A Two-Step Approach for the Acquisition of Individual Tree Outlines using Paid Crowdsourcing

DAVID COLLMAR[1], VOLKER WALTER[1], MICHAEL KÖLLE[1] & UWE SÖRGEL[1]

Abstract: In this paper, we aim to investigate the acquisition of tree outlines from image data using paid crowdsourcing. A common approach for improving data quality in crowdsourcing is multiple acquisition followed by integration, which can be done, for example, by majority voting. Based on this, we present a two-step approach to obtain individual tree outlines from orthophotos. A first crowd campaign is used to segment and detect relevant areas. Here, the dataset is annotated by several crowdworkers with a subsequent integration by majority voting. This is followed by a second campaign in which the tree outlines from the previously selected relevant areas are collected by multiple crowdworkers by means of polygons. These polygons are then integrated into one single polygon per tree using a pixel-based voting approach. Depending on the threshold, an improvement in Intersection over Union (IoU) of up to 13 percentage points can be achieved, with a standard deviation reduced by up to 40%.

1 Introduction

Due to the success of machine learning (ML), there is an enormous demand for training data (STONEBRAKER & REZIG 2019). While much effort is put into machine learning algorithms, only a fraction of the effort is spent on training data preparation (STONEBRAKER & REZIG 2019). Training data are needed in large quantities, and low-quality training data cannot be compensated even by the best ML algorithms (WHANG & LEE 2020). Therefore, high quality training data are essential and the collection of such data is an important research topic. A well-known approach for collecting training data, potentially at high quality, is crowdsourcing (WHANG & LEE 2020).
There are many publications that deal with the use of crowdsourcing for generation of training data. According to SARALIOGLUA & GUNGOR (2020), a majority of studies in the field of remote sensing use crowdsourcing for the acquisition of training data for image classification or data acquisition, and potential contributions through crowdsourcing "enable artificial intelligence to develop very quickly" (SARALIOGLUA & GUNGOR 2020).
In general, crowdsourcing comes in two types: Voluntary crowdsourcing and paid crowdsourcing. Voluntary crowdworkers are mostly motivated by intrinsic factors: No monetary or comparable compensations are expected, instead, crowdworkers are voluntarily participating in campaigns because it might be their hobby or general interest (HOSSAIN 2012; RYAN & DEZI 2000). Due to the difficulty of planning such campaigns, paid crowdsourcing can be an alternative. Here, extrinsic motivation is achieved through monetary compensation, which is one of the main motivations of users on paid crowdsourcing platforms (HOSSAIN 2012; PILZ & GEWALD 2013; YUEN et al. 2012).

[1] University of Stuttgart, Institute for Photogrammetry, Geschwister-Scholl-Str. 24D, 70174 Stuttgart, Germany, E-Mail: [David.Collmar, Volker.Walter, Michael.Koelle, Uwe.Soergel]@ifp.uni-stuttgart.de

 DOI: 10.24407/KXP:1841079707

Paid crowdsourcing can be used to generate training data in a fast and easy manner (SARALIOGLUA & GUNGOR 2020), at low cost (HIRTH et al. 2011) and with high quality of data – provided certain principles are followed (CHANDLER et al. 2013; CHEUNG et al. 2017). Quality improvements through task design may be included (ALLAHBAKSH et al. 2013; ZHANG et al. 2016), seeking "to guide the labelers to provide high quality labels" (ZHANG et al. 2016).
Crowdsourcing can be further enhanced by improvement of quality after data collection (ZHANG et al. 2016). This can be done by repeated labeling, where different workers are labeling the same dataset with subsequent integration. This approach is based on the theory of "Wisdom of Crowds", which follows the principle that "the output of the crowd can be greater than the sum of its parts" (CHANDLER et al. 2013). In addition to a potential increase in quality through multiple acquisitions (ZHANG et al. 2016), this also reduces the impact of so-called "satisficers", who are trying to maximize their earnings with minimal effort (CHANDLER et al. 2013) and thereby compromise quality. Paid crowdtasks can be published on platforms such as Amazon MTurk or Microworkers.com, which handle the recruitment and payment of workers. Crowdworkers registered on such platforms can freely select their job from those offered (HIRTH et al. 2011).
Crowdsourcing has been increasingly used in a wide variety of fields, including geospatial sciences: Both HADI et al. (2022) and SARALIOGLUA & GUNGOR (2022) use voluntary crowdsourcing for a crowd-based large-scale land classification using high-resolution satellite images, while PUTTINAOVARAT et al. (2022) apply it to assess flood damages. WALTER et al. (2020) extracted trees from 3D point clouds based on paid crowdsourcing, and WALTER et al. (2021) presented an approach to collect vehicles from point clouds by means of crowdsourcing using a two-level approach.
We investigate the potential of using the advantages of crowdsourcing for an efficient and cost-effective collection of high-quality tree polygons for later use as training data for machine learning in this paper. The rest of the paper is organized as follows: We first motivate and describe the developed methodology in sections 2 and 3. Section 4 presents the used dataset, while section 5 describes the performed experiments. Section 6 discusses the data analysis in detail, while section 7 ends this paper with the conclusion.

2 Preliminary Investigation

The success of crowd tasks depends strongly on task complexity: A too complicated task cannot be solved in a reliable way by means of crowdsourcing (ALLAHBAKHSH et al. 2013), which is why the crowdworkers' ability should be tested beforehand for tasks that might be challenging (CHANDLER et al. 2013). Since we are interested in high-quality annotations of individual tree outlines, which might be a rather complex issue, we followed this suggestion. Therefore, a test acquisition campaign was performed: We presented an image tile containing multiple trees of different size and shape (13 complete and 3 partial trees) to 25 crowdworkers on the crowdsourcing platform Microworkers.com. The crowdworkers were instructed to annotate all visible trees as precise as possible using individual polygons. The results are shown in Figure 1.

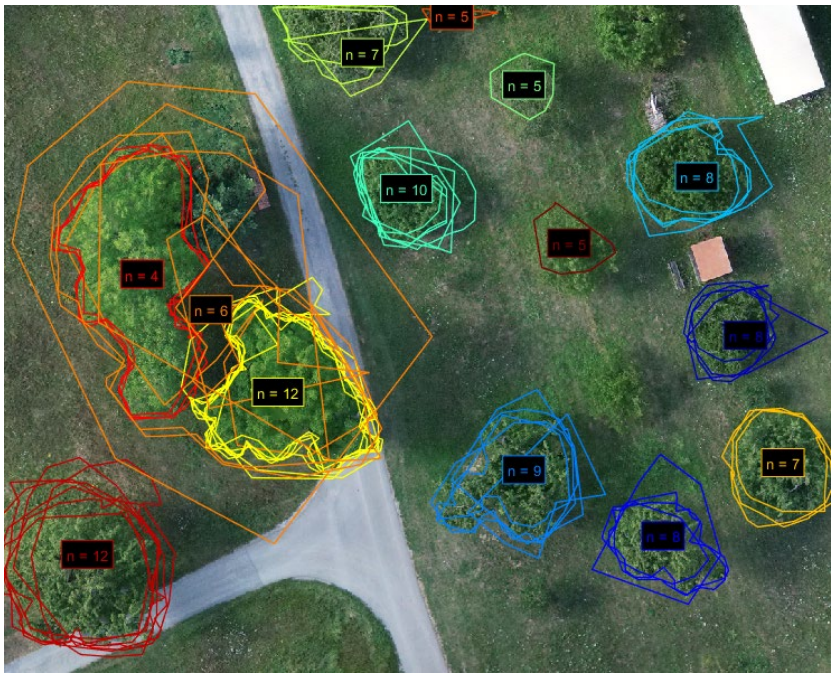

Fig. 1: Filtered crowd acquisitions on a sample image tile for 25 workers, color-coded by cluster

This task led to highly heterogenous results. In total 151 trees were annotated, resulting in only approximately 6 annotations per crowdworker. After automatically removing invalid polygons and spam acquisitions, 114 polygons were left, leading to an effective 4.56 annotations per crowdworker of vastly different quality. The annotations were not only unprecise, which can be a common issue in crowdsourcing for various reasons (CHEUNG et al. 2017), but also incomplete (multiple trees not annotated at all), incorrect (two trees in central left part of the image annotated as one tree) and inhomogeneously distributed (different amount of annotations per tree). The reason for this high heterogeneity seems to lie in the task itself: Workers are effectively doing two steps at once (localization and annotation of trees via polygons), leading to a too complex and time-consuming microtask: Microtasks have a typical working time of some minutes and typically low salary (HIRTH et al. 2011). The results are not surprising and in line with the observations of FINNERTY et al. (2013): Having crowdworkers perform a dual-task can increase the cognitive demand of the crowdworker too much and therefore significantly decrease performance.

In order to reduce the complexity of the task and ensure homogeneity of results, we choose to split the dual-task into two separate steps. Both steps are designed to be processed by the crowd, keeping the manual work at a minimum. The first step serves to identify relevant areas using a grid approach, whereas the second step aims to capture the tree outlines with precise polygons. Those two separated steps both consist of a clear task definition and adequate granularity, making it possible to conduct quality control through task design (ALLAHBAKHSH et al. 2013; ZHANG et al. 2016). Both steps also include a simple preparatory test which workers must pass. We hope to prevent insufficient effort responding through this method, following the recommendation of CHEUNG et al. (2017).

3 Methodology

3.1 Processing pipeline

The identification of relevant areas containing trees is realized by placing a grid consisting of square-shaped grid cells over the image data. Users are then asked to annotate each grid cell individually, having only the options to choose between “tree” and “no tree” – effectively performing a binary segmentation. This principle can be seen in Figure 2a. The shown 8×8 grid is chosen for demonstration purposes only. For real use cases, the cell size should be chosen much smaller, resulting in more grid elements (e.g. 50×50).

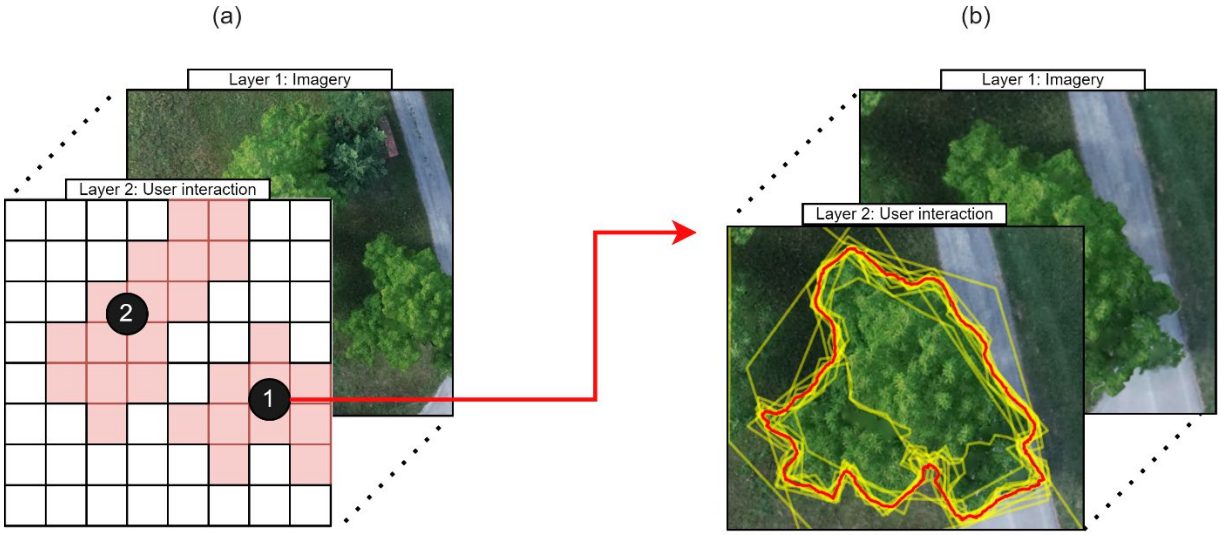

Fig. 2: Used two-step approach. (a) Step 1: Identification of relevant areas, (b) Step 2: Acquisition of precise tree outlines

Since we are using multiple labeling and are working on a binary classification, we can integrate our results using majority voting, a simple but effective method, as long as more than half of the workers provide correct results (ZHANG et al. 2016).

The integrated result is a downsampled segmentation map, which can be used for further processing. With this approach, we not only save costs and time, but also further simplify the task. Downsampling is possible since we are only interested in coarse tree positions and no pixel-precise segmentation is required for this approach. The degree of downsampling depends on the selected cell size.

Using the generated, downsampled segmentation map, tree clusters can easily be identified as closed areas. Each cluster can contain either a single tree or an area consisting of multiple trees. We are only interested in areas containing single trees, since we want to reduce complexity for the next task, following the suggestions of FINNERTY et al. (2013) to keep tasks simple. Therefore, a method for the distinction of areas containing only single trees from those containing multiple trees is required. Although there is no general solution for this, it is possible to draw conclusions about the number of trees in a selected area based on boundary conditions such as perimeter, area

or aspect ratio. In the example in Figure 2a, two clusters are found (indicated by numbers 1 and 2). In Figure 2, cluster 1 is taken for further demonstration of our approach.
Following the identification of relevant cluster areas, smaller image tiles containing only a single relevant cluster, therefore containing only individual trees, are cut out. These image tiles are then used as input for the second step, where precise tree outlines are annotated by means of polygons (see Figure 2b).

3.2 Integration of raw acquisitions

Multiple polygons of different quality are collected per tree, as can be seen in Figure 2b. Since only a single polygon per tree is desired as output, an integration operation must be performed. While the integration of classifications or even simple geometric shapes such as circles or rectangles can be easily solved, there is no trivial solution for more complex geometric shapes such as irregular polygons.
However, a fairly simple strategy can be used to integrate the polygons: A majority vote for each pixel of the RGB input data. Since users were annotating the polygons on pixel-level from the RGB data, we can convert all collected polygons to raster data. If the cell size of the raster representation is chosen to match the pixel size, there is only a small loss in information, such as the original positions of vertices. After the raster conversion, a pixelwise integration based on a binary vote can be performed: The number of different user labels for each raster cell is counted. If that number falls below a certain threshold, the pixel is omitted from the integrated raster shape, otherwise this pixel is included in the integrated raster shape.
This approach is shown in a very simplified form in Figure 3. For this figure, we chose the value 8 as the threshold for the binary voting, since this corresponds to an absolute majority for n=15, i.e., the number of crowdworkers who annotated one tree. As a result, any pixel with 8 or more acquisitions belongs to the integrated raster shape (Figure 3b). The integrated raster shape (Figure 3c) is then converted back into a polygon geometry, providing us with a tree outline polygon acquired through crowdsourcing.

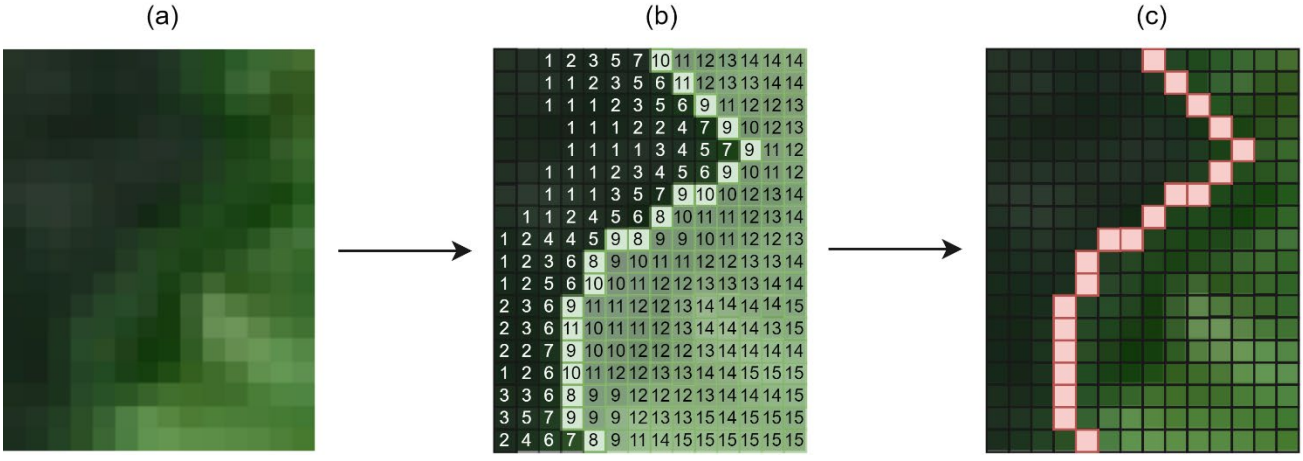

Fig. 3: Visualization of raster integration. (a) Section of image tile, (b) Majority vote on pixel basis with highlighted threshold areas, (c) Integrated polygon outline

4 Dataset

Our approach was tested for a designated test area near the town of Erligheim in southwestern Germany. The area mainly contains cherry orchards and covers an area of about 50,000 m². The imagery was taken using a DJI FC6310R camera together with a DJI Phantom 4 RTK. The flight height was 100 m and the flight was conducted in July 2022 at about 10 a.m. under slightly cloudy skies. The raw imagery was processed using Agisoft Metashape in order to generate an orthomosaic, with a resulting GSD of approx. 2.8 cm. Border areas of the orthomosaic were cut out due to distortions, resulting in an orthomosaic of 8000×8000 pixels (around 224×224 m). A small section of the used dataset can be seen in Figure 4. For later processing, the orthomosaic is divided into 60 image tiles, each consisting of 1000×1000 pixels (approx. 28×28 m).

Fig. 4: Section of the test area near Erligheim, mainly consisting of cherry orchards

5 Experiments

Two crowd campaigns were performed, one for each work step. Users were only able to participate once per campaign.

5.1 First Crowd Campaign

Each of the 60 image tiles in our dataset was presented to 6 different crowdworkers, resulting in a total of 360 crowd acquisitions. Users were paid $0.15 per 3 acquisitions, for a total cost of $18 for the whole dataset, or $0.30 per image tile.

5.2 Second crowd campaign

Processing the first crowd campaign led to a total of 125 individual trees that were detected. Smaller image sections, only containing these single trees, were created. These were presented to 15 different crowdworkers per tree, resulting in 1,875 total acquisitions at a cost of $0.10 per 3 trees, for a total cost of $62.5, or $0.50 per tree.

6 Evaluation

6.1 First crowd campaign: Coarse tree positions

Ground truth and integrated values can be compared directly for all grid cells. No ground truth was available, therefore reference data were collected by experts within our institute.
All raster elements can be divided in the categories “true positive” (TP), “false positive” (FP), “true negative” (TN) and “false negative” (FN), see OLSEN & DELEN (1997). Using these categories, the values for precision, recall and F1 score can be calculated for a quantification of quality of the integrated classification. Table 1 shows these results, both before and after the integration.

Table 1: Mean precision, recall and F1 score of all image tiles before and after integration

	Before integration	After integration
Mean Precision	87,73%	89,01%
Mean Recall	66,86%	85,25%
Mean F1	79,11%	87,73%

The relatively good average precision of around 88% remains mostly unchanged through the integration. We do only see little improvement, but, more importantly, there is no decrease either. On the other hand, the rather low recall value of around 67% increases up to 85%, resulting in a significantly improved F1 score overall. Thus, we can conclude that a multi-fold acquisition combined with element-wise majority voting may only have a small effect on worker accuracy, but can significantly increase data completeness. Figure 5 shows the distribution of F1 scores before and after the integration and illustrates the impact of integration on the F1 score.

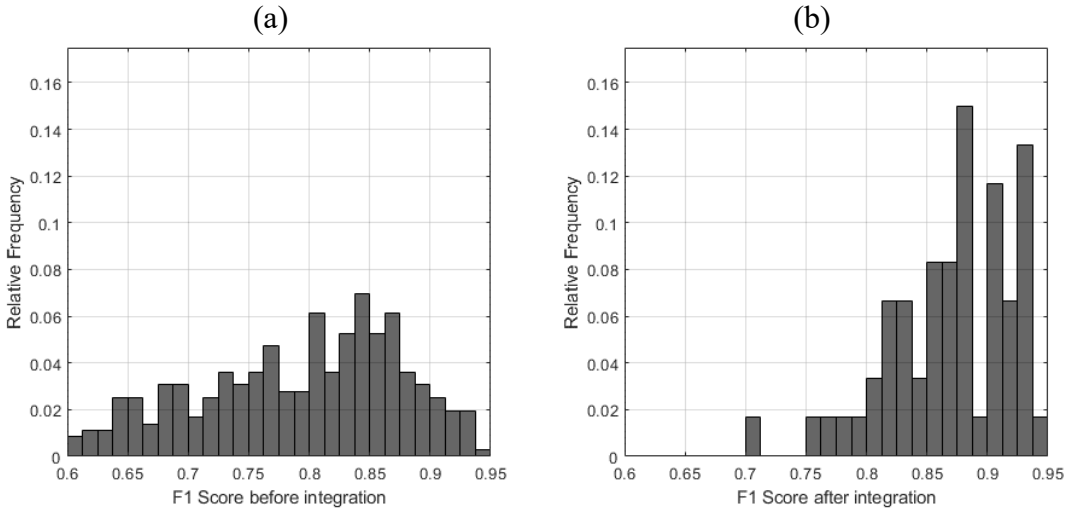

Fig. 5: Distribution of F1 scores. (a) Before integration and (b) after integration

6.2 Second crowd campaign: Precise tree outlines

For the evaluation of the polygon acquisitions, a ground truth is required as well. Since no independent reference data were available here either, we again had to rely on an expert-generated ground truth.

For each polygon acquisition, we calculated the Jaccard index to the respective ground truth. The Jaccard index calculates the similarity between two samples by calculating the Intersection over Union (JACCARD 1901), with 1 being a perfect similarity and 0 being no similarity at all. The mean calculated Jaccard index of all crowd acquisitions before any integration measures was 0.691, indicating an average similarity of 69.1% between acquisition and reference. This relatively low value can be attributed to the fact that no prior filtering of noisy data was performed, and crowd acquisitions in general tend to be noisy, visible in Figure 6b.
The integration approach described in section 5.2 is performed with all 125 trees of the dataset. An example can be seen in Figure 6, including all raw acquisitions (Figure 6b) and the integrated result (Figure 6c), using 8 as the binary vote threshold.

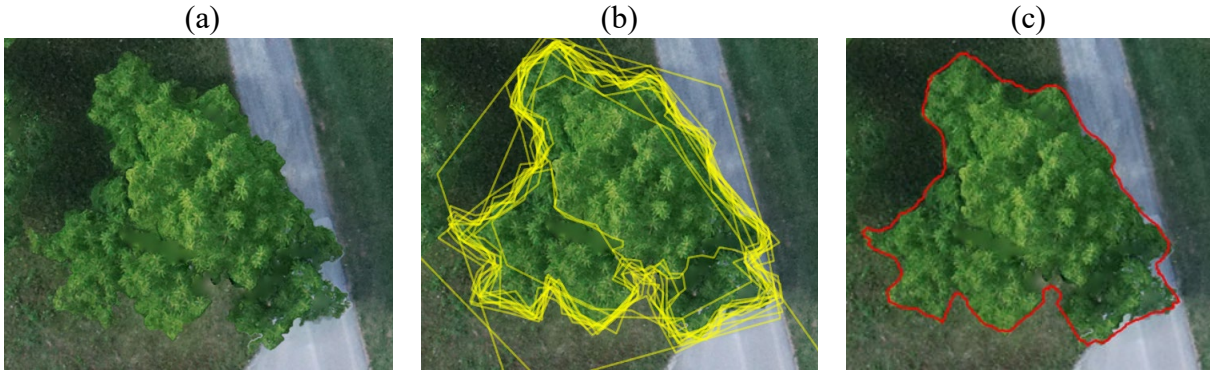

Fig. 6: Integration steps. (a) Raw image, (b) 15 Crowd acquisitions, (c) Integrated result

Figure 6 illustrates the effectiveness of our approach. While Figure 6b contains multiple noisy acquisitions, the integrated result consists of a clear outline, apparently free of noise, as can be seen in Figure 6c.
The Jaccard index can be calculated for all integrated polygons and then compared to the pre-integration values per tree, directly indicating the quality improvement achieved.
However, before integration, the binary vote threshold needs to be defined. Figure 3 in section 3.2 suggests that the integrated result is dependent on the threshold value. To show the influence of this binary vote threshold, we calculated all integrations for all 125 trees with all 15 possible values, their respective Jaccard index compared to the ground truth per tree, and the mean values. The result is shown in Figure 7.
As can be seen, best results are obtained with a threshold value of 7 with a resulting IoU of 0.834, compared to an IoU value of 0.691 without any integration measures. In this case, picking a value lower than 3 or higher than 11 will result in worse quality than not integrating the results at all. This can be explained in a fairly straightforward manner: Choosing the threshold value too low will include more acquisitions for the calculation of the integrated shape, even noisy ones, resulting in a decline in quality due to an increase of size of the integrated polygons: More grid cells outside the actual tree geometry are assigned to belong to the integrated polygon. Choosing the threshold too high on the other hand will exclude those grid cells not labeled by a vast majority of workers, leaving only a small pool of grid cells and resulting in a decline in quality due to an effect similar

to an erosion. The presumption that the quality of the integrated result is dependent on the picked binary vote threshold can therefore be confirmed.

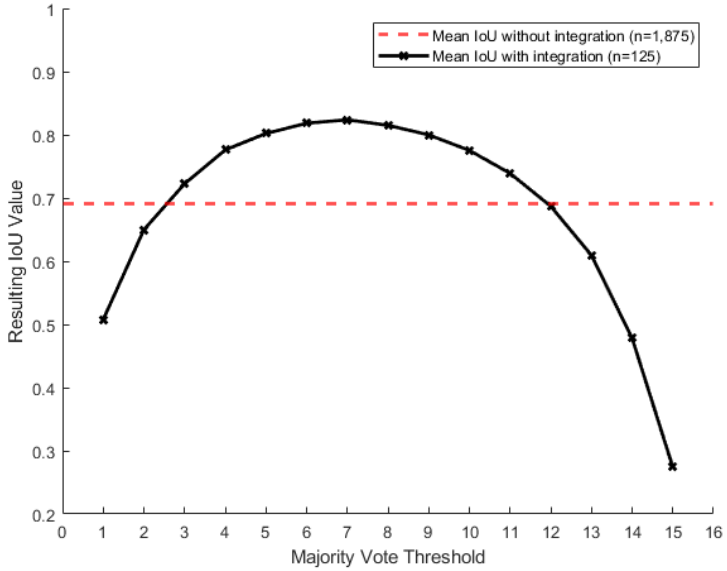

Fig. 7: Mean intersection over union (IoU) values of all trees for different binary vote thresholds

In our case, best IoU values are achieved with a vote threshold of 7, indicating the best improvement through integration. Not only does the mean IoU value increase by around 14 percentage points from 0.691 to 0.834 through the integration, the standard deviation of all observed values also reduces from a value of 0.198 before integration to a value of 0.119, equaling a reduction of around 40%. The distribution of the data values improves by a large margin, as can be seen in the histogram in Figure 8.

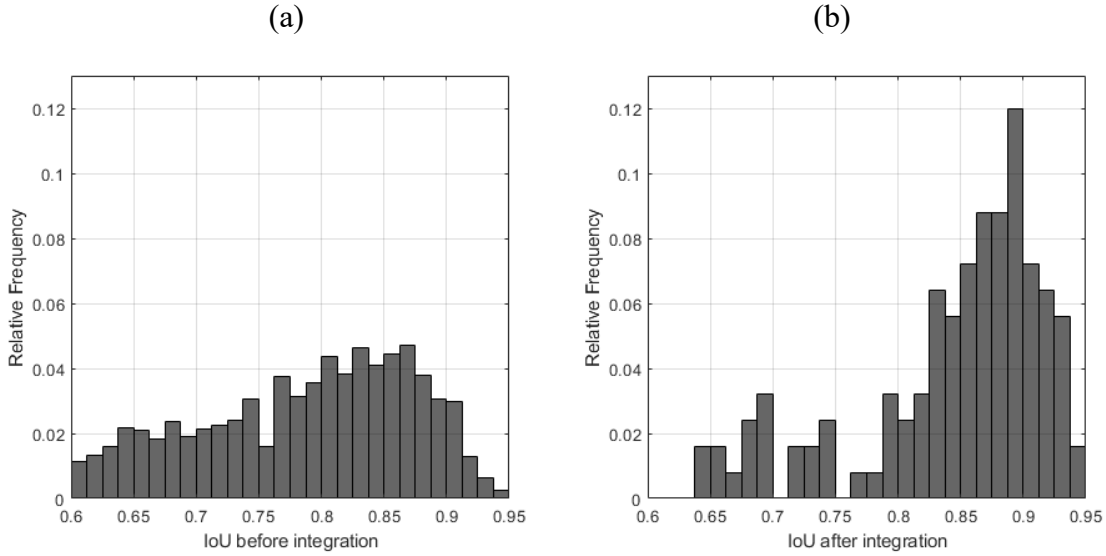

Fig. 8: Distribution of IoU values. (a) Before integration and (b) after integration

7 Conclusion

Using majority voting as integration measure, we were able to acquire a high-quality tree segmentation map with an average F1 score of nearly 88%, confirming that noisy data can be improved through repeated acquisitions with subsequent integration. Furthermore, we were able to prove this concept to be true even for irregular polygons when using a raster integration approach. The integrated polygons improve in quality, which can be confirmed not only visually, but also by Intersection over Union values.

The integrated polygons consist of a vast number of vertices and have complex geometries even when converted back to vector data. If smooth data are desired, line smoothing operations could be considered for cleaning up geometries. The high number of vertices could lead to either a more accurate geometry, since more details may be included, or a reduced accuracy due to higher noise. Further observations are required in order to investigate the influence of vertices number on the general geometry and to determine if there is a relationship with the binary vote threshold parameter.

8 Acknowledgement

Partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2120/1 – 390831618.

9 References

ALLAHBAKHSH, M., BENATALLAH, B., IGNJATOVIC, A., MOTAHARI-NEZHAD, H. R., BERTINO, E. & DUSTDAR, S., 2013: Quality control in crowdsourcing systems: Issues and directions. IEEE Internet Computing, **17**(2), 76-81.

CHANDLER, J., PAOLACCI, G. & MUELLER, P., 2013: Risks and rewards of crowdsourcing marketplaces. Handbook of human computation, Springer Verlag, New York, 377-392.

CHEUNG, J. H., BURNS, D. K., SINCLAIR, R. R. & SLITER, M., 2017: Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. Journal of Business and Psychology, **32**(4), 347-361.

FINNERTY, A., KUCHERBAEV, P., TRANQUILLINI, S. & CONVERTINO, G., 2013: Keep it simple: Reward and task design in crowdsourcing. Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI, 1-4.

HADI H., YOWARGANA, P., ZULKARNAIN, M. T., MOHAMAD, F., GOIB, B.K., HULTERA, P., STURN, T., KARNER, M., et al., 2022: A national-scale land cover reference dataset from local crowdsourcing initiatives in Indonesia. Scientific Data, **9**(1), 1-15.

HIRTH, M., HOßFELD, T. & TRAN-GIA, P., 2011: Anatomy of a crowdsourcing platform-using the example of microworkers.com. 2011 Fifth international conference on innovative mobile and internet services in ubiquitous computing, 322-329.

HOSSAIN, M., 2012: Users' motivation to participate in online crowdsourcing platforms. 2012 International Conference on Innovation Management and Technology Research, 310-315.

JACCARD, P., 1901: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat, **37**, 547-579.

OLSON, D. L. & DELEN, D., 2008: Advanced Data Mining Techniques, Springer Verlag, 1st edition, 138.

PILZ, D. & GEWALD, H., 2013: "Does money matter? Motivational factors for participation in paid- and non-profit-crowdsourcing communities.". Proceedings of the 11th International Conference on Wirtschaftsinformatik (WI), **1**, 577-592.

PUTTINAOVARAT, S., SAELIW, A., PRUITIKANEE, S., KONGCHAROEN, J., CHAI-ARAYALERT, S. & KHAIMOOK, K., 2022: Flood Damage Assessment Geospatial Application Using Geoinformatics and Deep Learning Classification. International Journal of Interactive Mobile Technologies, **16**(21).

RYAN, R. M. & DECI, E. L., 2000: Intrinsic and extrinsic motivations: Classic definitions and new directions. Contemporary educational psychology, **25**(1), 54-67.

SARALIOGLU, E. & GUNGOR, O., 2020: Crowdsourcing in remote sensing: A review of applications and future directions. IEEE Geoscience and Remote Sensing Magazine, **8**(4), 89-110.

SARALIOGLU, E. & GUNGOR, O., 2022: Crowdsourcing-based application to solve the problem of insufficient training data in deep learning-based classification of satellite images. Geocarto International, **37**(18), 5433-5452.

STONEBRAKER, M. & REZIG, E., 2019: "Machine learning and big data: What is important?". IEEE Data Eng. Bull., **42**.4, 3-7.

WALTER, V., KÖLLE, M. & YIN, Y., 2020: Evaluation and optimisation of crowd-based collection of trees from 3D point clouds. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, **4**, 49-56.

WALTER, V., KÖLLE, M., COLLMAR, D. & ZHANG, Y., 2021: A two-level approach for the crowd-based collection of vehicles from 3D point clouds. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, **4**, 97-104.

WHANG, S. E. & LEE, J. G., 2020: Data collection and quality challenges for deep learning. Proceedings of the VLDB Endowment, **13**(12), 3429-3432.

YUEN, M. C., KING, I. & LEUNG, K. S., 2012: Task recommendation in crowdsourcing systems. Proceedings of the first international workshop on crowdsourcing and data mining, 22-26.

ZHANG, J., WU, X. & SHENG, V. S., 2016: Learning from crowdsourced labeled data: a survey. Artificial Intelligence Review, **46**(4), 543-576.