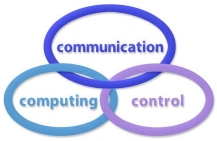
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 16, Issue: 2, Month: April, Year: 2021
Article Number: 4112, https://doi.org/10.15837/ijccc.2021.2.4112



CCC Publications

A New Semantic-Based Tool Detection Method for Robots

W.B. Chen, C. He, W.Z. Chen, Q.L. Chen, P.L. Wu

Wenbai Chen*, Chao He, Weizhao Chen, Qili Chen
School of Automation
Beijing Information Science & Technology University
Beijing 100192, China
*Corresponding author: chenwb03@126.com
1026853315@qq.com
670924002@qq.com
qilichen@hotmail.com

Peiliang Wu
School of Information Science & Technology
Yanshan University
Qinhuangdao 066004, China
peiliangwu@ysu.edu.cn

Abstract

Home helper robots have become more acceptable due to their excellent image recognition ability. However, some common household tools remain challenging to recognize, classify, and use by robots. We designed a detection method for the functional components of common household tools based on the mask regional convolutional neural network (Mask-R-CNN). This method is a multi-task branching target detection algorithm that includes tool classification, target box regression, and semantic segmentation. It provides accurate recognition of the functional components of tools. The method is compared with existing algorithms on the dataset UMD Part Affordance dataset and exhibits effective instance segmentation and key point detection, with higher accuracy and robustness than two traditional algorithms. The proposed method helps the robot understand and use household tools better than traditional object detection algorithms.

Keywords: Functional Components, Mask-R-CNN Network, Tool Classification, Functional Semantics.

1 Introduction

Tool use marks the beginning of human civilization, and cognitive function is the prerequisite for using tools. Cognitive science research has shown that the functional cognition of tools in humans is a process in which structural information on the tools is obtained by various sensory organs using primarily stereo vision. The information on the functional components of tools is processed by the brain and organically combined to represent the tool [1]. In the process of human cognition, a model of a tool is typically constructed by learning (the idea of clustering and supervised learning) the semantic

features and functions of similar tools. In recent years, Li's research team [2, 3] and Saxena's research team [4] almost unanimously agreed that functional cognition is crucial for detecting semantic features in objects and scenes and a vital component of human cognition. Functional semantic information can provide a more meaningful semantic description than a name or spatial location; it plays an important role in human activities, is used in daily language, and deeply profoundly people's understanding and interactions. In robot research, the development of machine intelligence has been based on imitating human abilities. At present, home service robots can imitate human perception, although there are limits, such as the ability of the robot to interact with people naturally and provide a pleasant service. The important reason is that the robot's cognition of tools is not comprehensive and in-depth. The research on service robot tool cognition has gone through two stages, namely, the perception stage (recognition and positioning) and the cognitive stage (natural interaction). In the perception stage, the tool representation is primarily based on the underlying features, such as color, texture, and depth. The features extracted in this stage are suitable for tool detection and location. However, due to the lack of consideration of the semantics of objects and scenes, the robot cannot recognize the service environment in the same or similar manner as human beings, making it difficult to have natural interactions with people and to achieve satisfactory service. In the cognitive stage, the middle-level and high-level semantic information of objects are considered. By adding this semantic information to the traditional perceptual information, we establish a process for learning the functional components of tools similar to the human cognition process. Therefore, robots can understand and use the tools using semantic information and imitating the human cognitive mode.

2 Releated Work

At present, the relationship between the robot and the tool and its components can be obtained and analyzed using various sensors, but information at the semantic level has to be obtained using semantic tags. However, it is difficult for the robot to obtain the semantic tags in a natural environment. In recent years, research has emerged on active functional cognition methods based on reasoning and learning [2, 4]. Scholars have used functional methods to analyze functional representation and perform tool classification, such as three-dimensional (3D) CAD models of chairs and other objects [5]. Subsequently, numerous studies focused on the detection of object grabbing points in two-dimensional (2D) images based on apparent features [6, 7, 8]. With the emergence of red, green, blue depth (RGB-D) sensors, such as Optrima and Kinect, fast and low-cost 3D data acquisition has become possible, leading to novel research results in tool function detection. Grabner et al. detected a surface suitable for sitting in 3D data [9], Kjellstrom et al. classified the function of tools from a video [10], and Zhu et al. proposed a task-oriented object model for learning construction patterns [11]. Hassan et al. integrated human, object, and environmental aspects for attribute modeling and achieved accurate functional detection [12]. Kemp et al. detected a tool tip that can be grasped by a robot [13]. Mar et al. determined the correct grasping method for grasping a tool [14], and Lenz et al. used a sparse autoencoder (SAE) to detect the grab position of an object [15]. Redmon detected the object's grasping position based on a deep learning method [16]. Myers et al. used a structured random forest (SRF) and the superpixel-based hierarchical matching pursuit (S-HMP) to detect seven functional components of common household tools, namely, grab, cut, scoop, contain, pound, support, and wrap-grasp [17]. Abelha et al. used a model-based approach to find the best alternative tool [18]. Based on ontology and semantics, Worgotter constructed six component primitives to describe the functional components of tools, including contain, cut, hit, hook, poke, and sift and applied them to the functional analysis of tools [19]. In the above-mentioned research, Refs. [15, 16] only focused on the grasp component of the tool but did not consider other functions. Reference [17] analyzed seven different functional components and their detection. Reference [18] regarded the tool as a single component and developed tool models and alternative tool searches based on 3D visual data. The same kind of tools may have different shapes, materials, and colors; however, they can be regarded as organic components of several main functional components. In this study, we focus on the bottom-up cognitive mode of human beings and construct an edge feature representation of the functional components of the tool. The tool representation and model of the functional semantic combination

of the components is established in the high-level semantic space. We perform tool classification and similarity assessment of different tools based on the tool function. We use a deep learning framework, i.e., mask regional convolutional neural network (Mask R-CNN), to implement the model for detecting the functional components of the tool. Mask R-CNN accurately detects the target in the image and generates high-quality segmentation results for each instance. The method is more accurate for target detection than a single detection network. In addition, due to the pixel-level annotation provided in model training, sufficient information can be obtained, and the target and background can be classified precisely. Therefore, the Mask R-CNN algorithm provides high accuracy.

3 Methods

The detection model to obtain the functional components of the tool is divided into an offline training stage and an online detection stage. The framework of the proposed method is illustrated in (Figure 1).

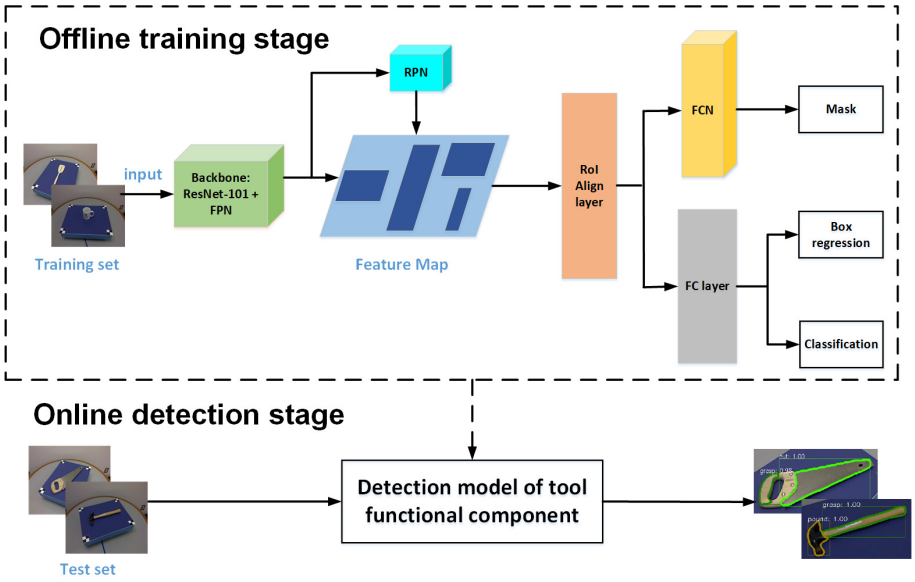


Figure 1: The framework of the proposed method

- **Offline training stage**. First, we perform a deep convolution operation on the training image to extract the feature map of the target image. We divide the original image into multiple anchor boxes using the region proposal network (RPN) and perform classification and regression analysis to obtain the region of interest (ROI) with a high score. Then, we perform two operations on the feature map. The first is to input the feature map of the corresponding size obtained from the feature extraction stage into a fully-connected neural network for classification and regression. The second is to input the feature map into the mask branch module and classify the feature maps by pixel to obtain the segmentation results. Finally, training is used to obtain the final detection model for the tool's functional components.
- **Online detection stage**. We input the target image into the trained detection model. The corresponding feature data are obtained from the deep convolutional network and the RPN and input into different network modules, such as the fully-connected neural network and the Mask network, to obtain the test results.

3.1 The Detector of the Tool's Functional Components

The Mask-R-CNN framework is used for detecting and positioning the functional components of common household tools to facilitate the recognition of the tools by the home service robot. The

Mask-R-CNN detection framework has a MASK module for object detection and functional instance segmentation to classify the tool's functional components. The system model is based on the concept of the Faster-R-CNN, which is also a two-stage detector. The first stage is the RPN, which creates ROIs. The second stage consists of feature extraction in the ROIs proposed in the first stage and subsequent three-step processing, i.e., classification, border regression, and segmentation. The features of the two stages are extracted by the underlying backbone network and are shared to increase the prediction speed. In the offline training stage, several processing operations are performed on the input image using the following modules:

- **Backbone Network**: The ResNet network is used to extract the features of the input image to obtain the feature map of the entire image.
- **Feature Pyramid Networks (FPN)**: The FPN network is used to extract different levels of target features; large targets are detected with deep features, and small targets are detected with shallow features.
- **Region Proposal Network (RPN)**: A sliding window strategy is used, and the original image is divided into multiple anchors. Classification and bounding box regression are performed on all anchors, and the RoI with the highest score is extracted. Non-maximum suppression is performed, and the results are input to the ROI Align Layer for feature extraction.
- **Region of Interest (ROI) Align**: The features in the ROI selected by the RPN are extracted from the feature map proposed by the backbone network (the value of the floating-point position is obtained by nonlinear interpolation), and the fixed-size feature map of each ROI is obtained.
- **Classification and Regression**: The ROI selected by the RPN is classified, and bounding box coordinate regression is performed. The classification module uses the Softmax activation function.
- **Segmentation**: The Mask network branch uses the FCN network to segment each RoI into K output units with the size of m^2, K is the number of categories, m is the length and width of the module. Each category corresponds to a segmentation module (to minimize the competition between the classes in the segmentation task). The Sigmoid activation function is used for each pixel.

3.2 Backbone Network

The ResNet50-FPN network is used as the network architecture of the component detection model. ResNet has a residual network structure. ResNet provides high accuracy due to the depth of the network layers. The structure of ResNet is shown in (Figure 2). ResNet50 contains 50 convolution operations and 4 residual blocks.

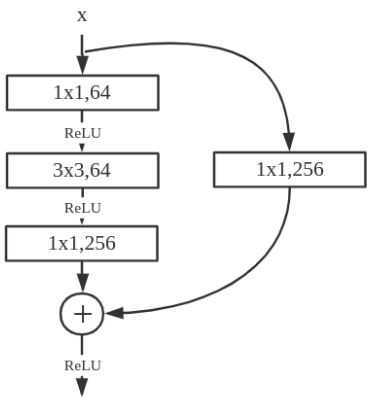


Figure 2: The structure of ResNet

The ResNet is combined with the FPN for multi-scale detection. The FPN is a well-designed multi-scale detection method. The FPN consists of three types of connections: bottom-up, top-down, and horizontal connections. The schematic diagram is shown in (Figure 3). The structure integrates the characteristics of each level to obtain sufficient semantic and spatial information.

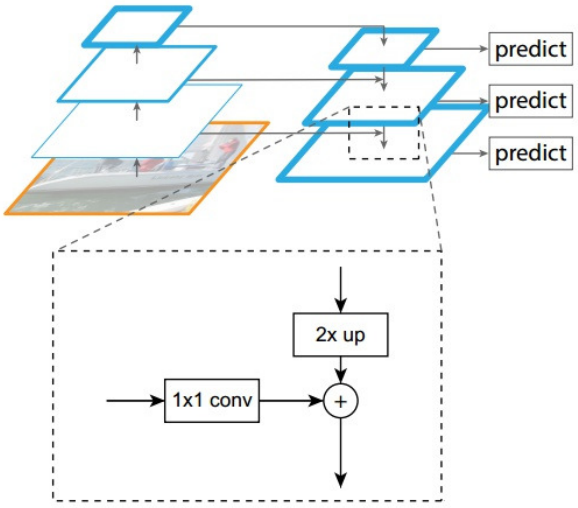


Figure 3: The schematic diagram of the FPN

The feature extraction network ResNet50+FPN is shown in (Figure 4). The image information is input into the residual network for feature extraction, and the FPN and the ResNet are combined to perform multi-scale feature extraction.

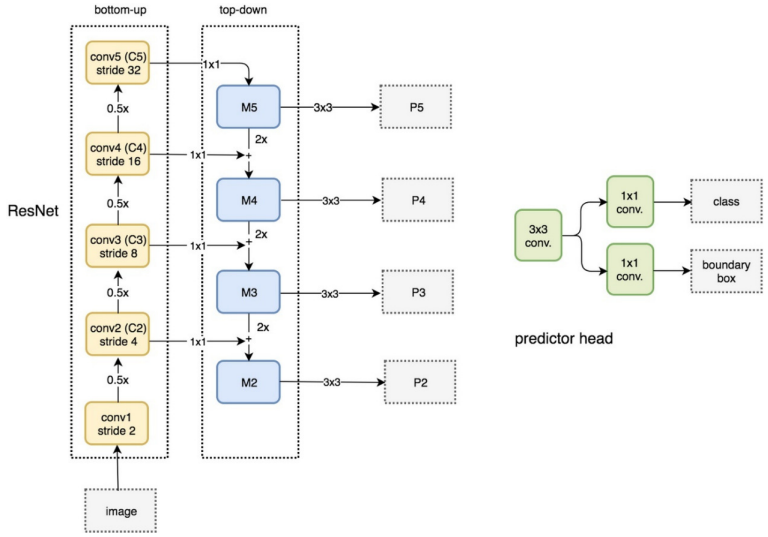


Figure 4: The structure of ResNet50+FPN

3.3 Mask Branch

This system model uses ResNet101+FPN as a feature extraction network to achieve state-of-the-art performance. In addition, we use ROIAlign instead of ROIPooling to improve the pooling operation. An interpolation process is used (bilinear interpolation to 14*14 and pooling to 7*7) to minimize the misalignment problem caused by direct sampling using only pooling. Finally, we add the Mask branch to achieve semantic segmentation. The Mask-R-CNN framework uses multi-task loss, total loss (L), classification loss (L_{cls}), prediction box loss (L_{box}), and mask loss (L_{mask}). The classification loss is defined as the cross-entropy loss function, the regression loss function is defined as the SmoothL1 loss

function, and the segmentation loss function is defined as the average binary cross-entropy loss. The schematic diagram of the loss functions is shown in (Figure 5).

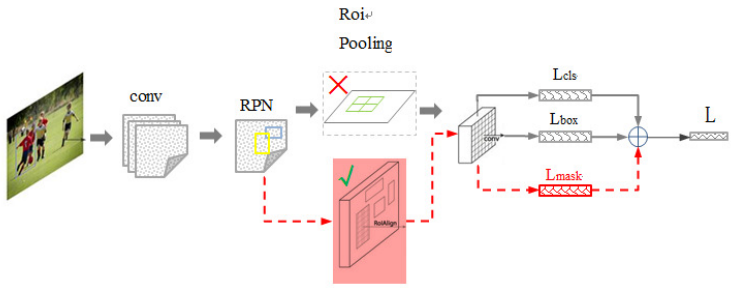


Figure 5: Schematic diagram of the loss definition of the Mask R-CNN

The Mask R-CNN is based on the Faster-R-CNN, but the feature extraction uses the ResNet-FPN architecture, and the Mask prediction module is added. There are three main differences between the Mask R-CNN and the Faster R-CNN framework:

(1) The improved ResNet-FPN structure is adopted in the basic feature extraction network, and the multi-layer feature map is conducive to the detection of multi-scale objects and small objects.

(2) The ROI Align method is proposed to replace ROI Pooling. Since ROI Pooling has two rounding operations, the extracted features deviate from the initial regression position; this case is referred to as a misalignment, which affects the accuracy of the detection and segmentation tasks. Therefore, Mask R-CNN does not use the rounding operation but retains all floating-point numbers. It obtains the values of multiple sampling points through bilinear interpolation and performs the maximum pooling operation on the values of multiple sampling points to obtain the final value.

(3) After obtaining the features of the ROI, the Mask module is added to perform classification and regression and predict the pixel category.

4 Experiment and Results

4.1 Dataset

This experiment uses the University of Maryland (UMD) Part Affordance dataset [17], which is a relatively complete dataset for the functional components of tools. The RGB-D information of 17 household tools, including kitchen and gardening tool, are included, namely turners, trowel, tenderizer, spoon, shovel, shear, scoop, scissors, saw, pot, mug, mallet, ladle, hammer, cup, and bowl. Each tool is depicted in nearly 300 different perspectives, for a total of more than 30,000 sets of RGB-D data. About one-third of the tools have been marked for component functionality. The ratio of training data to test data is about 4:1. (Table 1) lists seven functional components of tools and their descriptions.

Table 1: Seven functional components of tools in the UMD Part Affordance dataset

Function	Description
grasp	It can be grasped by hand. (handle)
cut	It can be used to separate objects. (blade)
scoop	It has a curved surface for collecting soft materials. (spoon)
contain	It has a deep cavity to hold the liquid. (bowl)
pound	It can be used to strike other objects. (hammerhead)
support	It can accommodate loose flat parts. (scraper)
wrap-grasp	It can be held by hand and palm. (cup, outer surface)

4.2 Model Training and Loss Function

The Pytorch framework was used in our experiments, and the basic code was the Mask R-CNN Benchmark from Facebook Research [20, 21]. We used the pre-trained ResNet-50 model for initialization, and the pre-training weights were the training weights of Mask R-CNN on the COCO dataset. The GPU was an NVIDIA TITAN Xp with 10 GB of memory. The training and test images had a size of 768×768. During training, the batch size was 2, the initial learning rate was 0.0025, and the weight attenuation was 0.0001. The maximum number of iterations was 72000. The training process required 10 hours. The loss curves obtained during the training process are shown in (Figure 6). (a), (b), (c), and (d) show the total loss L, the classification loss L_{class}, the prediction box loss L_{box}, and the Mask branch loss L_{mask}. The value of the loss function L and $L_{\text{class}} + L_{\text{box}} + L_{\text{mask}}$, in the Mask R-CNN was minimized, and the optimal model was used to predict the functional components of the tool. The trained model was applied to predict and analyze with test data.

The loss function of the Mask R-CNN is defined as:

$$L = L_{\text{class}} + L_{\text{box}} + L_{\text{mask}} \tag{1}$$

where $L_{\text{class}} + L_{\text{box}}$ are identified the same as in the Faster R-CNN:

$$L_{\text{class}} + L_{\text{box}} = \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}} \left(p_i, p_i^*\right) + \frac{1}{N_{\text{box}}} \sum_i p_i^* L_1^{\text{smooth}} \left(t_i - t_i^*\right) \tag{2}$$

$$L_{cls} \left(\{p_i, p_i^*\}\right) = -p_i^* \log \left(1 - p_i^*\right) - \left(1 - p_i^*\right) \log \left(1 - p_i^*\right) \tag{3}$$

The L_{mask} is the average binary cross-entropy loss:

$$L_{\text{mask}} = -\frac{1}{m^2} \sum_{1 \le i,j \le m} \left[y_{ij} \log y_{ij}^k + \left(1 - y_{ij}\right) \log \left(1 - y_{ij}^k\right) \right] \tag{4}$$

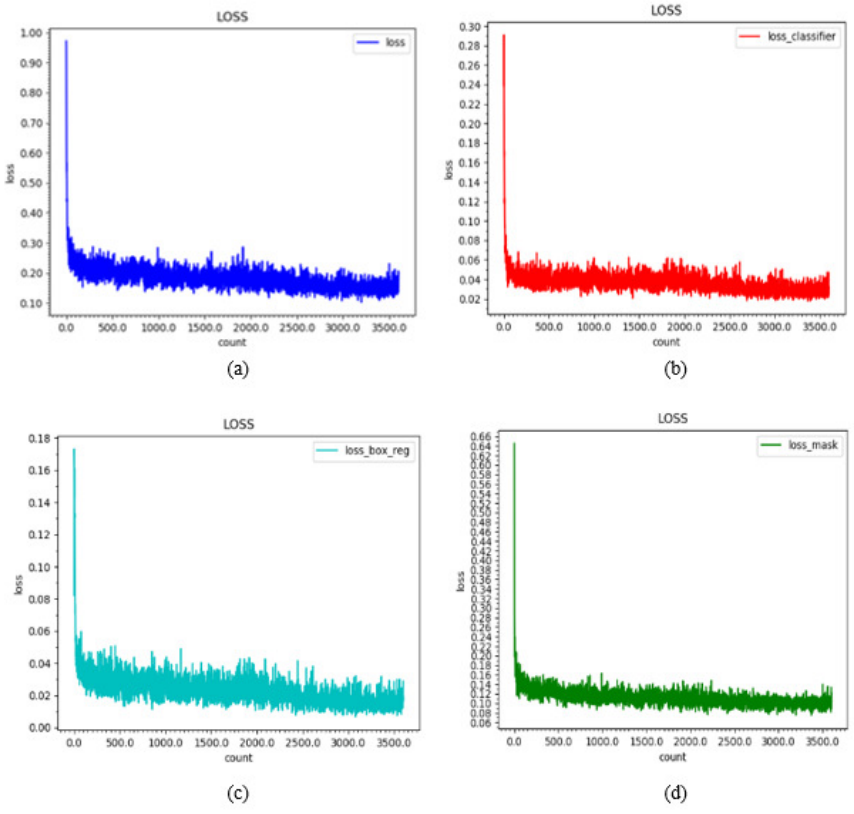


Figure 6: The loss curves

4.3 Results

In this study, 80% of the 20186 images in the tool dataset were used as the training set and 20% as the test set. The model performance evaluation index was the mean average precision (mAP). The AP (IoU=0.50) of each functional component of the model on the test set is shown in (Table 2).

Table 2: Detection accuracy and segmentation accuracy of the Mask R-CNN model

type	Bbox	segm
Grasp	0.7931	0.8155
Cut	0.9375	0.9586
Scoop	0.9973	0.9973
Contain	0.9886	0.9886
Pound	0.9925	0.9735
Support	1.0000	1.0000
Warg-grasp	1.0000	1.0000
mAP	0.9584	0.9622

The test results indicate that the proposed model detects the functions of the tool's components accurately, with an mAP of more than 90%. Therefore, the model is suitable for use in home service robots to detect common household tools. We used traditional machine learning algorithms and deep learning-based detection algorithms to detect the functional components of common household tools. The accuracy of the component detection algorithm based on the deep learning framework Mask-R-CNN is higher than that of traditional machine learning algorithms, such as the functional component detection algorithm based on SRF; however, the former requires more computing resources than the latter. The detection accuracies of different methods are listed in (Table 3).

Table 3: Detection accuracies of different methods

Type	Proposed method	SRF	S-HMP	SRF + coarse-to-fine	RGB-D part-based
grasp	0.815	0.314	0.367	0.554	0.50
cut	0.958	0.285	0.373	0.224	0.57
scoop	0.997	0.412	0.415	0.573	0.37
contain	0.988	0.635	0.810	0.605	0.68
pound	0.973	0.429	0.643	0.511	0.23
support	1.000	0.481	0.524	0.489	0.49
warg-grasp	1.000	0.666	0.767	0.787	0.36

(Figure 7) (a) shows the detection result of Ref. [22], (Figure 7) (b) shows the detection result of Ref. [23], and (Figure 7) (c) shows the experimental result of the proposed method. The objective of the detection methods in Refs. [22] and literature [23] is to detect the functional component of the tools. In (a), the functional components of the tools from left to right are contain, cut, grasp, pound, scoop, support, and wrap-grasp. If the tool has this feature, the position is highlighted in the detection result. In (b), the functions are the same as (a). The background color is blue, the target component is red, and the other functional components are yellow. The detection of functional components of the tools in this study focuses on the home service robot's cognition of the tools. The objective is to provide the service robot with an understanding of the tool, including its usage method and the functions of various components. Thus, the goal of this experiment is the detection of the functional components in a tool. As shown in (c), all the functional components of the tools, such as spatula, hammer, soup ladle, fruit knife, mug, and shovel (from left to right), are detected correctly. The detection results depict the segmentation results of the functional component instances and their classification, the borders, and the confidence score. Thus, the proposed algorithm exhibits higher accuracy than the other two algorithms and has advantages in terms of practicability.

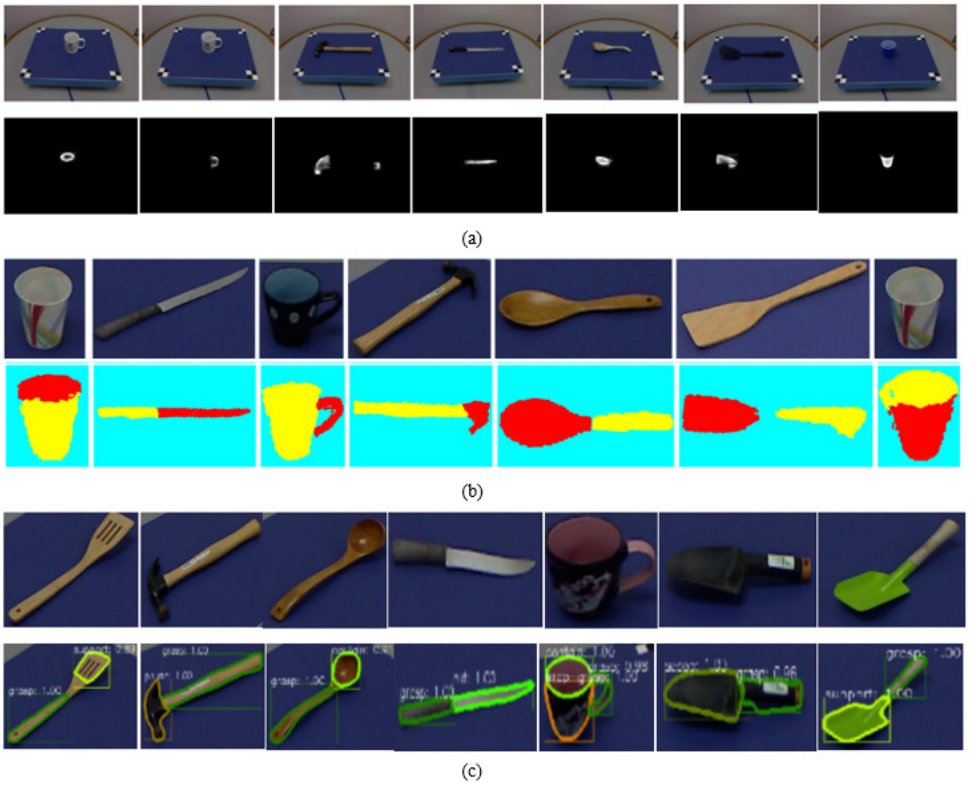


Figure 7: The results of the functional component detection and instance segmentation results of different algorithms

5 Conclusion

We proposed a method for detecting the functional parts of common tools using the Mask-R-CNN network for use in intelligent robots. The method extracted features from the ResNet+FPN network layer, performed classification and regression of the target frame using the RPN layer and the ROI Align layer, and included a module to obtain the semantic meaning of the segmented region after full convolution for up-sampling to detect the functional components of various tools. The multi-task loss function used in the network training of the algorithm included the classification loss, target frame regression loss, and mask segmentation loss. The value of the loss function was reduced in the learning process until the global optimal solution was obtained. The research results showed that the proposed method could identify the categories of commonly used household tools and their functional components efficiently and accurately, indicating that a household service robot would be able to detect and understand the usage of commonly used tools. The algorithm had higher accuracy and robustness than the two traditional machine learning algorithms. In a future study, we will optimize the detection algorithm to recognize common household tools in multiple complex scenes with multiple objects at different viewing angles and different light intensities.

Acknowledgment

This work is supported by Beijing Natural Science Foundation (Grant No. 4202026) and the National Key R & D Program of China (Grant No. 2018YFB1308300).

References

[1] Gibson, J. J. (1977). The theory of affordances. *Hilldale*, USA, 1(2).

[2] Zhu, Y., Fathi, A., Fei-Fei, L. (2014, September). Reasoning about object affordances in a knowl-

edge base representation. *In European conference on computer vision* (pp. 408-424). Springer, Cham.

[3] Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., Fei-Fei, L. (2014). Affordances provide a fundamental categorization principle for visual scenes. *arXiv preprint arXiv:1411.5340*.

[4] Koppula, H. S., Saxena, A. (2014, September). Physically grounded spatio-temporal object affordances. *In European Conference on Computer Vision* (pp. 831-847). Springer, Cham.

[5] Stark, L., Bowyer, K. (1994). Function-based generic recognition for multiple object categories. *CVGIP: Image Understanding*, 59(1), 1-21.

[6] Bohg, J., Kragic, D. (2009, June). Grasping familiar objects using shape context. In *2009 International Conference on Advanced Robotics* (pp. 1-6). IEEE.

[7] Saxena, A., Driemeyer, J., Ng, A. Y. (2008). Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, 27(2), 157-173.

[8] Stark, M., Lies, P., Zillich, M., Wyatt, J., Schiele, B. (2008, May). Functional object class detection based on learned affordance cues. In *International conference on computer vision systems* (pp. 435-444). Springer, Berlin, Heidelberg.

[9] Grabner, H., Gall, J., Van Gool, L.(2011, June). What makes a chair a chair?. *In CVPR 2011* (pp. 1529-1536). IEEE.

[10] Kjellström, H., Romero, J., Kragić, D. (2011). Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, 115(1), 81-90.

[11] Zhu, Y., Zhao, Y., Chun Zhu, S. (2015). Understanding tools: Task-oriented object modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2855-2864).

[12] Hassan, M., Dharmaratne, A. (2015, November). Attribute based affordance detection from human-object interaction images. In *Image and Video Technology* (pp. 220-232). Springer, Cham.

[13] Kemp, C. C., Edsinger, A. (2006, June). Robot manipulation of human tools: Autonomous detection and control of task relevant features. In *Proc. of the Fifth Intl. Conference on Development and Learning* (Vol. 42).

[14] Mar, T., Tikhanoff, V., Metta, G., Natale, L. (2015, November). Multi-model approach based on 3D functional features for tool affordance learning in robotics. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)* (pp. 482-489). IEEE.

[15] Lenz, I., Lee, H., Saxena, A. (2015). Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5), 705-724.

[16] Redmon, J., Angelova, A. (2015, May). Real-time grasp detection using convolutional neural networks. In *2015 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1316-1322). IEEE.

[17] Myers, A., Teo, C. L., Fermüller, C., Aloimonos, Y. (2015, May). Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1374-1381). IEEE.

[18] Abelha, P., Guerin, F., Schoeler, M. (2016, May). A model-based approach to finding substitute tools in 3d vision data. In *2016 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 2471-2478). IEEE.

[19] Schoeler, M., Wörgötter, F. (2015). Bootstrapping the semantics of tools: Affordance analysis of real world objects on a per-part basis. *IEEE Transactions on Cognitive and Developmental Systems*, 8(2), 84-98.

[20] Massa, F., Girshick, R. (2018). maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. *Accessed*: Apr, 29, 2019.

[21] [Online]. Available: https://github.com/facebookresearch/maskR-CNN-benchmark

[22] Peiliang Wu, Ben He, Lingfu Kong. (2017) A classification method of household daily tools based on functional semantic combination of components. *Robots*, 39(06): 786-794.

[23] Lakani, S. R., Rodríguez-Sánchez, A. J., Piater, J. (2019). Towards affordance detection for robot manipulation using affordance for parts and parts for affordance. *Autonomous Robots*, 43(5), 1155-1172.

Journal's webpage: http://univagora.ro/jour/index.php/ijccc/

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:

Chen W.B., He C., Chen W.Z., Chen Q.L., Wu P.L. (2021). A New Semantic-Based Tool Detection Method for Robots, *International Journal of Computers Communications & Control*, 16(2), 4112, 2021.

https://doi.org/10.15837/ijccc.2021.2.4112