Deep learning object detection as an assistance system for complex image labeling tasks

Max Leimkühler
German Research Center
for Artificial Intelligence
max.leimkuehler@dfki.de

Laura Sophie Gravemeier
German Research Center
for Artificial Intelligence
laura.gravemeier@dfki.de

Tim Biester
German Research Center
for Artificial Intelligence
tim.biester@dfki.de

Oliver Thomas
Universität Osnabrück
oliver.thomas@uni-osnabrueck.de

Abstract

Object detection via deep learning has many promising areas of application. However, robustness and accuracy of fully automated systems are often insufficient for practical use. Integrating results from Artificial Intelligence (AI) and human intelligence in collaborative settings might bridge the gap between efficiency and accuracy. This study proves increased efficiency when supporting human intelligence through AI without negative impact on effectiveness in a fine-grained car scratch image labeling task. Based on the confirmed benefits of AI with human intelligence in the loop approaches, this contribution discusses potential practical application scenarios and envisions the implementation of assistance systems supported by computer vision.

1. Introduction

Object detection use cases range from support for autonomous driving [1] over medical diagnostics [2] to product quality control [3]. Existing object detection models can also be used for labeling and annotating images in various scenarios. But while some image labeling tasks are easily handled by object detection models on their own and generate adequate results [4], others present great challenges [5]. In case of complex applications, a sufficient robustness cannot be guaranteed, which is a necessary requirement for full automation [6]. For example, this is the case with high intra-class differences which means that objects of one class have very different characteristics. The same is true for low inter-class differences, where objects from different subordinate categories have only marginal visual differences. While typically being repetitive, these fine-grained classification tasks are challenging and time-consuming for humans as well. Therefore, efficiently assisting human labeling activities and harnessing the benefits of human-AI collaboration could prove useful [7, 8].

Consequently, this research study aims at providing further empirical evidence by investigating the use of an object detection model for labeling images where robustness and accuracy are insufficient for full automation. Instead, the results of the object detection model can be utilized as an assistance system for partly automated image labeling with an AI with human intelligence in the loop approach. This paper investigates a specific application scenario where images of car paint scratches from the context of a repair shop are analyzed. When receiving images of the damaged cars from their clients, the service staff must label and count the scratches to forward this documentation to the insurance company. While no adequate detection model exists to successfully automate this task, humans performing this task could still benefit from an assistance system. The concept of using object detection models as an assistance system is applicable to various types of complex labeling tasks. However, the practical relevance of implementing AI with human intelligence in the loop is dependent on its efficiency and effectiveness. This results in the following research question *(RQ)* addressed in this paper: *Can human-AI collaboration increase image labeling efficiency (RQ 1) and effectiveness (RQ 2) by partially automating complex labeling tasks via the results of deep learning object detection?*

An experiment was conducted to investigate this issue, where two sets of subjects were presented with un-labeled and pre-labeled images respectively.

2. Foundations

2.1. Deep learning object detection

Deep learning object detection, along with classification and semantic segmentation, is part of the field of computer vision and is one of the more complex image processing tasks [5, 9, 10]. In deep learning object detection, the task is to localize and classify objects. Objects that can belong to different classes have

URI: https://hdl.handle.net/10125/70650
978-0-9981331-4-0

to be marked on the image with so-called bounding boxes [11]. Large labeled data sets are required to train robust deep learning detection models [12]. This is especially the case for complex use cases where there is a high inter-class difference or a low intra-class difference [5]. As mentioned before, high inter-class difference means that the objects within a class have very different characteristics. This is the case, for example, with scratches. These can be fine scratches as well as large scratched areas with different shapes. Both, for humans and Artificial Intelligence (AI) such a fine-grained detection problem is still a challenge [13]. The same problem occurs when there is a low intra-class difference. This means that the objects of different classes are very similar, for example when different bird species are to be distinguished from each other [5, 14].

As already mentioned, a large number of labelled training images are necessary, especially for complex applications. In many practical scenarios such a large data set is not available and deep learning architectures tend to overfit [15]. In these cases, transfer learning is a suitable method. In transfer learning a deep neural network, that was pre-trained on large data sets with a supervised machine learning approach, can be used and adapted to a specific use case by domain-specific fine-tuning [16]. The new use case may differ significantly from the original use case. Transfer learning offers the advantage that fewer labelled images are needed and the tendency of overfitting is not as distinct [15].

2.2. Human-AI collaboration

AI changes business sustainably. At this point however, AI doesn't show sufficient performance in many use cases when solving problems independently [17, 18, 19]. In these cases, approaches featuring human-computer collaboration are advisable. In human-computer collaboration, at least one person and one computer agent, in our case an AI, work together to achieve a common goal [20]. The idea behind the collaboration is that AI and humans have different abilities and strengths. Where AI is more analytical, consistent, fast, efficient and geared towards pattern recognition as well as probabilistic analysis, humans are more intuitive. Their strengths lie in flexibility, transfer performance, empathy, creativity and common sense [21, 22]. In line with a collective intelligence perspective that underlines the potentials for synergy between AI and human agents [23], studies show that the best results are achieved when AI and humans collaborate. In this context, AI can increase the physical capacity of workers, expand cognitive skills and replace them in low-level tasks. Therefore, human-AI collaboration can change business processes on different levels, such as speed, scalability or decision making [24].

One example of human-AI collaboration is AI with human intelligence in the loop. This approach is typical for business use cases. In these applications, the AI provides decision support in the form of predictions or recommendations. Human feedback is used to reach final decisions either in general or when AI results are inconclusive [25]. Human-AI collaboration settings can contain various forms of interaction between AI and human agent, ranging from verifying yes-no questions [26] over multistep interaction cycles [8, 27] or complex feedback [28]. When applying collaborative approaches of this nature, the efficiency and effectiveness of human decisions can be increased by AI [21, 29], while still profiting from the quality of human decision-making.

One research field that can benefit from implementing AI with human intelligence in the loop is computer vision [7, 27, 28, 30, 31]. E.g. for computer vision classification tasks, it could be shown that for a fine-grained classification problem the collaboration of humans with an AI classification model results in improved accuracy. In addition, less human interaction is required [7, 27]. Leveraging AI in computer vision use cases may be supported by AI-based digital assistance, which exist on the continuum between independent AI decisions and human autonomy. In this context, as of yet unanswered research questions on design guidelines and the acceptance of AI-based digital assistance is posed, while the orientation on the maxim of "ethics-by-design" is demanded [22]. While some research exists on utilizing computer vision approaches in assistance systems [32, 33, 34], there is a need to further investigate the conjunction of computer vision, assistance systems and AI applications with human intelligence in the loop. This paper will therefore provide further evidence on the efficiency and effectiveness of such human-AI collaboration systems in a computer vision use case. Proof of computer-vision-based digital assistants' performance is a fundamental basis for their practical application as well as future scientific contributions on drivers for adoption and acceptance in business use cases.

3. Research approach

In this paper, the research focus regarding complex labeling tasks is operationalized via labeling car scratches. Entailing multiform scratch manifestations and aiming at a precise label for the scratch location, this labeling task is not trivial and object detection model performance not yet sufficient for automation.

3.1. Development of deep learning object detection model for scratch detection

For training a deep learning object detection model to detect scratches tensorflow-gpu was used. As a data base, 2,323 2D images are provided, which contain a total of 6,744 instances of scratches. The images were provided by a medium-sized car repair shop for research purposes. The images were manually labelled by experts and inspected by a researcher. In this context, labeling means that the scratches on the images are marked with so-called bounding boxes. Care was taken to ensure that the data set contains a wide variety of perspectives, exposure conditions, types and sizes of scratches as well as cars. This is to counteract overfitting.

Due to the relatively small data set, a transfer learning approach was applied. The model used was Faster R-CNN ResNet50 pretrained on COCO dataset [35]. In total 249,629 training steps were performed until the total loss converged. Training was done with a constant learning rate of 0.0003 (solver: Momentum).

This object detection model is used in the following experiment to pre-label the images. A prediction is displayed by the model if the forecast probability of scratch detection is greater than or equal to 0.5. The inferences were performed before the experiment started and are only shown in the experiment afterwards. In figure 1 the process is visualized: inputs, outputs and actions are shown. Inputs and outputs are visualized by regular rectangles while actions are visualized by rounded rectangles.

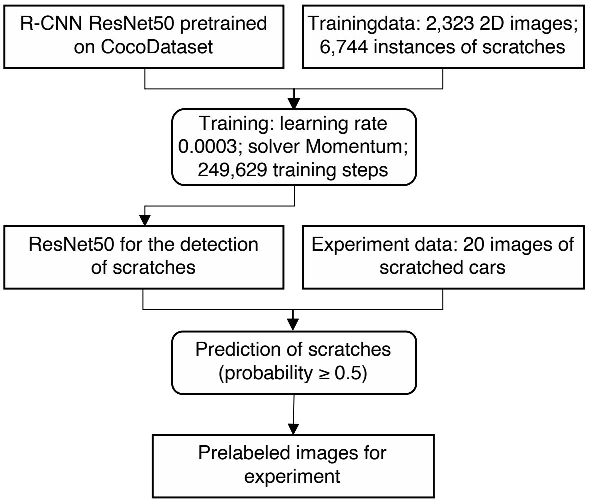

R-CNN ResNet50 pretrained on CocoDataset
Trainingdata: 2,323 2D images; 6,744 instances of scratches
Training: learning rate 0.0003; solver Momentum; 249,629 training steps
ResNet50 for the detection of scratches
Experiment data: 20 images of scratched cars
Prediction of scratches (probability ≥ 0.5)
Prelabeled images for experiment

Figure 1: Process of creating and applying predictive object detection model

3.2. Experiment design

To provide an empirical basis, the overall *RQ* was addressed by an experiment comparing the effectiveness and efficiency of subjects completing the labeling task. While one group was labeling without assistance (group A), the other group was correcting images that were pre-labeled by the deep learning object detection model (group B). Based on the *RQ*, three statistical hypotheses are derived:

H1: The *efficiency* of image labeling tasks increases with a higher degree of assistance, i.e. subjects correcting pre-labeled images (group B) need less time than subjects labeling car scratches without assistance (group A).

H2: The *effectiveness* of image labeling tasks increases with more assistance, i.e.
H2a: Subjects labeling car scratches with assistance (group B) show higher precision than subjects correcting un-labeled images (group A).
H2b: Subjects labeling car scratches with assistance (group B) show higher recall than subjects correcting un-labeled images (group A).

A total of 30 subjects took part in the experiment and the assignment to the groups A and B was random. The average age of the subjects in group A is 26.67 years and in group B 26.47. In Group A, 53% of the participants are female and 47% male. 47% of the participants from group B are female and 53% are male. After a short test phase for acclimatization, all subjects were presented with the same 20 images of car scratches (these pictures were not part of the training data set). The experimental design, shown in table 1, had one group label the scratch position without assistance, while the other group corrected images that were pre-labeled by the object detection model.

Table 1: Experiment overview

Group	*Ex-perimental condition*	*Level of auto-mation*	*Level of human in-volvement*	*n*
A	subjects labeling images without assitance	No auto-mation	High	15
B	subjects correcting pre-labeled images	Partial auto-mation	Medium	15

The subjects were instructed to label all visible scratches with rectangles while minimizing the amount of undamaged marked area. One exemplary labeling task before subjects' intervention is depicted in figure 2 for group A and in figure 3 for group B.

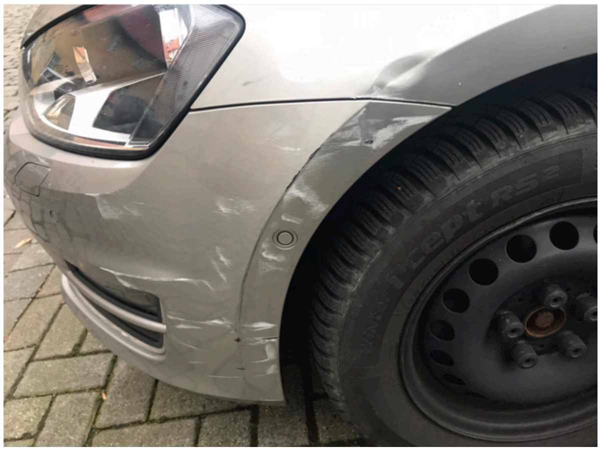


Figure 2: Experiment setup for group A with unlabeled images

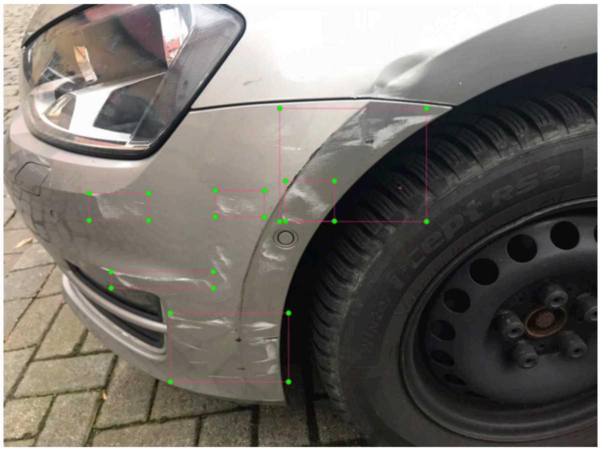


Figure 3: Experiment setup for group B with pre-labeled images

In the experiment the open source graphical image annotation tool LabelImg was used to annotate the images (https://github.com/tzutalin/labelImg). The following steps have to be performed by the test persons to create bounding boxes to mark the scratches. These steps must be repeated for each scratch, both for test persons from group A and B.

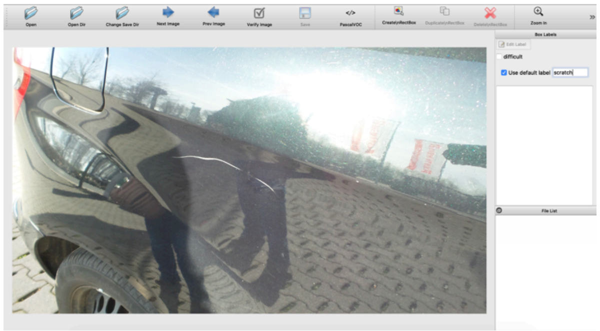
1. Push Button “Create RectBox“
2. Pull up the bounding box with the cursor

Once the labeling process has been completed and the test subjects believe that there are no more scratches to be marked on the image, the following steps have to be performed to proceed to the next image:

1. Push Button “Save“
2. Push Button “Next Image“

In addition, test persons from group B have to review the predictions from the object detection model. If the prediction is seen as correct from the test person, no manual steps have to be performed. If the pre-labeled scratch is considered completely wrong, it can be deleted by the test person using the "Delete RectBox" button. If the annotation is too small or too big, the bounding box has to be changed in size via cursor by the test person. If the bounding box needs to be moved, this can also be done with the cursor.

Before the experiment, test persons of both groups were instructed in the use of the program and had the same 5 training images to learn how to use the program (group B again had predictions of the deep learning object detection model, group A did not). In addition, the test persons were shown concrete examples of scratches and how these must be marked so that the test persons have a consistent understanding of a scratch. Figure 4 shows the graphical interface of LabelImg.



□ difficult

☑ Use default label scratch

Figure 4: Graphical image annotation tool LabelImg (example from Group A)

3.3. Evaluation procedure

In the context of the experiment, efficiency was operationalized by capturing the processing time of the subjects. In order to make the time measurement comparable, the measurement was automated. For this purpose, a python script was written, which monitored the interactions of the test person with LabelImg. After completion of the experiment the exact time needed for the label task was calculated. The time is displayed in decimal minutes. To measure effectiveness, two indicators were considered: how many of the labeled areas were correctly identified scratches (precision) and how many of the existing scratches were correctly identified (recall). A confusion matrix can be set up to calculate precision and recall for each test person.

Table 2: Confusion matrix

		Scratch	*No Scratch*
Actual	*Scratch*	True Positive (TP)	False Negative (FN)
	No Scratch	False Positive (FP)	-

Predicted

Precision and Recall [36] are calculated as described below:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Based on a groundtruth independently validated by two researchers, every reported scratch by the subject is classed as either a match (TP) or mismatch (FP). A TP is identified for each scratch classification that has an intersection over union (IOU) greater or equal to 0.5 [8, 37] with a corresponding label in the groundtruth. If duplicate detections are present, only the detection with the greatest IOU is considered in the evaluation and the duplicates are not taken into account. The remaining supposed scratches identified by the subjects are treated as FP. Respectively, scratches in the groundtruth that are not identified as true positives are treated as FN. This procedure for identifying the quality of the markings of the test persons is based on the procedure for measuring the quality of object detection models described in [9].

4. Results

4.1. Descriptive statistics

The object detection model by itself had a precision of 53,33 % and recall of 42,85 % on our 20 test images, which is deemed insufficient for full automation in active operation. Table 3 shows the confusion matrix, which is the basis for the calculation of precision and recall (double detections were not taken into account, as in the evaluation of the test persons' results; only the detection with the greatest IOU is considered). In 24 cases the object detection model identifies a scratch and was correct with this decision, as the confusion matrix shows. 21 objects were erroneously marked as scratches from the model and in 32 cases the model falsely did not detect scratches.

Table 3: Confusion matrix for predictions of deep learning object detection model

		Scratch	*No Scratch*
Actual	*Scratch*	24	32
	No Scratch	21	-

Predicted

To describe the efficiency of the test persons, time was measured and calculated as described above. Figure 5 shows a histogram to compare the efficiency of group A and B. The figure shows the tendency that on average, subjects of group B needed less time to perform the experiment than group A.

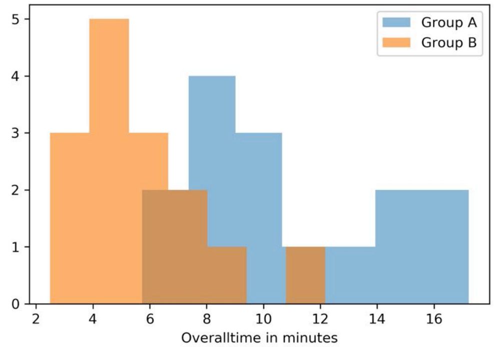


Figure 5: Efficiency of group A and B in comparison

To calculate precision and recall, a confusion matrix was created for each subject. In the next step precision and recall were calculated. In figure 5 the recall of each subject of both group A and group B is shown in a histogram. Recall indicates how many of the scratches were correctly identified. The histogram visualizes that the variance in group A is greater than in group B. This means that in group A there are test persons who recognize only very few scratches while other test persons in group A recognize comparatively many scratches. In group B this variance of the recall measure is less visible.

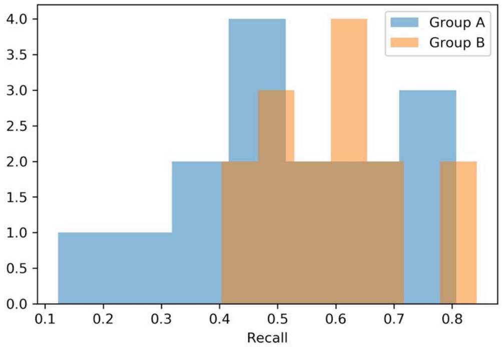


Figure 6: Recall of group A and B in comparison

In figure 7 the precision of each subject of group A as well as group B is shown in a histogram. The precision is a calculated measure that indicates how many of the set labels are actually scratches. It can be seen that the precision in group A has a larger variance than in group B.

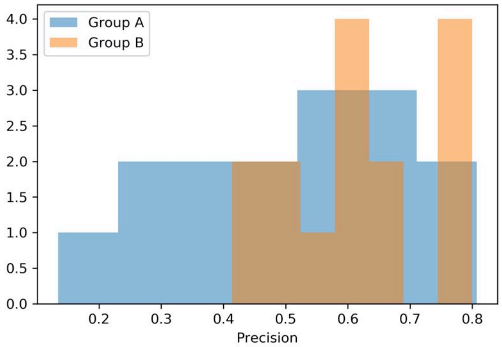


Figure 7: Precision of group A and B in comparison

Table 4 shows the results of the experiment, comparing the descriptive statistics for model performance and both groups. It turns out that on average, group A (only human interaction) took 5.29 minutes longer to label the images than group B (AI with human intelligence in the loop). Since the workflow can be fully automated when using an object detection model, it is not taken into account when evaluating efficiency.

Table 4: Results of experiment

	Object detection model	*Group A*	*Group B*
Mean overall time	-	10,79 minutes	5,50 minutes
Mean precision	53,33 %	52,67 %	61,48 %
Mean recall	42,86 %	51,22 %	59,48 %

Based on the descriptive data, Group B has higher scores in effectiveness, i.e. precision and recall, than Group A and the object detection model. On average, precision is 8.81 percentage points higher and recall is 8.26 percentage points higher for Group B in comparison to Group A. Likewise when compared to the model performance, Group B has higher scores in effectiveness in terms of both precision (+8.15 percentage points) and recall (+16.62 percentage points). In contrast, the descriptive data from group A shows relatively little difference in precision between the object detection model and human performance without assistance (-0,66 percentage points). Recall from group A increases in comparison to object detection model (+8,36 percantage points).

4.2. Statistical analysis of the hypotheses

As noted above, the descriptive data shows a lower overall time in the group working with pre-labeled images. A statistically significant difference between the groups was found both by using a mann-whitney-u-statistic ($U = 20$, $p < .001$) and a one-sided unpaired t-statistic (t (28) = 4.46, $p < .001$, $d = 1.63$). Subjects working on pre-labeled images are faster with a processing time of at least 3.27 minutes less (95%-CI[3.27, inf]). The Hypothesis *H1* can therefore be accepted.

H2, on the other hand, has to be rejected. Comparing the precision between both groups (*H2a*), a significant difference was found neither by using a two-sided unpaired t-statistic (t (28) = -1.53, $p = .14$) nor a mann-whitney-u-statistic ($U = 84.5$, $p = .13$). The same applies to recall (*H2b*), where neither t-test (t (28) = -1.38, $p = .18$) nor mann-whitney-u-test were significant. Even though descriptive data shows higher scores in precision and recall for the annotation of pre-labeled images, this is not proven as a statistically significant difference.

5. Discussion

Our empirical findings confirm results from previous research: efficiency is significantly increased by using an AI with human intelligence in the loop approach compared to no automation (*RQ 1*). On average, the processing time in this use case was cut in half, which means an improvement on a practically relevant scale. At the same time, while not increasing along with efficiency as expected (*RQ 2*), effectiveness is not negatively affected. Therefore, supplementing the manual labeling task with an AI with human intelligence in the loop approach proves beneficial in this use case. In the following, other practical application scenarios are discussed that should be further investigated in future research.

5.1. Limitations and observations

First, the robustness of the evaluation metric measuring efficiency should be examined further, as it can lead to blurring when evaluating the marking of scratches. Because in some cases, scratches can be interpreted in different ways, the use of the IOU metric and the IOU-dependent measures precision and recall can be problematic. For example, using the objective, technical calculation of whether a match is present or not based on the established groundtruth may lead to a different result than a subjective interpretation of the same marking based on a visual inspection. In figure 8 the context is visualized. While interpretation 1 would lead to a TP, interpretation 2 would lead to one TP (because one bounding box has an IOU greater than 50 %), but one FP at the same time. Therefore, interpretation 1 leads to precision of 1.0 and recall of 1.0, while interpretation 2 leads to precision of only 0.5 and recall of 1.0. Subjectively, however, a differentiation of the accuracy between the two interpretations is difficult and both solutions could be accepted as correct.

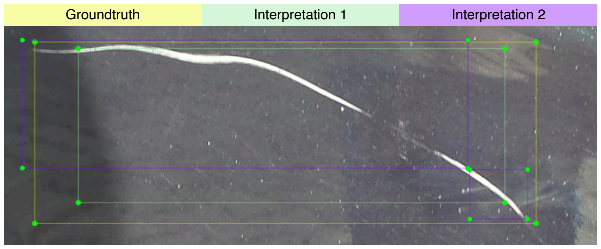


Figure 8: Different interpretations of a scratch

Nevertheless, it should be noted that the same approach was applied to the evaluation of all groups (group A and B as well as the object detection model), thus ensuring comparability and not limiting the basic conclusions in terms of their implications.

Furthermore, the choice of the IOU threshold must be discussed. In the literature, the common value for the IOU is 50 % [8, 37]. In order to further validate our statements, we investigated how a change in the IOU affects precision and recall. From a descriptive point of view, group B performs better in terms of precision and recall when the IOU is small to medium (IOU < 75 %). For an IOU greater than 75 % group A performs better in terms of precision and recall (see figure 9 and figure 10). In other words, if an exact localization is needed, humans without assistance might perform better at that specific task. However, these results have not been confirmed by statistical testing and can only serve as hypotheses for further research projects.

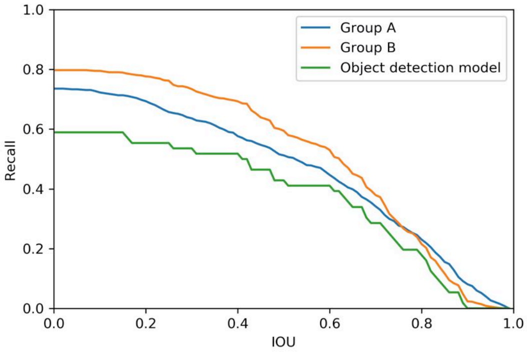


Figure 9: Recall-IOU curve

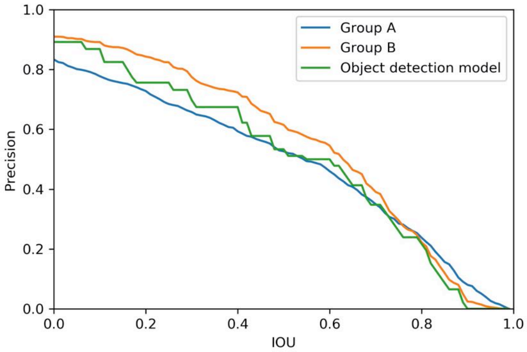


Figure 10: Precision-IOU curve

The second limitation affects both efficiency and effectiveness. During the execution of the experiment, two different ways in which test subjects were handling the task and interacting with the labeling program could be observed. While one group of subjects took a lot of time in annotating the scratches and operated the program carefully, the other group of test persons performed the task relatively quickly and operated the program with greater confidence. But since the assignment of subjects to group A and group B was randomized, the described different types of subjects (slow and cautious vs. fast and confident) were present in both groups and should not affect the results. However, it should be noted that the subjects were relatively young. As age might be related to technical affinity, this might have skewed the results. To investigate the influence of age on labeling performance with and without pre-labeled images, further evidence would be necessary. As the compared groups were randomized and similar in age distribution however, the observed data can still be interpreted meaningfully.

5.2. Implications

Three concrete implications are discussed in this chapter.

The first implication is the **transfer and generalization** of our results. Using the scratch detection as an exemplary use case, we demonstrated increased efficiency with unchanged effectiveness of an AI with human intelligence in the loop system for complex image labeling tasks. In all conceivable applications where a fine-grained image analysis problem exists and both the class and the position of objects within an image must be determined, this approach could mean an improvement in efficiency compared to humans completing the task without any automation. Possible application scenarios are assistance systems in medicine. In order to use computer vision object detection models on patients, extremely high robustness is required [38]. This is where the approach presented here can be useful, as it has been shown that there is no significant decrease in the system's effectiveness. The medical sector can benefit greatly from supplementing expert decisions with computer-vision-based assistance systems, as full automation in this domain is only conceivable for especially high levels of robustness.

Another example of AI with human intelligence in the loop can be found in [39] and is also from the health sector. When classifying X-ray-images with respect to establishing a diagnosis, it is shown that humans can learn from being in the loop and reconsider their decision. Even if in this application no object detection model but a classification model is used, the idea that humans can learn from the human-AI collaboration can be transferred to our presented AI with human intelligence in the loop system. By the additional indication of the location, the decision of the classification seems even more explainable. This makes the decision more comprehensible for the user of such a clinical decision support system, which is an important success factor for systems of this kind [40]. This is the second implication that is to be discussed: **human learning from being in the loop**. It is conceivable that in the long run, with the right suggestions from the AI, people will recognize more scratches through the repetitive task and learn from the AI in this sense or reconsider their decisions.

This mechanism should also be analyzed in reverse: how can AI learn from humans? The third implication regards **human intelligence with AI in the loop systems**. In a so-called interactive machine learning approach, the model is continuously improved through direct feedback from the user. Interactive machine learning enables the user to interact with the training process [41, 42]. In [43] for example, an interactive learning approach was successfully implemented using the example of semantic segmentation. With this approach the user can focus on the fast creation of training data. Since the major effort when developing deep learning-based computer vision is in labeling, it seems promising to use our demonstrated approach to label images for an interactive machine learning approach. This can be a way to develop robust deep learning object detection systems when large amounts of labeled images are needed but not available. In this scenario, an object detection model can be trained on few labeled images and then be used in an AI with human intelligence in the loop system to collaboratively label new images to increase the training data. This can be a solution to develop a customized object detection model more efficiently in one or more loops.

6. Conclusion and future research

This contribution gives some empirical evidence of the benefits from human-AI collaboration. In this specific use case of fine-grained image labeling, an AI with human intelligence in the loop approach reduces processing time with no negative implications for task effectiveness. The transferability of these results to other scenarios should be investigated. A wide variety of computer vision applications with insufficient performance of object detection models from different domains can be taken into consideration. As mentioned in the discussion, the methodological approach regarding the robustness of metrics and IOU calculation could be further refined.

First and foremost, however, the integration of computer vision applications with human-AI collaboration into assistance systems in real-world corporate contexts needs to be examined. Not only is the overall complexity of tasks and context information increased in practical applications, but user acceptance and system usability are key adoption factors. It would be interesting to investigate to what extent the use of the assistance system influences the user in his decision, e.g. for repetitive tasks and occurring fatigue. Therefore, a transfer of this approach into practical use cases is intended. To this end, the perspective of potential future users should be consulted in order to derive recommendations for action and implementation.

7. References

[1] Simhambhatla, R., K. Okiah, S. Kuchkula, and R. Slater, "Self-Driving Cars: Evaluation of Deep Learning Techniques for Object Detection in Different Driving Conditions", *SMU Data Science Review 2*(1), 2019.

[2] Litjens, G., T. Kooi, B.E. Bejnordi, et al., "A Survey on Deep Learning in Medical Image Analysis", *Medical Image Analysis 42*, 2017, pp. 60–88.

[3] Yang, J., S. Li, Z. Wang, and G. Yang, "Real-Time Tiny Part Defect Detection System in Manufacturing Using Deep Learning", *IEEE Access 7*, 2019, pp. 89278–89291.

[4] LeCun, Y., Y. Bengio, and G. Hinton, "Deep learning", *Nature 521*(7553), 2015, pp. 436–444.

[5] Zhao, B., J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation", *International Journal of Automation and Computing 14*(2), 2017, pp. 119–135.

[6] Wei, X.-S., J. Wu, and Q. Cui, "Deep Learning for Fine-Grained Image Analysis: A Survey", *arXiv:1907.03069 [cs]*, 2019.

[7] Branson, S., C. Wah, F. Schroff, et al., "Visual Recognition with Humans in the Loop", In K. Daniilidis, P. Maragos and N. Paragios, eds., *Computer Vision – ECCV 2010*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, 438–451.

[8] Russakovsky, O., L.-J. Li, and L. Fei-Fei, "Best of both worlds: Human-machine collaboration for object annotation", *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE (2015), 2121–2131.

[9] Liu, L., W. Ouyang, X. Wang, et al., "Deep Learning for Generic Object Detection: A Survey", *International Journal of Computer Vision 128*(2), 2020, pp. 261–318.

[10] Garcia-Garcia, A., S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A Review on Deep Learning Techniques Applied to Semantic Segmentation", *arXiv:1704.06857 [cs]*, 2017.

[11] Guo, Y., Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M.S. Lew, "Deep learning for visual understanding: A review", *Neurocomputing 187*, 2016, pp. 27–48.

[12] Simonyan, K., and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *arXiv:1409.1556 [cs]*, 2015.

[13] Yang, S., L. Bo, J. Wang, and L.G. Shapiro, "Unsupervised Template Learning for Fine-Grained Object Recognition", In F. Pereira, C.J.C. Burges, L. Bottou and K.Q. Weinberger, eds., *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, 3122–3130.

[14] Zhang, N., J. Donahue, R. Girshick, and T. Darrell, "Part-Based R-CNNs for Fine-Grained Category Detection", In D. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars, eds., *Computer Vision – ECCV 2014*. Springer International Publishing, Cham, 2014, 834–849.

[15] Donahue, J., Y. Jia, O. Vinyals, et al., "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", *arXiv:1310.1531 [cs]*, 2013.

[16] Girshick, R., J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", *arXiv:1311.2524 [cs]*, 2014.

[17] Dellermann, D., A. Calma, N. Lipusch, T. Weber, S. Weigel, and P. Ebel, "The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems", (2019).

[18] Zheng, N., Z. Liu, P. Ren, et al., "Hybrid-augmented intelligence: collaboration and cognition", *Frontiers of Information Technology & Electronic Engineering 18*(2), 2017, pp. 153–179.

[19] Kamar, E., "Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence", *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, AAAI Press (2016), 4070–4073.

[20] Terveen, L.G., "Overview of human-computer collaboration", *Knowledge-Based Systems 8*(2–3), 1995, pp. 67–81.

[21] Dellermann, D., P. Ebel, M. Söllner, and J.M. Leimeister, "Hybrid Intelligence", *Business & Information Systems Engineering 61*(5), 2019, pp. 637–643.

[22] Maedche, A., C. Legner, A. Benlian, et al., "AI-Based Digital Assistants: Opportunities, Threats, and Research Perspectives", *Business & Information Systems Engineering 61*(4), 2019, pp. 535–544.

[23] Peeters, M.M.M., J. van Diggelen, K. van den Bosch, et al., "Hybrid collective intelligence in a human–AI society", *AI & SOCIETY*, 2020.

[24] Wilson, J., and P. Daugherty, "Collaborative Intelligence: Humans and AI Are Joining Forces", *Harvard Business Review*, 2018.

[25] Holzinger, A., "Interactive machine learning for health informatics: when do we need the human-in-the-loop?", *Brain Informatics 3*(2), 2016, pp. 119–131.

[26] Batchelor, O., and R. Green, "Object detection for Verification Based Annotation", *2019 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, IEEE (2019), 1–6.

[27] Wah, C., S. Branson, P. Perona, and S. Belongie, "Multiclass recognition and part localization with humans in the loop", *2011 International Conference on Computer Vision*, IEEE (2011), 2524–2531.

[28] Parkash, A., and D. Parikh, "Attributes for Classifier Feedback", In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato and C. Schmid, eds., *Computer Vision – ECCV 2012*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, 354–368.

[29] Agrawal, A., J. Gans, and A. Goldfarb, *Prediction machines: the simple economics of artificial intelligence*, Harvard Business Review Press, Boston, Massachusetts, 2018.

[30] Deng, J., J. Krause, and L. Fei-Fei, "Fine-Grained Crowdsourcing for Fine-Grained Recognition", *2013 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE (2013), 580–587.

[31] Wah, C., G.V. Horn, S. Branson, S. Maji, P. Perona, and S. Belongie, "Similarity Comparisons for Interactive Fine-

Grained Categorization", *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE (2014), 859–866.

[32] Gao, J., Y. Yang, P. Lin, and D.S. Park, "Computer Vision in Healthcare Applications", *Journal of Healthcare Engineering 2018*, 2018, pp. 1–4.

[33] Neges, M., S. Adwernat, M. Wolf, and M. Abramovici, "3D Geometry Recognition for a PMI-Based Mixed Reality Assistant System in Prototype Construction", In A. Burduk, E. Chlebus, T. Nowakowski and A. Tubis, eds., *Intelligent Systems in Production Engineering and Maintenance*. Springer International Publishing, Cham, 2019, 3–11.

[34] Nguyen, V., H. Kim, S. Jun, and K. Boo, "A Study on Real-Time Detection Method of Lane and Vehicle for Lane Change Assistant System Using Vision System on Highway", *Engineering Science and Technology, an International Journal 21*(5), 2018, pp. 822–833.

[35] Lin, T.-Y., M. Maire, S. Belongie, et al., "Microsoft COCO: Common Objects in Context", *arXiv:1405.0312 [cs]*, 2015.

[36] Davis, J., and M. Goadrich, "The relationship between Precision-Recall and ROC curves", *Proceedings of the 23rd international conference on Machine learning - ICML '06*, ACM Press (2006), 233–240.

[37] Everingham, M., L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge", *International Journal of Computer Vision 88*(2), 2010, pp. 303–338.

[38] Yamashita, R., M. Nishio, R.K.G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology", *Insights into Imaging 9*(4), 2018, pp. 611–629.

[39] Abdel-Karim, B.M., N. Pfeuffer, G. Rohde, and O. Hinz, "How and What Can Humans Learn from Being in the Loop?: Invoking Contradiction Learning as a Measure to Make Humans Smarter", *KI - Künstliche Intelligenz 34*(2), 2020, pp. 199–207.

[40] Bussone, A., S. Stumpf, and D. O'Sullivan, "The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems", *2015 International Conference on Healthcare Informatics*, IEEE (2015), 160–169.

[41] Amershi, S., M. Cakmak, W.B. Knox, and T. Kulesza, "Power to the People: The Role of Humans in Interactive Machine Learning", *AI Magazine 35*(4), 2014, pp. 105.

[42] Stumpf, S., V. Rajaram, L. Li, et al., "Toward harnessing user feedback for machine learning", *Proceedings of the 12th international conference on Intelligent user interfaces - IUI '07*, ACM Press (2007), 82.

[43] Fails, J.A., and D.R. Olsen, "Interactive machine learning", *Proceedings of the 8th international conference on Intelligent user interfaces - IUI '03*, ACM Press (2003), 39.