Human-AI Collaboration for Improving the Identification of Cars for Autonomous Driving

Edwin Gamboa[1,*], Alejandro Libreros[1], Matthias Hirth[1] and Dan Dubiner[2]

[1]User-centric Analysis of Multimedia Data Group, TU Ilmenau, Ehrenbergstraße 29, Ilmenau, 98693, Germany
[2]Scalehub Gmbh, Heidbergstraße 100, Norderstedt, 22846, Germany

Abstract

Large and high-curated training data is required for Artificial Intelligence (AI) models to perform robustly and reliably. However, training data is scarce since its production normally requires manual expert annotation, which limits scalability. Crowdsourced micro-tasking can help to overcome this challenge, as it offers access to a global workforce that might enable high-scalable annotation of visual data in a cost-time effective way. Therefore, we aim to develop a workflow based on Human-AI collaboration that shall enable large-scale annotations of image data for autonomous driving systems. In this paper, we present the first steps towards this goal, in particular, a Human-AI approach for identifying cars. We assess the feasibility of this collaboration via three scenarios, each one representing different traffic and weather conditions. We found that crowdworkers improved the AI's work by identifying more than 40% of the missing cars. Crowdworkers' contribution was key in challenging situations in which identifying a car depended on context.

Keywords

Human-AI collaboration, Crowdsourcing, Micro-tasking, Autonomous driving, Anonymous annotation

1. Introduction

Autonomous driving is one of the most promising approaches to support smart mobility by reducing the associated risks of human behavior and driving fatigue [1]. Key enablers for autonomous driving systems are sets of sensors installed in the vehicle to monitor the vehicle's environment. Then, prediction and estimation models use the sensor data to understand the current driving situation and decide upon appropriate actions. The models must be highly accurate and have low processing time to minimize the risks of threatening road actors' lives [2]. Supervised learning outperforms classical identification algorithms in this field of application [3]. However, a supervised identification model needs large amounts of training data to later identify objects in a robust, accurate, and reliable way. A high-accurate model for identifying objects in the street must consider different scenarios such as rain, sun, sunset, night, and seasons, and each of them with particular settings related to, e.g., luminosity and reflectance. Still, the availability of public, accurate, reliable, and, especially, massive data sets is scarce for particular objectives, and existing data sets do not meet high-scale purposes, therefore, learning from those data is difficult [4]. Hence, machine learning models perform poorly in high-scale cases, leading to severe limitations that make object identification for autonomous driving still an open problem [5].

In this paper, we present our first steps towards a Human-AI collaboration to enable fast and highly reliable labeling of camera images in the context of autonomous driving. We find that the image data and the required labels exhibit domain-specific challenges, and we illustrate how to consider these challenges in the design of the crowdsourcing workflow. An AI model supports the crowdworkers with pre-annotations of the images to reduce their workload and cope with a large amount of data. The workflow is evaluated in a user study with crowdworkers who annotated almost 400 real-world images. Our results show that the workflow combines the strengths of automated pre-annotation and manual human refinement using scalable, public micro-tasking.

2. Related Work

For the past decade, attempts have been made to explore better ways to combine human-computer approaches to optimize image annotation [6]. Cheng et al. [7] have classified automatic image annotation as generative model-based, nearest neighbor-based, discriminative, tag completion-based, and deep learning-based. One common limitation of automatic image annotation is that those methods suppose availability of annotations, i.e., they address the problem of having different probability

HIL-DC2022: ACM CIKM 2022 Workshop Human-In-The-Loop Data Curation, October 22, 2022, Atlanta, Georgia
*Corresponding author.
edwin.gamboa@tu-ilmenau.de (E. Gamboa);
jose.libreros@tu-ilmenau.de (A. Libreros);
matthias.hirth@tu-ilmenau.de (M. Hirth);
dan.dubiner@scalehub.com (D. Dubiner)
0000-0002-5037-5279 (E. Gamboa); 0000-0002-5434-5464 (A. Libreros); 0000-0002-1359-363X (M. Hirth); 0000-0001-8077-0387 (D. Dubiner)

(a) Daylight city. (b) Nightlight city. (c) Rainy highway.

Figure 1: Sample images of the investigated scenarios.

distributions in the confidence of a set of identifications. Despite recent advances, the lack of trustworthiness of machine learning models has been shown [8]. Thus, the problem of retrieving missing objects is still open. To address this gap, manual annotations have been used, but this approach is limited for scalability purposes due to the scarce availability of experts. In this context, crowdsourcing has the potential to enable high-scalable annotations and produce reliable training data for AI models [9, 10, 8]. Heim [11] presents a cost-time analysis of manual segmentation for organs with experts and crowdworkers. Results show that domain experts achieved approximately 0.1 segmentations per hour vs. 35 segmentations from crowdworkers during the same time. Similarly, different works have employed crowdsourcing for the annotation of large datasets [12, 13]. Also, Boorboor et al. [14] showed how quality can be maximized in the case of lung nodule detection, and Hu et al. [8] have demonstrated that crowdsourcing might reduce the identification bias in challenging scenes. Nevertheless, crowdsourced micro-tasking implies challenges related to the variance in annotation quality, which is mainly related to the workers' lack of domain knowledge [9, 11]. Thus, a collaboration between AI and crowdsourcing might be feasible for addressing these issues as demonstrated in the medical field. However, to the best of our knowledge, this collaboration has not been studied in the context of autonomous driving considering different driving and weather scenarios.

3. Problem Statement

One of the main problems with the annotation of images for autonomous driving is the diversity of scenarios that may emerge. The driving situation can be highly different depending on the street environment, i.e., a highway or a narrow street inside a city, and vary in terms of, e.g., available driving space, number and type of other road users, available signs, and traffic lights. Additionally, numerous environmental factors such as lighting and weather conditions have to be considered. Considering this high diversity of scenarios, it seems likely that there are cases in which an AI delivers better results than crowdsourcing workers and vice versa. In the following, we will show this with concrete examples and illustrate the advantages of collaboration between AI and crowdworkers in this use case. We employ three self-collected videos representing different, typical street scenarios to assess the performance of the collaboration. A sample frame of each video is shown in Figure 1. First, a *Daylight city* video (Figure 1a), in which light conditions are ideal, but the image contains a lot of objects typical of a big city. Second, a *Nightlight city* video (Figure 1b) of a small city, in which light conditions are most challenging. Lastly, a *Rainy highway* video (Figure 1c), in which traffic is smooth, crowds of cars are infrequent, but the visual quality is affected by the rain. We randomly selected 399 frames, 133 from the daylight video, 133 from the rainy highway, and 133 from the nightlight video for our evaluation.

4. Study Design

This section presents the design process of the annotation task, the steps that crowdworkers performed when accessing it, and the process to evaluate the Human-AI collaboration.

4.1. Task Design

Fully annotating a video in the context of autonomous driving is rather complex, since such a task requires annotating different objects, e.g., cars, pedestrians, traffic signs, and other obstacles, frame by frame. Our first goal is to identify the main challenges of the annotation task itself and address the multi-object annotation problem later. Thus, we initially concentrate on the annotation of cars only. This annotation process can be further decomposed into a three-steps task, i.e., (1) Crowdworkers identify cars not detected by the AI, (2) crowdworkers identify wrong AI- and crowd-based annotations, and (3) crowdworkers fix the wrong annotations.

In this paper, we focus on the first step. We decided to request crowdworkers to use bounding boxing for the annotation instead of other methods like polygon enclosing, or free drawing to reduce workload. Other, more sophisticated, techniques like marking background/foreground via simple clicks were discarded since it might lead to high heterogeneity in the results [9]. We decided to use YOLOv3 [15] for the pre-annotation of the images since it has demonstrated high performance for traffic contexts with low computational cost. Also, YOLO tends to predict fewer false positives than other state-of-the-art object identification architectures like R-CNN, using pre-trained models [16].

We designed the task's instructions following guidelines for crowdsourcing and usable texts. We used illustrated instructions minimizing visual complexity [17], together with short sentences using simple English [18, 19, 13]. Also, we included examples of wrong and right annotations [11, 17]. The instructions and the User Interface (UI) annotation mechanisms were iteratively improved using the Crowdsourced Thinking Aloud Protocol method as proposed in [20].

4.2. Task Procedure

Training. As recommended by different works [18, 9], training tasks should be included to bring crowdworkers closer to the task domain and filter unreliable workers out. In particular, gold standard data can be used in which different complexity cases are trained.

In the training task, we show crowdworkers three randomly selected images, with different complexity levels. The complexity levels depended on the number of cars to be annotated, the amount of AI annotations, and the presence of cars that are hard to identify, e.g., very distanced or partially visible cars. Each training task includes additional hints relevant to the current frame and based on the workers' performance, e.g., highlighting missing cars after each try until all expected cars are annotated. Once the training task is successfully passed, crowdworkers can complete the annotation task. Quick instructions are visible during the whole process and crowdworkers can go back to the detailed instructions anytime they want.

Main Task Crowdworkers have to annotate five, randomly selected frames. We asked them to draw boxes around that the system, i.e., YOLO, did not find. To make the completion criteria clear, we ask them to annotate a maximum of 10 cars. To annotate only relevant cars in each frame, the crowdworkers should consider the following conditions: (1) The box should contain a car and fit its size. (2) Each box should contain only one car. (3) The box should contain a big enough car, i.e., the car's height is greater than 5% of the frame height.

When no cars are found, the worker can continue to the next frame. Annotated boxes that are too small, i.e., less than 5% of the frame height, are red highlighted in the task UI. If the crowdworker does not resize the small annotations, the system informs the worker and deletes the boxes. Before annotating each frame the workers are shown a 2-seconds video containing the 10 preceding frames. The goal of this video is to give context and support decision-making in case a crowdworker is not sure whether an object is a car. This video can be replayed anytime during the annotation.

4.3. Evaluation Procedure

Two experts manually inspected all frames to assess the quality of the YOLO annotations and the contribution of the crowdworkers to the annotation quality. The number of correct and incorrect YOLO identifications, the number of missing identifications, and the number of correct and incorrect crowdworkers' identifications were registered. Using the expert annotations, we calculate precision, recall, and F1-score to get more rigorous information about the behavior of each model.

5. Evaluation

We collected the crowdworkers' annotations via the Amazon Mechanical Turk platform on July 12, 2022. The crowdworkers could carry out the annotation tasks as many times as desired. In total, 14 crowdworkers annotated all frames in 1 hour and 16 minutes.

In the rest of this section, we present the results of our study in three main parts. First, we analyze YOLO's performance in terms of the identified cars in the frames. Then, the contribution of the crowdworkers to YOLO's work is assessed. Finally, we combined the identifications carried out by both YOLO and the crowdworkers and assessed the performance of this collaboration.

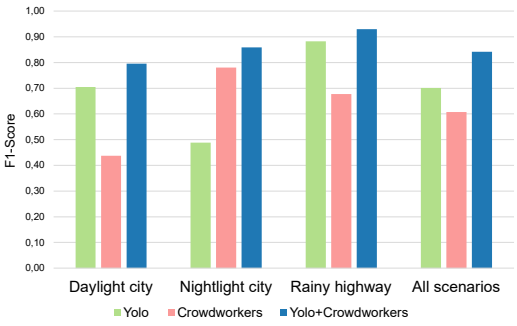

Figure 2: Performance of the identifications made by YOLO, the crowdworkers, and both combined. The crowdworkers' performance is based on the cars not identified by YOLO.

5.1. YOLO Performance

We found that the YOLO's best performance is achieved in the *Rainy highway* scenario. In this case, YOLO reaches a precision of 0.97 and managed to identify 81% of the cars, with an F1-Score of 0.88. Meanwhile, a moderate performance is observed in the *Daylight city* scenario, in which only 56% of cars are identified (Precision=0.95), resulting in an F1-Score of 0.70. Finally, the most challenging scenario for YOLO is the *Nightlight city*. In this case, only 32% of the cars were identified although a precision of 0.99 is achieved. This behavior leads to an F1-Score of 0.49. Analyzing YOLO's performance by combining *all scenarios*, we observe rather moderate results in the number of identified cars. Although most of YOLO's identifications were actually cars (Precision=0.96), YOLO identified only 55% of the cars correctly. Resulting in an F1-Score of 0.70 as shown in Figure 2.

YOLO's performance suggests a rather conservative behavior, in which only most certain cars are identified, thus achieving high precision, but not identifying a high proportion of cars, maybe due to difficult or untrained context conditions, e.g., crowds of cars, low brightness, too small cars, etc. Our results also confirm YOLO's difficulty to find cars in night conditions as in [21].

5.2. Crowdworkers' Performance

The crowdworkers' contribution is studied by considering only the cars that were not identified by YOLO since they received pre-annotated frames. We found that crowdworkers perform better in the *Nightlight city* scenario. In this case, they reached a precision of 0.97 and identified 65% of the missing cars. Thereby, resulting in an F1-Score of 0.78. In the *Rainy highway* scenario, the crowdworkers' precision decreased to 0.75 although they managed to identify 61% of the missing cars (F1-Score=0.68). In this case, the false positives resulted from trucks or construction vehicles identified as cars by the crowdworkers. The most challenging scenario for crowdworkers was the *Daylight city*. Here, the precision was 0.92, but the workers only identified 29% of the missing cars, which reduces the F-Score to 0.43. In this case, we observed that crowdworkers tend to skip objects that are in the middle of car crowds, e.g., in lines of parked vehicles. When analyzing *all scenarios* combined, similar to YOLO, the crowdworkers' precision was high, i.e., 0.92, but they managed to identify only 45% of the missing cars, which leads to an F-Score of 0.61.

Similar to YOLO, the crowdworkers' performance seems to be modest. The biggest issues for crowdworkers were finding missing cars in crowded scenarios, and avoiding annotating other types of vehicles as cars. The second issue is less critical since in a driving situation this is actually desired.

5.3. Collaboration Performance

To assess the performance of the proposed collaboration, we combine the identifications made by YOLO with those from crowdworkers. The best results for the collaboration are in the *Rainy highway* scenario, in which the share of identified cars increased to 93%, achieving a 12-percentage-point increase. Here, precision decreased slightly to 0.93, while the F1-Score increased to 0.93. This is somehow expected since the YOLO results were already really good. In contrast, the *Nightlight city* scenario received the most significant contribution from crowdworkers. In this case, the share of identified cars increased to 76%, meaning that 44% of the cars were identified by crowdworkers. The precision of the collaboration decreased again to 0.98, but the F1-Score was significantly increased to 0.86. This confirms again the ability of crowdworkers to make decisions, where an AI might be not enough trained. Finally, the *Daylight city* scenario remains the most challenging since the identified cars rate increased to 69%, i.e., 13-percentage-point after the crowdworkers' participation. The precision also decreased a little bit to 0.94, however, the F1-Score increased to 0.79. The results for *all scenarios* combined showed that the collaboration increased the share of identified cars in all frames to 75%. Thus, the crowdworkers contributed 20% of all the cars to be identified. Although the precision decreased to 0.95, the F1-Score increased to 0.84. The decrease in precision can be due to the non-cars vehicles annotated by crowdworkers.

6. Discussion and Conclusion

The success of autonomous driving vehicles relies heavily on well-trained AI models used to understand the current driving situation and take appropriate actions. To train such models, an extensive amount of labeled data is required. In this work, we studied the feasibility of a Human-AI collaboration via crowdsourcing for car identification as the first step towards a scalable pipeline for creating such labeled data. For this, we employed YOLOv3 to pre-annotate frames of three different scenarios that exhibit different image quality and traffic conditions. Then, we asked a group of crowdworkers to refine the AI-achieved annotations via a micro-task.

Our results showed that YOLO performed effectively in a *rainy highway* scenario, in which the cars are driving in two directions and no crowds of cars are observed in a frame. A more moderate performance was observed in a *daylight city* scenario that constantly exhibited dense crowds of multi-direction parked and moving cars, i.e., implying different perspectives and proximity. However, YOLO's performance was rather low in a *nightlight city* scenario in which poor light conditions represent an additional constraint. Thus, it confirms the limitations of

AI models in challenging contexts such the city scenarios. On the other hand, the crowdworkers obtained the best results in the worst YOLO scenario, i.e., *nightlight city*, contributing almost half of the car identifications and demonstrating their ability to make decisions based on the scene's hints. In the case of the *rainy highway*, the crowdworkers retrieved a significant amount of remaining cars, which were normally the most distant cars. Lastly, the *daylight city* scenario also represented a challenge for the crowdworkers. This might be related to the effort required to find partially hidden cars in dense parking locations.

The results show that a Human-AI collaboration might be feasible and scalable to save human effort by having pre-annotated data and reacting to untrained or challenging scenarios by taking advantage of crowdworkers' ability to make decisions based on context. Nevertheless, to achieve fully annotated frames further mechanisms should be investigated. For instance, a AI active learning scheme using crowdworkers contribution, and the inclusion of more crowdworkers per frame. Additionally, automatic active learning for frequent crowdworkers can be AI-supported, under a personalized training scheme based on their behavior. Finally, further steps for the detection and fixing of wrong identifications, e.g., as proposed in [22], and for addressing multi-object scenarios should be investigated.

Acknowledgments

This work was carried out under the project *Segmentation of visual media (Computer Vision) for cloud-based processing* co-financed by the program ProFIT Brandenburg of the Ministry of Economic and European Affairs of the State of Brandenburg in Germany and the European Regional Development Fund.

References

[1] S. Davies, Interconnected sensor networks and decision-making self-driving car control algorithms in smart sustainable urbanism, Contemp. Readings L. & Soc. Just. 12 (2020) 88. doi:10.22381/CRLSJ122202010.

[2] T. Brell, R. Philipsen, M. Ziefle, Suspicious minds? – users' perceptions of autonomous and connected driving, Theoretical Issues in Ergonomics Science 20 (2019) 301–331. URL: https://www.tandfonline.com/doi/abs/10.1080/1463922X.2018.1485985. doi:10.1080/1463922X.2018.1485985.

[3] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, Deep Learning for Computer Vision: A Brief Review, Computational Intelligence and Neuroscience 2018 (2018) 1–13. doi:10.1155/2018/7068349.

[4] H. Su, J. Deng, L. Fei-Fei, Crowdsourcing annotations for visual object detection, Uniwersytet śląski (2012) 40–46. URL: https://collaborate.princeton.edu/en/publications/crowdsourcing-annotations-for-visual-object-detection. doi:10.2/JQUERY.MIN.JS.

[5] H. Ning, R. Yin, A. Ullah, F. Shi, A Survey on Hybrid Human-Artificial Intelligence for Autonomous Driving, IEEE Transactions on Intelligent Transportation Systems 23 (2022) 6011–6026. doi:10.1109/TITS.2021.3074695.

[6] L. Wenyin, S. T. Dumais, Y. Sun, H. Zhang, M. Czerwinski, B. A. Field, others, Semi-Automatic Image Annotation., in: Interact, volume 1, 2001, pp. 326–333.

[7] Q. Cheng, Q. Zhang, P. Fu, C. Tu, S. Li, A survey and analysis on automatic image annotation, Pattern Recognition 79 (2018) 242–259. doi:10.1016/j.patcog.2018.02.017.

[8] X. Hu, H. Wang, A. Vegesana, S. Dube, K. Yu, G. Kao, S.-H. Chen, Y.-H. Lu, G. K. Thiruvathukal, M. Yin, Crowdsourcing Detection of Sampling Biases in Image Datasets, in: Proceedings of The Web Conference 2020, ACM, New York, NY, USA, 2020, pp. 2955–2961. doi:10.1145/3366423.3380063.

[9] A. Carlier, A. Salvador, X. Giró-i Nieto, O. Marques, V. Charvillat, Click'n'Cut: Crowdsourced Interactive Segmentation with Object Candidates, in: 3rd International ACM Workshop on Crowdsourcing for Multimedia (CrowdMM), Orlando, Florida (USA), 2014. URL: http://dx.doi.org/10.1145/2660114.2660125. doi:10.1145/2660114.2660125.

[10] X. Wang, L. Mudie, C. J. Brady, Crowdsourcing: An overview and applications to ophthalmology, 2016. doi:10.1097/ICU.0000000000000251.

[11] E. Heim, Large-scale medical image annotation with quality-controlled crowdsourcing (2018). URL: http://archiv.ub.uni-heidelberg.de/volltextserver/id/eprint/24641. doi:10.11588/HEIDOK.00024641.

[12] M. Amgad, H. Elfandy, H. Hussein, L. A. Atteya, M. A. T. Elsebaie, L. S. Abo Elnasr, R. A. Sakr, H. S. E. Salem, A. F. Ismail, A. M. Saad, J. Ahmed, M. A. T. Elsebaie, M. Rahman, I. A. Ruhban, N. M. Elgazar, Y. Alagha, M. H. Osman, A. M. Alhusseiny, M. M. Khalaf, A.-A. F. Younes, A. Abdulkarim, D. M. Younes, A. M. Gadallah, A. M. Elkashash, S. Y. Fala, B. M. Zaki, J. Beezley, D. R. Chittajallu, D. Manthey, D. A. Gutman, L. A. D. Cooper, Structured crowdsourcing enables convolutional segmentation of histology images, Bioinformatics 35 (2019) 3461–3467. doi:10.1093/bioinformatics/btz083.

[13] S. Ørting, A. Doyle, A. van Hilten, M. Hirth, O. Inel, C. R. Madan, P. Mavridis, H. Spiers, V. Cheplygina, A Survey of Crowdsourcing in Medical Image Analysis, 2019. doi:10.15346/hc.v7i1.1.

[14] S. Boorboor, S. Nadeem, J. H. Park, K. Baker, A. Kaufman, Crowdsourcing lung nodules detection and annotation, in: Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications, volume 10579, International Society for Optics and Photonics, SPIE, 2018, pp. 342–348. URL: https://doi.org/10.1117/12.2292563. doi:10.1117/12.2292563.

[15] J. Redmon, A. Farhadi, YOLO v.3, Tech report (2018) 1–6. URL: https://pjreddie.com/media/files/papers/YOLOv3.pdf.

[16] J. Du, Understanding of Object Detection Based on CNN Family and YOLO, Journal of Physics: Conference Series 1004 (2018) 012029. doi:10.1088/1742-6596/1004/1/012029.

[17] S. Khanna, A. Ratan, J. Davis, W. Thies, Evaluating and improving the usability of Mechanical Turk for low-income workers in India, in: ACM Symposium on Computing for Development, ACM DEV '10, Association for Computing Machinery, New York, NY, USA, 2010, pp. 1–10. doi:10.1145/1926180.1926195.

[18] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, P. Tran-Gia, Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing, IEEE Transactions on Multimedia 16 (2014) 541–558. doi:10.1109/TMM.2013.2291663.

[19] S. Krug, Don't make me think!: Web & Mobile Usability: Das intuitive Web, mitp Professional, MITP Verlags GmbH & Company KG, 2018. URL: https://books.google.de/books?id=e-VIDwAAQBAJ.

[20] E. Gamboa, R. Galda, C. Mayas, M. Hirth, The Crowd Thinks Aloud: Crowdsourcing Usability Testing with the Thinking Aloud Method, in: HCI International 2021 - Late Breaking Papers: Design and User Experience, Springer International Publishing, Cham, 2021, pp. 24–39. doi:10.1007/978-3-030-90238-4{_}3.

[21] C. Tung, M. R. Kelleher, R. J. Schlueter, B. Xu, Y.-H. Lu, G. K. Thiruvathukal, Y.-K. Chen, Y. Lu, Large-Scale Object Detection of Images from Network Cameras in Variable Ambient Lighting Conditions, in: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2019, pp. 393–398. doi:10.1109/MIPR.2019.00080.

[22] C. Tessier, F. Dehais, Authority Management and Conflict Solving in Human-Machine Systems., Aerospace Lab (2012) p–1.