Crowd-sourced information on building façades - A comparative study on the use of commercial and non-commercial crowdsourcing platforms

Robert Hecht
Leibniz Institute of
Ecological Urban and
Regional Development,
Dresden, Germany
r.hecht@ioer.de

Tim Wendt
Virtual City Systems
Berlin, Germany
twendt@virtualcitysystems.de

Martin Behnisch
Leibniz Institute of
Ecological Urban and
Regional Development
Dresden, Germany
m.behnisch@ioer.de

Abstract

The aim of the paper is to investigate the potential of crowdsourcing in order to obtain information on buildings and their façade properties, especially the window and door area ratio, based on given street view imagery. A commercial and a non-commercial crowd-sourcing platform is tested and the quality of the derived information is assessed in terms of accuracy using reference data. For this purpose, images of building facades together with corresponding tasks was provided on the commercial platform Amazon Mechanical Turk and a non-commercial platform Zooniverse. The results indicate qualitative differences between using commercially and non-commercially platform. A systematic underestimation of the area usable for BIPV could be observed by using for both platforms.

Keywords: Geocoded images, building façades, BIPV, energy systems, crowdsourcing, data quality

1 Introduction

In the context of the energy transition detailed information on buildings play an important role in assessing the suitability for the installation of renewable energy systems. There are already well-developed workflows for deriving the potentials for equipping roofs with photovoltaic (PV). However, the façade surfaces are also suitable for an architectural integration of PV so-called building-integrated photovoltaics (BIPV). Compared to other European countries the potential of BIPV in Germany is only minimally exploited (Montoro et. al. 2009). Therefore, an appropriate information basis must be created in order to determine the potential of the buildings.

3D building data provide only a rough assessment of the potential for BIPV, as the proportion of windows and doors is typically unknown. If the mean proportion of a certain category of building were known, it would be possible to make more precise overall estimates (e.g. absolute estimates in kWh/m²).

On the other hand, façade information is implicitly given in a variety of geocoded images in social media, commercial service providers and photo sharing platforms (FlickR, GoogleStreetView, Mapillary, Wikimapia etc.). If this information could be made explicit, at least the potential would be known for some of the buildings. One way to extract this information from images is crowdsourcing, which has already proven itself in many fields of application as an appropriate method of obtaining information from images. One example is the identifiation of blood cells affected by malaria (Mavandadi et al. 2012), the estimation of land cover based on Google Earth images (See et al. 2013) or even the collection of information on buildings based on street view images (Hecht et al. 2017). Current developments show that crowdsourcing can also go beyond classification (annotation, labeling). Hillen and Höfle (2015), for example, developed a prototype called Geo-reCAPTCHA for digitizing building footprints. Related work in connection with the extraction and use of 3D or building information are described in Goetz and Zipf (2013) and Herfort et al. (2018).

In our study we want to test the use of crowdsourcing platforms for the acquisition of façade information using 3D buildings and geocoded street images. Thus we look at how exactly the building façade information can be captured by crowdsourcing and whether there are differences between crowdsourcing platforms that provide for financial rewards (commercial) and non-financial rewards (non-commercial).

2 Crowdsoucing and Platforms

2.1 Crowdsourcing

Geographic information captured by citizens is playing an increasingly important role. In recent years, various terms have emerged from various disciplines that describe this process, namely crowdsourcing, citizen science, collaborative mapping or crowd-sourced information itself, such as Volunteered Geographic Information (VGI) or user-generated content (UGC). The form of data collection can be very different. According to See et al (2016), crowdsourced data can be captured either actively as part of a crowdsourcing system/campaign (e.g. OpenStreetMap, Wikimapia) or passively by mapping existing crowdsourced data collected for other purposes (e.g. mobile data, location-based social media content). Furthermore, the types of information (e.g. labels vs. geometry) or the forms of motivational strategies (gamification, financial rewards etc.) can vary.

In our context, we prefer the term crowdsourcing as a kind of participatory online activity, in particular the process of a voluntary undertaking of specific tasks (Estellés-Arolas and Ladrón-de-Guevara 2012). Crowdsourcing first appeared in Howe (2006) describing a business practice, which is now an attractive way to efficiently collect data from non-specialists over the Internet that are almost as good as expert labels (Snow et al. 2008). The idea of using online users for the purpose to label images goes back to Luis von Ahn who designed the ESP game and further developed reCAPTCHA (von Ahn et al. 2008), a system to verify humanity and simultaneously assisting the digitization of books by solving complex OCR problems with crowdsourced labels.

2.2 Crowdsourcing Platforms

Today there is a large number of different crowdsourcing platforms on which clients (company, organization, group or individual persons) can publish a tasks to be solved and on the other hand potential contractors (crowd) can solve this task. Basically, the platforms can be divided into commercial and non-commercial platforms. In the case of commercial platforms, users are rewarded on a monetary basis for completed tasks, thus providing additional incentives for task processing. Other motivating factors play a role in non-commercial platforms, such as thematic interest, entertainment, recognition or charity. Prominent examples for commercial crowdsourcing platforms are Amazon Mechanical Turk, Clickworker or Crowdflower. Examples of non-commercial crowdsourcing platforms are Crowdcrafting, Zooniverse, or Samasource.

A commercial and a non-commercial crowdsourcing platform were selected for the experimental study. These are presented in more detail in the following sections.

2.2.1 Amazon Mechanical Turk

Amazon Mechanical Turk (MTurk) was created by Jeff Bezos (Pontin 2007) in 2005 and is one of the most popular commercial crowdsourcing platforms on which users (workers) are motivated by financial rewards. Employers can publish microjobs (so-called Human Intelligence Tasks (HITs)) and design the framework conditions at their own discretion. On the other hand, workers can then research existing jobs and carry them out for a cash benefit determined by the employer. The employer can define how much time the completion of the task may take, how long the project is active and how much money the worker receives for the completion of a task (Chen et al. 2011). According to Buhrmester et al (2011) MTurk offers all necessary elements for successful implementation of crowdsourcing projects. The strengths of this platform are the open market design and a large number of workers.

2.2.2 Zooniverse

Zooniverse was chosen as a non-commercial citizen science platform where voluntary users are motivated by social incentive mechanisms in order to contribute to different research projects. The Zooniverse platform that exists today has evolved from the “Galaxy Zoo” project in which more than 75,000 volunteers participated. The data obtained were used for 50 publications (Smith et al. 2013). As a result of the popularity of this first project, Zooniverse has become one of the most important platform for different research domains (biology, history, art, medicine, etc.) and counts more than 1.1 million registered volunteers around the world (Cox et al. 2015).

3 Method

The following figure 1 shows a schematic representation of the examination approach. There is a fundamental differentiation between the activities of the project manager (preparation and statistical analysis) and the interpretation carried out by the users (crowdsourcing). In the preparation phase, the image data sets are generated, the task design is defined and implemented and the reference data required for validation is collected. Following crowdsourcing, statistical analysis is carried out using descriptive statistics and external validation.

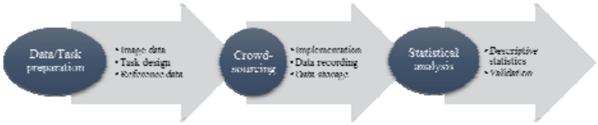


Figure 1: Workflow

3.1 Data/Task preparation

3.1.1 Image data

In this step, the image data is acquired and prepared for further processing. In our experimental study we had access to an internal image database of non-residential buildings from a research project (ENOB:dataNWG). The database contains façade photos of non-residential buildings of different usages from the federal states of North Rhine-Westphalia and Thuringia. Each building has a usage category and opens up the possibility of a differentiated consideration of BIPV potentials. The database contains several images taken from different viewing angles. For this reason, the images were preselected manually. The most important criterion for the selection was a complete and (in the case of several buildings in the picture) unambiguous representation of the façade. For data protection reasons, the images were subsequently processed and all license plates, faces and other indications of the citizens' place of residence have been blurred. The enhanced images could then be integrated into the commercial and non-commercial crowdsourcing platform.

3.1.2 Task Design

The task design is an important step, since the instructions and the user interface have an influence on user´s perception of the task and thus influence the quality of the results (Finnerty et al. 2013). The task is therefore to be kept as simple as possible in order to ensure good solubility regardless of the origin and culture of the users.

The aim is determine the potential of the façade area for BIPV. The BIPV potential of a façade is usually the windowless and doorless part of the façade. Users are therefore asked to estimate the proportion of window and door areas in %. This was achieved by querying the window and door area proportion via response options in 10 % steps (selection buttons). In order to increase the quality of the information and to measure the dispersion among the users, each image is interpreted by 10 different users.

However, the user could also indicate that an interpretation was not possible. These cases were not taken into account in the statistical evaluation of the results. The task was implemented together with instructions on the crowdsourcing platforms MTurk and Zooniverse (see user interface in Figure 3 and 4).

3.1.3 Reference data

For data validation, reference values has been gathered using Adobe Photoshop by determining the number of pixels of the window and doors in relation to the total number of pixels in the total façade area. This procedure was performed for a sample of 25 % of the image data set (186 images in total). Figure 2 shows an example of reference data collection. In a first step, the total façade area was determined by defining a binary mask. In a second step, the windows and door areas were erased from the binary mask.

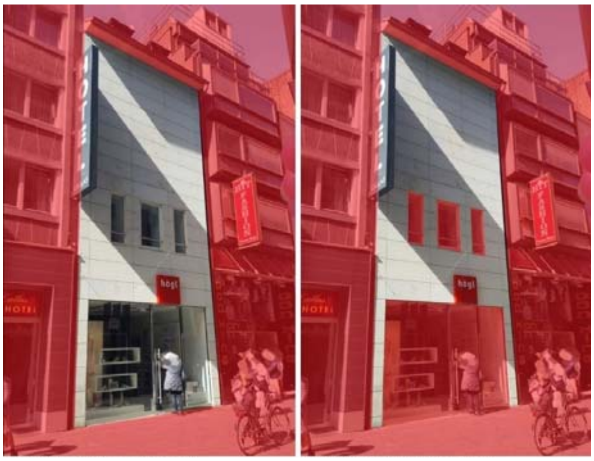
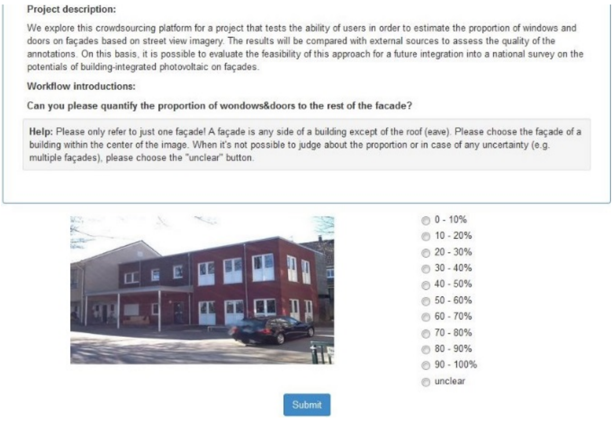


Figure 2: Reference data collection: façade area (left) and façade without windows and doors (right).

The reference values were stored in a database and serve as a basis for comparison during the validation phase. It should be noted that geometric image distortions due to perspective and acquisition system can lead to small deviations of the measured reference values from reality during the reference data collection, which are, however, negligible in the context of the study.

3.2 Crowdsourcing

In this crowdsourcing step, the actual recording and storage of the data takes place on the basis of the tasks implemented on the respective platforms. Both platforms offer a web interface to implement the tasks and design the user interfaces (Figure 3 and 4). The Zooniverse platform requires the Zooniverse community to be convinced of the project. To attract as many volunteers as possible, the project must be convincing, easy to understand, and transparent. In addition, the project should have appropriate project supporters in order to increase its credibility, relevance and dissemination. Only after a successful test phase and approval a project can be officially launched. In our case the Zooniverse project was in a test mode only. According to Zooniverse, the number of images was too small to become an official Zooniverse project. However, this mode was sufficient to assess the quality of the classifications based on a testing crowd.



Project description:

We explore this crowdsourcing platform for a project that tests the ability of users in order to estimate the proportion of windows and doors on façades based on street view imagery. The results will be compared with external sources to assess the quality of the annotations. On this basis, it is possible to evaluate the feasibility of this approach for a future integration into a national survey on the potentials of building-integrated photovoltaic on façades.

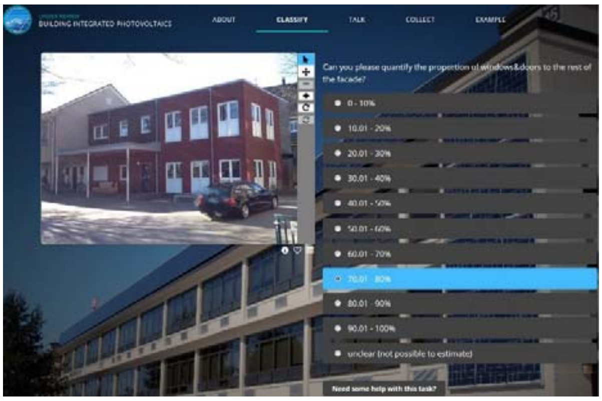
Workflow introductions:

Can you please quantify the proportion of wondows&doors to the rest of the facade?

Help: Please only refer to just one façade! A façade is any side of a building except of the roof (eave). Please choose the façade of a building within the center of the image. When it's not possible to judge about the proportion or in case of any uncertainty (e.g. multiple façades), please choose the "unclear" button.

○ 0 - 10%
○ 10 - 20%
○ 20 - 30%
○ 30 - 40%
○ 40 - 50%
○ 50 - 60%
○ 60 - 70%
○ 70 - 80%
○ 80 - 90%
○ 90 - 100%
○ unclear

Figure 3: Implemented user interface (MTurk)



BUILDING INTEGRATED PHOTOVOLTAICS
ABOUT
CLASSIFY
TALK
COLLECT
EXAMPLE

Can you please quantify the proportion of windows&doors to the rest of the facade?

● 0 - 10%
● 10.01 - 20%
● 20.01 - 30%
● 30.01 - 40%
● 40.01 - 50%
● 50.01 - 60%
● 60.01 - 70%
● 70.01 - 80%
● 80.01 - 90%
● 90.01 - 100%
● unclear (not possible to estimate)

Need some help with this task?

Figure 4: Implemented user interface (Zooniverse)

3.3 Deskriptive Statistics

After the crowdsourcing process, descriptive statistics are calculated that provide an initial overview. These are the number of images, number of categories, number of annotations, number of annotators, annotations per image, and the number of annotations per annotator.

3.4 Validation

In the final validation step, the quality of crowd-sourced information is measured and evaluated. For this purpose, the mean value of all responses (estimated window and door area proportions) of an image was calculated and compared with the respective reference value. This allows the errors to be quantified and investigated.

4 Results

4.1 Descriptive Statistics

Table 1 gives an overview of the descriptive statistics of the data obtained with MTurk and Zooniverse. With MTurk all annotations could be recorded within 59 minutes. With Zooniverse the required number of 10 annotations per image could not be reached after 50 days. This is due to the fact that the project ran in test mode with a limited number of users (testing community). According to the Zooniverse community, the desired 10 annotations per image would have been achieved in real operation mode. Significantly more users (annotators) took part in the Zooniverse project, but each user carried out only 8 annotations on average. MTurk's commercial platform generated a larger number of approximately 56 annotations per annotator. The financial incentive seems to lead to users specialising and doing the same job several times.

	MTurk	Zooniverse
Duration	59 min	50 days
No. of images	743	743
No. of categories	11	11
No. of annotations	7430	3808
No. of annotators	132	476
Annotations per image (mean)	10,0	5,1
No. of annotations per annotator (mean)	56,3	8,0

Table 1: Descriptive statistics of the data from MTurk and Zooniverse

4.2 Validation

Table 2 shows the results of the external validation. The results of the MTurk users are compared with the Zooniverse users based on the mean values for the percentage of window and door area. In addition, the mean difference and the standard deviation (SD) of the differences are presented. The SD values document the dispersion around the mean value.A direct comparison of the scatterplots show the differences between MTurk and Zooniverse (Figure 5 and 6). The mean values of the responses (windows and door area ratio) in % of each image are plotted against the corresponding reference values in %. A fitted regression line shows the linear relationship. As a result, the relationship between the reference value and the estimated value from crowdsourcing is stronger in the case of Zooniverse.

	MTurk	Zooniverse
Mean (reference in %)	27,16	27,16
Mean (crowd in %)	57,36	41,80
Difference (crowd - reference) in p.p.	30,20	14,64
Difference (SD)	13,21	10,42

Table 2: Results of the external validation (windows and door area ratio in %)

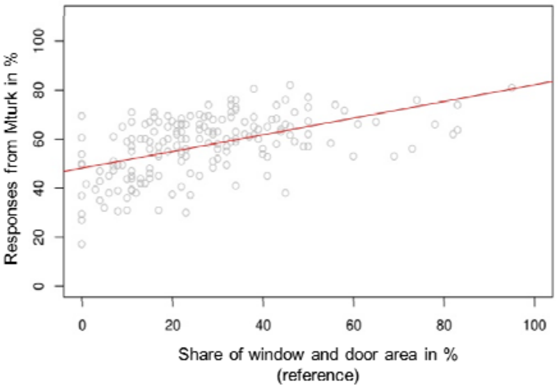


Figure 5: Crowdsourced data from MTurk and reference data

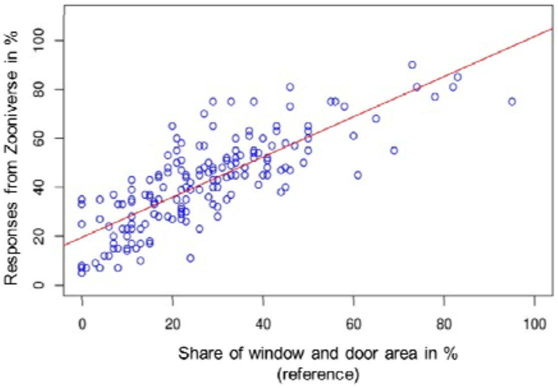


Figure 6: Crowdsourced data from Zooniverse and reference data

A systematic overestimation of the window and door area ratio of the façades could be observed for both platforms. Figure 7 and 8 show the positive and negative deviations based on the annotator responses and the calculated mean values of each image. However, the higher the proportion of windows and doors on the façades, the smaller the deviation. This peculiarity could be observed on both platforms, whereby the deviation is smaller using Zooniverse. This systematic correlation represents a new, useful insight and is suitable for the development of a correction model.

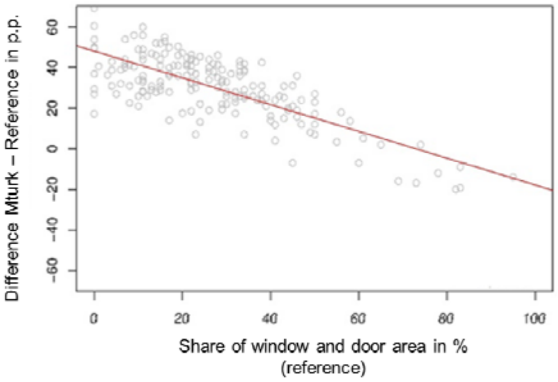


Figure 7: Deviations of the responses using Mturk

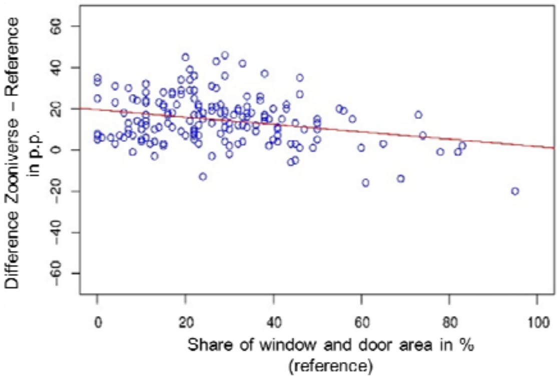
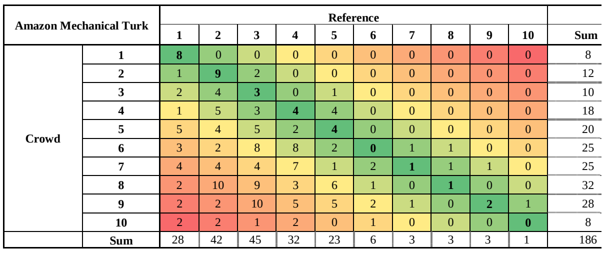


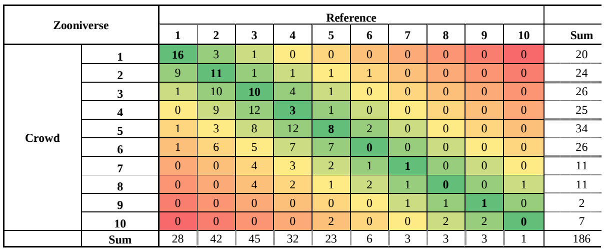
Figure 8: Deviations of the responses using Zooniverse

The confusion matrix can be used to show the magnitude of the deviations of the answers from the reference in %. In a previous step, the values of the deviations (in %) were discretized into 10 classes (class "1": 0 – 10%, class "2": 10 – 20% etc.). Figure 9 and 10 show the magnitudes using MTurk and Zooniverse. Here, too, it can be observed that much more misassignments occur when using the MTurk platform in contrast to Zooniverse.



Amazon Mechanical Turk		Reference										
		1	2	3	4	5	6	7	8	9	10	Sum
Crowd	1	**8**	0	0	0	0	0	0	0	0	0	8
	2	1	**9**	2	0	0	0	0	0	0	0	12
	3	2	4	**3**	0	1	0	0	0	0	0	10
	4	1	5	3	**4**	4	0	0	0	0	0	18
	5	5	4	5	2	**4**	0	0	0	0	0	20
	6	3	2	8	8	2	**0**	1	1	0	0	25
	7	4	4	4	7	1	2	**1**	1	1	0	25
	8	2	10	9	3	6	1	0	**1**	0	0	32
	9	2	2	10	5	5	2	1	0	**2**	1	28
	10	2	2	1	2	0	1	0	0	0	**0**	8
	Sum	28	42	45	32	23	6	3	3	3	1	186

Figure 9: Confusion matrix using MTurk



Zooniverse		Reference										
		1	2	3	4	5	6	7	8	9	10	Sum
Crowd	1	**16**	3	1	0	0	0	0	0	0	0	20
	2	9	**11**	1	1	1	1	0	0	0	0	24
	3	1	10	**10**	4	1	0	0	0	0	0	26
	4	0	9	12	**3**	1	0	0	0	0	0	25
	5	1	3	8	12	**8**	2	0	0	0	0	34
	6	1	6	5	7	7	**0**	0	0	0	0	26
	7	0	0	4	3	2	1	**1**	0	0	0	11
	8	0	0	4	2	1	2	1	**0**	0	1	11
	9	0	0	0	0	0	0	1	1	**1**	0	2
	10	0	0	0	0	2	0	0	2	2	**0**	7
	Sum	28	42	45	32	23	6	3	3	3	1	186

Figure 10: Confusion matrix using Zooniverse

4.3 Discussion

The work presents initial insights into the differences of crowd-sourced information on building façades using two crowdsourcing platforms that are very different in nature. In our experiment, obviously better data could be obtained with the non-commercial platform Zooniverse than with MTurk. However, it remains to be clarified whether the quality is the same in the operative mode (full Zooniverse project). In our study, only a Zooniverse testing community was available, which may have more experience and thus the quality of Zooniverse could be overestimated.

With the financial incentives of the commercial platform MTurk delivers extremely fast results. In our experiments we paid the crowd per annotation. Mao et al. (2013) showed that with well-designed financial incentive mechanisms one is able to trade quality for speed. Further optimization of the payment mechanism could therefore lead to an improvement in results using a commercial platform.

5 Conclusion and Outlook

Crowdsourcing has great potential for the efficient annotation of georeferenced image data. In this paper, the use of crowdsourcing was tested for estimating the share of window and door area on façades using commercial and non-commercial crowdsourcing platforms. The implemented task was the same for both platforms. The following main findings were obtained from the comparison with reference data:

- Higher Accuracies of the crowdsourced information could be observed using the non-commercial platform Zooniverse compared to the commercial platform MTurk
- A systematic overestimation of the window/door area could be observed for both platforms.
- The deviations of the measured ratio decreases with increasing window/door area portion.

The results allow an initial assessment of the ability of crowdsourcing to derive facade information. The next step should be to examine and find out the causes and explanations for the differences. For this purpose, the contributors must be examined in more detail and further investigations and analyses needs to be done. On the one hand, this concerns further investigations related to the causes of misinterpretations, e.g. due to poor quality of the visual material, didactics of the task, unclear definition of the façade, background and qualification of the users. It should also be checked whether the building types plays a role and, for example, an estimation of the photovoltaic potential only makes sense for certain categories.

Furthermore, it should be investigated whether the quality of the results can be improved by means of a post-processing using filtering techniques. For example, annotators with low trustworthiness could be excluded, thus improving the quality of the information. For this purpose, suitable quality measures must be developed to quantify trustworthiness. In addition, a corresponding correction model could be developed based on the systematic correlation of window and door areas and the deviation from the reference value.

With regard to the image data used, the suitability of alternative data sources should be investigated in the future. For example, VGI platforms such as Wikimapia or Mapillary can be a good alternative data source. However, their suitability for use can be investigated in the future.

6 Acknowledgement

This experimental study was part of the project "Development of a prefabricated standard BIPV façade for selected building categories in Germany" (Standard-BIPV, Fkz: 0324063C) funded by the Federal Ministry of Economics and Technology (BMWi). We thank the colleagues for their support in implementing the project and for their valuable advice in developing the task design. We also like to thank our colleagues in the project "Primary data collection for recording the structure and energy quality of non-residential buildings in Germany" (ENOB:dataNWG, Fkz: 03ET1315), also funded by BMWi for providing the image data from the screening (pilot phase). Thanks also go to Prof. Burghardt from the Institute of Cartography (Dresden University of Technology) for his valuable advice and support. Finally, we thank Grant Miller and the Zooniverse testing community for the valuable feedback during the implementation.

References

Buhrmester, M., Kwang, T. and Gosling, S. D., 2011: Amazon’s Mechanical Turk: A New Source of inexpensive, Yet High-Quality, Data? Perspectives on Psychological Science **6** (1), 3-5.

Chen, J. J., Menezes, N. J. and Bradley, A. D., 2011: Opportunities for Crowdsourcing Research on Amazon Mechanikal Turk. Seattle.

Cox, J.; Oh, E. Y.; Simmons, B.; Lintott, C.; Masters, K.; Greenhill, A.; Graham, G.; Holmes, K., 2015: Defining and measuring success in online citizen science: A case study of Zooniverse projects. Computing in Science & Engineering **17**, 28-41.

Finnerty, A., Kucherbaev, P. Tranquillini, S. and Convertino, G. (2013). Keep it simple: Reward and task design in crowdsourcing. In *Proc. of SIGCHI*, 2014, ACM.

Goetz, M. & Zipf, A. (2013): The Evolution of Geo-Crowdsourcing: Bringing Volunteered Geographic Information to the Third Dimension. In: Sui, D., Elwood, S. & Goodchild, M. (eds.): Crowdsourcing Geographic Knowledge. Volunteered Geographic Information (VGI) in Theory and Practice. Springer, Dordrecht.

Hecht, R.; Kalla, M. & Krüger, T., 2017: Crowd-sourced data collection to support automatic classification of building footprint data. ICC 2017: Proceedings of the 2017 International Cartographic Conference.

Herfort, B., Höfle, B. & Klonner, C. (2018): 3D micro-mapping: Towards assessing the quality of crowdsourcing to support 3D point cloud analysis. ISPRS Journal of Photogrammetry and Remote Sensing **137**, 73-83.

Hillen, F. & Höfle, B., 2015: Geo-reCAPTCHA: Crowdsourcing large amounts of geographic information from earth observation data. International Journal of Applied Earth Observation and Geoinformation **40**, 29-38.

Howe, J., 2006: The rise of crowdsourcing. Wired Mag 14, 1–4.

Mao, A.; Kamar, E.; Chen, Y.; Horvitz, E.; Schwamb, M.E.; Lintott, C.J. & Smith, A.M., 2013: Volunteering Versus Work for Pay: Incentives and Tradeoffs in Crowdsourcing. In *Proc. of HCOMP ’13*.

Mavandadi, S.; Dimitrov, S.; Feng, S.; Yu, F.; Sikora, U.; Yaglidere, O.; Padmanabhan, S.; Nielsen, K. & Ozcan, A., 2012: Distributed Image Analysis and Diagnosis through Crowd-Sourced Games: A Malaria Case Study, PLoSOne **7** (5)

Montoro, D. F.; Vanbuggenhout, P. & Ciesielska, J., 2009: Building Integrated Photovoltaics: An overview of the existing products and their fields of application. Sunrise (Hrsg.)

Pontin, J., 2007: Artifical Intelligence, With Help From the Humans. The New York Times (March 25, 2007).

See, L.; Comber, A.; Salk, C.; Fritz, S.; Van der Velde, M.; Perger, C.; Schill, C.; McCallum, I.; Kraxner, F. & Obersteiner, M., 2013: Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts, PLoSOne **8** (7)

Smith, A. M., Lynn, S. & Lintott, C. J., 2013: An Introduction to the Zooniverse. AAAI Technical Report Cr-13-01. 103

Snow, R., O’Connor, B., Jurafsky, D. and Ng, A. J. (2008). Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), ACM Press, 254–263.

von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science, 321(5895), 1465–1468.