ICIC Express Letters
Part B: Applications
Volume **14**, Number **7**, July **2023**

ICIC International ©2023 ISSN 2185-2766
pp. **701–708**

RESEARCH ON TOOLS AFFORDANCE COGNITION FOR HOUSEHOLD SERVICE ROBOT

Peiliang Wu[1,2,*], Zeshuo Wu[1], Chongchong Liu[1], Youyuan Qu[1]
Bingyi Mao[1,2] and Wenbai Chen[3]

[1]School of Information Science and Engineering
[2]The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province
Yanshan University
No. 438, West Hebei Avenue, Qinhuangdao 066004, P. R. China
{ wuzeshuo; moke }@stumail.ysu.edu.cn; { quyouyuanyingying; ysdxmby }@163.com
[*]Corresponding author: peiliangwu@ysu.edu.cn

[3]School of Automation
Beijing Information Science and Technology University
No. 12, Xiaoying East Road, Qinghe, Haidian District, Beijing 100192, P. R. China
chenwb@bistu.edu.cn

Received November 2022; accepted February 2023

Abstract. *Functional detection refers to the potential use of identifying objects and is the basis for robot perception and operation. In this paper, based on deep learning and AffNet-DR and OSAD algorithm, the affordance representation method and affordance cognitive training method of the tool are designed to meet the affordance cognitive requirements of daily tools for family service robots. The system can identify the affordance of daily tools in real time. It has a simple user interface, can mark the affordance of the tool in real time, and has a preliminary understanding of the unseen tools. Meanwhile, in the face of different input devices and weak platform computing, it also satisfies the requirements of simple cross-platform.*
Keywords: Affordance, Deep learning, AffNet-DR, OSAD

1. **Introduction.** Affordance detection refers to identifying the potential action possibility of objects in the image, which is an important ability of robot perception and operation. Adding utility to robotic tasks has been a particularly slow area of research; one of the main reasons is the abstraction of concepts. Finding the affordance relationship suitable for object and environment is a challenging problem, which has always been of great interest to the robot industry.

Gibson, a cognitive psychologist, proposed the affordance [1] which is used to indicate that human beings can dig out the hidden value of things in certain environments. The problem of visual affordance detection has been studied for decades [2] for perceiving action intent and being able to infer visual apocalypse from images/videos of humans and objects. For example, the affordance of scissors is to make people cut paper and cotton, because it has a handle for people to hold, and a sharp staggered double edge can cut some flat objects. Although this kind of utility requires people's subjective learning ability, objectively speaking, the commonly used affordance of object tools in life can be classified and counted. A few methods infer visual affordances for grasping with simple grippers [3] and explore non-robotics affordances. Recent work has shown that visual models can help focus attention for a pick and place robot [4]. The global search method has good detection effect on functional components, the real-time monitoring cannot be achieved due to the processing performance of the robot in the short term, and the real-time performance

DOI: 10.24507/icicelb.14.07.701

cannot be guaranteed on ordinary computers or robots. Mo et al. [5] proposed a point-wise affordable labeling framework with object kernel convolutional networks to handle various object-object interaction tasks. Mandikal and Grauman [6] proposed a closed-loop dexterous grasping method for object-centric visual inspiration. Qi et al. [7] proposed a self-supervised collection of experiences by predicting spatial affordance maps to clarify which parts of a scene are navigable.

Affordance cognition is widely used in various service scenarios, which improves the service efficiency of the agent. Existing robots can perform the set procedures, but have no ability to understand the tools and supplies in the family. However, if robots can be promoted in the family, it is bound to need to have knowledge and analysis of the daily working environment. In this paper, we develop a real-time affordance interaction recognition system for household service robots to identify the affordance of daily tools in the home environment. In Section 2, we give the description of the problem, detailing the two models. The system is based on the OSAD [8] and AffNet-DR [9] models, and the speed and accuracy of the two models are compared on the affordance of daily tools (see Section 3). The fourth section introduces the overall structure of the system, and conclusions and outlook are given in Section 5.

2. Problem Statement and Preliminaries.

The cognition and use of tools is an important part of a household service robot to perform tasks. The cognitive research on family common tools and the affordance of tools is to endow family service robots with the ability to recognize and use tools through artificial intelligence. If robots can be promoted in the family, it is bound to need the ability to understand and analyze the daily working environment. They need to understand the affordance of tools and their components, and better provide more accurate daily services, such as cutting apples, stirring and pouring water. Robots not only need to use specific tools in specific scenes, but also need to identify the affordance of their corresponding parts. For example, grasping a knife from the handle affords it to cut with the edge of a knife.

The paper studies and implements a system that can identify family daily tools. Based on AffNet-DR algorithm and OSAD algorithm, we realized the system of tools' affordance detection. The AffNet-DR algorithm jointly detects and locates the candidate regions in the image to segment the affordance graph. Through collaborative learning, OSAD can capture the common features between objects with the same potential implications, and learn to adapt well to perceived invisibility. Through comparing AffNet-DR and OSAD algorithm models, the speed and accuracy of the two algorithm models applied to the affordance of daily tools are explored.

2.1. AffNet-DR.

The AffNet-DR model combines Fast R-CNN and Mask R-CNN algorithms. Firstly, the model is trained on the COCO2017 training set and then migrated, so the model can converge quickly. Figure 1 shows the segmentation results of object affordance by AffNet-DR method.

2.2. OSAD.

OSAD firstly detects and estimates human action purpose, and then converts it to help detect the same functionality in all candidate support images. Through joint learning, OSAD can capture the common features between objects with the same functionality, and has good adaptability for sensing unknown functionality.

OSAD [8] has three components in implementation: Purpose Learning Module (PLM), Purpose Transfer Module (PTM), and Collaboration Enhancement Module (CEM). The goal of the PLM module is to estimate the purpose of action from the human-object interaction of image species. The PTM module transfers the action purpose to the query image through the attention mechanism to enhance the relevant features. CEM captures the intrinsic features between objects with common functionality to learn better affordance perception.

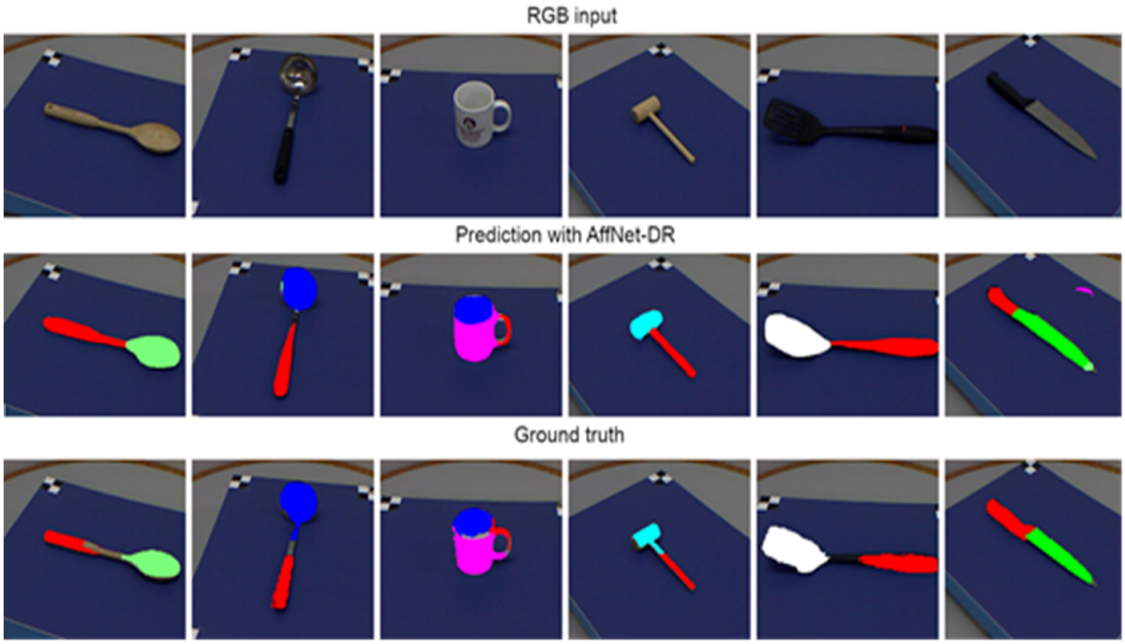

FIGURE 1. The results of affordance segmentation based AffNet-DR

3. Main Results.

3.1. **AffNet-DR.** AffNet-DR is trained on UMD Affordance dataset, which is generated by domain randomization in Unity game engine. The format of the dataset is RGB image, which contains 150 tools. The dataset provides a large and diverse set of daily tools, in which different types of tools may have the same function. Tools use scenarios include kitchens, workshops, gardens, etc., and one tool can be used in different scenarios. In recognition rate, AffNet-DR model has a positive answer to the question of 'can identify'. Figure 2 shows the tool affordance segmentation diagram of rule placement, which can segment the affordance well. When the confidence level is set to 0.9 as shown in Figure 2, the recall rate will be very low so that it cannot be used. In Figure 2, there are eight daily tools, namely transparent glass, mark cup, scissors, butter knife, knife, white plastic spoon, steel spoon, and fork. The knives, white plastic spoons and steel spoons are not identified, and the perspective and light affect the recognition rate. Secondly, in the identified tool, the transparent glass was identified wrongly. The transparent glass has only the grip part and the cup mouth, and there is no cup handle part, while AffNet-DR identifies it as a mark cup for two reasons. One is the shape of the cup is an octagonal

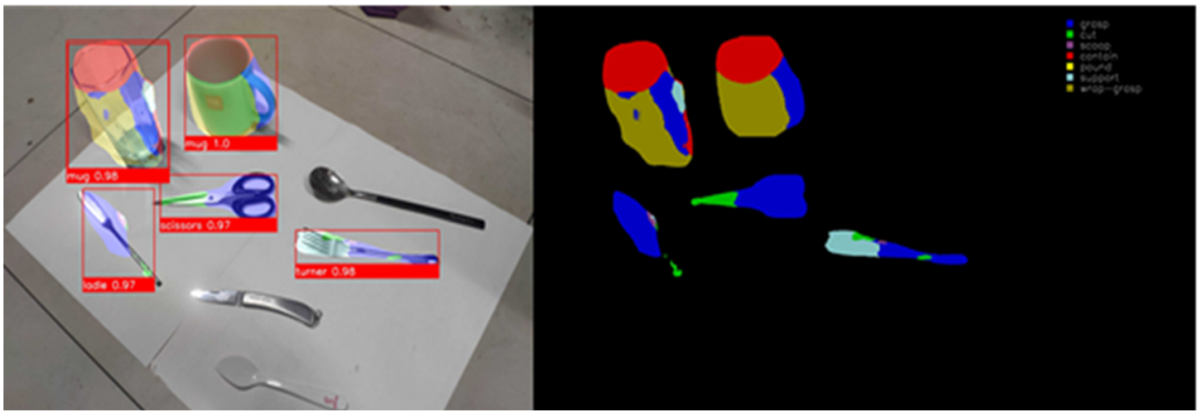

FIGURE 2. Confidence 0.9

cylinder, and all the cups in the data set are cylindrical, which cannot be identified. Second, this cup is transparent. Due to the reflection and refraction of light, there are problems in recognition, and it is identified as a mark cup.

Figure 3 shows the confidence set to 0.8. In this figure, only two are not identified. And in the two tools not identified, transparent glass has not been wrongly identified. Besides, the knife and steel spoon are correctly identified compared to the above Figure 2. However, at the same time, a fork is identified as a shovel and a fork, which is a problem of recognition style. It may be caused by the too similar fork and shovel in the data set, or may be the problem of the angle and light of the test image. Although the confidence level was adjusted to 0.1, the white plastic spoon in the lower right corner was still not identified. It may be because the similarity between the white background and the white spoon is too high, the segmentation of the AffNet-DR model cannot distinguish boundaries more accurately. On the right side of the image annotation, in the affordance annotation of the tool, the handle of the tool is not well connected to the other parts. For example, butter knife almost recognizes a whole as a function of cutting, but only a third of the previous affordance of butter knife is the function of cutting. The boundary between the holding part of the scissors and the staggered double-edged, and the boundary between the holding part of the knife and the cutting part need further optimization, but on the whole, it can be accurately identified.

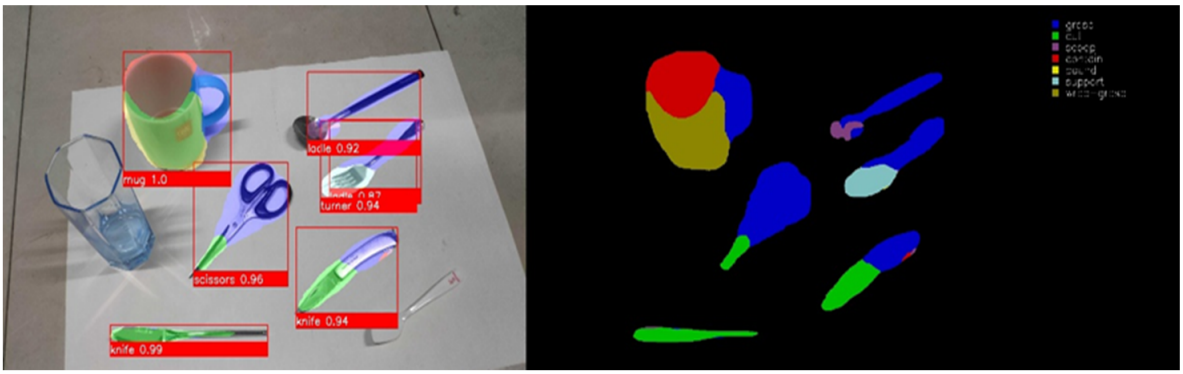

FIGURE 3. Confidence 0.8

Figure 4 shows the confidence level set to 0.7. When the confidence level is 0.7, the AffNet-DR model can identify the affordance of all tools, but the transparent glass is not better than the first two cases, and it is regarded as a mark glass. Therefore, from the comparison of the three figures, confidence is a parameter that changes greatly without changing the model. Here, from the actual test, confidence is set between 0.7 and 0.9; too low or too high confidence is not good for recognition.

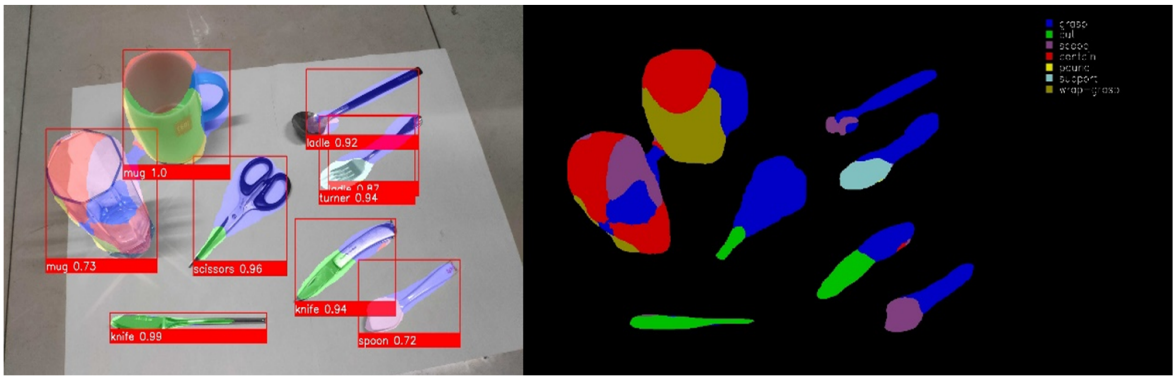

FIGURE 4. Confidence 0.7

3.2. **OSAD.** OSAD is more suitable for the classification and search of a large number of photos. The data set used for OSAD model training is Purpose-driven Affordance Dataset v2 (the second edition of Purpose-driven Functional Dataset, hereinafter referred to as PADv2), which contains a total of 30,000 images of 39 kinds of functions and 103 kinds of object categories. OSAD identifies the affordance of the items in the picture, and classifies them according to bounce, packaging, full, support, kick, lie, play, push-pull, ride, cover, wear, etc.

As shown in Figure 5, OSAD considers the affordance of these balls to be bounce, i.e., objects that can be bounced, such as basketball and volleyball, to be elastic and well classified. At the same time, this recognition is a kind of comparative recognition. If there is another sphere such as table tennis or tennis, it can be identified as an object with bounce characteristics. OSAD is a large-scale model, which has the ability to identify a large number of objects. Even in the model comparison, OSAD is not very suitable for the topic of the paper, that is, the affordance identification of commonly used tools in the family. However, from the perspective of a system, if only using the AffNet-DR model, it is very single. The tools in the family are more than the specific types identified in the AffNet-DR model. If a system can only identify the functional properties of specific items in a simple scene, the system is not perfect. It is necessary to add OSAD model as a supplement.

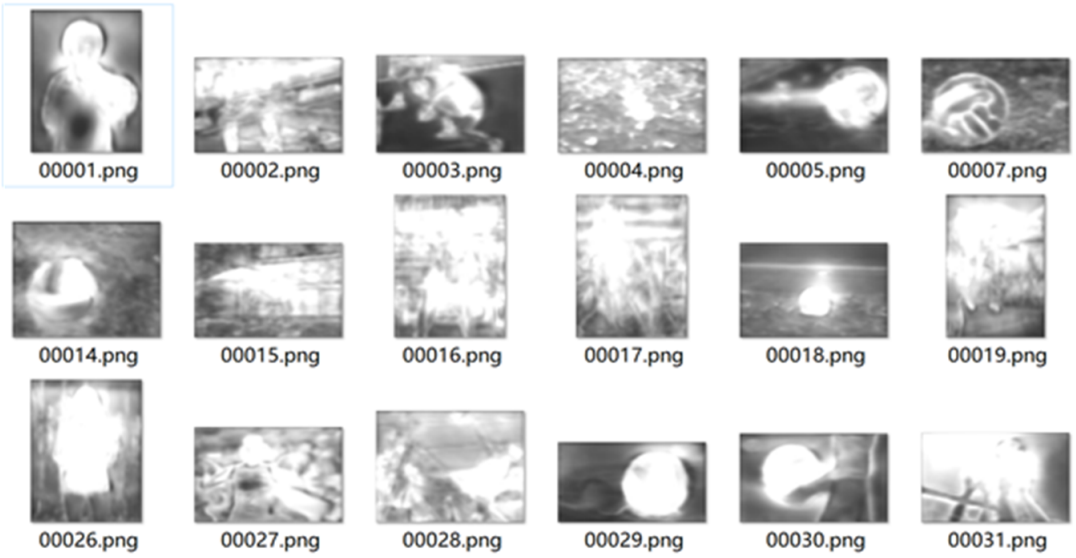

FIGURE 5. The identification of OSAD model

4. **Control Design.** The primary basic function of the system is the recording function. There are two kinds of recording functions of the image, photography and video. The main function design of the system is divided into two parts, namely OSAD part and AffNet-DR part. For robots, the camera needs to recognize the input image in real time, then perform affordance marking, and finally output the result to the screen. This is the most direct and important function that the robot can identify the object, which is the main function of both parts. Secondly, from the user's point of view, users hope that this system can have some manual functions to identify the existing images for testing, so it is also very important for image recognition. Similarly, users also need to recognize a recorded video, so it is also necessary to recognize a video and mark each frame of the video with affordance. The specific design is shown in Figure 6.

Figure 7 shows one of our system identification results. The left side of the main interface is the window for identifying the image, where the identification image and the identification result are placed. On the right side of the main interface, changes to system

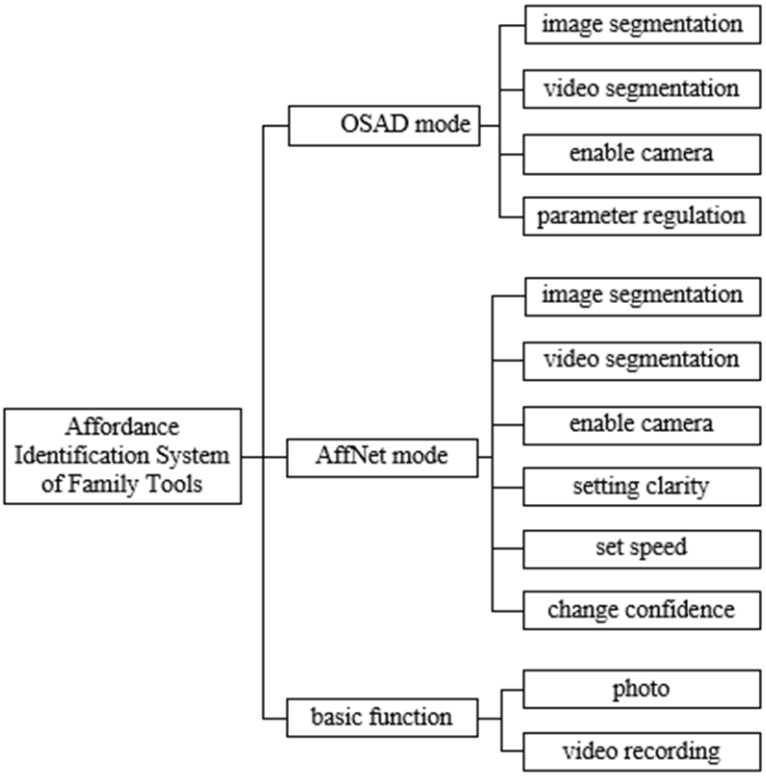

FIGURE 6. The system design

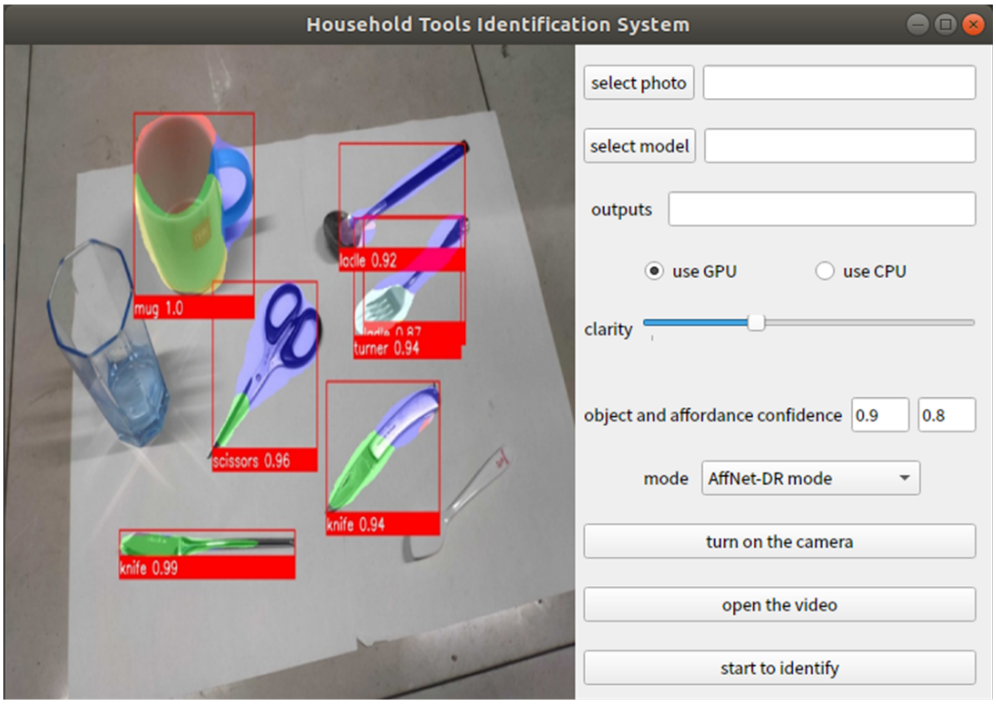

FIGURE 7. The identification of our system

parameters are placed. If the resolution of the image to be identified is too large, it may exceed the memory. The display memory of the display card is valuable. The display memory capacity range of the consumption-level display card is 4-8 GB. If the image pixels identified are more than 10 million, graphics memory must not be enough. However, the memory at the consumption level is relatively large, and the dual-channel 32 GB memory is also common and economical, so some high-resolution images can be identified by CPU

at the expense of recognition speed. Next are two similar features, adjusting speed and clarity. The adjustment of speed is the adjustment of recognition speed, and the adjustment of clarity is the adjustment of the clarity of the output window. The definition has an upper limit at a certain speed, and the recognition speed is always limiting the definition. The improvement of clarity will inevitably lead to the improvement of resolution, and the improvement of resolution will require the increase of time and memory index, so these two parameters cannot be obtained simultaneously.

Confidence greatly affects the recognition effect. If the identified item does not reach the confidence threshold, it will not be marked. From the previous design, it can be seen that there is a suitable confidence interval, and it is necessary to adjust the confidence interval reasonably. It is appropriate to test the confidence not less than 0.7 and not more than 0.9.

The system provides two recognition modes, namely, the AffNet-DR model and the OSAD model, for family function recognition. AffNet-DR has a recognition rate of more than 0.9 for simple scenes of single object, so it is more suitable for simple scenes. If the scene is complex, the average recognition rate is about 0.65, and can be clearly marked on the image. This model only recommends tools for identifying daily use, and it rarely correctly identifies unknown situations.

Above all, OSAD is more powerful, and it has the ability to recognize almost all affordance, which is determined by its algorithmic principles. OSAD can capture the common features between objects with the same affordance, and has good adaptability for sensing unknown functionality. OSAD can only classify it as a class of functionality. This recognition is the overall recognition of an object. It can identify the affordance similarity between the ping-pong racket and the golf club, but it cannot identify that the handles of the ping-pong racket and the golf club are the same. Therefore, this algorithm emphasizes the overall affordance recognition of an object. Therefore, this algorithm is especially suitable for classifying a large number of images, fast and has the ability to identify unknown, perhaps by adding a large number of images, to achieve unexpected results.

5. Conclusions.

In this paper, our study from the cognitive aspects of robots develops a system that can identify the affordance of household daily tools. The system can real-time identify the daily affordance of the family system. It has a simple user interface and can mark the affordance of the tool on the screen in real time. From the perspective of robot empowerment, this recognition system enables the robot to identify and understand the affordance of common household tools. Most of the previous recognition systems can only identify the affordance of tools within the model training. Meantime, OSAD model is used as a supplement to AffNet-DR, so that the robot has the ability to identify the affordance of new tools and enable them to have a preliminary understanding before tools that have not been seen. Finally, in a next step we would like to combine affordance recognition with real robots, so that real robots can use tools through affordance recognition.

Acknowledgment. This work is partially supported by the National Key R&D Program of China (2018YFB1308300), National Natural Science Foundation of China (U20A20167), Natural Science Foundation of Hebei Province (F202103079), Innovation Capability Improvement Plan Project of Hebei Province (22567626H), Beijing Natural Science Foundation (4202026), Natural Science Foundation of Xinjiang Uygur Autonomous Region Grant No. 2022D01A59, and Xinjiang Uygur Autonomous Region University Scientific Research Project (Key Natural Science Project) Grant No. XJEDU2021I029.

REFERENCES

[1] J. J. Gibson, *The Ecological Approach to Visual Perception*, Houghton Mifflin, 1979.
[2] T. Nagarajan, Y. Li, C. Feichtenhofer and K. Grauman, EGO-TOPO: Environment affordances from egocentric video, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.160-169, 2020.
[3] Z. Hou et al., Affordance transfer learning for human-object interaction detection, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.495-504, 2021.
[4] H. Wu, Z. Zhang, H. Cheng, K. Yang, J. Liu and Z. Guo, Learning affordance space in physical world for vision-based robotic object manipulation, *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp.4652-4658, 2020.
[5] K. Mo, Y. Qin, F. Xiang et al., O2O-Afford: Annotation-free large-scale object-object affordance learning, *Proc. of Conference on Robot Learning*, pp.1666-1677, 2022.
[6] P. Mandikal and K. Grauman, Learning dexterous grasping with object-centric visual affordances, *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp.6169-6176, 2021.
[7] W. Qi, R. T. Mullapudi, S. Gupta and D. Ramanan, Learning to move with affordance maps, *International Conference on Learning Representations (ICLR)*, 2020.
[8] W. Zhai, H. Luo and J. Zhang, One-shot object affordance detection in the wild, *Int. J. Comput. Vis.*, vol.130. pp.2472-2500, 2022.
[9] F. Chu, R. Xu and P. A. Vela, Learning affordance segmentation for real-world robotic manipulation via synthetic images, *IEEE Robotics and Automation Letters*, pp.1140-1147, 2019.