN associates all samples in mini-batch so that the network does not learn some specific features from a training sample.

Using data augmentation technology, which includes the following: (1) Disrupting the training samples and sorting them randomly and (2) Carrying out preconditioning before each mini-batch is input to the network, such as randomly clipping, translating and scaling.

Experimental Setup

Based on Pytorch, an open-source Python machine learning library, all models are trained using the supercomputer center established at the University of Shanghai for Science and Technology. In order to fairly compare the performance among models, the basic configurations of the models are the same. A consistent ratio was used for all models. The ratio of training set, validation set and test set is 60%:20%:20%. The initial value of learning rate is set to 0.001, and the attenuation rate is 0.01. The batch size is set to 8 which represents the number of images put into the network for training each time. SGD is used as the optimizer for network training.

When training on the own data set, firstly freezing the parameters of the backbone in the pre-trained model and then training other parameters for 50 epochs. An epoch means that the whole data set is sent to the network for training once. After that, we release the limitation of parameters of the backbone and train all the parameters so as to fine-tune the parameters for another 50 epochs. This kind of training strategy improves the convergence speed and training efficiency of the network.

Evaluation and Results

The recall measures the ability of finding the positives. The precision represents the proportion of parts that the detector considers to be positives and indeed positives to all results considered to be positives. Their value range is 0 to 1. The Formulas 5 and 6 are their calculation methods, where the definition of TP, FP, TN and FN is shown in [Table 3](#).

Table 3. The description of TP, FP, TN and FN.

Confusion matrix		The truth value	
		Positive	Negative
The predicted value	Positive	TP	FP
	Negative	FN	TN

$$recall = TP / (TP + FN) \quad (5)$$

$$precision = TP / (TP + FP) \quad (6)$$

When the detector outputs a large number of boxes and covers all GT boxes, the recall will reach 1, but the detection effect is not good. Similarly, the precision of 1 does not mean the best detection effect. AP is a balanced indicator of the precision and the recall. With the recall as the horizontal axis and the precision as the vertical axis, P-R curve is drawn. Draw a line segment to the left starting from each peak point, until which intersects with the vertical line of the previous peak point. The area of the region enclosed by the line segment and the coordinate axes equals AP value. The mAP, the mean of AP for all categories, is one of the important evaluation indexes for multi-class object detection. The problem studied in this paper is a multi-class issue, so mAP is used to evaluate the model performance. Test the models using the test set and get the results, as shown in Table 4, retaining two decimal places. In order to better compare the performance, Figure 4 shows the P-R curves of the six categories in the models.

The backbone has a strong impact on the network performance. Compared with ResNet50, MobileNetV2 is specially designed for the lightweight network, whose feature extraction ability is limited. The mAP of Faster R-CNN with MobileNetV2 (0.55) as the backbone is 76% of that with ResNet50 (0.72). YOLOv4 owns the highest mAP. In YOLOv4, the AP of face-skin is the highest, followed by upper limbs, hands and lower limbs. Notably, they are

Table 4. Test results on test set.

Model	Backbone	AP						mAP
		Face	Foot	Hand	Lowerlimb	Trunk	Upperlimb	
Faster R-CNN	ResNet50	89.22%	50.74%	76.42%	75.58%	63.44%	78.48%	0.72
Faster R-CNN	MobileNetV2	80.79%	37.69%	56.39%	50.26%	42.38%	65.09%	0.55
YOLOv3	DarkNet-53	89.18%	47.35%	73.47%	73.84%	58.21%	77.45%	0.70
YOLOv4	CSPDarknet53	94.31%	55.76%	82.57%	81.55%	69.07%	84.30%	0.78
CenterNet	Hourglass	90.60%	41.49%	63.13%	61.21%	72.41%	64.49%	0.66

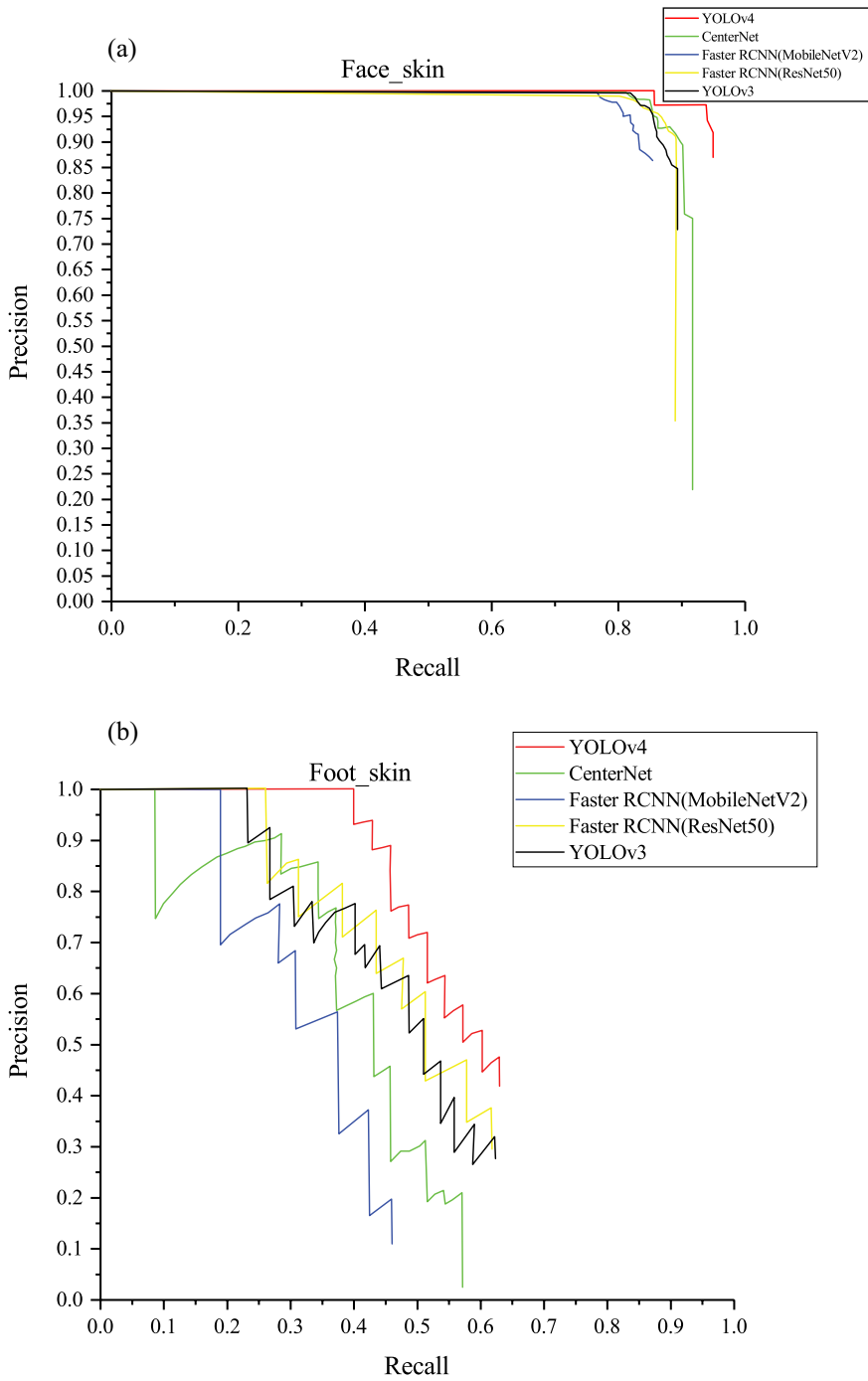


Figure 4. P-R curves of the six categories in the models.

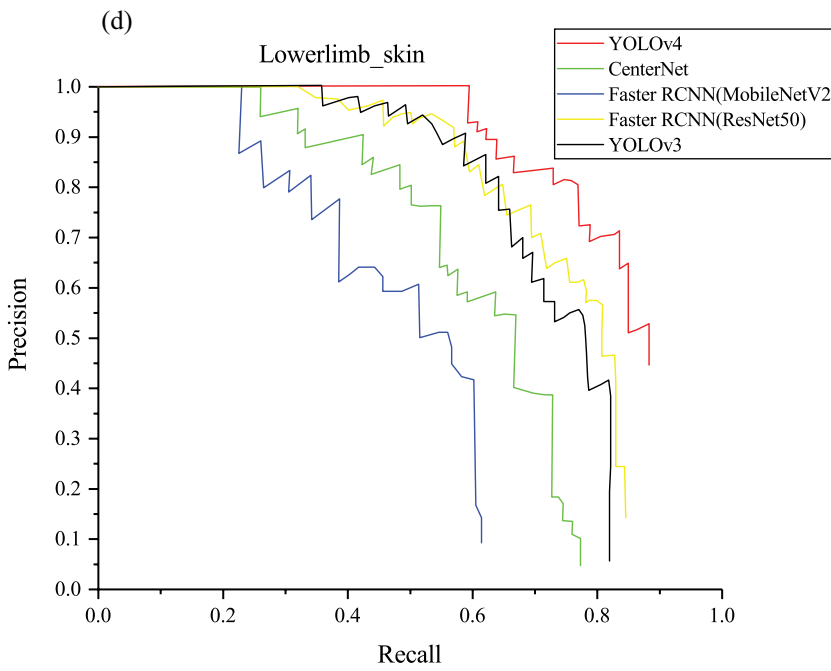
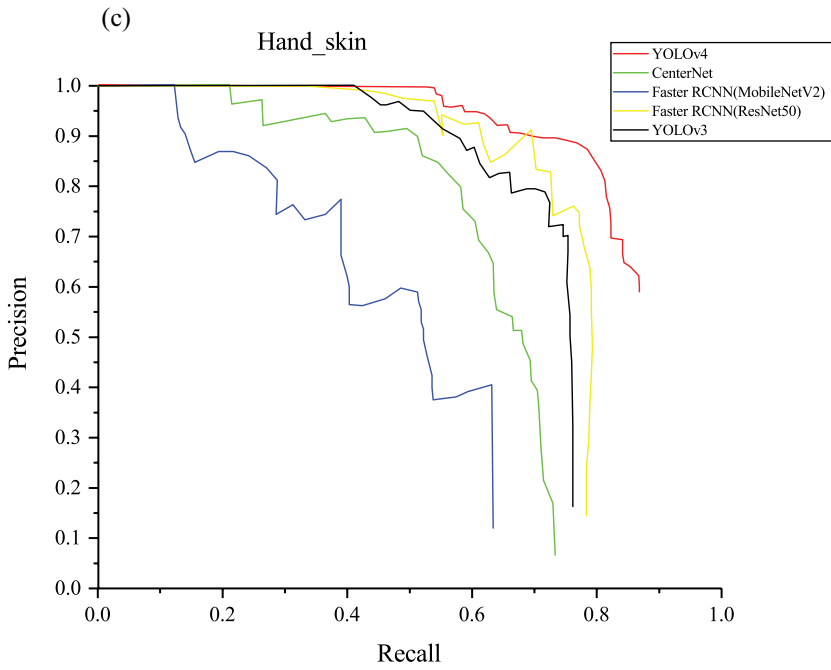


Figure 4b. Continue

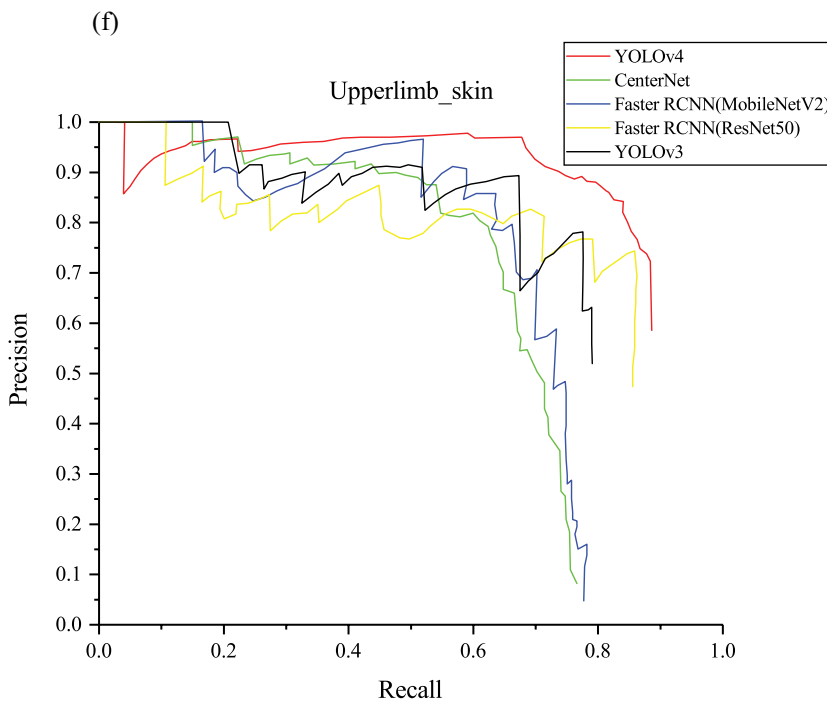
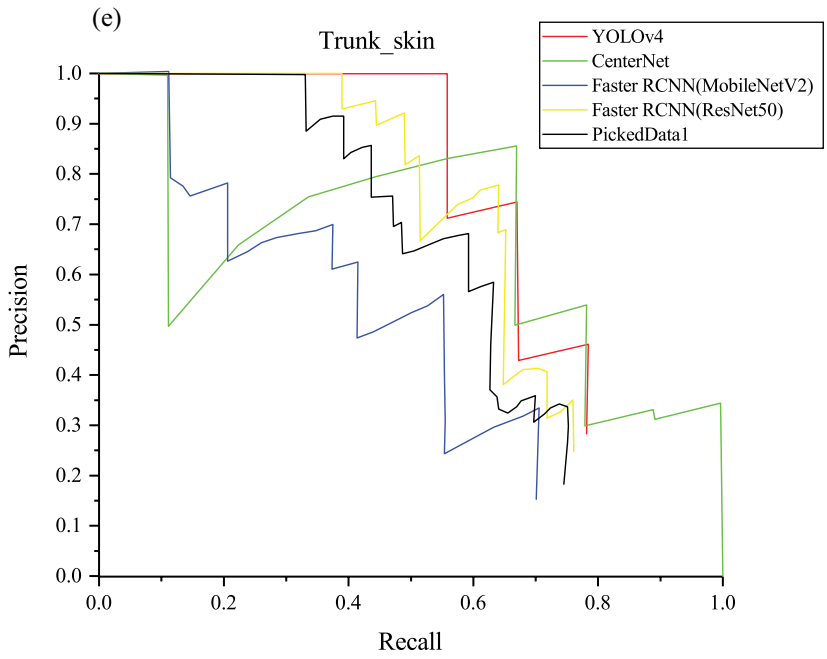


Figure 4c. Continue

all over 80%. The AP of trunk and feet are relatively low. In addition, CenterNet produces the highest AP for the trunk and YOLOv4 has the highest AP for other categories in all models.

Discussion

Due to personal privacy, the sample size of the trunk is comparatively small. In some cases, the people in the picture wear clothes to cover sensitive parts, as shown in the lower right corner of Figure 5. Lack of samples causes the learned features to be less useful for identifying the trunk. The part of the human feet occupies a small region carrying limited information in the whole range of the human body, as shown in Figure 5. The resolution of small objects is further reduced along with repeated down-sampling, resulting in further weakened feature information or even loss. In summary, the detection effect of trunk and



Figure 5. The example images for the trunk and feet problems.

foot is barely satisfactory by nature. Due to the fact that CenterNet has the best effect in detecting the trunk, we suspect that it may be due to the special coding method of CenterNet. However, the detection effect of CenterNet in other categories is not satisfactory. Overall, YOLOv4 is still the best model in our study.

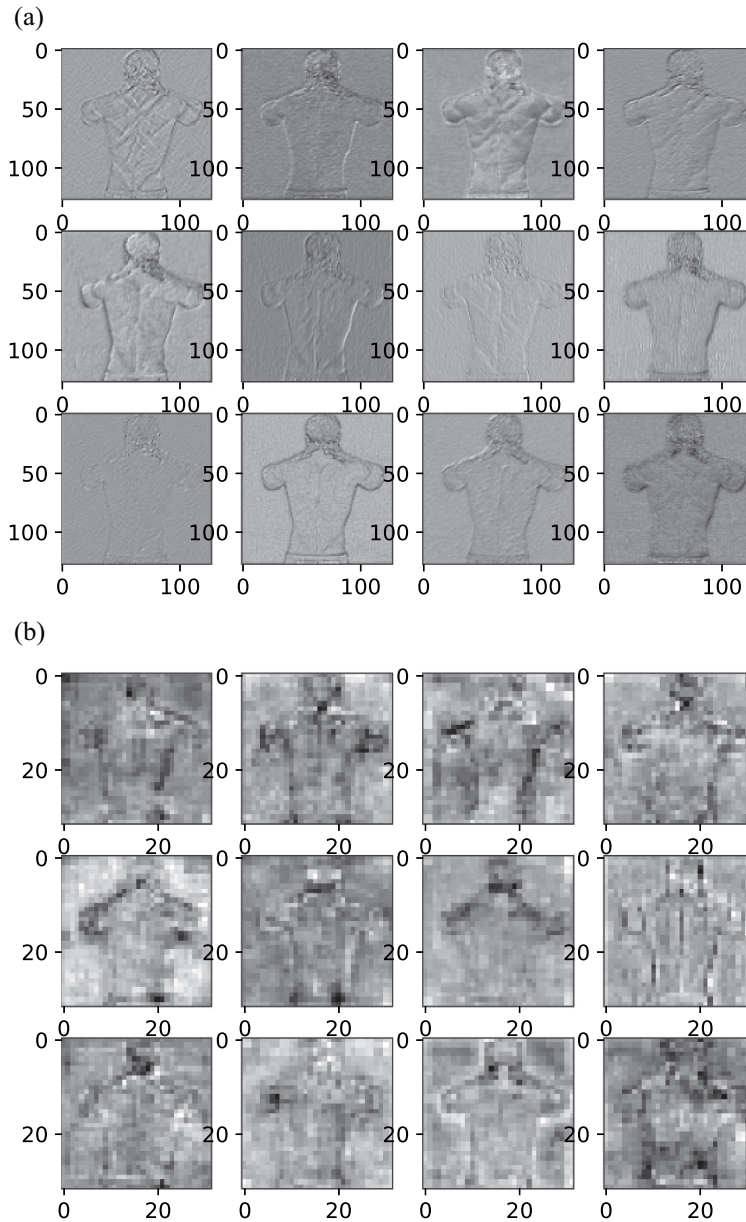


Figure 6. The feature map visualization results.

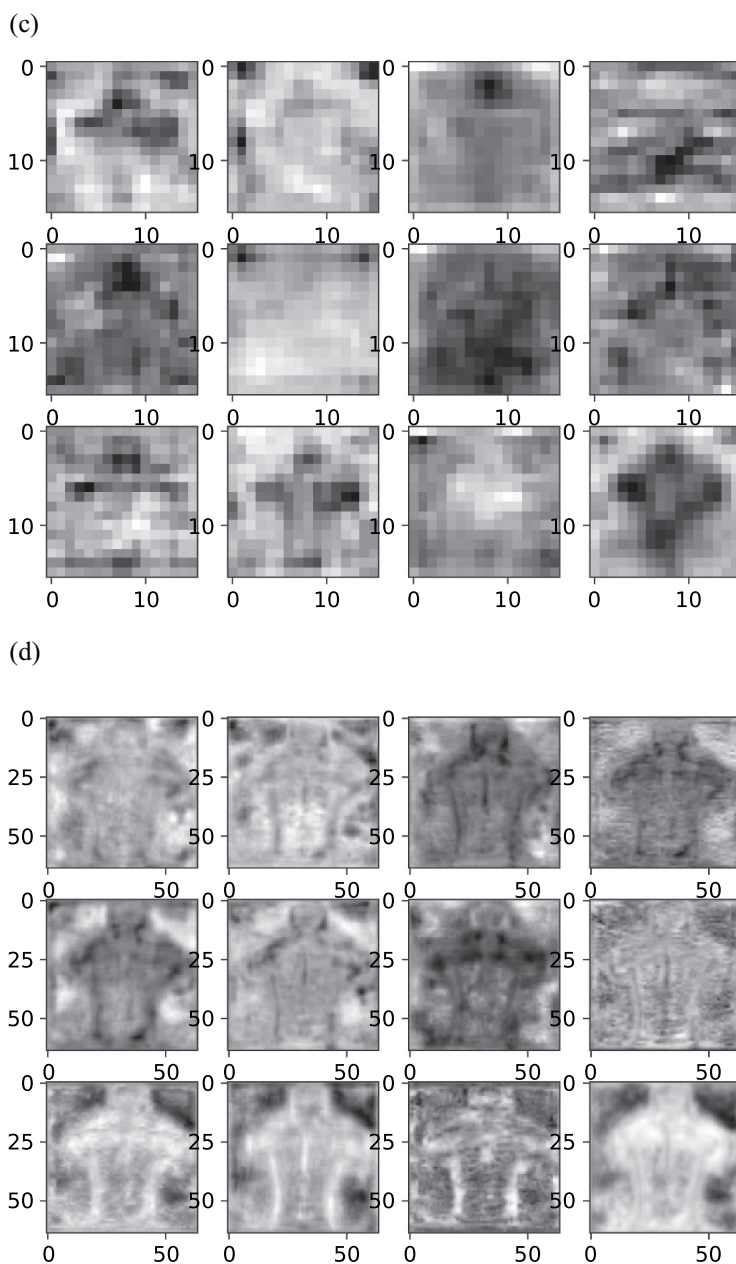


Figure 6b. Continue

As everyone knows, the neural network interpretability is weak. Generally, it depends on the test and experience to determine why the algorithm is effective or invalid. To observe the feature maps of the convolution layers, we visualize them. The shallow convolution layers extract the detailed features close to the original image data. With progressively deeper layers, the extracted effective features are less and

less and become more and more abstract. Considering the large number of convolution layers, we typically chose a portion of the visualization graphs for display, as shown in Figure 6, which takes the first 12 feature maps of some convolution layers. We can identify faintly what the specific object is in the shallow feature map in Figure 6(a), while we cannot do this in the deep feature maps which extract highly abstract features in Figure 6(c). The convolution layers shown in Figure 6(b) are located after Figure 6(a) and before Figure 6(c), which reflects the process from concrete features to abstract features extracted to a certain extent. The highlighted part in Figure 6(d) reflects the information concerned by the convolution layer. It can be seen that some convolution layers pay attention to the edge of the object, some pay attention to the interior of the object, and some pay attention to the background information.

Conclusion

With the help of the camera and the manipulator, as well as the relationship between their coordinates, we can control the robot arm to complete the bathing task on the human skin, identifying the skin is a key process. The experimental study found that YOLOv4 has the best recognition effect, and mAP reaches 78%, which is acceptable in practice. It is feasible to use object detection algorithm to detect the skin in bathing task. Then, we can increase the number of pictures in the data set, train the YOLOv4 model again to improve its detection effect, and deploy the model to realize the skin detection.

We can also add some auxiliary measures to ensure that the manipulator will not move to non-target areas, such as adding the tactile sensor. Owing to the safety requirements, the mechanical arm movement is finely controlled according to the feedback of the tactile sensor. After moving to the designated position, the manipulator can complete the bathing task performing pre-designed movement mode, such as the vertical or horizontal scrubbing. The main contributions of this paper are as follows

Verifying the feasibility and effectiveness of the object detection algorithm in performing the bathing task;

Conducting experimental investigations on the performance of four object detection algorithms under two classification standards for skin detection, and the performances are evaluated.

We put forward the tentative idea about the future work of the robot autonomous bath system.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Natural Science Foundation of China under Grant [62073224]; and the National Natural Science Foundation of China under Grant [61903255].

ORCID

Ping Li  <http://orcid.org/0000-0003-0361-4895>

References

- Bochkovskiy, A., C.-Y. Wang, and H.-Y. M. Liao. 2020. YOLOv4: Optimal speed and accuracy of object detection. *arXiv Preprint arXiv* 2004.10934.
- Chandra, S., S. Tsogkas, and I. Kokkinos. 2015. Accurate human-limb segmentation in RGB-D images for intelligent mobility assistance robots. International Conference on Computer Vision Workshop: 436-42. doi: [10.1109/ICCVW.2015.64](https://doi.org/10.1109/ICCVW.2015.64).
- Duan, K., S. Bai, L.-X. Xie, H.-G. Qi, Q.-M. Huang, and Q. Tian. 2019. Centernet: Keypoint triplets for object detection. International Conference on Computer Vision: 6568-77. doi: [10.1109/ICCV.2019.00667](https://doi.org/10.1109/ICCV.2019.00667).
- Dunlop, D. D., S. L. Hughes, and L. M. Manheim. 1997. Disability in activities of daily living: Patterns of change and a hierarchy of disability. *American Journal of Public Health* 87 (3):378-83. doi:[10.2105/AJPH.87.3.378](https://doi.org/10.2105/AJPH.87.3.378).
- Fang, G., N. M. Kwok, and G. Dissanayake. 2013. Skin colour detection using the statistical decision theory. *Advanced Materials Research* 694-697:1891-95. doi:[10.4028/scientific.net/AMR.694-697.1891](https://doi.org/10.4028/scientific.net/AMR.694-697.1891).
- Fotouhi, M., M. H. Rohban, and S. Kasaei. 2009. Skin detection using contourlet-based texture analysis. Fourth International Conference on Digital Telecommunications: 59-64. doi: [10.1109/ICDT.2009.18](https://doi.org/10.1109/ICDT.2009.18).
- Girshick, R. 2015. Fast R-CNN. International Conference on Computer Vision: 1440-48. doi: [10.1109/ICCV.2015.169](https://doi.org/10.1109/ICCV.2015.169).
- Girshick, R., J. Donahue, T. Darrell, and J. Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *Computer Vision and Pattern Recognition* 580-87. doi:[10.1109/CVPR.2014.81](https://doi.org/10.1109/CVPR.2014.81).
- He, K.-M., X.-Y. Zhang, S.-Q. Ren, and J. Sun. 2015a. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. International Conference on Computer Vision: 1026-34. doi: [10.1109/ICCV.2015.123](https://doi.org/10.1109/ICCV.2015.123).
- He, K.-M., X.-Y. Zhang, S.-Q. Ren, and J. Sun. 2015b. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (9):1904-16. doi:[10.1109/TPAMI.2015.2389824](https://doi.org/10.1109/TPAMI.2015.2389824).
- He, K.-M., X.-Y. Zhang, S.-Q. Ren, and J. Sun. 2016. Deep residual learning for image recognition. *Computer Vision and Pattern Recognition* 770-78. doi:[10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).

- Hussain, N., M. A. Khan, M. Sharif, S. A. Khan, A. A. Albeshar, T. Saba, and A. Armaghan. 2020. A deep neural network and classical features based scheme for objects recognition: An application for machine inspection. *Multimedia Tools and Applications* 1-23. doi:10.1007/s11042-020-08852-3.
- Ioffe, S., and C. Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning* 1:448–56.
- Kawulok, M., J. Kawulok, and J. Nalepa. 2014. Spatial-based skin detection using discriminative skin-presence features. *Pattern Recognition Letters* 41:3–13. doi:10.1016/j.patrec.2013.08.028.
- Khan, M. A., K. Muhammad, M. Sharif, T. Akram, and V. H. C. D. Albuquerque. 2021c. Multi-class skin lesion detection and classification via teledermatology. *IEEE Journal of Biomedical and Health Informatics*. doi:10.1109/JBHI.2021.3067789.
- Khan, M. A., M. S. Sarfraz, M. Alhaisoni, A. A. Albeshar, S. Wang, and I. Ashraf. 2020. StomachNet: Optimal deep learning features fusion for stomach abnormalities classification. *IEEE Access* 8:197969–81. doi:10.1109/ACCESS.2020.3034217.
- Khan, M. A., M. Sharif, T. Akram, R. Damaseviciu, and R. Maskeliunas. 2021d. Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization. *Diagnostics* 11 (5):1–26. doi:10.3390/diagnostics11050811.
- Khan, M. A., N. Hussain, A. Majid, M. Alhaisoni, S. A. C. Bukhari, S. Kadry, Y. Nam, and Y.-D. Zhang. 2021b. Classification of positive COVID-19 CT scans using deep learning. *Computers, Materials & Continua* 66 (3):2923–38. doi:10.32604/cmc.2021.013191.
- Khan, M. A., T. Akram, Y.-D. Zhang, and M. Sharif. 2021a. Attributes based skin lesion detection and recognition: A mask RCNN and transfer learning-based deep learning framework. *Pattern Recognition Letters* 143:58–66. doi:10.1016/j.patrec.2020.12.015.
- Khan, M. A., Y.-D. Zhang, and M. Sharif. 2021. Pixels to classes: Intelligent learning framework for multiclass skin lesion localization and classification. *Computers & Electrical Engineering* 90:106956. doi:10.1016/j.compeleceng.2020.106956.
- Law, H., and J. Deng. 2018. Cornernet: Detecting objects as paired keypoints. *European Conference on Computer Vision* 11218:765–81. doi:10.1007/978-3-030-01264-9_45.
- LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521 (7553):436–44. doi:10.1038/nature14539.
- Lin, T.-Y., P. Dollár, R. Girshick, K.-M. He, B. Hariharan, and S. Belongie. 2017. Feature pyramid networks for object detection. *Computer Vision and Pattern Recognition* 936–44. doi:10.1109/CVPR.2017.106.
- Liu, S., L. Qi, H.-F. Qin, J.-P. Shi, and J.-Y. Jia. 2018. Path aggregation network for instance segmentation. *Computer Vision and Pattern Recognition* 8759–68. doi:10.1109/CVPR.2018.00913.
- Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. 2016. SSD: Single shot multibox detector. *European Conference on Computer Vision* 9905:21–37. doi:10.1007/978-3-319-46448-0_2.
- Misra, D. 2019. Mish: A self regularized non-monotonic neural activation function. *arXiv Preprint arXiv* 1908.08681.
- Nadian, A., and A. Talebpour. 2011. Pixel-based skin detection using sinc function. *IEEE Symposium on Computers & Informatics*. doi: 10.1109/ISCI.2011.5958934.
- Newell, A., K.-Y. Yang, and J. Deng. 2016. Stacked hourglass networks for human pose estimation. *European Conference on Computer Vision* 9912:483–99. doi:10.1007/978-3-319-46484-8_29.
- Pattnaik, G., V. K. Shrivastava, and K. Parvathi. 2020. Transfer learning-based framework for classification of pest in tomato plants. *Applied Artificial Intelligence* 34 (13):981–93. doi:10.1080/08839514.2020.1792034.

- Rashid, M., M. A. Khan, M. Alhaisoni, S.-H. Wang, S. R. Naqvi, A. Rehman, and T. Saba. 2020. A sustainable deep learning framework for object recognition using multi-layers deep features fusion and selection. *Sustainability* 12 (12):5037. doi:10.3390/su12125037.
- Rashid, M., M. A. Khan, M. Sharif, M. Raza, M. M. Sarfraz, and F. Afza. 2019. Object detection and classification: A joint selection and fusion strategy of deep convolutional neural network and SIFT point features. *Multimedia Tools and Applications* 78 (12):15751–77. doi:10.1007/s11042-018-7031-0.
- Redmon, J., and A. Farhadi. 2018. YOLOv3: An incremental improvement. *arXiv Preprint arXiv* 1804.02767.
- Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. 2016. You only look once: Unified, real-time object detection. *Computer Vision and Pattern Recognition* 2016:779–88. doi:10.1109/CVPR.2016.91.
- Ren, S.-Q., K. He, R. Girshick, and J. Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems* 28: 91–99.
- Rezatofighi, H., N. Tsoi, J. Y. Gwak, A. Sadeghian, I. Reid, and S. Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. *Computer Vision and Pattern Recognition* 658–66. doi:10.1109/CVPR.2019.00075.
- Tan, W. R., C. S. Chan, P. Yogarajah, and J. Condell. 2012. A fusion approach for efficient human skin detection. *IEEE Transactions on Industrial Informatics* 8 (1):138–47. doi:10.1109/TII.2011.2172451.
- Uijlings, J. R. R., K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders. 2013. Selective search for object recognition. *International Journal of Computer Vision* 104 (2):154–71. doi:10.1007/s11263-013-0620-5.
- Wang, C.-Y., H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh. 2020. CSPNet: A new backbone that can enhance learning capability of CNN. *Computer Vision and Pattern Recognition Workshops* 1571–80. doi:10.1109/CVPRW50498.2020.00203.
- Werle, J., and K. Hauer. 2016. Design of a bath robot system – User definition and user requirements based on International Classification of Functioning, disability and health (ICF). *IEEE International Symposium on Robot and Human Interactive Communication*: 459–466. doi: 10.1109/ROMAN.2016.7745159.
- Wu, X.-W., D. Sahoo, and S. C. H. Hoi. 2020. Recent advances in deep learning for object detection. *Neurocomputing* 396:39–64. doi:10.1016/j.neucom.2020.01.085.
- Zhang, S.-F., C. Chi, Y.-Q. Yao, Z. Lei, and S. Z. Li. 2020. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. *Computer Vision and Pattern Recognition* 9756–65. doi:10.1109/CVPR42600.2020.00978.
- Zheng, Z.-H., P. Wang, W. Liu, J.-Z. Li, R.-G. Ye, and D.-W. Ren. 2020. Distance-IOU loss: Faster and better learning for bounding box regression. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (7):12993–3000. doi:10.1609/aaai.v34i07.6999.
- Zhu, J.-Q., and C.-H. Cai. 2011. Region growing based high brightness skin detection. 10th International Symposium on Signals, Circuits and Systems. doi: 10.1109/ISSCS.2011.5978652.
- Zlatintsi, A., A. C. Dometios, N. Kardaris, I. Rodomagoulakis, P. Koutras, X. Papageorgiou, P. Maragos, C. S. Tzafestas, P. Vartholomeos, K. Hauer, et al. 2020. I-Support: A robotic platform of an assistive bathing robot for the elderly population. *Robotics and Autonomous Systems* 103451. doi:10.1016/j.robot.2020.103451.