ELSEVIER

Pattern Recognition Letters

A benchmark image dataset for industrial tools

Cai Luo[a,*], Leijian Yu[b], Erfu Yang[b], Huiyu Zhou[c], Peng Ren[d,*]

[a] *Department of Mechanical and Electronic Engineering, China University of Petroleum (East China), Qingdao, China*
[b] *Department of Design, Manufacture & Engineering Management, University of Strathclyde, Glasgow, UK*
[c] *Department of Informatics, University of Leicester, Leicester, UK*
[d] *Department of Information and Control Engineering, China University of Petroleum (East China), Qingdao, China*

ABSTRACT

Robots and Artificial Intelligence (AI) play an increasingly important role in manufacture. One of the tasks is to identify tools in the scene so that the tools can be applied to different assembly purposes. In the AI community, many datasets have been generated and deployed to train robots to recognize individual items, however, these datasets are scene-specific and lack generic background. In this paper, we report our dataset contains photos of 8 objects types that would be easily recognized by qualified workers. This is achieved by gathering images of common tools in a typical factory. The ground truth categories of our dataset are manually labeled by experienced workers, which would be worthy evaluation tools for the intelligence industrial systems. The equipment used and the image collection process are discussed, along with the data format. The mean average precisions range from 64.37% to 78.20%, which bring the possibility for future improvement. The dataset is ideal to evaluate and benchmark view-point variant, vision-based control algorithm for industry robots. It is now public available from https://github.com/tools-dataset/Industrial-Tools-Detection-Dataset.

1. Introduction

Every day, industry workers use variety tools for their daily duties, like cutting steel plate, tightening screw, hammering a nail, or measuring length, as shown in Fig. 1. By virtue of training and memory, workers can effortlessly identify a tool and know its function. They are also able to choose suitable tools for different needs. While in the machine world, robots are still struggling to acquire the ability to pick correct instruments for assigned tasks through their visual sensors [20,28,41]. As robots like SCHAFT, Atlas, Valkyrie and REEM-C begin to manipulate standard tools and equipments commonly available in industrial environment, ranging from small screw drivers to full-size vehicles [6,32,34,38]. The proliferation of AI embodied in robots increases the needs for these humanoid machines can work with their own hands, so they can take the tasks from repairing satellites to working in a remote factory without human intervention [21,24]. It appears clear that for dealing with such complex scenarios, robust and efficient object detection algorithms are very important. Deep learning related methods has make great success in other field [2,3,15,35,37,39,40]. However, for deep learning methods, training datasets play the vital roles [4,16]. So, in order to identify different tools successfully, specifically designed datasets are needed.

To advance object recognition research in industry, we introduce a dataset for Industrial Tools Detection (ITD). It appears clear that it would bring great possibilities for robots to use a wide variety of instruments if they could distinguish these tools in factories or construction sites [23,24,33]. The dataset detailed in this paper is introduce to identify tools at the level of usages, and provide precise predictions for a robot to interact within the industry scenarios. Furthermore, the dataset is a challenging benchmark to evaluate view-point variant, vision-based control algorithms for industry robots.

The main contribution of this paper are as follows:

- We present a new large-scale object dataset, which consists of 8 object categories, 24 common industrial tools overall and multi distinct views of each tool. The dataset provides hand-labeled ground truth for more than 11,000 RGB images.
- We evaluate state-of-the-art object detection algorithms on ITD and define benchmark as baseline references for developing future new algorithms.
- Dataset and code from this work are available on-line at: https://github.com/tools-dataset/Industrial-Tools-Detection-Dataset.

[*] Corresponding authors.
E-mail addresses: luo_cai@upc.edu.cn (C. Luo), pengren@upc.edu.cn (P. Ren).

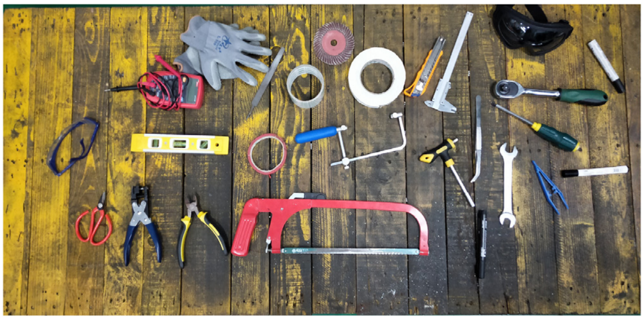


Fig. 1. Industrial tools. Variety industrial tools for workers' daily duties have been chosen for our dataset based on the purpose of evaluating view-point variant, vision-based control algorithms for industry robots.

Table 1
Comparison among MS-COCO, PASCAL VOC, ImageNET, TAS, HRSC2016, DOTA, Cornell grasping and ITD.

Datasets	Instances	Objects of interest
MS-COCO	123,287	Nature objects
PASCAL VOC	21,503	Nature objects
ImageNET	349,319	Nature objects
TAS	1319	Aerial targets
HRSC2016	2976	Aerial targets
DOTA	188,282	Aerial targets
Cornell grasping	1035	Daily tools
ITD	11,000	**Industrial tools**

2. Related works

The purpose of choosing a suitable tool is to fulfill a task goal as quickly as possible [14]. The problem of picking and using tools has been widely studied in robotics, computer vision, artificial perception and psychology for many years and will be hot topics for the next decades as well. Many efforts have been dedicated to detect geometric characteristic of tools and how to handle the items correctly and firmly. They often assume prior information of objects shape and general location. Some of them also need the assistance of affordance labels or predefined markers to accomplish the manipulation tasks.

The creation of ground truth image and video datasets helped stimulated a flood of interest in the related areas. Large datasets like MS-COCO [25] is the de facto standard evaluation instrument for object detection. For the object categories classification, the PASCAL VOC [11,18] and ImageNET [10,31] are always in the datasets list of researchers. These datasets have proven to be very good performance test fields for computer vision algorithms in natural scenes.

In the field of tools detection, recognition and manipulation, datasets have been play a critical role as an algorithm assess. However, such successes have been slow to industrial field imagery due to the scarcity of optimal annotated datasets for tools in industrial environments. Unlike common daily objects, the collection and classification of industrial tools are much more difficult. Workers in the factories will need some special trainings in order to know the correct usage of tools [7]. They will need another several years to get the experiences to figure out how to choose the most suitable ones according to the tasks. Furthermore, the detection of industrial items are highly dependent on contextual information, which means the items in the datasets should be in their natural environments. Datasets like TAS [19], HRSC2016 [27] and DOTA [36] only contains large items like vehicles, planes and ships that are difficult to manipulate by robots. Some pioneering works have grounded the tool handling in a constrained testing samples [7,13]. Deep learning method has been applied by Ian et al. to solve the grasp problem by using a dataset which containing several daily tools. Kuan et al. proposed an affordance learning approach for tool manipulation through pre-selected objects. When it comes to general industrial tools, such as hammers, wrenches or saws, researches are normally depend on their own testing sets. All these datasets are short in the number of tools varieties, which prevent them from being widely usable.

Our target is to simulate all possible situation of intelligent industrial systems. When collecting data, we gather the most common posture of the tools and place them in the location where they may found normally. Next, we analyze the properties of ITD in comparison to several other popular datasets. These include MS-COCO, PASCAL VOC 2012 and Cornell grasping dataset. Each of these datasets varies significantly in numbers of tools categories and quantities of images. MS-COCO was created to detect and segment of items occurring in their natural context. PASCAL VOC focuses on object detection in natural images. They both have at least 20 different categories, such as person, animals, aeroplane, chair and monitor. But none of them include the tools, especially industrial tools. Cornell grasping dataset has the largest number of categories in previous common tools datasets. The comparison results can be seen in Table 1. Note that ITD surpass Cornell grasping

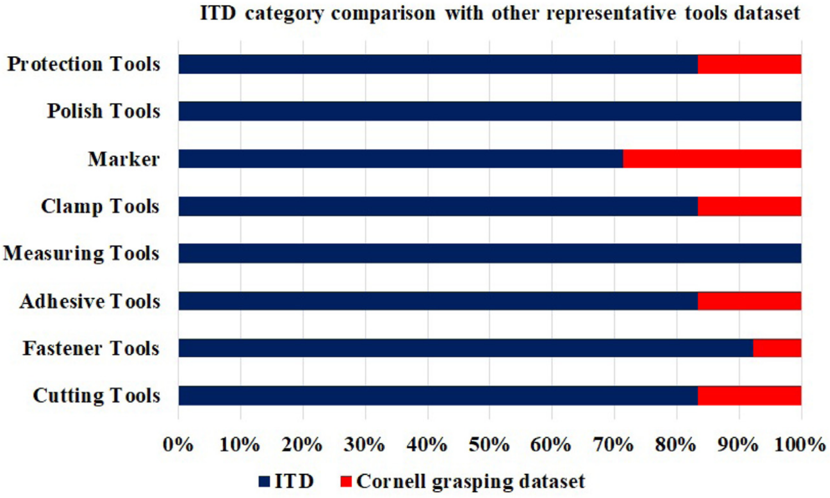
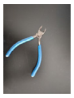


Fig. 2. Category Comparison. We perform an evaluation comparison between ITD and Cornell grasping dataset and responding quantity of items.

Table 2
The categories and usages of the tools in ITD Dataset.

Category	Sample image	Name	Affordance
Cutting Tools		Scissor Utility Knife Puncher Nipper plier	Cut or separate small amounts of a material from the work piece by means of shear deformation This can be accomplished by single-point tools or multi-point tools
Fastener Tools		Open-end Wrench Torque wrench Hex wrench Screw driver	Provide grip and mechanical advantage in applying torque to turn objects or affixes multi objects together and the joints can be dismantled without damaging the joining components
Adhesive Tools		Pressure-sensitive tape Water activated tape Heat sensitive tape	Bind items together and resists their separation through non metallic substance applied to one surface
Measuring Tools		Multi-meter Vernier scale Air level	Measure a physical quantity This may require one-hand or two-hand operation
Clamp Tools		Plastic tweezers Flap tip clamp	Hold or pick-up items tightly together to be easily handled with the fingers
Marker		Permanent marker pen Waterproof marker	Draw or highlight notices on items They can be water-proof, dry-erase, or permanent.
Polish Tools		Machine file Sand paper	Smooth a workpiece's surface by rubbing it or using a chemical action
Protection Tools		Safety goggle Weld eye protector Glove	Enclose or protect body from injury or harmful contacts include physical, electrical, heat or chemicals

dataset not only in tools category numbers, but also in total number of tools, as shown in Fig. 2.

In our datasets, we strive to collect images rich in classification, illumination and localization. ITD collected 24 daily industry tools. 8 categories are chosen, including cutting tools, fastener tools, adhesive tools, measuring tools, clamp tools, marker, polish tools and protection tools, as shown in Table 2. Fig. 1 shows the examples of these tools. Compared with previous datasets, ITD can aid intelligent industrial systems specifically.

3. Industrial tool detection dataset

This section presents how the ITD Dataset are selected. And what are the hardware and software used for the data collection are also described.

3.1. Object categories

As robots begin to manipulate standard tools and equipments available in industry scenarios, they will need to identify the tools and know the usages of them. This is achieved by gathering images of common tools in a typical factory through a computer vision and artificial intelligence study from September 2017 to May 2018. The dataset contains photos of 8 objects types that would be recognized by a qualified worker. The dataset has been collected in five distinct scenarios in factory, workshop, assembly line, and construction site scenarios characterized as shown in Fig. 3. When people or industrial robots work in a factory, they are often in a moving state, which can results in view angle change, motion blur, illumination and clutter background. We specially designed dynamic scenes in factory environments to collect data.

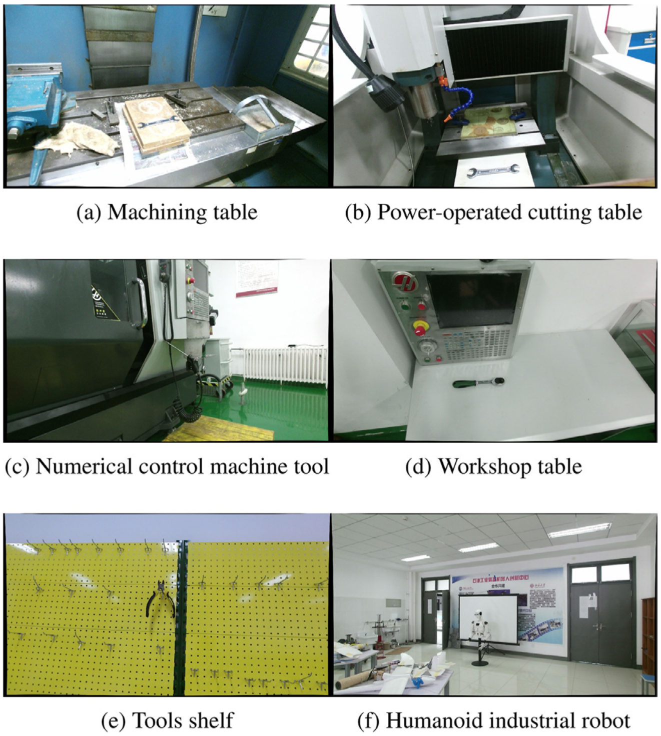

(a) Machining table
(b) Power-operated cutting table
(c) Numerical control machine tool
(d) Workshop table
(e) Tools shelf
(f) Humanoid industrial robot

Fig. 3. Tools in different industrial scenes. ITD contains a wide variety of object categories in different industrial environments. We strive to collect images rich in classification, illumination and localization.

3.2. Dataset format

Data was collected using a kinect 2.0 sensor [22] delivering 30 RGB-D frames per second at a resolution 1024 × 575 pixels + 512 × 424 depth frames. Since the items are relatively small, we collected data at the distance between 5 m and 1 m. Items are placed in their usual posture and environment and the camera point-of-view is that of the worker eyes. The worker was required to walk smoothly around the item while the camera was kept facing the target item consistently.

In order to compute the intrinsic arguments of the camera, we used a calibration checkerboard with known size. The dimension of

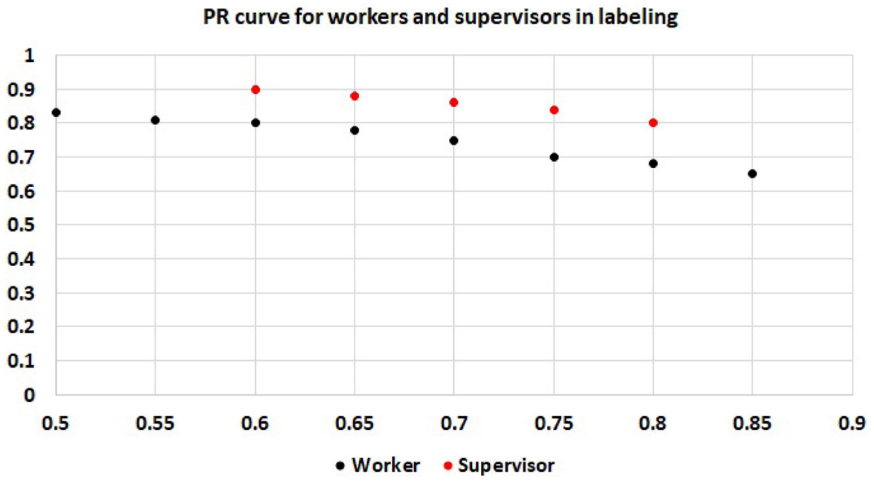


Fig. 4. Precision and recall rate of labeling. 8 workers ranging in experience years from 1 to 10 were hired to label tools in ITD dataset. We assessed the category labeling tasks by comparing to dedicated supervisors. We analyzed precision and recall of five senior workers (managers and supervisors from factories) with the results obtained from the front-line workers.

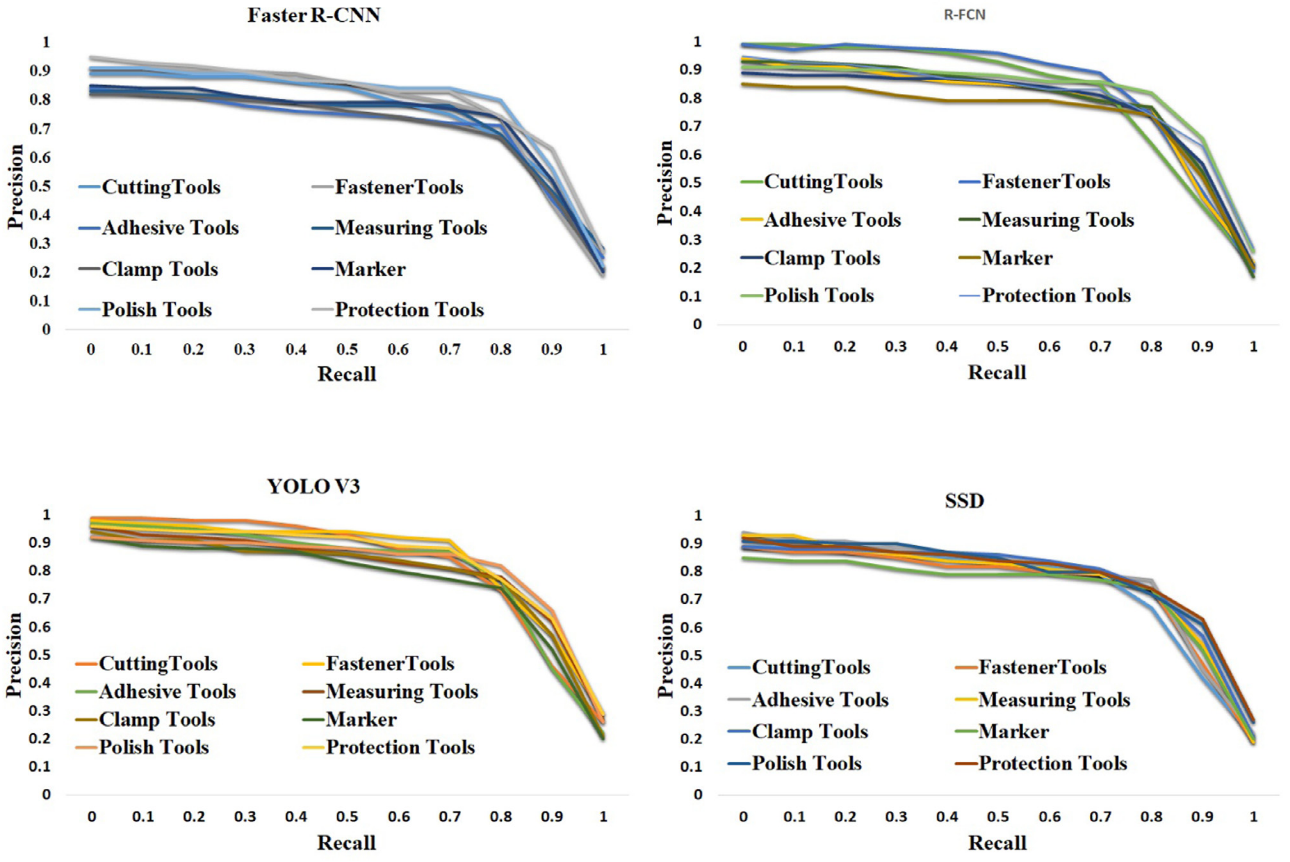


Fig. 5. Precision and recall curves of 4 detection methods. The experiments have been conducted on a PC with a 2.40 GHz Intel(R) Xeon(R) CPU E5-2620 CPU, a GTX TITAN X GPU and 128GB memory. As we can see from the results exhibited, performances in clamp tools, marker and measuring tools are suboptimal.

the checkerboard is 9 squares × 7 squares, whereas the length of each square is 3 cm. The calibration parameters and OpenCV tools used for calibration are also included in the dataset.

3.3. Ground truth

8 workers with Mechanical Engineer Certificate ranging in experience years from 1 to 10 were hired to label every tools they saw in inside and outside factories. For a given tool, a worker was asked to identify the tool's name, the category it belonged to and the possible usage. This task took a total of ∼200 worker hours to complete. We assessed the category labeling tasks by comparing to dedicated supervisors. We analyzed precision and recall of five senior workers (managers and supervisors from factories) with the results obtained from the front-line workers. The true positives(TP), false positives(FP) and false negatives(FN) are defined as following [5]:

1. TP means the positive labeling that are categories as the positive class,

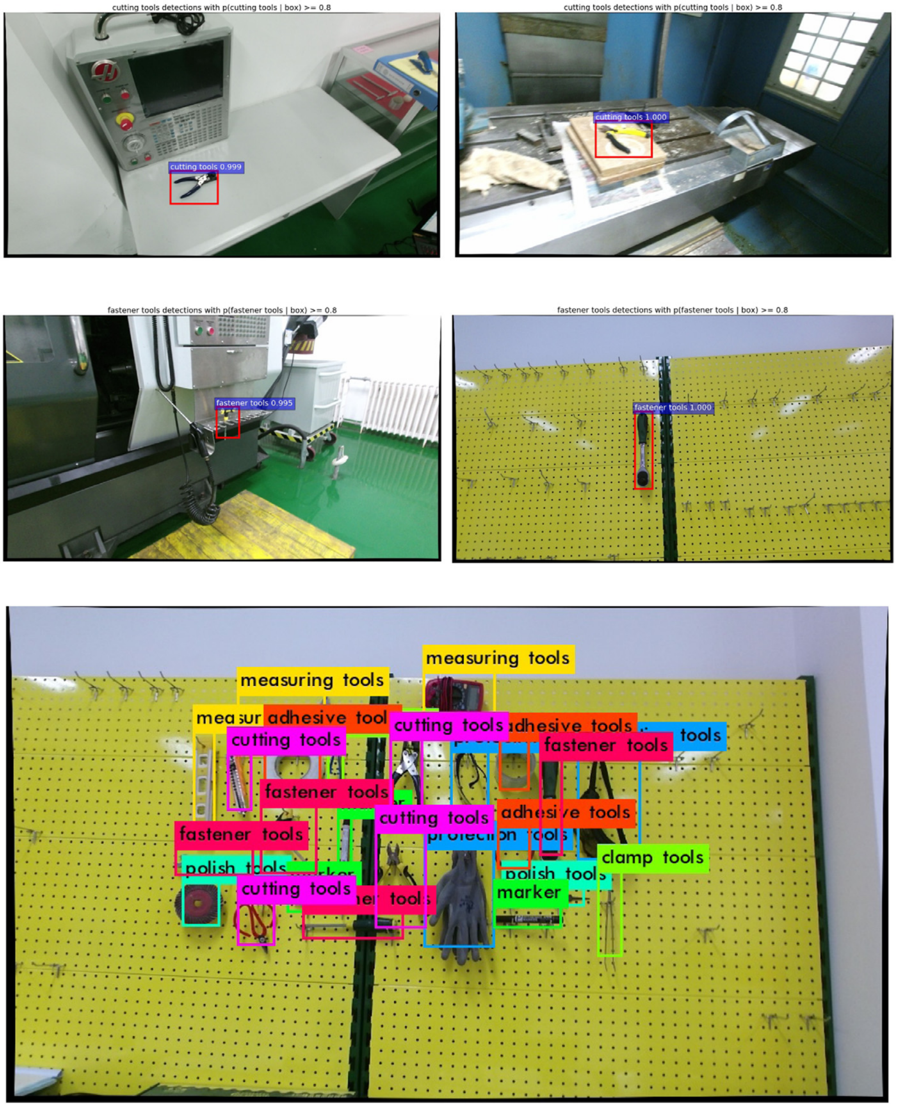


Fig. 6. Detection results in single and multi tools scenes. The conclusion presents that tools features can be easily affected by clutter background and dynamic environmental illumination. The image blur caused by the worker moving also make the performance fall short. This implies the defects of current detection methods and extensive efforts have to be dedicated according to the industrial requirements.

2. FP stands for the negative labeling that are categories as the positive class,
3. FN denotes the positive labeling that are categories as the negative class.

The precision and recall rate are computed by:

$$Precision = \frac{TP}{TP+FP}$$
$$Recall = \frac{TP}{TP+FN}, \tag{1}$$

The results can be seen in Fig. 4. It shows that the front-line workers have high recall rate than the senior workers. The labeling results are provided as ground truth in order to evaluate different vision-based target detection algorithms.

4. Experiments and discussion

4.1. Object recognition evaluation

We evaluate state of the art object detection algorithms on ITD dataset. We carefully choose the Fast Region-based Convolutional Network(Faster R-CNN) [17,30], Region Fully Convolutional Networks (R-FCN) [8], You Only Look Once (YOLO) V3 [29] and Single Shot MultiBox Detector(SSD) [26] as our benchmark methods for they have been widely used in object detection. We first briefly describe all these representations we have used for assessment.

4.1.1. Faster R-CNN

Faster R-CNN is a hybrid of deep convolutional network and region detector. The deep convolutional network combines a Region Proposal Network (RPN) and an object detection network [30]. The quality of detector is improved by using sparse object proposals. The whole image will be processed through conventional and max polling layers in order to produce a conventional feature map. A fixed length feature vector will be extracted by the region of interest pooling layer from the feature map. The features can be used for faster inference by classification and bounding-box regression.

4.1.2. R-FCN

The detection strategy of R-FCN consists of region proposal and region classification [8]. The candidate regions are extracted by the Region Proposal Network. R-FCN ends with a position-sensitive

Table 3
Numerical results of baseline models evaluated with ground truth on Faster R-CNN, R-FCN, YOLO V3 and SSD methods over the ITD dataset.

	Faster	R-FCN	YOLO	SSD
Cutting tools	70.12	72.65	85.56	69.51
Fastener tools	58.61	63.64	71.81	53.43
Adhesive tools	81.75	81.80	88.43	80.93
Measuring tools	60.53	63.64	83.41	61.66
Clamp tools	61.31	63.64	66.26	60.65
Marker	62.45	63.58	67.83	60.92
Polish tools	50.76	54.52	73.18	48.32
Protection tools	69.41	72.71	89.11	68.18
mAP	64.37	67.02	78.20	62.95

region of interest pooling layer. By cropping features from this last layer prior to prediction, R-FCN model could achieve similar accuracy to Faster R-CNN with less running time.

4.1.3. *YOLO V3*

YOLO applies an end-to-end single convolutional neural network that divides the image into regions, bounding boxes and region probabilities [29]. By examining the entire image during the training procedure, it get the contextual information and the knowledge of surroundings.

4.1.4. *SSD*

Single Shot MultiBox Detector (SSD) approach uses a single feed forward convolutional network that procedures bounding boxes collection and anchor offsets without requiring a pre-proposal classification [26].

4.2. *Protocol*

4.2.1. *Protocal for holdout validation*

We spitted the dataset by categories into training (50%), validation (25%) and testing (25%) sets randomly. We adopted the PASCAL Visual Object Challenge mean average precision (mAP) evaluation metrics [12]. The mAP is calculated by (2):

$$mAP = \frac{\sum_{n=1}^{C} AvgPrecision(n)}{C}$$

$$AvgPrecision = \sum_{l=1\ldots N} P(l)\triangle Recall(l), \qquad (2)$$

where C denotes the number of categories, $P(l)$ and $\triangle Recall(l)$ denote the precision value at every threshold and change in the recall respectively.

A detection is marked correct when the intersection size of the bounding boxes of the trial and the ground truth is more then half the size of their union. The numerical results (AP) of baseline models evaluated with ground truths are shown in Table 3. For its performance in skewed datasets [9], the precision and recall (PR) curve is also used as a valuable analytical tool for assessment.

4.2.2. *Protocal for 4-fold cross-validation test*

To further validate the ITD dataset, the 4-fold cross-calidation test were carried out, which ensures that every image is tested once to prevent any bias error [1]. The dataset is divided by categories into 4 subsets (25% each) randomly. Every subset will works as the test dataset once, while the other three subsets are used as training and validation dataset. To be specific, when the subdataset is secleted to train the model, 30% of images in subset will be used as validation dataset to fine-tune the model hyperparameters. And every model will be trained and tested four times to validate the proposed ITD dataset.

4.3. *Results*

The experiments have been conducted on a PC with a 2.40 GHz Intel(R) Xeon(R) CPU E5-2620 CPU, a GTX TITAN X GPU and 128GB memory. Fig. 5 shows the PR curves for Faster R-CNN, R-FCN, YOLO V3 and SSD methods over the ITD dataset and Fig. 6 shows the single and multi tools detection results in different industrial scenes.

4.3.1. *Comparison between different tools*

As we can see from the results exhibited in Table 3, performances in clamp tools, marker and measuring tools are suboptimal, which attribute to their relatively small and may easily blocked by tools holder and grippers. Items like cutting tools, adhesive tools and protection tools, present good results partly due to their large size and difficult to be covered. YOLOv3 leads to the best accuracy, followed by R-FCN. The mAP results of SSD is lower than the others. The random crop approach used by the SSD data augmentation method may cause the consequence.

Table 4
The performance of Faster R-CNN, R-FCN, YOLO V3 and SSD over 4-fold cross validation on the ITD dataset.

Test fold	Method	AvgPrecision(%)							
		Cutting	Fasterner	Adhesive	Measuring	Clamp	Marker	Polish	Protection
1st fold	Faster	70.32	59.31	81.66	59.78	61.01	62.01	51.34	68.88
	R-FCN	72.51	62.69	81.12	68.08	67.97	62.76	54.41	72.68
	YOLO	86.31	70.31	87.56	82.76	66.11	66.68	72.89	89.47
	SSD	70.08	54.21	79.78	60.66	59.88	60.07	48.76	69.12
2nd fold	Faster	70.21	59.98	81.87	59.31	61.32	62.41	51.51	68.92
	R-FCN	72.33	62.97	81.65	68.45	68.02	62.56	54.62	72.72
	YOLO	86.75	70.87	87.43	82.81	66.15	66.72	72.81	89.71
	SSD	69.79	54.66	80.02	60.76	59.93	60.26	48.78	69.04
3rd fold	Faster	69.93	59.21	81.32	59.82	60.89	61.93	51.46	68.56
	R-FCN	73.08	62.12	80.97	67.91	67.78	62.83	54.21	72.55
	YOLO	86.82	70.42	87.88	82.21	66.09	66.81	72.63	89.32
	SSD	70.43	54.68	79.87	60.79	59.12	59.89	48.61	69.53
4th fold	Faster	71.13	60.08	82.02	59.44	61.12	62.30	51.21	68.31
	R-FCN	72.23	62.45	81.43	68.22	67.45	62.78	54.61	72.18
	YOLO	86.12	69.89	87.33	82.88	65.93	66.53	72.46	90.03
	SSD	69.92	54.13	79.91	60.43	60.09	59.87	48.80	68.95
Average	Faster	70.40	59.65	81.72	59.59	61.09	62.16	51.38	68.67
	R-FCN	72.54	62.56	81.29	68.17	67.81	62.73	54.46	72.52
	YOLO	86.50	70.37	87.55	82.67	66.07	66.69	72.70	89.63
	SSD	70.06	54.42	79.90	60.66	59.76	60.02	48.74	69.16

4.3.2. Comparison between different methods

The curves demonstrated in Fig. 5 indicate that YOLO V3 is superior to other approaches. It is probably due to the improvement of predication strategy. YOLO extracts features at 3 different scales [29]. The change allows the method to get more meaningful information from small size objects. However, speed results show the different trend, the R-FCN algorithm is 52,989*s*, while YOLO v3 algorithm is 771,072*s*. These approaches will degrade in industrial tools detection for relatively small training instances. It figures that for tools detection in industrial environments, those methods should ameliorate accordingly.

4.3.3. Comparison between different scenes

By analyzing the detection results of each scene (examples shown in Fig. 6), the conclusion presents that tools features can be easily affected by clutter background and dynamic environmental illumination. The image blur caused by the worker moving also make the performance fall short. This implies the defects of current detection methods and extensive efforts have to be dedicated according to the industrial requirements.

4.3.4. Comparison through 4-fold cross-validation test

By adopting the 4-fold cross-validation method, the performance of each model over the ITD dataset is demonstrated in Table 4. In general, YOLOv3 still outperforms the other three detection methods. The different results between each categories are mainly caused by tools with different features. And the same categories get similar results among different test folds. It can conclude that there is also no huge bias error in ITD dataset.

5. Conclusion

We build a large-scale dataset for tools detection in industrial environments which is much more specialized and suitable than any other general datasets in this field. We also establish a benchmark for items detection in industrial scenes. We believe ITD will promote the development of tools detection algorithms in industry. We currently only label tools in general but labeling grasping places may also provide significant manipulation information that may be useful for industrial utilization. In the future, we intend to further extend the dataset in terms of categories and sample quantities.

Conflict of interest

There are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgment

The work of C. Luo was supported in part by the National Natural Science Foundation of China under Grant 61701541, in part by the Shandong Provincial Natural Science Foundation, China under Grant ZR2017QF003, and in part by the Fundamental Research Funds for the Central Universities under Grant 19CX02021A. The work of H. Zhou was supported in part by UK EPSRC under Grants EP/N508664/1, EP/R007187/1, EP/N011074/1, and Royal Society-Newton Advanced Fellowship under Grant NA160342. The work of P. Ren was supported in part by the Shandong Provincial Natural Science Foundation, China under Grant ZR2019MF019. The authors would like to thank the Oil Industry Training Center in China University of Petroleum for the indutrial tools anotation support.

References

[1] M.A. Al-antari, M.A. Al-masni, M.-T. Choi, S.-M. Han, T.-S. Kim, A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification, Int. J. Med. Inform. 117 (2018) 44–54.

[2] X. Bai, C. Yan, H. Yang, L. Bai, J. Zhou, E.R. Hancock, Adaptive hash retrieval with kernel based similarity, Pattern Recognit. 75 (2018) 136–148.

[3] X. Bai, H. Zhang, J. Zhou, Vhr object detection based on structural feature extraction and query expansion, IEEE Trans. Geosci. Remote Sens. 52 (10) (2014) 6508–6520.

[4] X. Bai, J. Zhou, A. Robles-Kelly, Pattern recognition for high performance imaging, 2018,

[5] A. Borji, D.N. Sihite, L. Itti, Salient object detection: a benchmark, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 414–429.

[6] C. Breazeal, B. Scassellati, Robots that imitate humans, Trends Cogn. Sci. 6 (11) (2002) 481–487.

[7] I.M. Bullock, T. Feix, A.M. Dollar, The yale human grasping dataset: grasp, object, and task data in household and machine shop environments, Int. J. Robot. Rese. 34 (3) (2015) 251–255.

[8] J. Dai, Y. Li, K. He, J. Sun, R-fcn: Object detection via region-based fully convolutional networks, in: Advances in Neural Information Processing Systems, 2016, pp. 379–387.

[9] J. Davis, M. Goadrich, The relationship between precision-recall and roc curves, in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 233–240.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, CVPR09, 2009.

[11] M. Everingham, S.M.A. Eslami, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: a retrospective, Int. J. Comput. Vis. 111 (1) (2015) 98–136.

[12] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[13] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, S. Savarese, Learning task-oriented grasping for tool manipulation from simulated self-supervision, (2018) arXiv:1806.09266.

[14] C. Fermüller, F. Wang, Y. Yang, K. Zampogiannis, Y. Zhang, F. Barranco, M. Pfeiffer, Prediction of manipulation actions, Int. J. Comput. Vis. 126 (2–4) (2018) 358–374.

[15] F. Gao, M. Fei, W. Jun, S. Jinping, Y. Erfu, Z. Huiyu, Visual saliency modeling for river detection in high-resolution sar imagery, IEEE Access (2018) 1000–1014.

[16] F. Gao, T. Huang, S. Jinping, W. Jun, H. Amir, Y. Erfu, A new algorithm of sar image target recognition based on improved deep convolutional neural network, Cogn. Comput. (2018).

[17] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

[18] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.

[19] G. Heitz, D. Koller, Learning spatial context: Using stuff to find things, in: European Conference on Computer Vision, Springer, 2008, pp. 30–43.

[20] L. Jamone, E. Ugur, A. Cangelosi, L. Fadiga, A. Bernardino, J. Piater, J. Santos-Victor, Affordances in psychology, neuroscience and robotics: a survey, IEEE Trans. Cogn. Dev. Syst. (2016).

[21] H.S. Koppula, R. Gupta, A. Saxena, Learning human activities and object affordances from rgb-d videos, Int. J. Robot. Res. 32 (8) (2013) 951–970.

[22] E. Lachat, H. Macher, M. Mittet, T. Landes, P. Grussenmeyer, First experiences with kinect v2 sensor for close range 3d modelling, Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 40 (5) (2015) 93.

[23] K. Lai, L. Bo, X. Ren, D. Fox, A large-scale hierarchical multi-view rgb-d object dataset, in: Robotics and Automation (ICRA), 2011 IEEE International Conference on, IEEE, 2011, pp. 1817–1824.

[24] I. Lenz, H. Lee, A. Saxena, Deep learning for detecting robotic grasps, Int. J. Robot. Res. 34 (4-5) (2015) 705–724.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: Single shot multibox detector, ECCV, 2016.

[27] Z. Liu, L. Yuan, L. Weng, Y. Yang, A high resolution optical satellite image dataset for ship recognition and some new baselines., in: ICPRAM, 2017, pp. 324–331.

[28] A. Myers, C.L. Teo, C. Fermüller, Y. Aloimonos, Affordance detection of tool parts from geometric features., in: ICRA, 2015, pp. 1374–1381.

[29] J. Redmon, A. Farhadi, Yolov3: an incremental improvement, (2018) arXiv:1804.02767.

[30] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, (2015) arXiv:1506.01497.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. (IJCV) 115 (3) (2015) 211–252, doi:10.1007/s11263-015-0816-y.

[32] O. Stasse, T. Flayols, R. Budhiraja, K. Giraud-Esclasse, J. Carpentier, J. Mirabel, A. Del Prete, P. Souères, N. Mansard, F. Lamiraux, et al., Talos: a new humanoid research platform targeted for industrial applications, in: Humanoid

Robotics (Humanoids), 2017 IEEE-RAS 17th International Conference on, IEEE, 2017, pp. 689–695.

[33] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, et al., The limits and potentials of deep learning for robotics, Int. J. Robot. Res. 37 (4–5) (2018) 405–420.

[34] V. Tikhanoff, U. Pattacini, L. Natale, G. Metta, Exploring affordances and tool use on the icub, in: Humanoid Robots (Humanoids), 2013 13th IEEE-RAS International Conference on, IEEE, 2013, pp. 130–137.

[35] C. Wang, X. Bai, S. Wang, J. Zhou, P. Ren, Multiscale visual attention networks for object detection in vhr remote sensing images, IEEE Geosci. Remote Sens. Lett. 16 (2) (2019) 310–314.

[36] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, Dota: a large-scale dataset for object detection in aerial images, in: Proc. CVPR, 2018.

[37] B. Xiao, E.R. Hancock, R.C. Wilson, Graph characteristics from the heat kernel trace, Pattern Recognit. 42 (11) (2009) 2589–2606.

[38] S.-J. Yi, S.G. McGill, L. Vadakedathu, Q. He, I. Ha, J. Han, H. Song, M. Rouleau, B.-T. Zhang, D. Hong, et al., Team Thor's entry in the darpa robotics challenge trials 2013, J. Field Robot. 32 (3) (2015) 315–335.

[39] Z. Yue, G. Fei, X. Qingxu, W. Jun, H. Teng, Y. Erfu, Z. Huiyu, A novel semi-supervised convolutional neural network method for synthetic aperture radar image recognition, Cogn. Comput. (2019) 1–12.

[40] H. Zhang, X. Bai, J. Zhou, J. Cheng, H. Zhao, Object detection via structural feature selection and shape model, IEEE Trans. Image Process. 22 (12) (2013) 4984–4995.

[41] Y. Zhu, Y. Zhao, S. Chun Zhu, Understanding tools: task-oriented object modeling, learning and recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2855–2864.