Enhancing Robot-Assisted WEEE Disassembly through Optimizing Automated Detection of Small Components

Ioannis Athanasiadis, Athanasios Psaltis, Apostolos Axenopoulos, and Petros Daras

The Visual Computing Lab - Centre for Research and Technology Hellas/Information Technologies Institute
{athaioan, at.psaltis, axenop, daras}@iti.gr

Abstract. Automated detection of small objects poses additional challenges, compared to bigger-sized ones, due to the former's limited resolution for extracting discriminative information. In such cases, even a slight misalignment between a candidate region and its ground truth target has a huge impact on their IoU which significantly increases the amount of noisy information. Given the fact that state of the art two-stage detection algorithms generate predefined shaped and sized candidate regions in pixel-level interval, the aforementioned misalignments are very likely to occur. In this work, a scalable object detection approach is introduced -specifically dedicated to small object parts- incorporating both learnable and handcrafted features. In particular, a set of simplified Gabor waveforms (SGWs) is applied to the raw data, ultimately producing an improved set of anchors for the region proposal network. These Gabor filters are further utilized generating a soft attention mask. Additionally, the interaction of a human with the object is also exploited by taking advantage of affordance-based information for further improvement of detection performance. Experiments have been conducted in a newly introduced device disassembly segmentation dataset, demonstrating the robustness of the method in detection of small device components.

Keywords: Small object detection, WEEE disassembly

1 Introduction

With the advent of industry 4.0, the role of robotics in the industrial environments has evolved. Traditional industrial robots have started being replaced by collaborative robots. The rationale behind this selection is, instead of using high-precision but also dangerous traditional robots in fully automatized processes, to exploit the ability of collaborative robots to coexist with humans in a fenceless way, in order to assist the latter in solving complex cognitive tasks. Some examples include automated parts assembling or disassembling. More specifically, in the context of a Waste Electrical and Electronic Equipment (WEEE) disassembly scenario, within an industrial WEEE recycling environment, the fully-robotised disassembly is not feasible due to the complexity and high variability

of devices. Thus, the role of a collaborative robot in assisting the human worker in detecting and removing hazardous components from the electronic devices is much appreciated. In this direction, Computer Vision is necessary to assist the robot's perception of the surrounding environment. Nevertheless, recognition of the small components to be disassembled is a challenging task, since current state of the art computer vision approaches fail to detect objects in very low resolution. In this work, a novel methodology is proposed for effective detection of low-resolution objects, which makes it suitable for automated detection of very small components in robot-assisted WEEE disassembly tasks.

2 Related Work

Object detection is the process of localizing and classifying the objects appeared on an image. Moreover, its numerous applications, ranging from self driving cars to medical image processing along with its importance in providing the machines with the ability to perceive the world, have attracted many researchers to this field. A plethora of approaches have been proposed for the object detection task that can be categorized into two broad groups, namely the two-stage and the one-stage methods, while there is a complementary set of algorithms that aims at enhancing the previous two. One of the first attempts to utilise Convolutional Neural Networks (CNNs) in object detection was the R-CNN [3] in which a number of class-agnostic candidate regions are proposed and fed to a CNN to extract a fixed-length feature descriptor for each region. Thereafter, a set of class-specific Support-Vector Machines (SVMs), classifies these regions based on their extracted descriptors. Built upon R-CNN success, the Fast R-CNN [2] targets the inefficiency of having to pass each of the candidate regions individually through the CNN by forward passing the input image to the network once, generating its feature map and applying ROI pooling for each of the candidate regions to extract their feature representations. Based on the previously mentioned methods, Faster R-CNN [12] introduced a trainable mechanism for the purpose of proposing candidate regions called Regional Proposal Network (RPN). Given a number of fixed shape and size regions, called *anchors*, the RPN distinguishes them between foreground and background before passing the former to the classifier. Mask R-CNN [4] extended the Faster R-CNN by adding an extra head for segmentation and replaced the ROI pooling with ROI align resulting in higher accuracy predictions. Guided Anchoring [16] proposed an approach detaching the hyper-parameterizing needed for the anchoring process. Additional classifiers are added in Cascade R-CNN [1] aiming at progressively increasing the IOU's of the proposed regions with the ground truth objects resulting in improved prediction performance. Prior-knowledge in interpreted in the object detection process by Reasoning R-CNN [17] which consists of two cascade classification levels, in the first one only visual information is considered, while the second one capitalizes on more informative feature descriptors complemented by high-level information as encoded by the reasoning module. In contrary to the R-CNN family methods in which the processes of

region proposing and region classification are done by discrete modules, in the one-stage methods the regions are generated and classified in a single pass manner. When it comes to one-stage object detection approaches YOLO [11] and the SSD [9] are the most indicative ones. Although this category of methods offers faster performance compared to the RPN-based one, they are limited in terms of accuracy due to having a high imbalance between positive and negative regions fed to the classifier, where the positive and negative terms refer to the presence and the absence of ground truth object respectively. A novel Focal loss has been proposed in [8] addressing that imbalance by having the ambiguous regions contribute more in the loss calculation, thus valuing the hard examples more than the easily classified ones. The potential of exploiting heuristic information for the purpose of sampling the generated anchors, that are more likely to include objects, are presented in [19] and [18] with promising results. Another method based on handcrafted features is [14], where the High Possible Regions Proposal Network, similarly to how RPN operates, proposes candidate regions given an additional feature map as generated by the application of a set of simplified Gabor wavelets (SGWs) on the input image. The use of context information is adapted in [13] focusing on detecting small objects, that the baseline one and two-stage methods struggle with. In [6] the Perceptual Generative Adversarial Network is introduced targeted at enriching the poor visual representation of small objects by super-resolving them. Although Deep NNs have proved to be quite powerful, given sufficient data, they still rely heavily on large datasets and informative visual representations. In the case of small objects specifically, often none of the previous requirements are met thus their detection frequently fails. The motivation behind this work is to effectively boost the performance of current state of the art object detection methods in detecting small objects by exploiting additional streams of information to make up for the poor visual quality small object posses. The main contributions of our work comprise the effective incorporation of handcrafted features into the object detection process through either an anchoring or a soft attention mechanism guided by SGWs. Finally, we have also tested the introduction of an additional input stream based on *object affordances*, with the objective of further improving the detection accuracy. The aforementioned term of object affordance, refers to all possible ways a specific object can be manipulated during its usage.

3 Our Approach

3.1 Overview

Most of the Computer Vision domains have been greatly benefited from the Deep Learning (DL) era. Regarding the object detection specifically, replacing the initial handcrafted feature extraction process with deep data-driven architectures, has shown considerable potential and displayed remarkable results. Although current state of the art object detection algorithms achieve sufficient object detection accuracy, they perform disproportionately better at detecting big and

medium sized objects compared to the smaller ones. The representation generated by the DL-based models, in the case of small objects, are mostly noisy and poor in terms of quality. The former relying heavily on rich feature representation for both object localization and classification combined with the lacking small objects visual information, results in unsatisfactory object detection. Nevertheless, there are applications in which detecting small objects is of high importance. The motive behind this work is to investigate various ways of increasing small object detection performance by exploiting additional information streams. A suitable baseline method is chosen which we progressively enhance through using both handcrafted feature as well as an additional stream of human-to-object interaction information. In the context of this work the detection performance is our main priority, thus we focus on enhancing the two-stage object detection algorithms, nevertheless our proposed method can be mildly modified in order to be applicable to one-stage approaches as well.

3.2 Two-Stage Detection pipeline

State of the art two-stage detection approaches are consisted of two discrete modules responsible for region proposing and classifying respectively. In the first stage, a set of regions are being validated on their objectness; namely the possibility of containing a ground truth object. The most confident regions, in terms of objectness, are passed to the second stage. In the second stage, a feature representation is extracted for each proposed region which is indicative of the area the latter occupy on the image. Finally each region is classified into one the available categories based on their extracted feature.

First Stage - RPN: Given that target objects may appear anywhere on the image, an anchoring scheme is deployed to generate a number of densely distributed anchors. These anchors are generated on a pixel basis across the feature map and have their size and shape defined by the hyperparameters of scale and ratio respectively. These parameters shall be fine-tuned based on the specific application through carefully considering the input image resolution, the potential shapes of the objects interested in detecting as well as their size relatively to the input. Finally, all the uniformly generated anchors constitute the *candidate regions* and are passed to the RPN. Each candidate region is labeled as foreground region if its intersection over union (IOU) with any ground truth object exceeds a predefined threshold and as background otherwise. Finally the RPN is trained to classify its input regions between foreground and background as well as refines those falling in the former category, in order to better fit their corresponding ground truth targets. The refined foreground regions compose the *proposed regions*.

Second Stage - Region Classification: For each proposed region a feature representation is extracted by pooling onto the feature maps that were previously used for regional proposing. Then these regions are classified into one of

the available categories and are further refined by class-aware regressors. The feature maps being shared among the two modules allow for region proposing and classifying modules to be trained simultaneously.

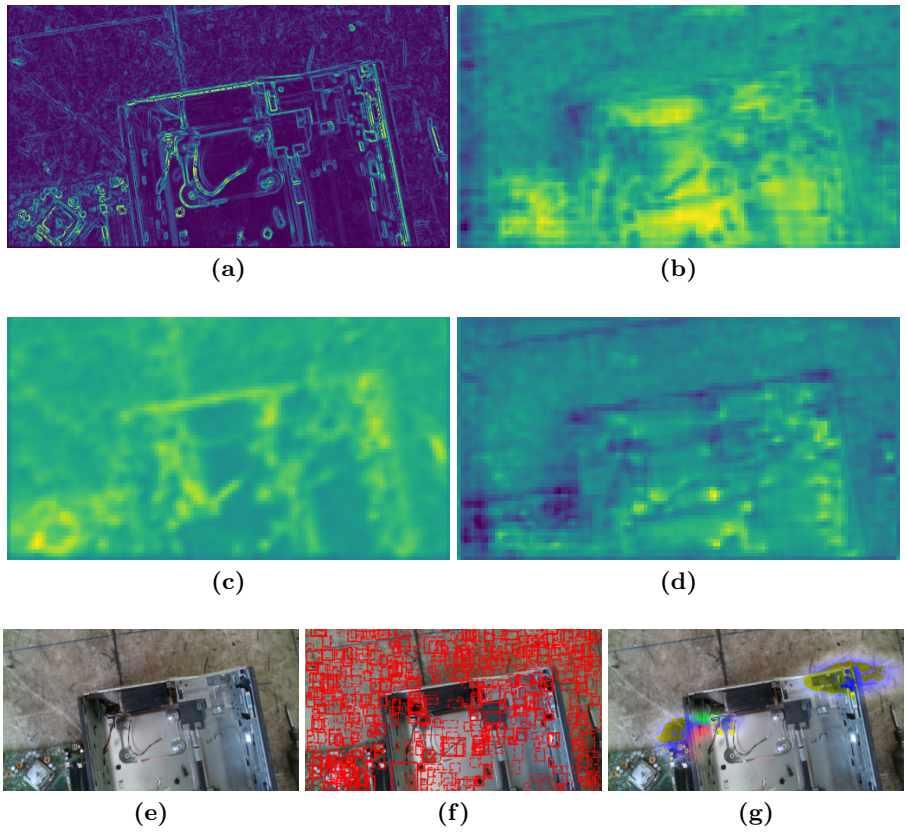

Fig. 1: **(a)** Gabor Feature. **(b)** F_s. **(c)** Gabor-driven Soft Attention. **(d)** Soft Attention gain **(e)** Input Image. **(f)** Edge Anchors. **(g)** Affordance Mask.

3.3 Background

Mask R-CNN: In cases where heavy overlapping between the relevant objects occurs, detecting them by their bounding boxes would result in high ambiguity thus instance segmentation is deemed to be more appropriate. For that purpose Mask R-CNN [4] was chosen as a base architecture for both its state of the art performance and its efficiency. Besides that, Mask R-CNN architecture having discrete feature maps for different object sizes through utilizing the Feature Pyramid Network (FPN) as proposed in [7], renders it even more appealing approach in cases where small sized object detection is required.

Cascade R-CNN: The method as described in [1] is applied on the baseline architecture, with the purpose of training higher *quality* classifiers. The term quality of a classifier refers to the IOU threshold, a proposed region needs to exceed to be considered as ground truth target, which the classifier was trained with. The problem with directly increasing the IOU threshold is that only a handful of candidate region would meet such strict IOU criterion resulting in insufficient training samples. In the context of each candidate being refined to better fit its target, the candidate regions are fed through multiple classifiers of increasing quality to progressively increase their IOU before being passed to the following classifier of higher quality.

3.4 Anchoring by SGWs:

EA-CNN$_P$: Although the uniform anchoring has proved to be quite effective in most cases, it is still limited to generating anchors in discrete pixel intervals with fixed shapes and scales resulting in misalignment between the anchors and their respective ground truth targets, which can be crucial in the case of small objects detection. In order to restrict these deviations, additional anchors are generated considering heuristic information which, unlike the densely distributed anchors, are not bounded by any extrinsic hyper parameters; thus tend to better align with the ground truth objects. In figure 3, we depict a candidate-target IoU comparison between the edge and the default anchors. In order to maintain efficiency while still improving in detection quality, only the small-sized heuristically generated regions, referred to as *edge anchors* (figure 1f), are considered; since the minor misalignments the bigger object exhibit have barely any effect on their detection. Inspired by [14], the input image is filtered by a set of simplified

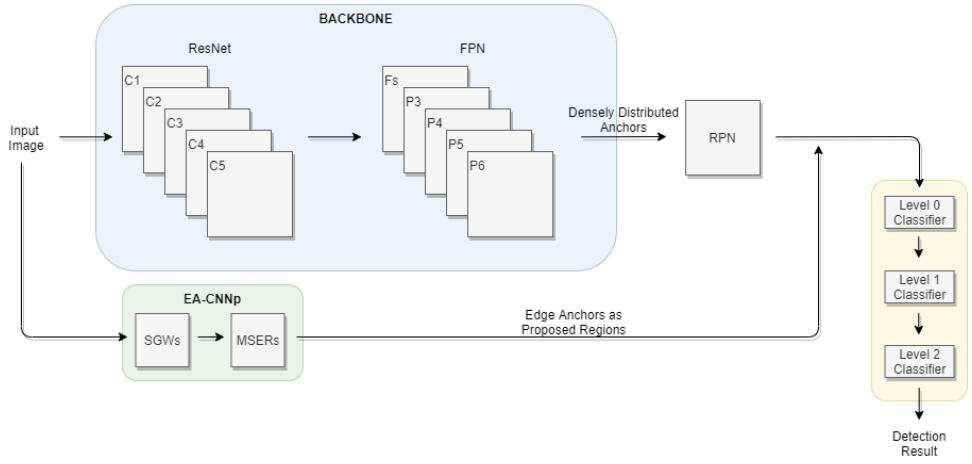

Fig. 2: The architecture combining Cascade with EA-CNN$_P$.

Gabor wavelets (SGWs) producing an edge-enhanced image, termed as *Gabor feature* (figure 1a). The MSER algorithm is then applied on that feature, to extract the edge anchors. Finally in this approach, all edge anchors are merged with the proposed regions and are fed jointly to the classification stage. The architecture described above combined with the cascade R-CNN is shown in figure 2.

Edge anchor IoU: **0.74**
Default anchor IoU: **0.54**

Edge anchor IoU: **0.74**
Default anchor IoU: **0.64**

Fig. 3: Displaying the anchors that align the most with the groud truth target. The green, red and blue regions refer to edge anchors, densely distributed (default) anchors and their corresponding ground truth targets respectively.

EA-CNN$_{\mathbf{C}}$: As a next step, we aim at integrating the edge anchors into the RPN. Due to the former being of varying scale and having continuous center coordinates, some modifications are required so as to be compatible with the RPN training procedure. In order for the FPN feature map to remain scale specific and have the bounding box regressors referring to identical shaped anchors, the edge anchors are refined to match the closest available shape and size configuration, as dictated by the scale and ratio hyper parameters. The issue of edge anchor centers not aligning with the pixel grid is addressed through rounding their centers along with upsampling the feature map with the purpose of lessening the quantization error. Although restricting the edge anchors into predefined shapes and sizes, partially opposes the sense of acquiring the best alignment possible between the candidate regions and the ground truth targets, having scale and shape consistent feature maps is of high importance for regression stability. In order to minimize the refinement edge anchors have to undertake to fit the predefined scale configurations, we introduce additional feature map dedicated to the edge anchors, called *edge* maps. These maps correspond to different scales relevant to small objects and are identical to the feature map of the first FPN level (F_s). After the modifications described above the RPN is able to evaluate regions given both edge and regular anchors as input. Based that on that, MSER is applied to both grayscale input image and its edge-enchanced version resulting

in more but less precise edge anchors. The proposed architecture is shown bellow in figure 4.

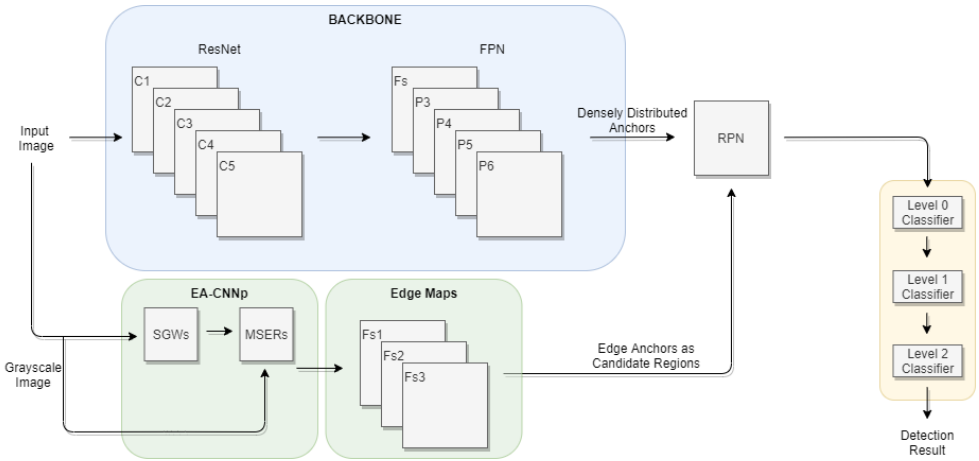

Fig. 4: The architecture combining Cascade with EA-CNN$_{\mathrm{C}}$.

3.5 Attention-based by SGWs:

Attention-SE: In the previously described approach, the Gabor feature ($\mathrm{G_f}$) is used to guide an additional anchoring mechanism targeted at small objects. Although the anchors generated by these methods seem to be reasonable, the pre-

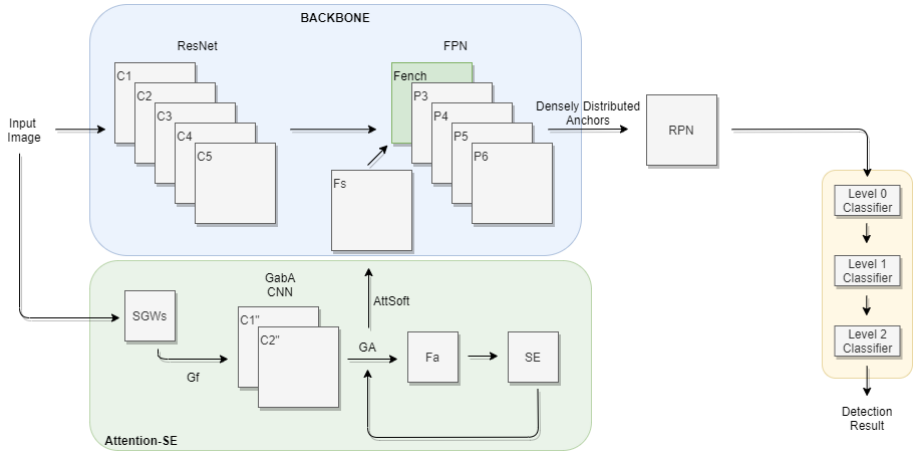

Fig. 5: The architecture combining Cascade with Attention-SE.

processing step required, does not allow for an end-to-end training. Therefore, we propose an architecture in which a soft attention mask $Att_{\text{soft}} \in \mathbb{R}^{\frac{M}{4} \times \frac{N}{4} \times D}$ is generated driven by the G_{f}, where D is a hyperparameter defining the depth of the FPN feature maps while M and N refer to the input image dimensions. More specifically, at first the $G_{\text{f}} \in \mathbb{R}^{M \times N \times 3}$ is passed through a CNN called $GabA$ resulting in $G_{\text{A}} \in \mathbb{R}^{\frac{M}{4} \times \frac{N}{4} \times D}$ as shown in the equation 1. Thereafter based on the the Squeeze and Excitation [5] approach, a depth-wise attention vector $SE \in \mathbb{R}^{1 \times D}$ is constructed by applying average global pooling onto the G_{A} and feeding the result to a fully connected layer $F_{\text{a}} \in \mathbb{R}^{D \times D}$ as shown in the equation 2. Afterwards, the soft attention mask is calculated through depth-wise multiplying the SE with the G_{a} as shown in equation 3 . Finally, based on the equation 4, an enhanced feature map F_{enh} is generated, which will be replacing F_{s} both during region proposing and classification stages. An example on how the Att_{soft} alters the base F_{s} is shown in figure 1d, where the difference between F_{enh} and the F_{s} is visualised. The proposed architecture can be seen in figure 5.

$$G_{\text{A}} = ReLU[GabA(G_{\text{f}})] \tag{1}$$

$$SE = Softmax[F_{\text{a}}(G_{\text{A}})] \tag{2}$$

$$Att_{\text{soft}} = ReLU[SE \otimes G_{\text{A}}] \tag{3}$$

$$F_{enh} = F_s \otimes Att_{\text{soft}} \tag{4}$$

Attention-SE enhanced by Human-Object Interaction (HOI) information: Motivated by the work presented in [15], which achieves increased object recognition performance by exploiting affordance-based knowledge, we incorporate a stream related to that kind of information. More specifically, we enhance the Attention-SE by using the method presented in [10], where potential interaction hotspots are predicted given a static image. A brief description on how their proposed method works is required, in order to better interpret its prediction results. The so-called hotspot prediction is achieved through three discrete stages. At first a typical action recognition model has been trained to predict the various action occurring in video sequences where objects of interest are being manipulated by the human. As a second step, an action anticipation model is trained to map a static image into its corresponding afforded actions. Finally in the third step, given the anticipated afforded action the HOI hotspots are generated by applying a feature visualization technique. An example of HOI hotspot prediction can be seen figure 1g. These HOI masks, as shown in the architecture of figure 6, are passed through a set of vanilla 2D convolutional layers in order to generate the *affordance feature maps*, where the term *affordance* corresponds to the way the human interacts with the object. Finally, while region proposing in based solely on visual information, during the region classification stage, the

ROI-pooling is applied to both regular feature maps corresponding to visual information and the feature maps related to the afforded action. The descriptors generated from these discrete information streams, are concatenated and fed to the classifier. Through this approach classification enhancement is achieved by capitalizing on both visual (sensor) and human-object (motor) information. Finally, the hotspots prediction are not limited in providing spatial information solely, since discrete colorization indicates different ways of human-object manipulations. The described architecture is shown in figure 6.

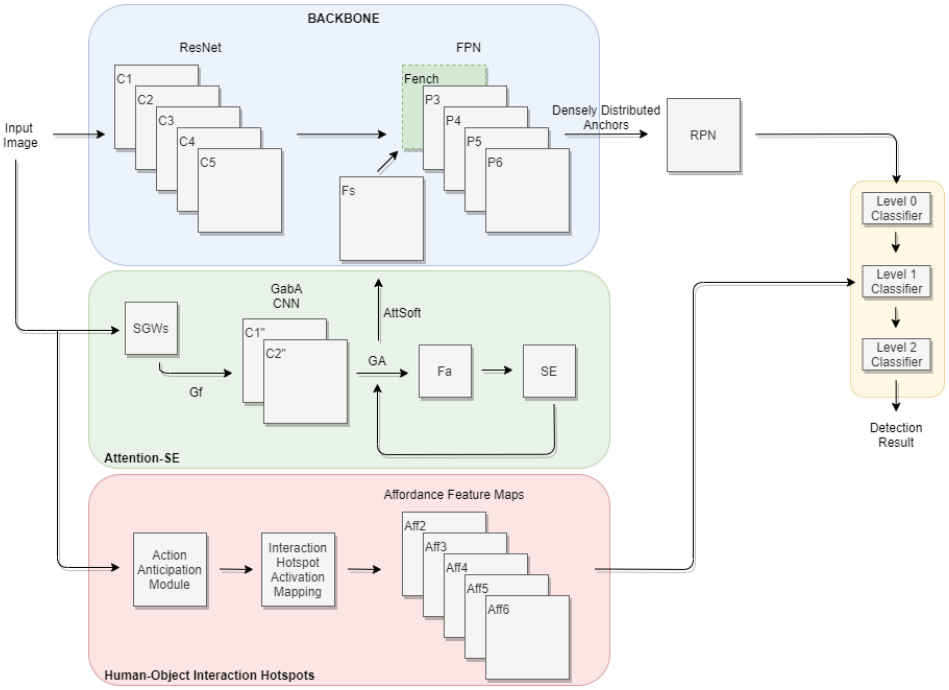

Fig. 6: Attention-SE enhanced by Human-Object Interaction (HOI).

4 Experiments

4.1 Dataset Construction

A set of WEEE disassembly procedures were recorded, in the form of video at the industrial environment. All the recordings were carried out through three cameras, two of which were in fixed position while the other one, being handheld, resulted in more informative views. Thereafter, a number of frames utilizing all three views, were manually annotated aiming at maximizing the variance in terms of unique annotations. A binary mask was defined for each WEEE

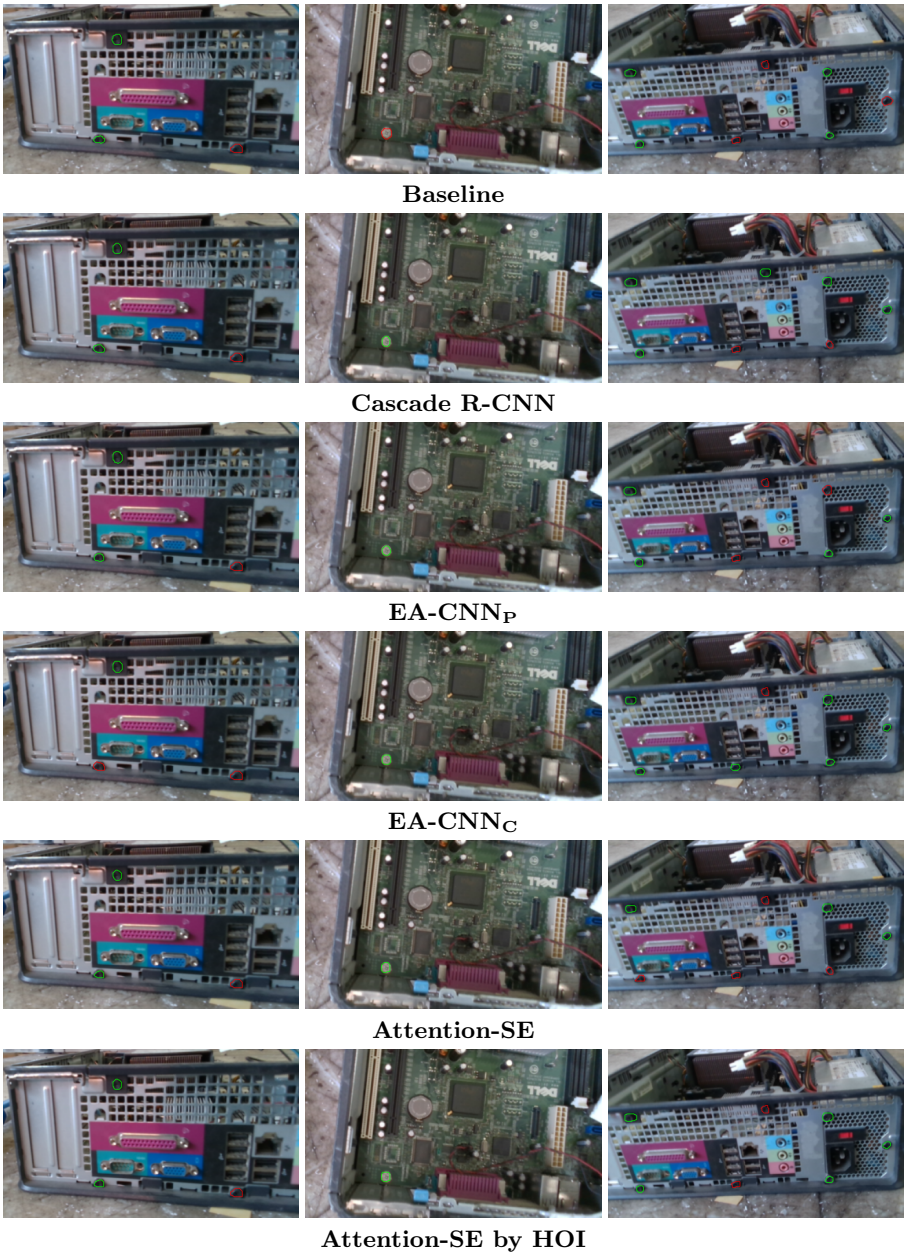

Fig. 7: Examples of screw detection performance of the various approaches. The green and red outlines indicate the successful and the failed screw detection respectively.

component as well as their corresponding WEEE device, distinguishing them from the background environment as well as other instances.

Dataset: Although, our dataset consists of multiple WEEE categories, in this work we focus on PC-Towers explicitly, since it was found to be the most challenging due to the presence of class ambiguities, high occlusions and significant small components. The overall annotated frames referring to the PC-Tower is 395 originated from five unique PC-Tower disassembly procedures. In order to avoid having similar looking frames during training and evaluation, we split the dataset on a procedure basis, that is, the training set is formed by getting the frames only from four disassembly session and leaving the remaining one for evaluation purposes. The dataset split as described previously, results in 325 and 70 of fully annotated frames in the training and validation sets respectively. Moreover, there are 23 unique components-categories with all of them being present in the training set while only 16 of them in validation set.

4.2 Evaluation metrics

To evaluate the performance of each of the modalities added, we report the standard COCO mean Average Precision (mAP) $AP_{0.5:0.95}$,$AP_{0.5}$, $AP_{0.75}$, and mean Average Recall (mAR) $AR_{0.5:0.95}$ metrics. Moreover, since we are aiming at small object detection specifically, we also report the metrics of AP_S and AR_S averaged over the $[0.5 : 0.95]$ IoU range thresholds.

4.3 Implementation Details

In all implementations, the ResNet-101 backbone was used for feature extraction combined with the FPN neck. The network's weights were initialized using a model pretrained on MS COCO dataset. The input images were resized such that they biggest dimension is 512 pixel wide, while their aspect ratio is retained. The Stochastic gradient descent (SDG) optimizer was used with a momentum value of 0.9. The model was trained for 400 epochs, using a mini batch size of one image, with an initial learning rate of 0.001 decayed by a factor of 3 at epoch 100 and 250. During the first 100 epochs the backbone layers were kept frozen while the whole network was trained thereafter. Non Maximum Suppression (NMS) was applied both during the proposing and classification stages with thresholds of 0.9 and 0.3 respectively. Regarding the uniform distributed anchors, they are generated using ratios of 0.5, 1, and 2 while their scales were set to $[10, 32, 64, 128, 256]$ targeting objects of various sizes. The RPN was set to propose 2000 candidate regions at training, while during testing 1000 regions are proposed and only the 100 most confident predictions are kept. Finally, data augmentation was applied to address the relatively small dataset, by applying random rotation as well as adding random motion blur noise during training.

Gabor-based Anchoring: Finally, for the purpose of generating the *edge* anchors, the regions produced by the MSER algorithm occupying an area larger than 15^{15} pixels or having their aspect ratio fall outside the $[0.5, 2]$ interval, are filtered out. Moreover, when the edge anchors are treated as candidate regions additional scales of 8,15 and 25 pixels are introduced.

Human-Object Interaction information: The visual information is encoded into a descriptor of 256 length while the ones referring to based on the Human-Object Interaction information have a length of 64.

Table 1: Object detection results on PC-Tower dataset.

Method	**Cascade**	$\boldsymbol{AP_{0.5:0.95}}$	$\boldsymbol{AP_{0.5}}$	$\boldsymbol{AP_{0.75}}$	$\boldsymbol{AP_S}$	$\boldsymbol{AR_{0.5:0.95}}$	$\boldsymbol{AR_S}$
Baseline	-	40.3	69.8	39.5	23.1	47.3	28.0
	✓	39.5	66.2	40.1	23.6	**48.8**	29.0
EA-CNN$_{\mathrm{P}}$	-	38.7	68.2	38.0	22.9	46.9	28.8
	✓	40.9	**70.5**	42.3	24.6	48.2	30.1
EA-CNN$_{\mathrm{C}}$	-	39.6	69.5	40.1	25.5	47.2	31.4
	✓	40.6	69.5	41.9	25.2	48.4	**32.3**
Attention-SE	-	**41.0**	70.1	41.2	24.0	46.4	31.0
	✓	40.7	69.5	**42.6**	25.0	48.3	28.4
Attention-SE by Affordance	✓	38.9	68.6	38.9	**25.7**	48.7	31.5

4.4 Results

Based on the results presented in Table 1, the best overall $AP_{0.5}$ is achieved when edge anchors are fed directly to the classification stage without any sort of refinement. On the other hand, when more strict IoU thresholds are required, Attention-SE is the most dominant method in terms of mAP. Comparing the methods utilising Gabor-based anchoring, superior small object detection is accomplished when edge anchors are integrated into the RPN module. Regarding the Attention-based approaches, the one capitalizing on the affordance modality performs better at detecting the small objects. Moreover, it is evident that the best performances have been achieved when the cascade architecture is deployed. Additionally, the qualitative displayed in figure 7, referring to screws, are indicative of how each of the proposed method can enhance the baseline in terms of small objects detection quality. Finally, our disassembly dataset consisted of highly variant views, greatly occluded and significantly small components arises challenges that affect the object detection performance. Although, the mAP metrics are relatively low compared to other object detection benchmarks, the detection performance is deemed to be reasonable considering the challenging nature of our dataset.

4.5 Conclusions

In this work we investigated various approaches aiming at boosting the small object detection. At first a set of handcrafted features are generated in order to guide an additional anchoring mechanism targeted specifically at small-sized objects. Applying such anchoring mechanism into the detection pipeline yielded promising results. Additionally, Attention-based approaches were also considered, making use of the previously mentioned features to generate a soft attention mask targeted at small objects. Moreover, build upon the proposed soft attention approach, we further enhance the object detection by taking advantage of an additional stream of information based on Human-Object interaction. Finally our proposed approach has generalization capabilities and is seamlessly applicable to any two-stage object detection approach.

4.6 Future Work

In the future, experiments are to be conducted on the whole disassembly dataset including all four WEEE devices. Moreover, our dataset will be further enriched by annotating additional disassembly sessions on different WEEE disassembly plant premises. Finally, although the authors of [10] state that relatively accurate interaction hotspots can be generated even in cases of unseen object categories, as a future work we consider training their proposed model on the WEEE domain in order to obtain more accurate hotspot interaction prediction and subsequently increase the detection performance.

Acknowledgment

This work was supported by the European Commission under contract H2020-820742 HR-Recycler.

References

1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: High quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
2. Girshick, R.: Fast r-cnn object detection with caffe. Microsoft Research (2015)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
4. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
5. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
6. Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., Yan, S.: Perceptual generative adversarial networks for small object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1222–1230 (2017)

7. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
8. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
9. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
10. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8688–8697 (2019)
11. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
12. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
13. Ren, Y., Zhu, C., Xiao, S.: Small object detection in optical remote sensing images via modified faster r-cnn. Applied Sciences **8**(5), 813 (2018)
14. Shao, F., Wang, X., Meng, F., Zhu, J., Wang, D., Dai, J.: Improved faster r-cnn traffic sign detection based on a second region of interest and highly possible regions proposal network. Sensors **19**(10), 2288 (2019)
15. Thermos, S., Papadopoulos, G.T., Daras, P., Potamianos, G.: Deep affordance-grounded sensorimotor object recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6167–6175 (2017)
16. Wang, J., Chen, K., Yang, S., Loy, C.C., Lin, D.: Region proposal by guided anchoring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2965–2974 (2019)
17. Xu, H., Jiang, C., Liang, X., Lin, L., Li, Z.: Reasoning-rcnn: Unifying adaptive global reasoning into large-scale object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6419–6428 (2019)
18. Zhang, J., Zhang, J., Yu, S.: Hot anchors: A heuristic anchors sampling method in rcnn-based object detection. Sensors **18**(10), 3415 (2018)
19. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: European conference on computer vision. pp. 391–405. Springer (2014)