ICWaste: Automatic Extraction of Waste Images through Computer Vision Analysis of Waste Collection Recordings

Kristian L. Tromborg*
Department of Architecture, Design and Media Technology, Aalborg University
Aalborg, Denmark
ktromb19@student.aau.dk

Rasmus S. B. Sørensen
Department of Architecture, Design and Media Technology, Aalborg University
Aalborg, Denmark
rsaren18@student.aau.dk

Abstract

Littering remains a significant environmental and financial concern. Current solutions to combat littering include preventive and proactive measures, but efforts have also gone into automating the process of collecting litter through robots, eliminating the costly need for human labour. However, the development of accurate waste detection for robotic cleanup is currently limited by insufficient training data. We propose a system leveraging volunteer litter collectors equipped with cameras mounted on their litter pickers, that can automatically extract waste images from recorded videos during litter collection. The system tracks when litter is being picked up and automatically saves an image of the litter as training data. The evaluation shows promising results for the system to be implemented into litter collectors' working routine, but the ability to track the many different litter picker designs requires further development, as they can differ greatly in appearance. Improvements can also be made to the process of extracting waste frames, since not all collected litter is extracted as training data in the tested videos.

Keywords: Littering, Data Collection, Image Data, Object Detection, Neural Networks, Computer Vision, YOLO

ACM Reference Format:
Kristian L. Tromborg and Rasmus S. B. Sørensen. 2024. ICWaste: Automatic Extraction of Waste Images through Computer Vision Analysis of Waste Collection Recordings . In . ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

As of 2022, Denmark is the country in the group of OECD (Organisation for Economic Co-operation and Development) countries with the highest average generation of municipal waste per capita, with the average danish citizen generating 787 kilograms of waste yearly [10]. Furthermore, it is estimated by the Danish Ministry of the Environment that the danish municipalities collect an estimated 700 tons of litter from nature every year, costing them around 500 million Danish Kroner (circa 67 million euros) [22]. Additionally, The Danish Society for Nature Conservation collected 100 tons of litter in 2023, totalling an estimated 800 tons of litter collected from nature in Denmark that year [24]. Collecting litter from nature is a costly effort, and not just for Denmark. A German study from 2020 found that German towns and municipalities pay an estimated 700 million euros to clean up public littering and dispose of public waste [17]. The issue of littering needs to be solved, not just for financial reasons, but also for the environmental impact of littering, which includes land and air pollution.

Means of combating the problem of littering, deal with preventive measures to reduce future littering, and proactive measures to remove current litter found in nature. Although research has found that certain preventive measures can reduce littering in both adults and children [17, 19], littering will still happen as indicated by the same studies. Proactive measures were successful by providing monetary rewards for the collection of litter [25, 27], but efforts have also gone into the development of automating litter collection through robots detecting and picking up waste [11, 18, 34]. While studies have looked into developing these robots, a big problem is the detection of the many different representations that garbage can take. To develop models that can accurately detect and segment waste into the proper sorting categories [23], a lot of training data is required. Collecting this data is a task requiring manual labour, which has been hard to consistently motivate people to do.

In this paper, we develop a solution utilising intrinsically motivated volunteer litter collectors' efforts in collecting litter, by building a solution that can extract waste image data seamlessly while the litter collectors are out picking up litter, by attaching a small camera to the litter picker, recording the process. Furthermore, we develop a web application allowing users to upload videos of these recordings for analysis, and reward their efforts. With this solution, we aim to increase the amount of available waste image data,

*Both authors contributed equally to this research.

Aalborg University, 2024, Denmark

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

while simultaneously helping incentivize proactive efforts to reduce litter in nature.

2 Background

Efforts to combat the issue of littering concern both collection of litter and preventative methods to reduce littering in the first place. Various different preventive approaches exist, such as charity bins that promise donations to charity when waste is disposed in them, and ballot bins where smokers can vote on "light-hearted questions" with their cigarette buts [32]. According to the article by ZeroWasteScotland, the charity bins reduced littering by 30% in the streets where they where placed and the ballot bins reduced cigarette butt littering by 8.9% [32]. While the article does not provide references to support their claims, Kataržyte et. al. further supports the ineffectiveness of ballot bins to significantly reduce cigarette butt littering [15]. Although preventive campaigns can help reduce littering to some extent, littering will still occur and measures are therefore needed to remove it. To help achieve this removal, a company called The Plastic Bank is incentivising picking up plastic litter by treating it as a currency, and is offering goods and services in exchange for plastic. The company has operating collection centers in Haiti and the Phillipines among others, where people can hand in plastic in exchange for credits that can be used for the goods they provide [27]. The Plastic Bank then process the plastic and sells it for reproduction to make a sustainable business out of plastic collection. Powers et. al. supports the claim that small monetary rewards can increase the motivation for people to pick up litter in unsupervised areas [25]. However, providing monetary rewards for litter collection is essentially paying people to pick up litter, something most countries are already doing and is part of what is costing Germany and Denmark so much money to clean up littering, as earlier mentioned.

Instead of having people clean up littering, efforts have also gone into automating the picking of litter through robots. Kulshreshtha et. al. [18] designed a robot to automatically detect and pick up garbage, which received a high detection rate (95.2%) of waste on a relatively small dataset, the TACO dataset [26], containing around 1500 annotated images. However, the detection was binary, meaning the robot was only able to detect whether something was waste or not, and not what type of waste it was, like plastic, glass, metal etc. Furthermore, the detection performance was not evaluated in a real-world test, and the proposed 95.2% detection rate was only based on a small selection of images from the TACO dataset. It is therefore not feasible to assume such a performance, when brought to real-world use. A similar case can be seen in a study by Gupta et. al. [11], where a robot is able to detect and classify three types of waste, namely paper, crushable plastic and non-crushable plastic, with a reported accuracy of 90%, also using the TACO dataset for training. The study does however not provide any information about how the evaluation was carried out or how much waste was collected in the evaluation, which again means that the results cannot be presumed as real-world performance. While these studies could potentially be promising as preliminary research, the factor of properly sorting the collected waste also has to be taken into account. Modern sorting requirements of waste have several different categories of waste, like the 9 categories seen in Denmark [23]. These requirements urge for detection models capable of performing more complex classifications than one [18] or few class-detection [11], for future potential robots and automatic waste sorting services, to be able to properly sort the collected waste. To do this, models need more image data of each category of waste, data which is sparse given that the TACO dataset currently only has 1500 annotated images, split between 28 super categories (60 sub-categories), and other open-source available datasets of waste image, either have less than 10000 images in total, or are web-scraped datasets, which generally are not of high quality since the scraped content is not quality controlled (see Figure 1). An overview of the open-source datasets we could find, can be seen in Table 1. The creators of the state-of-the-art object detection architecture, YOLO (You-Only-Look-Once) generally recommend at least 10000 object instances per class when training[36]. If we consider the 9 sorting categories from the Danish sorting requirements, this would mean at least 90000 annotated pieces of waste for training. Furthermore, the variety of the image data must be representative of the deployed environment [36], which can differ greatly in outside environments and in different parts of the world. This will be elaborated further in the coming section.

Figure 1. The figure show four examples of training data from the dataset "Waste Classification Data v2" found in Table 1. Web-scrapers can include misrepresentations of waste data and stock-photos without context, leaving it as a poor option despite the larger size of the datasets.

Name of Dataset	Data Amount
TrashCan 1.0	7.212
Trash-ICRA19	5.700
TACO	1.500
MJU-Waste v1.0	2.475
UAVVaste	772
Trashnet	2.527
WaDaBa	4.000
GLASSENSE-VISION	2.000
Cigarette butt dataset	2.200
Waste Class. Data v2	~27.500

Table 1. Table displaying some of the datasets from the review "Waste Dataset Review" [1]

As Table 1 shows, numerous datasets have been created for the purpose of waste classification, however they are usually tailored to their own specific purpose, which can range from underwater environments to home supplies. Notably, this paper focuses on waste in the wild which requires quite a big dataset because of the large amount of subcategories. The only dataset currently focusing of waste in the wild is the TACO dataset which consists of 1500 images. Not nearly enough to cover the amount of subcategories. Furthermore, it is worth noting that the biggest dataset in the table is a web-scraped dataset which often results in a large portion of the images being ineffective for machine learning purposes as they can be very inconsistent and in some occasions even affect training negatively (as can be seen in Figure 1). The significance of the variability of waste across different countries needs to be emphasised as well. The difference in appearance of everyday products around the world, means that the litter will also look different in many cases, which impacts the models trained to recognise the litter. A study by De Vries et. al. [8] found that several popular object classification models performed worse in regions where they had an under-sampled representation in their training data. This means that a classification model trained on data from e.g. Italy would not have the same accuracy when tested on data from Denmark.

Currently, various methods have aimed to create large datasets for training. One of the approaches is to generate synthetic training data [7, 31, 35]. This approach can be promising in generating data which e.g. is hard to ethically recreate such as potential crash scenario data for autonomous driving [7]. However, an issue stated with synthetic data, is the amount of unique textures required [31, 35], an issue which can be especially problematic in the field of waste images, where the texture of waste can differ significantly, both because of difference in appearance based on country [8], but also appearance affected by state of decay. These factors can make it difficult to generate data to accurately represent the many states and appearances waste can take, and also affect how much real-world data is needed already [8, 36] .

Another approach to acquiring training data, is to engage users in collecting it. Different methods have been tested, including using gamification to incentivize continuous usage. Features like leaderboards, points and rewards have proved to have initial success in keeping the users captivated [9]. However, long-term engagement studies have documented challenges, indicating a decline in user satisfaction over time [12, 16, 33].

One notable example of user-engagement through gamification elements is OpenLitterMap (OLM) [20]. OLM is a web and mobile application that facilitates geo-located litter images acquired through their users. The quality of the images are manually maintained by members of the OLM team. There is however a heavy reliance on a small group of top contributers. 88% of the total dataset consists of images taken by the top 10 litter collectors, which highlights the need to motivate users more effectively. Furthermore OLMs users report that sometimes picking up litter takes twice as long, as they have to take a picture of it with their phone and upload it while picking it up. OLM claims their dataset consists of over 486.000 images. It is however not publicly available which means other litter related projects are not able to benefit from it.

Another local initiative, the danish mobile application "SpotAffald", employs gamification elements in the form of unlockable badges and statistics regarding users monthly uploads [6]. Despite an initially low amount of downloads, its users have shown long term engagement by still uploading images over a year after its initial release. This has resulted in a significant amount of annotated images. To promote the application at its release the creators reached out to organisations dedicated to keeping the environment clean, along with promoting the application in five different Facebook groups. This indicates that focusing on users who are already intrinsically motivated, like volunteer litter collectors, to collect data may have a positive impact on long term user engagement.

Previous attempts at motivating continuous collection of waste image data have been scarce and mainly focused on using gamification elements to motivate usage, with the means of collecting data requiring the user to manually take pictures of the waste they find [6, 20]. As many of the active users of SpotAffald and OpenLitterMap are litter collectors using litter pickers to collect the waste after they have taken a picture of it, implementing a solution that works seamlessly with their waste collection could prove beneficial and less disruptive to their process than the act of manually taking a photo with their phone. These litter collectors, as well as other volunteer litter collectors, are already intrinsically motivated to go out collecting waste in nature. In this project,

we aim to build a solution integrated into the process of collecting litter, which automatically collects image data of the waste being picked up, without disrupting the flow of the litter collection process, allowing for easy and seamless collection of training data for future detection models.

3 System Design

The solution we propose will make use of a small camera attached to the litter picker, which can track when a user is picking up waste, and collect an image of said waste automatically. This integration removes the need for providing motivation to collect the data, as the litter collectors are already intrinsically motivate to collect litter, which we then capture images of. Additionally, we will design a web platform where litter collectors can upload the videos recorded of the collection process, and our system will automatically extract the image data from these videos. The web application will also provide extrinsic rewards for uploading the videos in an attempt to motivate continuous uploading from the users. This study will however mainly focus on the development and evaluation of the system in charge of automatically extracting the waste images from the videos, and a thorough evaluation of the motivational components of the web application is reserved for future work.

The process of collecting waste images should be integrated seamlessly into the workflow of the litter collector, without disrupting their regular process. As we are relying on our users own motivation to go out and pick up litter, the interactions they should have to make with the camera and system should be minimised so it does not negatively affect their motivation. Nevertheless, a few preparatory actions are required to setup the recording process. Firstly, the user will have to mount the camera on the litter picker and ensure that it is properly aligned with the picker's claws. Ideally, the camera should be centrally positioned between the claws to optimise tracking capabilities, and positioned in close proximity to ensure clear and sizeable depiction of waste within the captured images (see Figure 2). However, the system should be flexible enough to be able to track the claws at a tilted angle or at a slight distance, that is, if the camera is placed further up on the picker.

Figure 2. The litter picker with the camera mounted, ready for recording

3.1 Tracking the litter picker

When the camera has been mounted, the user can go about their usual routine of picking up litter while the camera records the process. Whenever they are finished they can stop the camera and once they get back home, upload it to the website where the system pipeline starts when its done uploading. An overview of this pipeline can be seen in Figure 3. At the start the system goes through it calibration phase

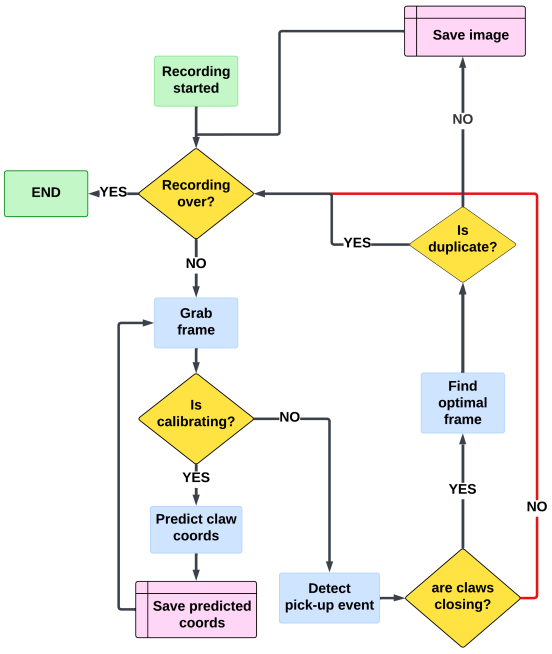

Figure 3. Flowchart of the overall system pipeline. Each frame of a given video is processed to track pick-up events and find the best frame showing a clear image of the waste in the image.

which takes 10 seconds. During these 10 seconds, it will try to locate the claws of the picker by predicting a bounding box of their positions in each frame of video and aggregate a position of each claw based on the predictions of all frames within the time-span. To predict the positions of the claws, we make use of the "You only look once"-model (YOLO), an object detection neural network architecture able to perform real-time predictions while still outperforming other object detection architectures [28]. A pre-trained YOLOv8 model is re-trained on a manually labelled dataset of around 800 images of the picker, collected from videos recorded of picking up litter. For future reference this model will be known as the "YOLOv8Claws" model. When the claws have been located in the frame, the system will be tracking their movement to determine when a piece of waste is being picked up. Once significant movement of the claws has been detected due to waste being picked up, the system will search through

the previous seven seconds of the video (which depending of the Frames Per Second of the video is either the last 210 frames at 30FPS or 420 frames at 60FPS) to try and find the most suitable frame of the waste to save as training data. The seven second period comes from initial testing during the development phase of the pipeline. This period should include all of the frames where the litter is visible to increase the chance of the system successfully saving an image of the litter, while also minimising how many frames the system should process. The process of finding the best possible image of the waste consists of several sub-processes which will be explained in the next section.

3.2 Finding the most optimal frame

A full flowchart of the process discussed in this section can be seen on Figure 4. For each of the 210-420 frames leading up to the detected pick-up event, another YOLOv8 model, which will be refferenced to as the "YOLOv8Trash" model, will try and find a piece of waste in the frame. To decide which frame is the optimal frame to save, some rules has been set for the system.

1. The bounding box of the litter should not be too close to edges of the frame, as it might mean saving an image where not all of the litter is in the frame.
2. The claws must not overlap with the bounding box of the litter, to make sure they do not cover the litter in the frame.
3. The claws should not be close to closing all the way as the most optimal image of the litter will be right before the claws close.

The bounding box of the predicted waste and corresponding frame number for each frame will all be saved for further processing. The YOLOv8Trash model is trained on a small dataset of around 2000 images of waste from the TrAAUsh dataset [21], collected via the application SpotAffald [6]. After the predicted bounding boxes have been saved, they are passed on to the next process comparing the coordinates of the bounding box for the waste against the coordinates for the claws of the picker to make sure the claws are not covering the waste in the image. Bounding boxes that are not covered, are then passed on to the next process. Here, the area of each bounding box is calculated to find the biggest box. The 10 frames with the biggest calculated areas are then passed on to the next process, where the motion blur of each frame is found using the variance of laplician function in the OpenCV library [3]. This function highlights regions of an image where there is a sudden change in intensity. It is usually used to detect edges in an image, but can also be used to detect the overall variance of an image. If an image is very blury there is going to be less edges and less of a variance in the image [29]. The frame with the least motion blur is then chosen as the most optimal frame to save as the waste image. Lastly, the waste image is checked against previously registered pick-up events to detect duplicate pick-ups, using the Image Hash library [5]. This feature was implemented due to the proposed web application offering rewards for picking up detected waste, to combat potential attempts at cheating the system to get more rewards. The final image is then saved as training data, and the pipeline starts over (see Figure 3). For a visual example of the system pipeline, see Figure 18.

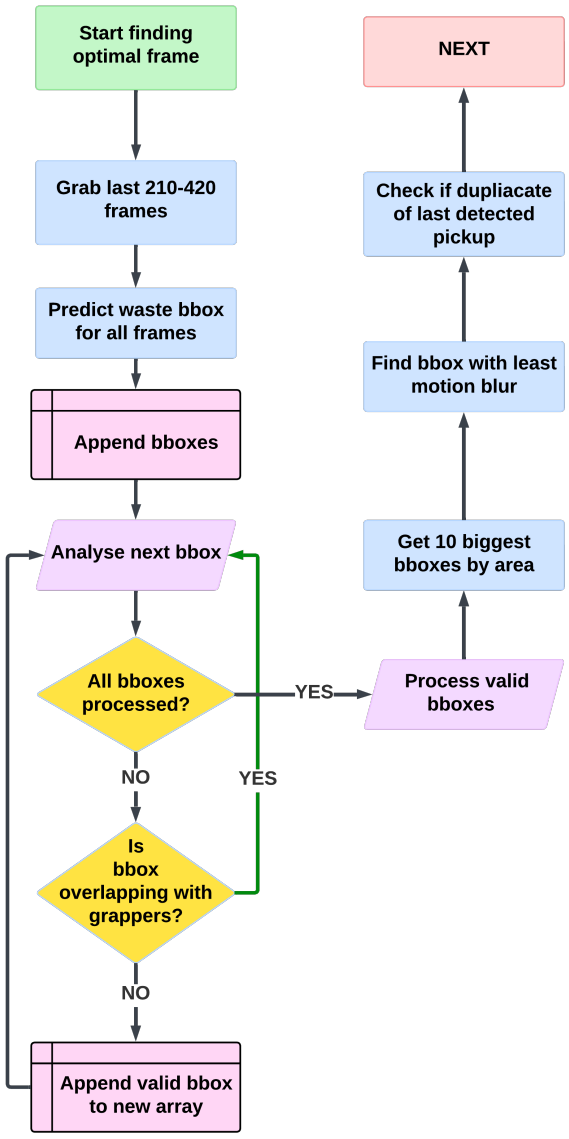

Figure 4. Finding the best frame to save includes several processes like find the frame where the piece of waste in the image is largest by area and most clear.

4 Webapp design

To provide a means of uploading video recordings of litter collections, a web application was developed. This was also designed as an introductory proposal to motivate litter collectors to record and upload the videos of the litter collection and further encourage litter collection. To support this motivation, the design of the web application draws on the proposed elements of motivation synergy described in a study by Amibile [2]. Here she describes the concept of "extrinsics in service of intrinsics", where she proposes that certain types of extrinsic motivation can combine synergistically with intrinsic motivation and are especially effective if initials levels of intrinsic motivation is high. For the extrinsic components to be "synergistic extrinsic motivators", they should, among others, support the user's sense of competence e.g. through recognition, feedback and rewards that re-affirm that competence [2].

The design of the web application includes feedback of the users efforts in the form of statistical data like timeline-graphs displaying how much waste the user has collected. Furthermore, we propose a point system rewarding the user with "waste points" for the waste they collect. These points can be used in a shop to buy coupons giving discounts to various shops. This was chosen because Kaiser et. al. [14] found that monetary rewards can provide sustained incentive as long as the rewards are not discontinued. These coupons are meant to provide such incentive, adding to the extrinsic motivation provided by the web application. Lastly, the design includes a global leaderboard of litter collectors efforts. While extrinsic gamification elements have issues with sustaining long-term motivation [12, 16, 33], it could in this context add to the sense of synergistic extrinsic motivation (recognition, feedback) proposed by Amibile [2]. Examples of the design components can be seen in Figure 5.

While we believe the motivational components in the design of the web application could provide further motivation to collect and upload videos of litter collections, the focus of this study remains on the implementation and evaluation of the system in charge of automatically extracting waste images from the litter collection videos. The design of the web application remains but an introductory proposal to how one could motivate data collection in this context, and further design considerations as well as an evaluation of the effectiveness of the motivational components described, remains future work.

5 Evaluation

The evaluation consisted of two separate parts. A system evaluation where the goal was to evaluate the efficiency of the system, and a usability test where the goal was to evaluate the process of setting up and recording videos with the litter picker. The overall goal was to examine how robust the system is to alternating setups, such as filming at different angles relative to the claws of the picker and testing how well the system works when setup by first-time users.

5.1 System Evaluation

The system pipeline consists of multiple sub-parts that needed their own evaluation to measure their efficiency. The sub-parts are as follows;

1. The initial stage following the calibration, involving the detection of a pick-up event. That is, how well can the system predict movement of the claws and detect when litter is being picked up? This was measured by calculating the precision, recall and accuracy of detected pick-up events.
2. When a pick-up event has been detected, the optimal frame must be found for saving. How well does the system segment the waste in the image? This was measured by calculating the IOU (intersection over union) of its predicted bounding boxes against manually labelled ones.
3. Lastly, the duplication detection is to be evaluated. How easy is it to "cheat" the system by picking up the same piece of waste consecutively? This was measured

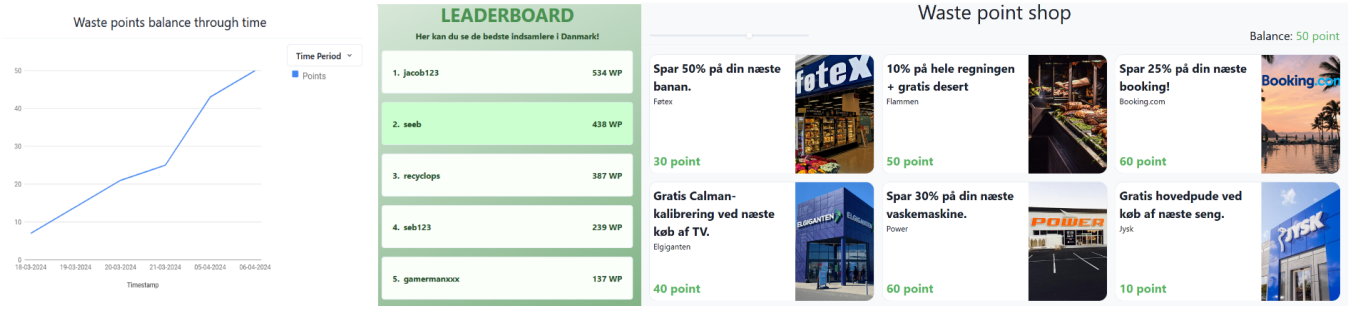

Figure 5. The right-most image in the figure shows a snippet of the proposed coupon shop. The discounts in the coupons are only meant to serve as examples. The middle image displays a leaderboard from the web app, and the left-most image, a graph of waste points awarded for uploads over time.

by looking at the percentage of images the system was able to correctly detect as duplicates.

Furthermore, it was necessary to evaluate how robust the system was to changes in the setup of the camera on the picker. As litter picker designs and cameras used for recording can differ and users of the system can interpret the process of setting up the camera for recording differently, all parts of the system pipeline was tested with different setup settings. Lastly, a system evaluation was performed on a longer video recorded by a new user. This user would not be biased by previous insight into how the system works, and had their own perception of the litter collection process.

5.1.1 Procedure. Multiple different videos were recorded, with multiple scenarios for the setup of the litter picker and its camera (see Figure 8 and 6);

- Two different litter picker designs were tested (Fig. 7).
- The camera was placed at two different angles, which are at a horizontal and a vertical level. (Fig. 6).
- The camera was placed at three different distances from the claws of the picker; Right in front of its' claws and 12cm and 25cm further up on the arm of the litter picker (Fig. 8).

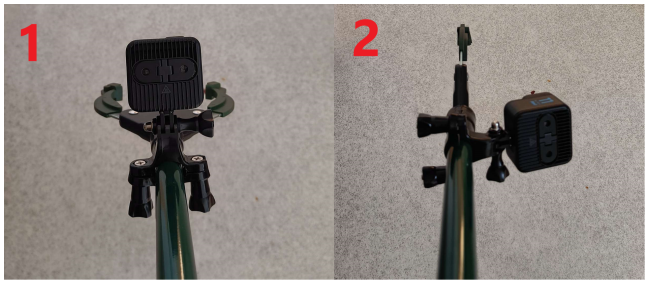

Figure 6. The figure display the two different angles tested in the evaluation. (Horizontal and vertical.

It is generally preferred that the camera is as close to the claws and the litter in the frame as possible, since this would provide a higher resolution of the waste in the extracted image, and the object detection models used will have higher performance since close-by objects are easier to detect than objects further away [13]. However, because mounting the camera is a manual process prone to user error, the system should be robust enough to handle recordings at different distances and angles. To represent all combinations of angles and distances, there should be six videos recorded per litter picker design, but since the claws of picker 2 cannot be turned, only 5 videos were recorded for this design. For each combination of settings, the same scenario was recorded; At a fixed location, at a fixed time of day (between 10:00-13:00), a video was recorded of a collection of 30 pieces of waste. The same 30 pieces were used for every video to further eliminate potential bias.

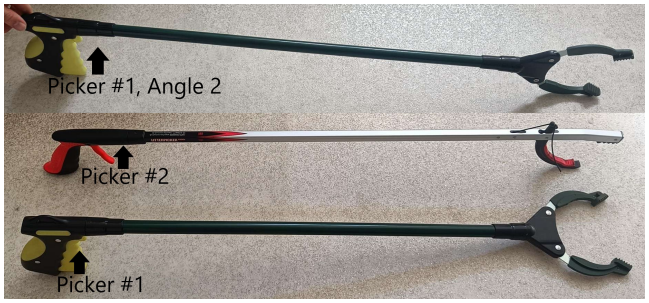

Figure 7. The two different litter picker designs used in the evaluation

Once all the different combinations of settings had been recorded, an "optimal setup" was determined based on the system performance on the different settings. With this optimal setup, a second video was then recorded with a different camera, to see if it could make a difference in the system performance. The first camera was a GoPro HERO11 Mini, shot in 1920x1080 resolution at 60 FPS. The second camera was an SJCAM C100+ shot in 2560x1440 resolution at 30 FPS. For each video, it was evaluated how well the system handles the tracking of pick-up events, by calculating the precision and recall of said events. For the extraction of waste frames, we evaluated how many frames were properly extracted out of the total 30 pieces of waste, and how well the predicted bounding boxes surrounding the waste fit, compared against a manually labelled ground truth. This was done by calculating the intersection-over-union (IOU) of the predicted bounding box and the ground truth bounding box. The precision of the predictions was also estimated at different IOU thresholds, where a predicted bounding box was rejected, if the IOU was lower than a given threshold.

The last evaluation on a new user was then conducted, and the user was asked to mount the camera in accordance with the best performing setup, using the best performing camera. They were not informed about how the system works, and were only asked to mount the camera, and collect litter for about 30 minutes.

5.2 Usability test

The system is built to be primarily used by volunteer litter collectors. Since anyone can become a volunteer litter collector, the system should be usable by anyone capable of collecting litter with a litter picker. One of the intentions behind the web application was also to provide a startup guide for new users, which was given to the users as a step by step document, to let them know how to setup the system and record in a way that the system can detect and track the claws of the litter picker during recording. The purpose

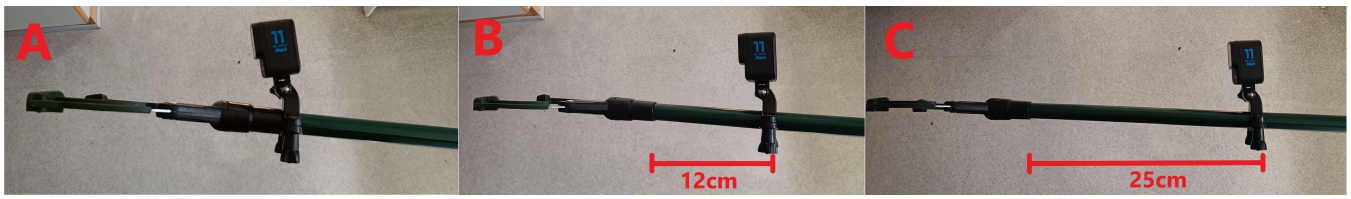

Figure 8. The figure displays the three different distances tested in the evaluation. The left-most image in the figure is where the camera is closest to the claws of the picker, and the right-most is where it is furthest away.

of this usability evaluation was therefore to determine how easy it is for anyone to follow the guide, get started with recording and subsequently, how well the system can calibrate and find the claws after anyone has made a recording of a given collection after following the startup guide.

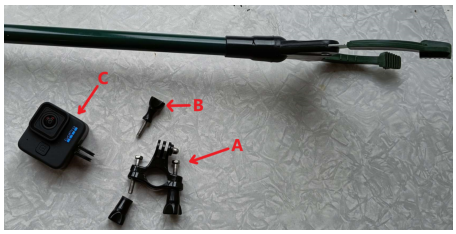

Figure 9. An image from the startup guide showing the different parts to collect on the litter picker

5.2.1 Procedure. The study was conducted on 11 people in the age group of 20-30 years (see Table 2). Participants where given a litter picker, a camera and a camera mount (see Figure 9) and were asked to follow the guide to get started with mounting the camera, starting a recording and picking up litter. After preparing the litter picker and recording a collection of 3 pieces of waste, they were then asked to answer a standardised System Usability Scale-survey (SUS) [4] evaluating the perceived usability of setting up the camera on the picker and properly recording a collection. The videos of each recording were then run through the system, to determine if it could properly calibrate and detect the pickup-events.

Average age	Male	Female
25.64	9	2

Table 2. Information about the participants for the usability test

6 Results

6.1 System Evaluation

The system evaluation revealed that the calibration phase is not yet robust enough to handle multiple litter picker designs, and different distances from the claws. The system was only able to calibrate at the distance closest to the claws, and only for the first litter picker design (Picker 1 in Figure 7). As the calibration must succeed for the rest of the system to work it means that the rest of the videos for picker 2 and the two other distances failed as a whole. Examples of how it failed to detect the claws can be seen in Figure 10. As the distance from the camera to the claws grow, the system seems to increasingly struggle at detecting the them.

Figure 10. Image of system failing to detect claws on medium distance.

This means there was only 2 setups left to evaluate before deciding on which one to use for the final system evaluation. These two setups are showcased in Figure 6, where the camera is close to the claws but the angle of the garbage picker differs. The different setups will be called setup A1 and A2 as shown in Figure 6 and 8.

	Setup A1	Setup A2
Litter detected	26	24
Closings detected	30	30

Table 3. The results from the videos where the calibration was successful. The setup name is a combination of the settings that can be seen in Figure 6 and 8

The results seen in Table 3 show that both setups were good at detecting when the claws were closing as they had a 100% detection rate. Furthermore there were 0 times where they detected the claws closing when they were not. Setup A1 however had a slight edge over setup A2 in litter it managed to detect, which means it will be the main setup going forward. With the optimal setup decided, the system performance was then evaluated on the two different cameras, with differing frame rate and resolution.

6.2 Optimal Setup: Tracking Pick-up Events

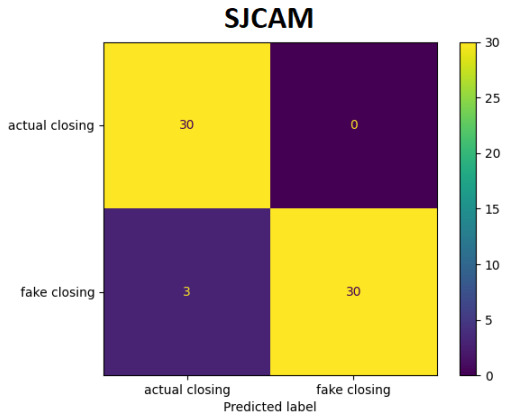

Figure 11. Confusion matrix of the systems predictions on claw-movement in the 30 fps video.

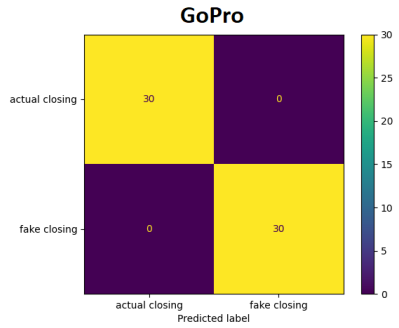

Figure 12. Confusion matrix of the systems predictions on claw-movement in the 60 fps video.

An overview of how well the system predicted grabber movement can be seen in Figure 11 and 12. Every time the system detects movement of the claws it will be classified as either a fake closing or an actual closing of the claws. In Table 4 the precision, recall and accuracy can be seen for the two cameras. The system performs well on both cameras but the GoPro does perform better than the SJCAM with a precision, recall and accuracy of 1.0 which means it correctly predicted the movement of the claws 100% of the times.

	SJCAM	GoPro
Precision	0.91	1.0
Recall	1.0	1.0
Accuracy	0.95	1.0

Table 4. The precision, recall and accuracy from the systems movement classification on the two different cameras.

In terms of amount of litter detected and correctly saved, in total 30 pieces of the same litter were picked up in each video. These results show that the GoPro again has a slide edge over the SJCAM, as the system correctly saved an image of the garbage 25 times, while the system only managed to do so 18 times (see Table 5).

6.3 Optimal Setup: Finding the best frame

The results of the optimal setup with different cameras show, that while the video shot with SJCAM have more accurately predicted bounding boxes than the video shot with the GoPro, it did properly extract less frames of waste in total (see Table 5). The GoPro on two occasions predicted a bounding box surrounding a different object in the image than the intended piece of waste. These mistakes negatively affected the average IOU with the GoPro, partially explaining the higher standard deviation, and why the median IOU differs more from the average, than with the SJCAM. This can also be seen in Table 5, where the average IOU for the GoPro without the two mis-detections, is both higher and with a smaller standard deviation.

	GoPro	SJCAM	GoPro, No outliers
Average IOU	0.786±0.236	0.896±0.110	0.852±0.097
Median IOU	0.880	0.931	0.883
Extractions (out of 30)	25	18	23

Table 5. Average and median IOUs for the predicted bounding boxes from the two cameras, as well as the amount of properly extracted waste images.

However, as can be seen in Figure 13, the precision at higher IOU thresholds dropped more in the GoPro video,

than in the SJCAM video, even without the two outliers. Implications of how resolution and frame rate can have affected performance, will be covered in the discussion.

As the system was able to better extract proper frames using the GoPro camera, and did not perform significantly worse than the SJCAM (outliers excluded), it was used as the recording camera in the last test involving a new user.

6.4 Optimal Setup: Testing on a New User

From the user test it quickly becomes clear that the system is not as robust when used on a video recorded by a person who does not know the system pipeline. In total, 77 pieces of litter were picked up during the 30 minute video and as can be seen on Figure 14 the claws were closed a total of 79 times, which means there were two times where the user closed the claws without picking up any litter. A total of 100 closings were wrongly predicted by the system to have happened during the video and it correctly predicted 239 fake closings.

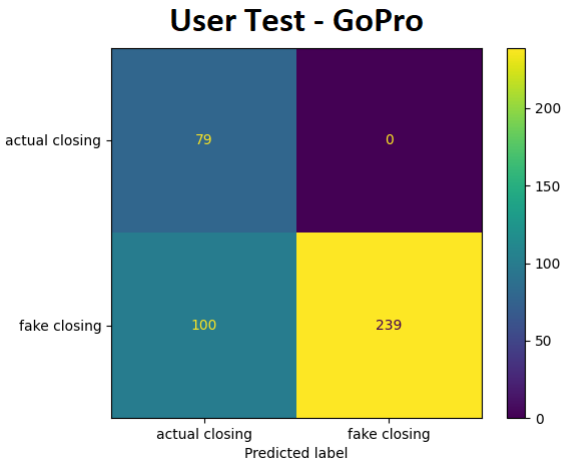

Figure 14. Confusion matrix of the systems prediction of grabber movement in the user test video.

As can be seen in Table 6 the recall remains the same as in previous tests having no times where the system predicted an actual closing as a fake closing. However due to a big increase in wrongly predicted actual closings were it should have predicted a fake closing the precision of this part of the system has fallen to 0.44.

Of the 77 pieces of litter that were picked up during the video, the system managed to correctly save images of 55, which means that part of the system has a success rate of 71.4%. A total of 79 images were saved from the videos as images of trash which means that 69.6% of the images saved were actual useful images of litter.

	GoPro
Precision	0.44
Recall	1.0
Accuracy	0.76

Table 6. Evaluation results of detecting pick-ups for the 30 minute video recorded by a new user.

6.5 Usability test

The results from the usability test showed that all participants were able to properly mount the camera and perform a recording of a waste collection, with only minor variations in the distance and angle of the placement of the camera mount. Nevertheless, the system was able to calibrate and find the claws of the picker in 100% of the videos and properly detect the pickup of the waste in all videos as well. However, in some videos where the participant had placed the camera further up on the picker, the system struggled to properly segment the waste in the image. As can be seen in Figure 15, the plastic to which the claws of the picker are attached is falsely detected as waste, resulting in an improper segmentation. While the left image is still usable since it is not obscured by the picker, the predicted bounding box is not, and the right image is partially obscured and cannot be used as training data.

	Average	Min.	Max.	Median
SUS-score	82.7 ±13.9	60	95	87.5

Table 7. Table of the results from the SUS-surveys

The SUS-scores from the survey were calculated per person according to the guidelines described in the original paper proposing the system usability scale [4] and an overview can be seen in Table 7. The average SUS-score of 82.7 (±13.9) corresponds to the letter grade A (see Figure 16), indicating acceptable usability, according to the interpretation guidelines proposed by Jeff Sauro [30]. The lowest scoring statement was "I found the system very cumbersome to use". Additionally, some participants mentioned that the picker with the go-pro mounted might be heavy to use for a prolonged period, and it was observed that two participants subconsciously used both hands to pick up litter. Three participants also expressed that they would only need the guide the first time they were setting up the camera for recording, as it was an easy procedure to remember.

7 Discussion

This study proposes a solution to cover an area of the growing need for big datasets in machine learning. This area being the problem of collecting images of litter. This section will discuss how well this solution managed to cover that area, by looking at the results of each part of the system.

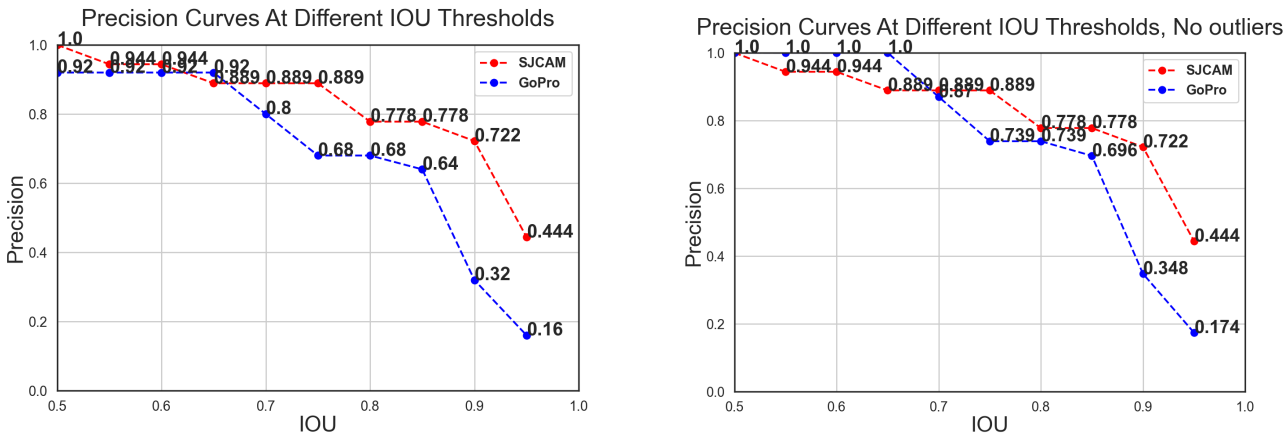

Figure 13. Precision scores for the predicted bounding boxes for the two cameras at different IOU thresholds, starting at 50%, iterating at intervals of 5 up to 95%

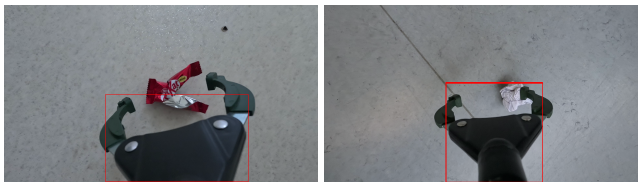

Figure 15. "Optimal frames" saved as training data from videos recorded during the usability test.

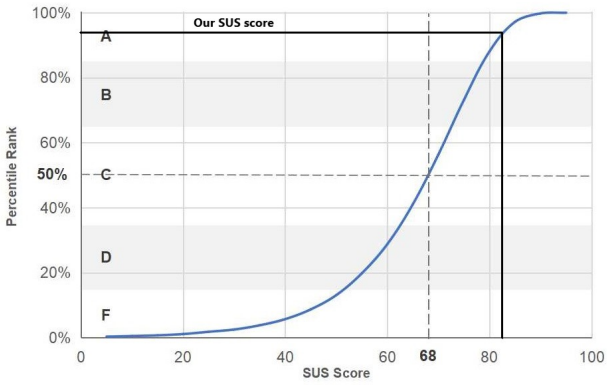

Figure 16. The calculated average SUS score for the 11 participants was 82.7, corresponding to the letter grade A as per the interpretation guide by Jeff Sauro [30].

7.1 System Usabillity

The results from the usability test indicates that setting up the system for recording, accompanied by a setup guide, is a comprehensible way to get started with recording. However, the lack of robustness of the systems' ability to find an optimal frame, in respect to the placement distance of the camera on the picker, means that even a perceived proper placement, can lead to improper results. As seen in Figure 15, the system failed to properly find the optimal frame to save, despite a proper calibration, and a perceived proper placement from the user. The further one strays from the proposed "optimal setup" in the system evaluation, the more prone the system is to faulty behaviour. The optimal setup does not include the plastic to which the claws are attached in the frame when recording, meaning it will not cause faulty segmentations. However, the system should be robust enough to not confuse parts of the picker with waste. To fix this, the model in charge of detecting the claws of the picker could include an additional class to represent not just the claws, but their connected parts as well. When locating the claws during the calibration phase, the system could also try to locate their connected parts and create a better area in which not to consider when making predictions during the segmentation phase.

Another finding from the usability test, was the perceived cumbersomeness of carrying the litter picker with the go pro camera mounted. While weighing only around 150 grams, the camera is heavier the further down on the litter picker it is placed. This can be problematic for longer use and especially since the determined optimal position of the camera is at the very end of the stick before the claws of the picker. One should therefore consider using as light a camera as possible. An option could also be to move the camera further up on the stick, but the further away it is moved, the smaller the detected waste will appear in the image. The images of waste should be of as high resolution as possible, and a lighter camera is therefore preferred to solve this issue.

Furthermore it is worth noting the environmental influence of conducting the usability test inside. The difference

in background and lighting can have affected the performance since it differs from the data used to train the models used for detecting the claws and segmenting the waste, data which was all captured outside under different weather conditions. The model used for segmenting waste is trained on the TrAAUsh [21] dataset, consisting only of images captured outside in natural environments, meaning it differs significantly from the images captured during the usability test. However, it is difficult to determine how big of an influence the environment could have had on the model detecting part of the picker as waste, since the other distances tested in the system evaluation, where that part is also visible, were not able to calibrate properly, so a pickup-event was never recorded in the videos, and an attempt at segmentation was never performed.

The results from the calibrations performed in the videos from the usability tests is in line with the results from the system evaluation, since the guide for the usability test was built around the optimal setup determined in the system evaluation. While there were differences in distance and angle when set up by the participants, they rarely differed as much as the different angles and distances seen in the system evaluation (see Figure 6 and 8). Here, the change in environment seemingly did not affect the models ability to detect the claws. The effect of a different environment should however be investigated further in regards to the calibration, as some people might perform the calibration inside, before going out to collect litter.

7.2 System Evaluation

The calibration phase of the system works well when the camera is close to the claws, however at different distances and other picker designs it begins to fail which questions the robustness of the calibration phase. The lack of ability to properly calibrate the system when recording with the second picker design and at distances further away from the claws, proves that more data is needed of various picker designs and various distances in order to properly locate the claws in future videos. Only 800 images were used for training the YOLOv8Claws model which is a very small dataset, so it was expected that there might be problems with the calibration when recording with litter picker designs that were significantly different from the ones used to train the model initially. However, with more data of several different picker designs at various different settings, the system should become better at locating and tracking the claws. This could be achieved by involving the user in the process, by having them annotate the claws in the first 10 seconds of the video when they upload it to the website. This would leave us with a wide range different pickers to train the model on, which would allow more flexibility in terms of picker designs and the position of the camera which we learned was important from the usability test. The idea of website annotations could be expanded on further by having a separate page on the website where users could also annotate the images of litter from videos they have uploaded.

As the calibration phase was not robust enough to handle significant changes in distance from the claws to the camera, or changes in the type of litter picker, finding the optimal setup was easily determined. However an optimal solution would be robust enough to handle these changes to accommodate both user needs and leave room for possible user failure. By user needs we mean users wanting to mount the camera further away from the claws to ease the weight and by user failure we mean leaving leeway for the user not needing to perfectly mount the camera for it to work. The YOLOv8Claws model was trained on 800 images taken from two different videos and finished training with an accuracy of 99%. The training data was all images taken from distance A on figure 8, which indicates that a potential solution to this problem is to simply retrain the YOLOv8Claws model on more images taken from different distances. These training sets should be easy to make as we already have videos of the claws at different distances from the evaluation. The same solution could potentially be used to make it more robust to different picker designs. From the evaluation it seems the angle of the camera have little effect on the system as a whole, as the results from A1 and A2 were similar (see Table 3). The distance from the camera to the claws of the picker and different picker designs, remains the biggest influence in the systems ability to calibrate and track the claws.

The evaluation of the system on the two different cameras, SJCAM and the GoPro, showed interesting results that had pros and cons for both choices. The SJCAM recorded in 30 fps which meant that during a recording, there were less frames where the litter was actually in the frame resulting in less potential savable frames. The system was during development only tested on a GoPro recording in 60 FPS. During the evaluation it quickly became clear that the rules set up for determining an "optimal frame" of the litter were too severe for a 30 FPS recording. Some of the images that were discarded as not an "optimal image" could still very well be used in a dataset, however since there was double the amount of frames to work with during development, the definition of the "optimal frame" was too strict for a 30 FPS recording. These rules being that the litter should be in a certain region of the frame and the claws should not take up a certain amount of the bounding box surrounding the litter. The calibration phase and the general tracking of the claws seems to be robust enough to handle both a 30 FPS recording and a 60 FPS recording as they both had very high precision, recall and accuracy if the camera is at distance A (see Figure 8). Throughout all of the system evaluation, the tracking recall was at 1.0 meaning it should in most cases never detect any false negatives, or fake closings when there

is an actual closing of the claws.

Regarding image quality there is a clear difference on the images saved from the 30 FPS videos and the 60 FPS videos. There is usually a greater amount of motion blur in the images from the 30 FPS videos. The system tries to combat this by finding the least blurry frame with the variance of laplician function, however a lot of times the images are still quite blurry whereas in the 60 FPS videos, the images are much less blurry and usually quite clear (see Figure 17). The argument for recording in 30 FPS simply comes down to the fact that they take a lot less time for the system to process.

Figure 17. The laplacian of two images from the different cameras. The higher the laplacian value, the more clear the image is.

The system evaluation on a new user showed quite different results than the previous evaluation, mainly because the users behaviour with the picker was different. For a significant portion of the recording the user has the claws closed enough that the system kept predicting that the claws were closing, resulting in the large amounts false positives. A positive thing to take away from this is that even though this was the case, the system only saved 77 images of "optimal frames" of litter, which could indicate that even though there is false positives it can detect that there is no litter in the frames and therefore not save any images. A possible solution to this problem could be to adjust how the system predicts claw movement, making it so the claws have to be more closed for it to actually predict that they are closing. This should however be done with a lot of testing and care as big pieces of litter could go undetected because the claws do not close fully on those. Of the 77 images the system saved, 55 were actual images of litter and the last 22 were images where the system had detected non litter objects as litter. This means it correctly saves about 71% of the litter picked up, which still leaves room for improvement. The YOLOv8Trash model is trained on 1150 images from the TrAAUsh dataset and it has 6 different classes, which means we do not meet the minimum recommendations of 1500 images per class or 10000 instances per class [36]. It can be speculated if having all these classes were even necessary in this context as maybe a binary classification model could have performed just as well requiring a smaller dataset. The classes would then be "litter" and "not litter".

A part of this whole project that has not been covered as much is the annotation of the images that is going to happen after videos have been uploaded and processed. This was originally thought to be a manual process done by a person, however there could be potential to automate this process using OpenAI's automatic speech recognition system Whisper. Every time a user picks up a peace of litter they could say the class of the litter out loud and the Whisper model would then be used to create timestamps in the video that could be matched with the systems detection of the claws closing.

8 Conclusion

This project sought out to create a sustainable method to automatic image gathering of litter, by utilising intrinsically motivated volunteer litter collectors' effort to clean the environment. To achieve this we created a website for litter collectors to upload recordings of them picking up litter as well as a system to automatically extract images of litter from that recording. The evaluation exposed some of the systems weaknesses, like its overall robustness. In its current state the system does not account enough for user behaviour like accidentally holding the claws semi closed during large parts of the recording. Furthermore it is not robust enough to handle different picker designs or different distances from the camera to the claws. However, the system does perform well when the camera is close to the claws and the right picker is used. With this setup the calibration phase has always been successful and the tracking of the claws for the rest of the recording is close to perfect when not accounting for human error. This could suggest that with enough data the YOLOv8Claws model could reach even better results. The systems' ability to save images shows promising results being able to consistently save images of over half of the litter being picked up in recordings. The evaluation showed that what type of camera and what type of video settings the video is recorded with has impact on the systems performance and the quality of the images. Overall the system performed to a satisfactory extent but has room for a lot of improvements going forward.

References

[1] Agnieszka Mikołajczyk (AgaMiko). 2022. *Waste datasets review.* https://github.com/AgaMiko/waste-datasets-review.

[2] Teresa M Amabile. 1993. Motivational synergy: Toward new conceptualizations of intrinsic and extrinsic motivation in the workplace. *Human resource management review* 3, 3 (1993), 185–201.

[3] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[4] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[5] Johannes Buchner. 2021. *ImageHash.* https://pypi.org/project/ImageHash.

[6] Christoffer Cæsar Fælled and Chatrine Elisabeth Larsen. 2021. SpotAffald: A Citizen Science Approach to Litter in Context. *Aalborg University, Department of Architecture, Design and Media Technology* (06 2021).

[7] Tim Dahmen, Patrick Trampert, Faysal Boughorbel, Janis Sprenger, Matthias Klusch, Klaus Fischer, Christian Kübel, and Philipp Slusallek. 2019. Digital reality: a model-based approach to supervised learning from synthetic data. *AI Perspectives* 1 (2019), 1–12.

[8] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. 2019. Does object recognition work for everyone?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 52–59.

[9] Elisa D. Mekler et al. 2017. *Towards understanding the effects of individual gamification elements on intrinsic motivation and performance.* https://www.sciencedirect.com/science/article/pii/S0747563215301229.

[10] Eurostat. Last Updated: 08/02/2024. *Municipal waste by waste management operations.* https://ec.europa.eu/eurostat/databrowser/view/env_wasmun/default/bar?lang=en.

[11] Shreya Gupta, HM Kruthik, Chaya Hegde, Shreya Agrawal, and SB Bhanu Prashanth. 2021. Gar-Bot: Garbage collecting and segregating robot. In *Journal of Physics: Conference Series*, Vol. 1950. IOP Publishing, 012023.

[12] Michael D Hanus and Jesse Fox. 2015. Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & education* 80 (2015), 152–161.

[13] Yu Hao, Haoyang Pei, Yixuan Lyu, Zhongzheng Yuan, John-Ross Rizzo, Yao Wang, and Yi Fang. 2023. Understanding the Impact of Image Quality and Distance of Objects to Object Detection Performance. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 11436–11442.

[14] Florian G Kaiser, Laura Henn, and Beatrice Marschke. 2020. Financial rewards for long-term environmental protection. *Journal of Environmental Psychology* 68 (2020), 101411.

[15] Marija Kataržytė, Arūnas Balčiūnas, Mirco Haseler, Viktorija Sabaliauskaitė, Laura Lauciūtė, Kseniia Stepanova, Cristina Nazzari, and Gerald Schernewski. 2020. Cigarette butts on Baltic Sea beaches: Monitoring, pollution and mitigation measures. *Marine Pollution Bulletin* 156 (2020), 111248.

[16] Jonna Koivisto and Juho Hamari. 2014. Demographic differences in perceived benefits from gamification. *Computers in Human Behavior* 35 (2014), 179–188. https://doi.org/10.1016/j.chb.2014.03.007

[17] Verband kommunaler Unternehmen. Berlin, 20.08.2020. *Einwegplastik und Zigarettenkippen in der Umwelt kosten Kommunen jährlich 700 Millionen Euro.* https://www.vku.de/presse/pressemitteilungen/archiv-2020-pressemitteilungen/einwegplastik-und-zigarettenkippen-in-der-umwelt-kosten-kommunen-jaehrlich-700-millionen-euro/.

[18] Medhasvi Kulshreshtha, Sushma S Chandra, Princy Randhawa, Georgios Tsaramirsis, Adil Khadidos, and Alaa O Khadidos. 2021. OATCR: Outdoor autonomous trash-collecting robot design using YOLOv4-tiny. *Electronics* 10, 18 (2021), 2292.

[19] Petra Lindemann-Matthies, Isabel Bönigk, and Dorothee Benkowitz. 2012. Can't see the wood for the litter: Evaluation of litter behavior modification in a forest. *Applied Environmental Education & Communication* 11, 2 (2012), 108–116.

[20] Seán Lynch. 2018. OpenLitterMap.com–open data on plastic pollution with blockchain rewards (littercoin). *Open Geospatial Data, Software and Standards* 3, 1 (2018), 1–10.

[21] Christoffer Cæsar Fælled Markus Löchtenfeld and Chatrine Elisabeth Larsen. 2021. *TrAAUsh - A data set of litter in context.* https://github.com/loechti/TrAAUsh.

[22] Miljøstyrelsen. [n. d.]. *Henkastet Affald.* https://mst.dk/erhverv/groen-produktion-og-affald/affald-og-genanvendelse/affaldshaandtering/affaldsfraktioner/henkastet-affald.

[23] Miljøstyrelsen. [n. d.]. *Strømlinet affaldssortering.* https://mst.dk/borger/affald-og-forurening/sortering-af-affald/stroemlinet-affaldssortering.

[24] Danmarks Natursfredningsforening. [n. d.]. *Affaldsindsamlingen.* https://www.affaldsindsamlingen.dk/om/resultater/.

[25] Richard B Powers, J Grayson Osborne, and Emmett G Anderson. 1973. Positive reinforcement of litter removal in the natural environment 1, 2. *Journal of Applied Behavior Analysis* 6, 4 (1973), 579–586.

[26] Pedro F Proença and Pedro Simoes. 2020. Taco: Trash annotations in context for litter detection. *arXiv preprint arXiv:2003.06975* (2020).

[27] The United Nations Environment Programme. 28.03.2018. *Is monetizing waste the secret to ending plastic pollution?* https://www.unep.org/news-and-stories/story/monetizing-waste-secret-ending-plastic-pollution?_ga=2.176825573.1981233233.1710239764-1432561996.1710239764.

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[29] Adrian Rosebrock. 2015. *Blur detection with OpenCV.* https://pyimagesearch.com/2015/09/07/blur-detection-with-opencv/.

[30] Jeff Sauro. 2018. *5 Ways to Interpret a SUS Score.* https://measuringu.com/interpret-sus-score/.

[31] Dominik Schraml. 2019. Physically based synthetic image generation for machine learning: a review of pertinent literature. *Photonics and Education in Measurement Science 2019* 11144 (2019), 108–120.

[32] Zero Waste Scotland. 21.02.2023. *Some of the best litter prevention campaigns from around the world.* https://www.zerowastescotland.org.uk/resources/some-best-litter-prevention-campaigns-around-world.

[33] Stefan Stepanovic and Tobias Mettler. 2018. Gamification applied for health promotion: does it really foster long-term engagement? A scoping review. In *Proceedings of the 26th European Conference on Information Systems*. AIS, 1–16.

[34] Jian Su, Yu Cao, Anqi Tang, Siyuan Wang, and Janet Dong. 2021. Design of litter collection robot for urban environment. In *ASME International Mechanical Engineering Congress and Exposition*, Vol. 85611. American Society of Mechanical Engineers, V07AT07A018.

[35] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 23–30.

[36] Ultralytics. [n. d.]. *Tips for Best Training Results.* https://docs.ultralytics.com/yolov5/tutorials/tips_for_best_training_results/.

A Appendinx A

Received 29 May 2024; revised TBD; accepted TBD

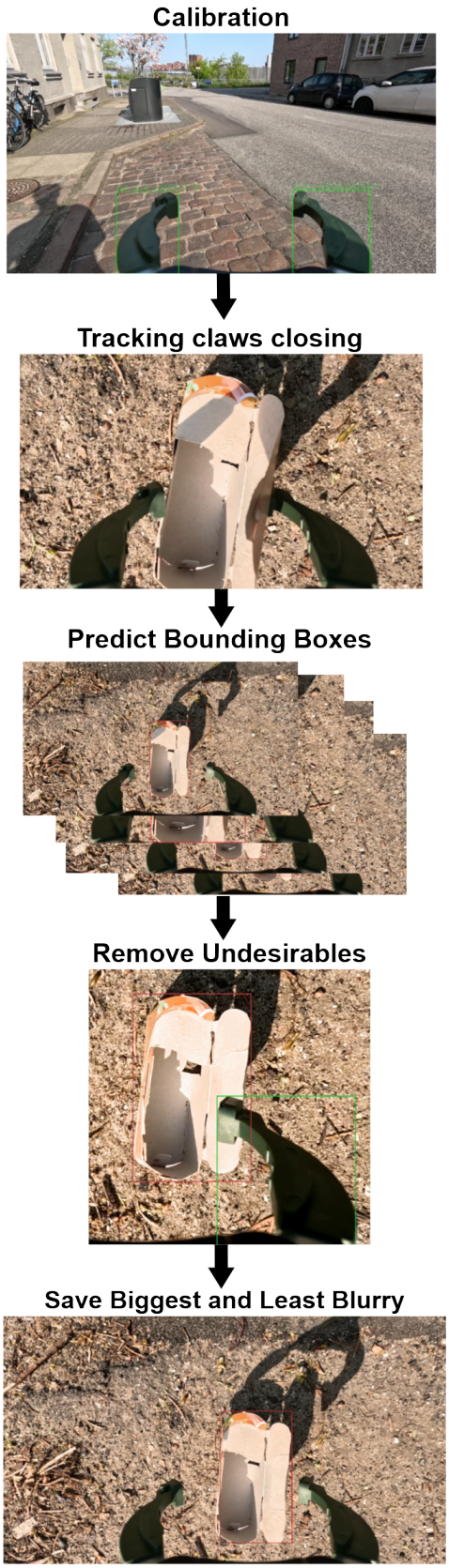

Figure 18. A visual example of the system pipeline