2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Distilling Location Proposals of Unknown Objects through Gaze Information for Human-Robot Interaction

Daniel Weber[1], Thiago Santini[1], Andreas Zell[1], and Enkelejda Kasneci[1]

***Abstract*— Successful and meaningful human-robot interaction requires robots to have knowledge about the interaction context – e.g., which objects should be interacted with. Unfortunately, the corpora of interactive objects is – for all practical purposes – infinite. This fact hinders the deployment of robots with pre-trained object-detection neural networks other than in pre-defined scenarios. A more flexible alternative to pre-training is to let a human teach the robot about new objects after deployment. However, doing so manually presents significant usability issues as the user must manipulate the object and communicate the object's boundaries to the robot. In this work, we propose streamlining this process by using automatic object location proposal methods in combination with human gaze to distill pertinent object location proposals. Experiments show that the proposed method 1) increased the precision by a factor of approximately 21 compared to location proposal alone, 2) is able to locate objects sufficiently similar to a state-of-the-art pre-trained deep-learning method (FCOS) without any training, and 3) detected objects that were completely missed by FCOS. Furthermore, the method is able to locate objects for which FCOS was not trained on, which are undetectable for FCOS by definition.**

I. Introduction

In today's modern world, interaction between human and machines is omnipresent, e.g. in the figure of Alexa and Siri. Moreover, the significant progress in augmented reality is also pushing the boundaries of the cooperation between human and robots. For instance, this emerging kind of human-robot interaction (HRI) has already helped to optimize manufacturing steps in production as well as been applied in factories [1] and for assembly guidance [2]. The great majority of such technological developments has been strongly fueled by machine learning methods, such as neural networks. The collection of huge databases allows us to train and continuously improve (deep) neural networks in order to fulfill challenging tasks. However, these use cases typically operate under the assumption that there are sufficient data sets for training available. But what if the training data is biased (e.g., geographically [3]) or not labeled, for example in many production processes – such as, the assembly of a recently developed electric engine of a car? Furthermore, in some application scenarios such as search and rescue work with unmanned aerial vehicles (UAVs) or the classification of medical images, there might be very few or no available training instances. For example for UAVs, aerial footage is simply difficult to obtain [4], whereas for medical images, storage is often prohibited due to patient privacy [5].

In addition, labeling data is a costly process due to the amount of human effort involved. Drawing a high quality bounding box in an image, including quality and coverage verification, can take a human from 7 up to 42 seconds per object [6], [7]. With multiple objects in a scene this can quickly add up to prohibitive amounts.

In this paper we address this challenge by connecting findings from two research areas: eye tracking and robotics. On the human side, we use the gaze modality to enable the exchange of information for a specific problem on the robot side, namely the detection of unknown objects. Our goal is to enable the deployment of a robot in a non-predefined scenario and to explain an interaction context to the robot, e.g., the class of an object after detection. That is, rather than using a neural network for object detection, we resort to the human gaze and want the robot to detect which object the human is looking at, even though its class is not yet known (see Figure 1). Moreover, interaction requires online operation – in contrast to post processing. To the best of our knowledge, this is the first work to combine the well known technique of selective search [8], which outputs thousands of class-independent object location proposals, with human gaze information to separate useful and useless areas of interest in a scene image. Thus, the proposed approach enables us to detect and process objects in an image without training but still in an efficient way. In summary, our most important contributions are:

Fig. 1: Without specialized pre-training, the robot does not know the objects in front of it. Nonetheless, through our proposed approach, the robot is capable of detecting these unknown objects based on gaze-based human-robot interaction without any training instances.

1) A novel method towards the deployment of robots in non-predefined scenarios.
2) We are the first to connect eye tracking and robotics to detect unknown objects without the usage of neural networks, alleviating training-data dependency.

[1]University of Tübingen, 72076 Tübingen, Germany

3) As a proof of concept we conduct an experiment and demonstrate the validity and feasibility of our method.

II. Related Work

Mapping gaze data from a head-mounted eye tracker with moving point of view, i.e. coordinate system, to a known reference frame is a well-known and open problem in current research. Most works, such as [9], [10], that were confronted with this issue solved it by using fiducial markers. Even though [11] additionally tested feature matching and achieved reasonable results, markers provided better stability and reliability at significantly less computational cost in all of their test cases. Apart from that, their purpose was to match a picture of an image displayed on a screen to a planar reference image, which was very similar to the one displayed on the screen. As described in [12], feature matching reaches its limit when applied to a three-dimensional target object. Accordingly, in our case it is more difficult to find and match features than with a simple painting or a poster, especially when the viewing perspectives differ significantly. In [13] the authors succeeded in mapping the gaze by utilizing velocity features. However, this was limited to the user looking at one of several pre-defined key points.

In recent years, object recognition has been one of the most intensively researched areas in computer vision. The availability of better hardware led to the emergence of deep neural networks as a go-to solution for object detection. YOLO [14], Mask R-CNN [15], SSD [16] and FCOS [17] are great examples of the extensive use of neural networks that constantly have been pushing the boundaries of object detection. These networks typically rely on fully supervised learning methods and the existence of large annotated data sets, such as PASCAL VOC [18], Microsoft COCO [19] and Imagenet [20]. This means that they do not generalize well and lack reliability on unknown domains [21].

Moreover, with increasing climate-related public awareness, there has been some research focusing on energy efficiency of neural networks [22] and its environmental impact [23]. [24] analyzed the power consumption of popular image classification models. Consequently, we follow the recommendation of [23] and prioritize a simple non-deep-learning approach instead.

A few works have already investigated the combination of eye tracking and computer vision tasks. The authors of [25] performed gaze guided object recognition by matching features around human fixations to features from known objects in a database. After a database was created, it was possible to classify an image, but not to determine the position of the object within the image. [26] concentrated on annotating images with bounding boxes. They utilized fixation points to extend existing training data with gaze information. Subsequently, a model was trained that predicted bounding boxes from the fixations while viewing an image. A strategy for superpixel segmentation with eye tracking data was proposed by [27]. Just like the previous method, training data was already required right from the start. In addition, both methods require multiple gaze points. In contrast, our method is able to operate with as few as one gaze point, thus being applicable in an online fashion.

In this paper, we build on existing work and benefit from collaborative working with a robot. In this context, eye tracking can play an important role and connect humans and robots in a natural and intuitive manner, offering an additional communication channel available even when traditional channels, such as speech and gestures [28], might not be available for HRI – e.g., during microsurgery [29]. We use the human gaze to enable a robot to interact with its unknown environment by letting it recognize objects we are looking at. Thereby, we bridge the gap between existing approaches for object detection and data independence with eye tracking.

III. Method

In this work, we propose finding pertinent and accurate location proposals of unknown objects through gaze information. This process can be thought of as three building blocks: 1) estimating the human partner's gaze in the robot's frame of reference, 2) generating location proposals for unknown objects, and 3) distilling the location proposals using the gaze information. Throughout this section, we assume the robot to be equipped with at least one camera.

A. Gaze Estimation

The most straightforward and inexpensive way of estimating the partner's gaze in the robot's frame of reference is by estimating the gaze directly through the robot's sensors – e.g., through appearance or model-based remote gaze estimation methods [30], [31]. However, this poses a key limitation as the partner must be facing the robot, severely limiting the perspectives from which gaze-based HRI can happen.

This limitation can be alleviated through multiple remote eye trackers distributed around the environment or the usage of a head-mounted eye tracker. However, in both cases, it is necessary to map the estimated gaze from the eye tracker frame of reference to the robot's. This transformation can be achieved in multiple ways, for example by 1) directly finding the eye tracker's pose in the robot's camera or vice versa, or 2) indirect co-location, by finding at least four corresponding points in images of the eye tracker's and robot's cameras[1].

In this work, we favor the usage of a head-mounted eye tracker due to the reduced costs (i.e., only a single eye tracker is required) and user constraints. Moreover, we employ fiducial markers [32] for co-location as these provide a robust and inexpensive solution to the gaze mapping issue that can be employed in traditional HRI scenarios such as in factories, care facilities, or individual homes.

B. Unknown Object Location Proposal

Location (or region) proposal methods consist of determining candidate object locations (e.g., bounding boxes, or

[1]By finding the plane defined by the these four points, one can estimate the pose of each camera relative to the plane and, thus, the pose of one camera relative to the other.

segmentation masks) that *might* contain an object. This task can be realized, for example, through segmentation [33], randomly-sampled boxes classification [34], jumping windows [35], and selective search [8]. Such methods are typically used as an alternative to exhaustive search for object detection to reduce the search space, speeding up the detection and reducing the associated computing costs.

The cardinality of the proposed locations set is, naturally, image-dependent but tends to be in the order of thousands. Normally, each location proposal is run through a pre-trained classifier to detect whether an object is present in it. However, many of these methods, such as the ones proposed by [8], [34], have a particularly interesting property: The proposed locations are *class-independent*. In other words, within the proposed regions there are objects that a computer vision system might not have been trained to identify – i.e., unknown objects. This begs the question: Can we identify pertinent locations from the set of proposals for interaction or to teach a robot about new objects in a natural way?

C. Distillation Through Gaze Information

In this work, we approach the task of identifying location proposals that are pertinent for a human-robot interaction from the full set of class-independent proposals by using gaze information from the human partner. This distillation process can be activated through multimodal interactions – e.g., through touch or voice. Nevertheless, we also envision an automatic approach in which the robot notices the human's gaze continously attending to a region where no known object has been identified yet.

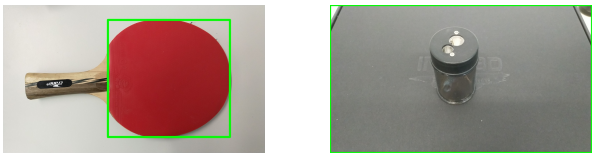
In order to obtain an initial set of candidate bounding boxes, we resort to selective search [8]. Selective search uses the segmentation method from Felzenszwalb and Huttenlocher [36] to analyze the intensity of the pixels of the image and perform segmentation. The segmented parts and groups of adjacent segments are then used to calculate and propose regions of interest. In other words, this algorithm-based approach combines the high recall of exhaustive search with the image guided sampling process of segmentation and outputs bounding boxes in a hierarchical order. The benefits here are two-fold: the method can capture all possible object locations and the region proposals are guided by the structure of the image, such as color, texture, size and shape, leading to a reduced number of proposed locations. In this paper, we will refer to the position with respect to the order in which the boxes appear in the output set of region proposals as position *index*.

Although the number of bounding boxes is reduced in comparison to an exhaustive search approach, this does not effect the high recall we need to ensure that we can find a suitable box for each object. Moreover, it is possible to further distill the regions into a smaller and more-pertinent set of proposals: Since we know that the gaze coordinate has to lie within the searched bounding box, we can employ this information as a filtering mechanism. Let $(x_1^{(i)}, y_1^{(i)}) \in \mathbb{N}^2$ be the lower left and $(x_2^{(i)}, y_2^{(i)}) \in \mathbb{N}^2$ be the upper right corners of the bounding box $B_i \in B$, where B is the full set of class-independent proposals. By tracking our gaze point $g = (x, y) \in \mathbb{N}^2$, we can distill a smaller subset $B_g \subset B$ of pertinent proposals from B:

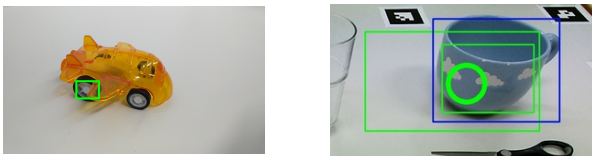
$$B_g \coloneqq \left\{ B_i \in B \,\middle|\, x_1^{(i)} \leq x \leq x_2^{(i)},\ y_1^{(i)} \leq y \leq y_2^{(i)} \right\}.$$

This subset B_g contains only bounding boxes that have an intersection with the object marked by the gaze point. As we will see later, to achieve satisfactory results, we are dependent on a high gaze-tracking accuracy and a robust gaze mapping.

Note that getting multiple (but hierarchically-sorted) bounding boxes proposals is not a disadvantage but an advantage in our use case. As previously mentioned in [8], an object can consist of different colors, multiple objects can have the same color, or the object could be indistinguishable from its background. In Figure 2 one can see that this could lead to problems if the detection fails in terms that the only proposed bounding box is not correct or the object is not detected at all.



(a) table tennis racket (b) sharpener



(c) toy (d) cup

Fig. 2: Objects can vary in shape and size, have different backgrounds and can consist of multiple colors. This may cause errors regarding the detection. The green boxes in the figure indicate proposed regions. In (a) the red part is proposed earlier, meaning the corresponding bounding box has a lower position index than the whole racket. (d) shows the first three proposals we receive for the blue cup. The first two (green) are not as accurate as the third (blue). Through interaction it is possible to communicate the preferred bounding box.

Moreover, we strive for a more human-like learning process, in the sense of an interaction between robot and human, similar to that of a human with another human. Multiple proposals also mean that we can decide to choose the second or third proposed and more accurate box instead of the first one (see Figure 2d). Interaction between robot and human makes these decisions possible and brings us closer to a natural learning process.

IV. Experimental Setup

In order to showcase a working proof of concept of the proposed application, we collected a session for a participant (one of the system's designers) with the whole system

working in real-time[2]. This session serves as basis for our initial evaluation of the system.

On a table, we placed different objects, including partially overlapping objects to some extent. To have a wide appearance range, we selected objects with distinct sizes, colors, and shapes. In Figure 3, one can see the robot and his view in front of the table with all objects he is supposed to detect. For the sake of simplicity and for later evaluation, we have used ordinary office and household items that are all part of the Microsoft COCO data set [19].

Fig. 3: With a Microsoft Kinect v2 the robot sees different objects on a table: Keyboard, scissors, cups, bottle, fork, knife, spoon, mouse and a small toy car.

As hardware, we used the first generation of Pupil Core [37], a head mounted eye tracker developed by Pupil Labs. Although Pupil Labs provides a software solution called *Pupil Capture* and *Pupil Player*, we decided to utilize *EyeRecToo* [38], an open-source software for real-time pervasive head-mounted eye-tracking. The main reasons were the calibration method *CalibMe* [39], the robust detection of ArUco markers, and slippage robustness [40]. *EyeRecToo*'s pupil tracking pipeline was set to use *PuRe* [41] / *PuReST* [42]. Our robot counterpart is a Scitos G5 from MetraLabs [43] equipped with a Microsoft Kinect for XBox One. We accessed the RGB channels of the Kinect v2 using *ROS* [44], *libfreenect2* [45], and *iai_kinect2* [46]. For the implementation, we make extensive use of the *OpenCV* [47] library.

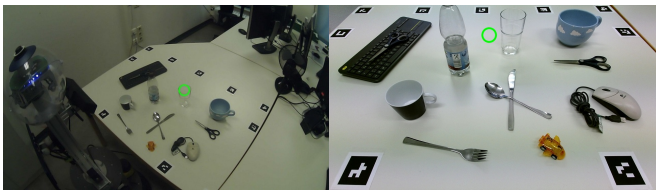
V. Evaluation

To establish reference ground-truth values for the object locations, we have employed the Fully Convolutional One-Stage Object Detector (FCOS) [17] trained on Microsoft COCO [19], using the ResNeXt-64x4d-101 backbone with deformable convolutions. This serves as a baseline representing a state-of-the-art object detection for supervised learning. Given an image viewed from the robot's perpective, the output of FCOS is shown in Figure 5a. It is worth noting that the bottle was detected twice; in this case, we opted to ignore the smaller inaccurate bounding box. Moreover, neither the knife that overlaps with the spoon nor the scissor placed on the keyboard are recognized by FCOS, despite all of these classes being present in the training data. Thus, we discuss these separately.

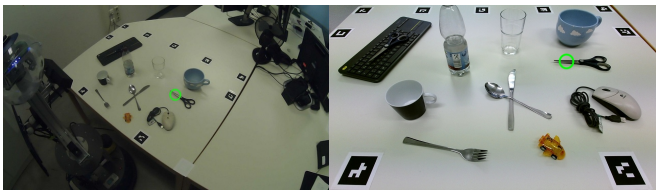
[2]Eye tracking and gaze mapping working at about 30 frames per second.

A. Qualitative Analysis

Both eye tracking and marker detection work in real time, as well as the subsequent gaze mapping. Therefore, our method is suitable for real-time human-robot interaction. As long as the accuracy in all three steps is high enough, the robot knows at any time where we are looking at. The human is even unrestricted in his movements. Figure 4 shows two attempts of pointing out an object to the robot. One was successful and the other one failed. Although the human from whom the gaze point in Figure 4a originated actually looked at the glass and his gaze was tracked correctly, the gaze point in the robot's view is not on the glass, i.e. the mapping procedure was problematic in this case. This exemplifies that enough markers have to be detected to guarantee reliable mapping and usability. This could be ensured, for example, by using more accurate markers such as infrared tokens. In addition, the tracking of the human gaze must work reliably to achieve satisfactory usability. Therefore, we have carefully calibrated the eye tracker to achieve the desired accuracy. During interaction, however, the device is likely to slip [48] such that slippage robustness is paramount.



(a) Failed



(b) Successful

Fig. 4: A failed and a successful attempt of mapping the human gaze (left) on the robot's view (right).

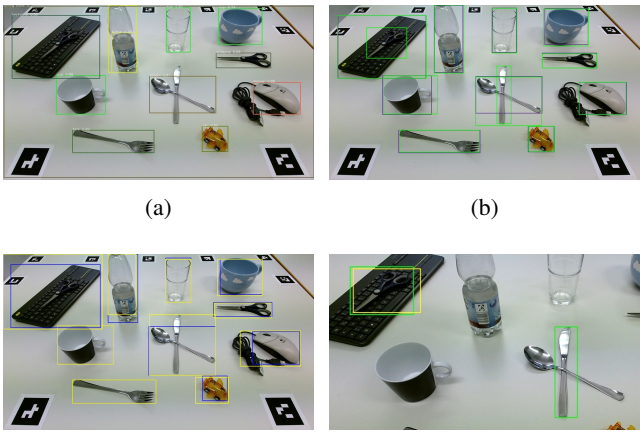
In contrast to the gaze mapping, the region proposal achieves real-time operation only at lower frame rates. The calculation of all the 2198 region proposals on our picture of the robot's view with a resolution of 1900x1080 took about 2.7 seconds with the "quality" method. Nonetheless, this is not a problem, as region proposal is not required for each frame but only sporadically. Once a correct bounding box for the intended object has been found, it can be tracked with well-known tracking algorithms like KCF [49] or CSRT [50].

B. Quantitative Analysis

To evaluate the efficiency of our method we compare the position indices of each bounding box within the complete hierarchical set of region proposals from the selective search algorithm with the indices we have distilled. Of course these boxes should not only be easy and fast to find but have to be accurate as well. For this reason, we need to investigate the similarity of the proposed boxes w.r.t. the ground truth. As measurement for accuracy, we calculate the Jaccard index $J(B_1, B_2)$, also known as Intersection over Union (IoU). This means that the closer the Jaccard index is to 1, the greater the similarity between the boxes. For object detection, if the Jaccard index is more than 0.5, a detection is typically considered correct [18]. Nevertheless, in general, a higher value is desirable. [51] provides a comparison of different values of the Jaccard index and describes 0.5 as very loose, 0.9 as very strict and 0.7 as reasonable compromise in between. Therefore, we set 0.7 as threshold and characterize bounding boxes with at least this value as "sufficient". This allows us to analyze whether the selective search algorithm is a good choice and provides region proposals that are accurate enough, i.e. sufficient, for our use case.

In Table I the Jaccard index between the boxes predicted by FCOS and the best box in our set of proposals is listed for each item. Note that the knife and scissor placed on the keyboard are omitted from Table I because they are not recognized by FCOS, which means we do not have any reference values for these items. We will discuss these items separately at the end of this section. Besides, depending on whether we want to consider the mouse cable or not, the values in Table I naturally change. Even though we can distill boxes for both cases, here we stick to the output of FCOS and only consider the mouse without cable. One must keep in mind, however, that the composition of the mouse from two sub objects leads to different predictions being made. For example, if the cable had not been bundled up, the relevant position index would indeed be smaller or the mean Jaccard index would be larger. In our test, the use of a second gaze point allowed the delimitation to boxes containing the cable and the mouse body alone. Compared to regular object detection, we are not bound to fixed ideas of objects but can vary the object's bounding box depending on the situation. It is worth noting that this is a good example where our method surpasses all pre-trained object detectors in terms of flexibility. In this particular case, there is not only one correct box, but two. *Gaze allows us to resolve such ambiguities*.

We can observe that the Jaccard indices with few exceptions are all above 0.85 and the vast majority is even above 0.9 (see column 4). This is highlighted visually in Figure 5b, which shows the bounding box detected by FCOS and by the proposed approach. With a mean value of the Jaccard indices of 0.919, the proposed boxes are highly relevant. This value can also be used as upper bound for the accuracy of our fast distillation. Also worth noting are the massively high indices of the respective best box in each category. Whereas



(a) (b) (c) (d)

Fig. 5: (a) shows the objects detected with FCOS. The confidence of the prediction can also be seen in Table I. (b) shows a comparison of the total best bounding box (green) with the ground truth given by FCOS (purple). (c) shows a comparison of the best bounding box among the first 15 proposals (blue) with earlier sufficient boxes (yellow). Note that these boxes are identical for the cup and the fork. The knife and scissors on the keyboard have been omitted as they are handled separately. (d) shows the best bounding boxes distilled for the knife and the scissor on the keyboard.

the position index of the best box of the bottle in the full set is the lowest at 292, the mean value of the position index is about 1315. Through distillation, we managed to improve that value to an average of 61.5. Figure 6 illustrates that the vast majority of the boxes in the full set of proposals has an Jaccard index below 0.1 and is therefore irrelevant.

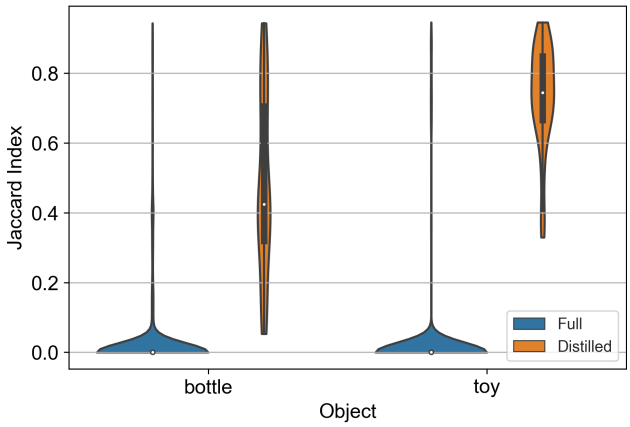


Fig. 6: The violin plot shows the distribution of the Jaccard indices for the full and the distilled set of bounding boxes using the example of the bottle and the toy.

As Table I and Figure 5c show, we do not have to find exactly the best boxes. With our fast distillation method, we were able to provide a proposal among the first 15 boxes of each distilled subset with at least 94 % accuracy to the best. That is, with an average accuracy of even 97.91 %, we were almost as accurate as the best possible box, with a much smaller position index. Therefore, we need much less communication with the robot to reach the desired box.

TABLE I: Comparison between the full and our distilled set of bounding boxes.

Item	FCOS	Best total		#Boxes	First sufficient			Best among first 15			Recall		Precision		F_1 score	
	Confidence	Index	IoU	Dist.	Index	IoU	Acc.[1]	Index	IoU	Acc.	Full	Dist.	Full	Dist.	Full	Dist.
Bottle	0.69	**292**[2]	0.943	98	1	0.851	90.24 %	**12**	0.943	100 %	1	1	0.012	0.265	0.023	0.419
Cup (black)	0.88	1748	0.863	221	**3**	0.828	95.94 %	**3**	0.828	95.94 %	1	1	0.029	0.29	0.057	0.449
Cup (blue)	0.74	1199	0.918	110	3	0.870	94.77 %	15	0.899	97.93 %	1	1	0.014	0.273	0.027	0.429
Fork	0.83	1873	0.945	34	**1**	0.939	99.37 %	**1**	0.939	99.37 %	1	1	0.013	0.824	0.025	0.903
Glass	0.82	1429	0.935	110	3	0.896	95.83 %	4	0.922	98.61 %	1	1	0.020	0.391	0.038	0.562
Keyboard	0.55	1839	0.988	189	1	0.751	76.01 %	9	0.981	99.29 %	1	1	0.046	0.529	0.087	0.692
Mouse	0.88	883	0.968	231	1	0.714	73.76 %	11	0.955	98.66 %	1	0.96	0.011	0.104	0.023	0.188
Scissor	0.68	1137	0.954	89	2	0.880	92.24 %	9	0.898	94.13 %	1	1	0.018	0.449	0.036	0.620
Spoon	0.61	670	0.727	118	**3**	0.720	99.04 %	**3**	0.720	99.04 %	1	1	0.002	0.034	0.004	0.066
Toy car	0.52	2079	0.946	68	3	0.712	75.26 %	5	0.909	96.09 %	1	1	0.021	0.662	0.040	0.797
Ø	0.72	1314.9	0.919	126.8	2.1	0.816	89.25 %	7.2	0.899	97.91 %	1	0.996	0.018	0.382	0.036	0.512

[1] Accuracy compared to the best box in the full set (ratio of the two Jaccard indices).
[2] Bold indices within the same line indicate identical boxes.

In addition, we have considered earlier sufficient boxes in the sense of boxes with a Jaccard index of at least 0.7. Figure 5c shows these bounding boxes along with the best boxes among the first 15 described above. Their position index is of course lower and, in our particular case, never higher than three. As highlighted in Table I, it is often sensible to fall back to earlier boxes. For instance, in the case of the blue cup, it is possible to reduce the position index from 15 to 3 while reducing the Jaccard index only by 0.03. In contrast, the toy car's position index is acceptable either way, and a significant drop in accuracy results if the position index is lowered from 5 to 3. In general, the average accuracy is about ten percent lower compared to the best possible box, but, as previously mentioned, it is found early since it is one of the first three proposals.

Considering that the sufficient boxes (Jaccard index > 0.7) are the relevant ones, we define a) *recall* as the ratio between relevant boxes retrieved by our method and all relevant boxes, as well as b) *precision* as the ratio of boxes retrieved by our method that are relevant. The per-object recall and precision are reported in Table I together with the F_1 score $(2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}})$, which is the harmonic mean of precision and recall. Whereas the recall remained virtually the same, the precision increased significantly due to the distillation using the proposed method. To be more specific, while on average not even $2\,\%$ of the boxes in the full set could be considered as sufficient, almost $40\,\%$ of the distilled boxes have a Jaccard index of at least 0.7. Moreover, the resulting mean F_1 value is about 14 times higher for the distilled sets compared to the full set.

Finally, we would like to discuss the objects that were not recognized by FCOS. The knife and the spoon lying on top of each other was a difficult task. Both FCOS and our proposed method struggled on this part. Although FCOS was not able to detect the knife at all, we at least managed to get a sufficient box with a high position index of 75. Figure 5d shows our best possible result. However, we needed several attempts to match the human gaze point with the knife since the knife's width is relatively small.

Detecting the dark blue scissor on the black keyboard was on the other hand quite easy in terms of mapping the gaze point on the object. Even though it was generally an even harder task with respect to the color of the background, unlike FCOS, the proposed method was able to find one sufficient and one best bounding box with the position index of 45 and 99, respectively. These bounding boxes were also relatively late to reach but far earlier than 635 and 1251, their indexes in the full set.

C. Limitations

Our method is based on a high accuracy in each partial step. This is an issue if we have either bad gaze tracking or mapping, which could result by too small or too few markers, low-quality hardware, or external disturbances. Even if the mapping part is accurate, a sloppy gaze estimation can lead to a gaze point that does not overlap with the object. With an inaccurate gaze point in the robot's view, accurate bounding box proposals are difficult and sometime impossible to distill. In this case, we have to repeat pointing out to the object.

Furthermore, with one exception (scissors on the keyboard), the proof of concept was carried out on a plain white table. Although we would expect more candidate boxes in less homogeneous settings, our experiments suggest that there would still be highly relevant boxes due to the high recall that is in the nature of the method.

VI. Conclusion

In this work, we have proposed and evaluated a novel method that enables the deployment of robots in non-predefined scenarios. The proposed method combines automatic object location proposals with human gaze to distill pertinent location proposals. Just by looking at an object and some human robot communication, we can find a bounding box with a Jaccard index of almost 0.9 compared to the ground truth. These boxes can then be used to quickly extend the robot's object detection neural network.

Out of thousands of possible region proposals, we successfully distilled useful object-independent bounding boxes, increasing the precision of the location proposals by over 21 times with virtually no recall loss. Despite challenging scenarios, our method was consistently applicable and does not need any training at all. Relative to a state-of-the-art object detector (FCOS) trained on the Microsoft COCO data set, we achieved an average Jaccard index of almost 0.9 for at least one box out of the first 15 proposals. Looking only at

the first sufficient box of each object, we observed an average accuracy of $89.25\,\%$ compared to the respective best possible box in the full set of proposals.

Since our gaze method significantly improved the position index of important bounding boxes compared to the large initial number of region proposals, it enables concentrating exclusively on relevant boxes. This allows the robot to find the intended object more quickly and to generally reduce the necessary communication, improving the human-robot interaction. In addition, we could find bounding boxes to objects that could not even been detected by FCOS.

In summary, our proposed method is therefore a broadly applicable and natural way to achieve unknown-object detection by a robot in HRI scenarios. However, a significant amount of work remains for future work as we plan to extend our proof of concept by evaluating our method with more participants and additionally investigate the impact of imperfect labels on the training of neural networks.

ACKNOWLEDGMENT

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645.

REFERENCES

[1] A. Y. Nee, S. Ong, G. Chryssolouris, and D. Mourtzis, "Augmented reality applications in design and manufacturing," *CIRP annals*, vol. 61, no. 2, pp. 657–679, 2012.

[2] L. Rentzos, S. Papanastasiou, N. Papakostas, and G. Chryssolouris, "Augmented reality for human-based assembly: using product and process semantics," *IFAC Proceedings Volumes*, vol. 46, no. 15, pp. 98–101, 2013.

[3] T. DeVries, I. Misra, C. Wang, and L. van der Maaten, "Does object recognition work for everyone?" in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 52–59.

[4] M. B. Bejiga, A. Zeggada, A. Nouffidj, and F. Melgani, "A convolutional neural network approach for assisting avalanche search and rescue operations with UAV imagery," *Remote Sensing*, vol. 9, no. 2, p. 100, 2017.

[5] J. Cho, K. Lee, E. Shin, G. Choy, and S. Do, "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?" *arXiv preprint arXiv:1511.06348*, 2015.

[6] D. P. Papadopoulos, J. R. R. Uijlings, F. Keller, and V. Ferrari, "Extreme clicking for efficient object annotation," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[7] H. Su, J. Deng, and L. Fei-Fei, "Crowdsourcing annotations for visual object detection," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[8] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[9] C.-M. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 83–90.

[10] V. Peysakhovich, F. Dehais, and A. Duchowski, "ArUco/gaze tracking in real environments," in *Eye Tracking for Spatial Research, Proceedings of the 3rd International Workshop*. ETH Zurich, 2018.

[11] M. Kalash, K. Singh, R. Eskicioglu, and N. D. Bruce, "Gaze-contingent interactive visualization of high-dynamic-range imagery," in *2016 IEEE Second Workshop on Eye Tracking and Visualization (ETVIS)*. IEEE, 2016, pp. 16–20.

[12] J. J. MacInnes, S. Iqbal, J. Pearson, and E. N. Johnson, "Wearable eye-tracking for research: Automated dynamic gaze mapping and accuracy/precision comparisons across devices," *bioRxiv*, p. 299925, 2018.

[13] R. M. Aronson and H. Admoni, "Semantic gaze labeling for human-robot shared manipulation," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 2019, pp. 1–9.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[15] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[17] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[18] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

[21] M. J. Wilber, C. Fang, H. Jin, A. Hertzmann, J. Collomosse, and S. Belongie, "Bam! the behance artistic media dataset for recognition beyond photography," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1202–1211.

[22] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, 2016.

[23] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," *arXiv preprint arXiv:1906.02243*, 2019.

[24] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," *arXiv preprint arXiv:1605.07678*, 2016.

[25] T. Toyama, T. Kieninger, F. Shafait, and A. Dengel, "Gaze guided object recognition using a head-mounted eye tracker," in *Proceedings of the 2012 ACM Symposium on Eye Tracking Research & Applications*. ACM, 2012, pp. 91–98.

[26] D. P. Papadopoulos, A. D. Clarke, F. Keller, and V. Ferrari, "Training object class detectors from eye tracking data," in *European conference on computer vision*. Springer, 2014, pp. 361–376.

[27] F. Xiao, L. Peng, L. Fu, and X. Gao, "Salient object detection based on eye tracking data," *Signal Processing*, vol. 144, pp. 392–397, 2018.

[28] R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel, "Natural human-robot interaction using speech, head pose and gestures," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, Sendai, Japan, September 28 - October 2, 2004*. IEEE, 2004, pp. 2422–2427. [Online]. Available: https://doi.org/10.1109/IROS.2004.1389771

[29] W. Fuhl, T. Santini, C. Reichert, D. Claus, A. Herkommer, H. Bahmani, K. Rifai, S. Wahl, and E. Kasneci, "Non-intrusive practitioner pupil detection for unmodified microscope oculars," *Comp. in Bio. and Med.*, vol. 79, pp. 36–44, 2016. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2016.10.005

[30] X. Zhang, Y. Sugano, and A. Bulling, "Revisiting data normalization for appearance-based gaze estimation," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. ACM, 2018, p. 12.

[31] O. Palinko, F. Rea, G. Sandini, and A. Sciutti, "Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 5048–5054.

[32] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable

fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.
[33] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3241–3248.
[34] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012.
[35] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *The IEEE 12th International Conference on Computer Vision (ICCV)*. IEEE, 2009, pp. 606–613.
[36] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, no. 2, pp. 167–181, Sept. 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000022288.19776.77
[37] Pupil Labs, https://pupil-labs.com/, 2019, accessed: 2020-02-09.
[38] T. Santini, W. Fuhl, D. Geisler, and E. Kasneci, "EyeRecToo: Open-source software for real-time pervasive head-mounted eye-tracking," in *12th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, 02 2017.
[39] T. Santini, W. Fuhl, and E. Kasneci, "CalibMe: Fast and unsupervised eye tracker calibration for gaze-based pervasive human-computer interaction," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 05 2017.
[40] T. Santini, D. C. Niehorster, and E. Kasneci, "Get a grip: slippage-robust and glint-free gaze estimation for real-time pervasive head-mounted eye tracking," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. ACM, 2019, p. 17.
[41] T. Santini, W. Fuhl, and E. Kasneci, "PuRe: Robust pupil detection for real-time pervasive eye tracking," *Computer Vision and Image Understanding*, vol. 170, pp. 40 – 50, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314218300146
[42] ——, "PuReST: Robust pupil tracking for real-time pervasive eye tracking," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ser. ETRA '18. New York, NY, USA: ACM, 2018, pp. 61:1–61:5. [Online]. Available: http://doi.acm.org/10.1145/3204493.3204578
[43] MetraLabs, https://www.metralabs.com/mobiler-roboter-scitos-g5/, 2019, accessed: 2020-02-09.
[44] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "ROS: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, no. 3.2. Kobe, Japan, 2009, p. 5.
[45] L. Xiang, F. Echtler, C. Kerl, *et al.*, "libfreenect2: Release 0.2," apr 2016. [Online]. Available: https://doi.org/10.5281/zenodo.50641
[46] T. Wiedemeyer, "IAI Kinect2," https://github.com/code-iai/iai_kinect2, Institute for Artificial Intelligence, University Bremen, 2014 – 2015, accessed: 2019-11-21.
[47] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.
[48] D. C. Niehorster, T. Santini, R. S. Hessels, I. T. Hooge, E. Kasneci, and M. Nyström, "The impact of slippage on the data quality of head-worn eye trackers," *Behavior Research Methods*, pp. 1–21, 2020.
[49] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 3, pp. 583–596, 2014.
[50] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6309–6318.
[51] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*. Springer, 2014, pp. 391–405.