Analyzing Workers Performance in Online Mapping Tasks Across Web, Mobile, and Virtual Reality Platforms

Gerard van Alphen, Sihang Qiu*, Alessandro Bozzon, Geert-Jan Houben
Delft University of Technology
gerard@valphen.com, {s.qiu-1, a.bozzon, g.j.p.m.houben}@tudelft.nl

Abstract

In online crowd mapping, crowd workers recruited through crowdsourcing marketplaces collect geographic data. Compared to traditional mapping methods, where workers physically explore the area, the benefit of using online crowd mapping is the potential to be cost-effective and time-efficient. Previous studies have focused on mapping urban objects using street-level imagery. However, they are specifically aimed at a single type of object, and only through web platforms. To the best of our knowledge, there is still a lack of understanding on how workers perform the mapping tasks through different platforms. Aiming to fill this knowledge gap, we investigate the worker performance across web, mobile, and virtual reality platforms by designing a multi-platform system for mapping urban objects using street-level imagery with novel methods for geo-location estimation. We design a preliminary study to show the feasibility of executing online mapping tasks on three platforms. The result demonstrates that the type of task and execution platform can affect the worker performance in terms of worker accuracy, execution time, user engagement, and cognitive load.

Introduction

Gathering geographic data, such as location and height, of urban objects can be of great importance, for instance for municipalities having planning and maintenance needs. Traditional mapping approaches recruit municipal workers to document each individual object by physically observing the environment, with clear limitations in time and cost. The emergence of street-level imagery services leads to novel urban mapping approaches. Such services, like Google Street View and Mapillary, can offer up-to-date panoramic street-level imagery of urban environments, with high spatial coverage. This can help address the problem of mapping urban objects by eliminating the need to physically be in the environment.

A time- and cost-efficient approach would be to use machine learning to recognize objects in images, which has been an active area of research, also for tasks such as cataloging urban trees (Li et al. 2015; Wegner et al. 2016; Li and Ratti 2018). However, machine learning does have some drawbacks, since it requires a large amount of high-quality training data to have the potential to be accurate. Even when enough training data is available, machine learning might still fail to acquire the geographic data when an object is obstructed or when it blends in with its background. In many cases human observers can easily recognize such scenarios. To reduce the labour intensive nature of physically mapping urban objects and offering high-quality training data for machine learning (or overcoming its drawbacks), microtask crowdsourcing offers a solution. Having online crowd workers execute a simple online web-based task for a reward, like annotating objects on street-level imagery, can provide data requesters with the required geographic data (Qiu et al. 2019; Saha et al. 2019).

Prior studies with regard to online urban mapping systems are implemented for task execution based on Web browsers. Previous work highlighted issues with workers' drop-off rate, and mentioned the potential of using mobile platform (Saha et al. 2019) or adding virtual reality (VR) support (Ma et al. 2018) to improve user engagement. To verify the suitability of such approaches, there is the need for empirical work aimed at analyzing the role of task execution platforms in the worker performance in terms of output quality, execution time, user engagement and cognitive task load. In this work, we will fill the knowledge gap by addressing the following research questions:

RQ1: *How can crowd mapping microtasks be executed via multiple platforms (web, mobile, and VR) using street-level imagery?*

RQ2: *How do the mapping platforms affect the worker performance in terms of output accuracy, execution time, worker engagement, and cognitive task load?*

Original contributions. We designed and developed a multi-platform online mapping application for urban object annotating using street-level imagery, using both techniques from previous research and novel approaches proposed by us. The platform supports a web-based, a mobile-based and a mobile VR-based task execution.

To analyze the worker performance across different platforms, we set up an experiment where we analyze the worker performance in terms of output accuracy, execution time, user engagement, and cognitive load across these platforms for different types of tasks. We demonstrated the feasibility of executing online mapping tasks following the "find-fix-verify" pattern (Bernstein et al. 2010), including object de-

*Corresponding author.

tection (find), data enrichment (fix) and verification (verify), on each of these platforms. Results show that the web platform is suitable for executing detection and enrichment related online mapping tasks, and the mobile platform is suitable for executing verification tasks. Furthermore, the VR platform shows the potential to be a promising mapping tool since it can produce bounding boxes with high accuracy and better engage crowd workers. Our findings have important implications for online crowd mapping task design.

Related Work

Crowd mapping

Crowd mapping is a relatively new area of research. Traditionally, mapping is done by having workers physically explore the area to be mapped. This is a time-consuming and cost-inefficient approach for municipalities (Bader et al. 2017; Goodchild 2007). The first examples of crowd mapping can be found in managing global crises and disasters, dating back as far as 2008, where the Ushahidi[1] crowdmapping platform was used to map post-election violence in Kenya (Meier 2012). Two years later, during the 2010 Haiti earthquake the Ushahidi platform was used to collect data about the location of events that happened as a result of the earthquake such as fires and collapsed buildings (Shahid and Elbanna 2015). The term crowd mapping starts appearing in (Rogers and Scholz 2011; Tobias 2011; Mora 2011). A definition is found in (Caminha and Furtado 2012), where they define the concept as "combining the aggregation of a Geographic Information System and crowd-generated content".

Automatic detection of urban objects

An feasible approach to crowd mapping is using an automated approach to detect object in street-level imagery. Services like Google Street View[2], Mapillary[3] and OpenStreetCam[4] offer worldwide street-level images with high coverage. In recent years multiple systems have been developed with the aim of cataloging urban objects (Li et al. 2015; Wegner et al. 2016; Li and Ratti 2018). For instance, the classification algorithm by (Wegner et al. 2016) uses multiple views to detect the geo-location and species of trees in an urban environment. Although their results are fairly accurate, they do indicate a number of false positives that are easily recognized by humans. False positives for example occurred when the object was occluded by another object or when the algorithm classified a telephone pole as a tree due to the visual similarities.

Based on the studied literature, it is evident that crowd mapping approaches can not yet be replaced completely by computer vision techniques, especially when high precision is required. Furthermore, these computer vision methods need to be trained by human annotated data.

[1]https://www.ushahidi.com/

[2]https://www.google.com/streetview/

[3]https://www.mapillary.com

[4]https://openstreetcam.org/

Crowd mapping using street-level imagery

Recent research investigated the use of crowdsourcing techniques based on street-level imagery. Using these images no longer requires (crowd) workers to be physically in the area, as they can virtually look around and collect the required data. A first feasibility study was done by (Hara, Le, and Froehlich 2012) where they looked at the possibility of using Google Street View to determine sidewalk accessibility issues. They concluded that untrained crowd workers can locate and identify sidewalk accessibility problems with relatively high accuracy (Hara, Le, and Froehlich 2012). More researchers published papers on combining street-level imagery and crowd mapping (Quercia 2013; Salesses, Schechtner, and Hidalgo 2013; Qiu et al. 2019). Furthermore, (Saha et al. 2019) started a study named Project Sidewalk — another system to determine sidewalk accessibility issues using Google Street View. This was the first pilot study at scale for a crowd mapping system using street-level imagery with 581 contributing users collecting 71,873 labels.

All of the existing crowd mapping systems found in the literature are developed on web-based platforms. To the best of our knowledge, there is no research has been done on executing mapping tasks on alternative platforms and how this would affect the worker performance.

Mapping Task Design

To ensure high output quality for the tasks executed by workers, a appropriate task design is required. The "find-fix-verify" pattern as proposed by (Bernstein et al. 2010) separates the responsibility of each task executed by the crowd. This pattern splits tasks into a series of generation and review stages, with the aim of improving the task output quality: identification, generation, and verification stages. In their research, they showed that this pattern could achieve high quality task output despite relatively high individual error rates. This means that even though there might be workers that execute certain tasks with errors, this will be fixed in different stages of the find-fix-verify pattern. For each of these types of tasks, a translation is made for crowd mapping which is discussed in the following sub-sections.

Find task. The find task is the first stage of the find-fix-verify pattern, which is the identification stage. For crowd mapping, this can be interpreted into the main objective being finding the object matching the description as specified by the task requester, using street-level imagery. The output of these tasks will be a list of geographic locations (latitude, longitude) of objects marked by the crowd worker.

To estimate the geo-location of an object, the worker is asked to mark the same object from two positions, since the latitude and longitude can be more precisely estimated by calculating the intersect of two lines through the object from both positions, as shown in Figure 1. Therefore, the quality of the output largely depends on whether a worker correctly or precisely annotates an object.

Furthermore, the output of a find task has to be verified by other workers, which will be done in the fix and verify tasks.

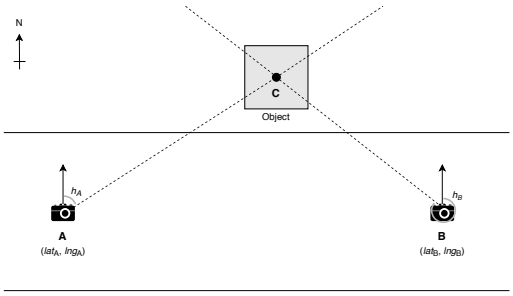


Figure 1: The geo-location of the targeted object C can be estimated by calculating the intersection of the "heading" lines of two positions A and B.

Fix task. The second part of the find-fix-verify pattern is the fix task. Based on the output of find tasks, fix tasks can be automatically generated. The aim of the fix task is to check if the annotated object is correct and, if so, to enrich the data generated at the find task. The worker is asked to draw a bounding box enclosing the annotated object, and to generate a list of labels describing the object. Furthermore, when the worker working on find tasks has made an error, this result can be discarded by the worker executing the fix task. The bounding box with user generated labels will be further verified in the next stage — the verify task.

Verify task. The final type of task serves as a quality control mechanism for the data generated at the find and fix tasks: the verification stage. For this task, the crowd worker will indicate whether the annotated object matches the description and whether the bounding box correctly encloses the object. Finally, the worker will check if the user generated labels are relevant for the given task. When a certain object went through the find and fix process and was finally verified at the verify task, the collected data is presented to the task requester. Optionally, a verify task can be executed by multiple workers, as it is also possible for workers to make errors during the verify task. The output will then be aggregated based on the majority of the votes for each verification part.

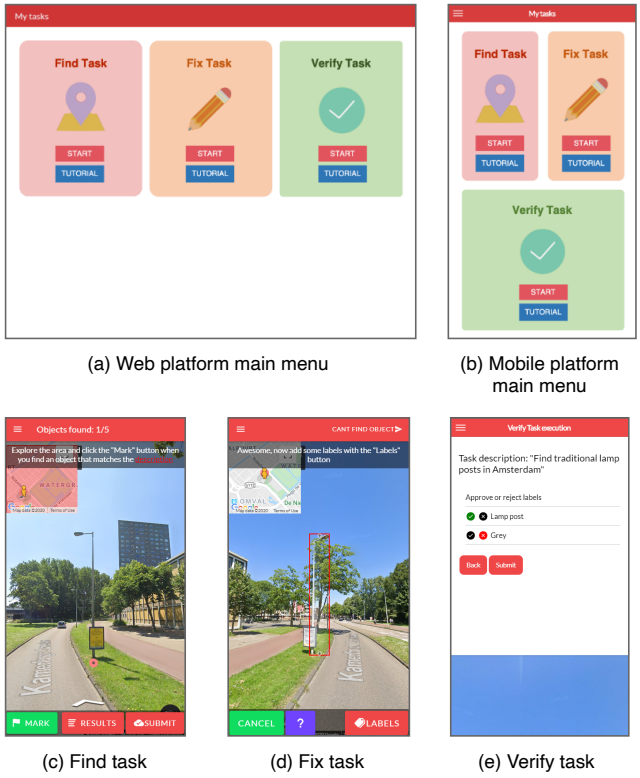
After the find-fix-verify process, the verified bounding box can be used to output a cropped image of the object which, combined with the labels, could help with not only creating urban object maps but also training machine learning models.

Mapping Platforms

An important part of this research is looking into different ways of executing crowd mapping tasks. To facilitate this, a number of task execution platforms are implemented. The reasoning behind these platforms is discussed below.

Web and mobile platforms

A web-based platform could be considered the traditional approach for executing online crowd mapping tasks. Implementing a web-based execution platform will give insight into the differences between a traditional approach and the more novel approaches of the other platforms.



(a) Web platform main menu (b) Mobile platform main menu

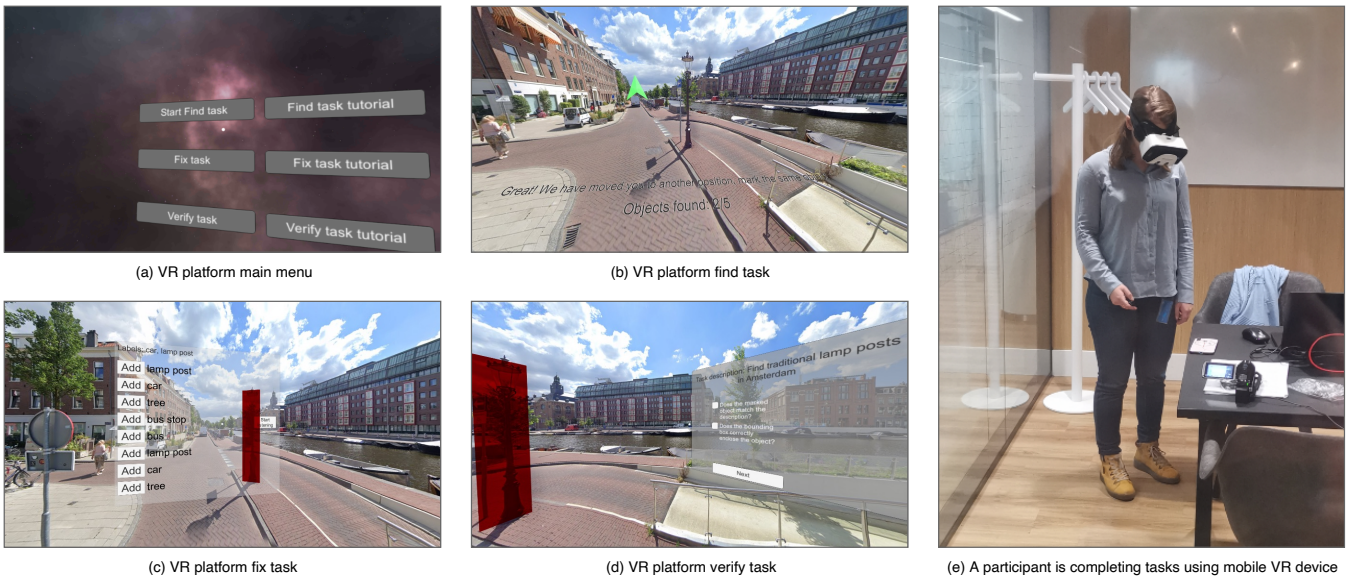
(c) Find task (d) Fix task (e) Verify task

Figure 2: The user interfaces of web and mobile platforms, including the main menu of (a) web platform and (b) mobile platform, and the task interfaces of both mobile and web platforms for (c) find task, (d) fix task, and (e) verify task respectively.

The mobile platform for crowd mapping task execution should not be overlooked since the growing number of people all over the world own a smartphone, also considering the fact that it has been used previously for crowdsourcing in other studies for tasks such as digitizing local-language documents (Gupta et al. 2012), surveys (Eagle 2009) and image tagging (Yan et al. 2009). The difference in interaction as compared to web could bring challenges and possibilities. A significant advantage of this platform is the fact that smartphones are portable, which means tasks could theoretically be executed wherever and whenever.

The web and mobile platforms are implemented using the Ionic framework[5], which is a cross-platform app development platform. Both platforms run on the same code base, but with customised presentations and interactions, as can be seen in Figure 2. In find tasks, a web user uses the mouse to click on a street-level image to identify an object, while a mobile user finds new objects by touching the screen. In fix tasks, the web user drags the mouse to draw a bounding box and type labels using the keyboard, while the mobile

[5]Ionic Framework. https://ionicframework.com/



(a) VR platform main menu

(b) VR platform find task

(c) VR platform fix task

(d) VR platform verify task

(e) A participant is completing tasks using mobile VR device

Figure 3: The gaze-based user interfaces of virtual reality (VR) platforms, including (a) the main menu, and the task interfaces for (b) find task, (c) fix task, and (d) verify task respectively. The mobile VR device is being used by a participant (e).

user drags the finger on the screen to draw a bounding box and uses built-in virtual keyboard for typing. In verify task, the user selects an option by either mouse-clicking (web) or screen-touching (mobile). The task instructions are always shown on the interfaces while the user is completing the task.

VR platform

In recent years, virtual reality (VR) has become increasingly popular as a growing number of consumer products have been released by companies such as Oculus and HTC. It offers a much more immersive experience, especially for gaming, as it gives players an impression that they are parts of the game environment. Inspired by this, we investigate the worker performance in crowd mapping tasks on a VR platform. However, current virtual reality headsets are still relatively expensive. An alternative is to use the screen and processing power of the mobile phone to create a virtual reality experience, namely Mobile VR. Projecting a stereoscopic image on the phone screen and using a pair of lenses is much more cost effective and accessible. These lenses with phone holders are widely sold. In addition, research has been done on using virtual reality for crowdsourcing tasks, such as the study by (Ma et al. 2018) where they showed the feasibility of executing crowdsourcing tasks using virtual reality. Research aimed specifically at crowd mapping still remains unexplored.

In our study, the VR application for online mapping is developed using the Unity software[6], which is a game development platform with tools to develop for mobile VR. The panorama image is fetched from the street level imagery provider and wrapped on the inside of a sphere in 3D space. The user is then placed at the center of this sphere, with the ability to look around, which could induce the effect of place illusion (Gonzalez-Franco and Lanier 2017), being defined as "the strong illusion of being in a place in spite of the sure knowledge that you are not there" (Slater 2009).

The VR platform is developed based a gaze-based UI[7], where user uses gazing (for a few seconds) instead of mouse-clicking or screen-touching. As shown in Figure 3, on the main menu, the user can select an option by gazing on a button. In find tasks, the user can see the task instructions on the ground and then identify an object by simply gazing on it. In fix tasks, the user can draw a bounding box by moving head and generate labels via voice-based input, or choosing labels from the auto-suggestions. In verify tasks, similarly, the user selects their desired option by gazing on the corresponding radio button.

Tutorials

To familiarize workers with the controls, a guided tutorial was implemented for each task type on each platform. In our study, users have to complete a predefined routine which covers each aspect of the task execution flow. This approach is commonly known as *onboarding*. As with Project Sidewalk (Saha et al. 2019), workers first have to complete the tutorial before they are allowed to start an actual task. This decreases the chance of workers making error caused by being unfamiliar with the task execution flow and interactions.

Experiment

To evaluate the worker performance across the three platforms, a preliminary study is designed and a qualitative anal-

[6]Unity. https://unity.com/

[7]https://designguidelines.withgoogle.com/cardboard/

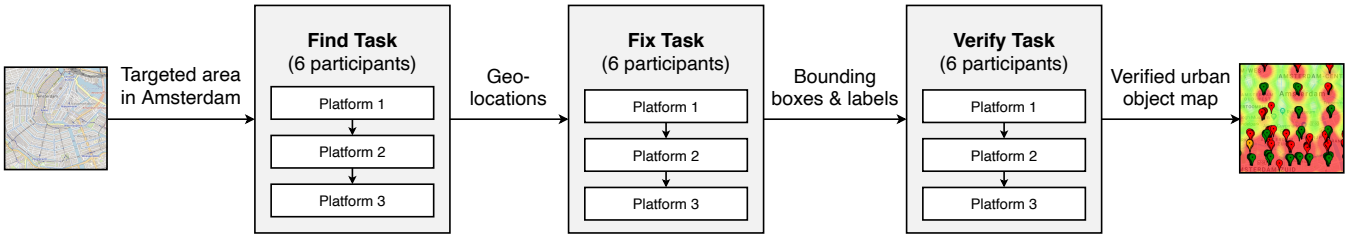


Figure 4: The procedure of the experiment follows the pattern of "find-fix-verify". For each task type, a worker must complete tasks across all three platforms (Web, Mobile and VR), but with different orders to avoid learning bias.

ysis is conducted. The plan for this experiment and the variables involved are discussed in the following sub-sections.

Experimental procedure

As shown in Figure 4, each worker is assigned to one of the three task types (find, fix or verify), and executes tasks on each of the three platforms (web, mobile, and mobile VR). Once a worker complete a task on a platform, we ask the worker to fill out a post-task survey about user engagement and cognitive load (meaning each worker needs to complete three post-task surveys for three platforms).

To prevent the learning bias, the order of the platforms used is different for each worker. For each task type there are $3! = 6$ different possible orders of execution. Therefore, we recruit 6 unique participants for each task type, which results in a total of 18 participants for three task types. We first invited 6 participants to complete the find tasks. The output (geo-locations) of find tasks were fed as the input to fix tasks, which are performed by another 6 participants. Finally, bounding boxes and labels produced in fix tasks were verified by another 6 participants in verify tasks.

All of the participants are unpaid and untrained volunteers recruited from the campus of our institute. We provided all the necessary devices for completing tasks, including a laptop, a smart phone, and a mobile VR headset (working together with the smart phone). Participants executed the mapping tasks under our guidance in a meeting room. The experiment has been approved by the human research ethics committee of our institute.

Evaluation metrics

To answer RQ2, the worker performance is evaluated by a range of different metrics.

Output accuracy. Since the goal of crowd mapping is to collect geo-location data for urban objects, it is important that the task output is accurate on each platform. Given that the controls differ on the three platforms, it is possible that the accuracy also differs. To evaluate this, the collected (and estimated) geo-locations of the urban objects are compared to a ground truth dataset for calculating the annotation error (in meters). Furthermore, the "correctness" of the provided bounding boxes and labels at the fix tasks, and the answers of verify tasks are manually checked.

Execution time (per object). In this study we measure the the execution time (in seconds), which takes the worker to find, fix, or verify an urban object. This does not include the time used to complete the tutorial or filling out the post-task surveys.

User engagement. To make user engagement quantifiable, workers are asked to fill out a User Engagement Scale Short Form (UES-SF). This form contains twelve questions looking at different factors, as described by (O'Brien, Cairns, and Hall 2018). Workers are asked to answer questions for these factors on a scale from 1 to 7. The UES score can be calculated using the approach proposed by (O'Brien, Cairns, and Hall 2018). A higher UES score represents a higher user engagement in the corresponding platform.

Cognitive load. To determine the perceived workload for a certain task type on a certain platform, workers are asked to fill out the NASA Task Load Index (TLX) form, giving insight into how mentally demanding the tasks are (Hart and Staveland 1988). The questions in the TLX form need to be answered on scale of 0 to 100, with 5-point increments. The lower TLX score is, the less cognitive task load a worker perceives.

Furthermore, workers are also be able to give additional comments on their experience.

Case study: urban lamp posts

Lamp posts form an integral part of a cityscape. In a large city like Amsterdam, thousands of lamp posts are spread around the city. In this study, we use urban lamp posts as targeted urban objects, and choose Amsterdam as the targeted area. The ground truth data in terms of the geo-location of lamp posts of Amsterdam is available at the municipality[8] and government websites[9].

Find task. Each worker starts at a location (randomly assigned from a pre-defined list) in Amsterdam with the first assigned platform, and then is asked to find at least five lamp posts in the targeted area. After finding (at least) five objects, the worker is asked to do the same at two other locations, on two other platforms respectively, giving a total of (at least) 15 annotated objects per worker for each platform.

Fix task. Each worker is asked to enrich the lamp post annotations (done by the workers from the find tasks), by adding labels on the state and surroundings of the lamp post

[8] https://data.amsterdam.nl/

[9] https://data.overheid.nl/

and drawing a bounding box around the entire annotated object.

Verify task. Each worker is asked to verify the bounding box and labels of lamp post annotations, where they must indicate whether the object indeed is a lamp post, whether the bounding box properly covers the object, and whether the given labels are correct.

Evaluation

Output accuracy

The output accuracy of the system was measured by three different approaches: geo-location estimation accuracy, data enrichment accuracy, and verification accuracy.

Find task — geo-location estimation. Each of the collected geo-locations during the experiments were compared to the ground truth dataset of street lighting in Amsterdam. The geo-location estimation accuracy is measured by calculating the distances (in meters) between the coordinates of the points in the dataset and the annotated objects.

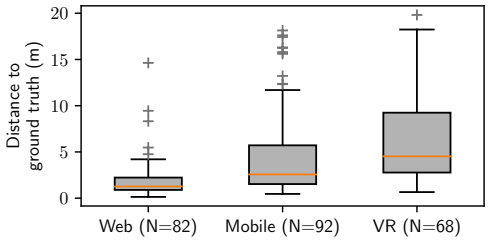


Figure 5: A boxplot of geo-location estimation accuracy measured by the distance between estimated geo-location and the ground truth (in meters).

In total 251 geo-locations were collected across the three platforms. A boxplot which shows the distribution of estimation accuracy is presented in Figure 5. All objects with a distance greater than 20 meters from the ground truth were considered invalid and omitted from the data (a total of 9 results across all platforms). The figure shows the web platform has the least estimation error, followed by the mobile platform and finally the VR platform with means $\pm$ standard deviation of 1.85 ± 2.08, 4.61 ± 4.64 and 6.58 ± 4.87 (m) respectively. For the web platform, 83% of the collected geo-locations were within 2.5 meters of the object in the ground truth. For each of the combination of platforms the difference in distance from the ground truth is of statistical significance (Wilcoxon rank-sum test, $p \leq 0.001$).

Fix task. For the fix task, worker are asked to enrich the objects found by the find task workers. Results are presented in Table 1. Here it becomes clear that the VR platform performs best for drawing bounding boxes with an accuracy of 88.4%. Most bounding boxes errors among all platforms were caused by not containing the entire object, but only a part of it. Therefore, the VR workers were able to be more accurate in drawing the bounding boxes. The labeling accuracy was roughly similar for the web and mobile platforms, but was worse on the VR platform (possibly due to voice recognition). We found that many VR workers only chose the suggested labels rather than adding new labels, since adding new labels on VR platform needs more effort compared to web and mobile platforms.

Table 1: Accuracy (% of correct bounding boxes and % of correct labels) of fix task output for each platform and task type.

Accuracy of enrichment	**Web**	**Mobile**	**VR**
% cor. bounding boxes	72.9	71.4	**88.4**
% cor. labels	**98.9**	96.2	83.6

Verify task. The results of fix tasks can be seen in Table 2. All platforms performed well in determining whether the object matched the description (web platform performs slightly better). In terms of verifying the bounding boxes, the VR platform outperforms the others, which is similar to the results of fix tasks. As for the label verification, all the platforms again performed similarly.

Table 2: Accuracy (% of correct object verification, % of correct bounding boxes verification, and % of correct label verification) of verify task output for each platform and task type.

Accuracy of verification	**Web**	**Mobile**	**VR**
% cor. object	**97.1**	92.9	90.5
% cor. bounding boxes	70.0	78.6	**81.0**
% cor. label	**90.5**	88.9	84.5

Execution time (per object)

For each task, the execution time (per object) is measured during the task execution. The timer starts when first starting the task and stops when an object is found, fixed or verified. The timer is then started again and stopped when another object is found/fixed/verified and so on. The average execution time across all the platforms and task types is shown in Table 3.

Table 3: Average execution time per object (in seconds, $\mu \pm \sigma$) for each platform and task type.

Task type	**Web**	**Mobile**	**VR**
Find	18 ± 15	$\mathbf{16 \pm 15}$	20 ± 19
Fix	25 ± 21	$\mathbf{19 \pm 13}$	41 ± 36
Verify	17 ± 17	$\mathbf{13 \pm 9}$	27 ± 19

The mobile platform had the shortest execution time for all the task types. Similarly, it was observed that the VR platform has the longest execution time for all the task types. An obvious outlier is the execution time of VR fix tasks, which

Table 4: User engagement scores (UES-SF) given by 18 workers across three platforms and three task types.

Find task		Fix task		Verify task	
Worker	**UES-SF score**	***Worker***	**UES-SF score**	***Worker***	**UES-SF score**
1	VR (5.8) > Web (4.7) > Mobile (4.2)	*7*	Web (4.8) > VR (4.7) > Mobile (4.1)	*13*	Mobile (5.5) > Web (4.6) > VR (2.9)
2	VR (4.4) = Mobile (4.4) > Web (3.5)	*8*	Web (5.3) = Mobile (5.3) > VR (4.2)	*14*	Mobile (6.2) > VR (5.7) > Web (5.0)
3	Web (5.3) > VR (5.1) > Mobile (3.6)	*9*	Web (4.6) = Mobile (4.6) > VR (4.5)	*15*	Mobile (5.7) > Web (5.3) > VR (4.3)
4	VR (5.1) > Web (5.0) > Mobile (4.6)	*10*	VR (4.1) > Web (3.9) > Mobile (2.5)	*16*	VR (5.1) > Mobile (4.4) > Web (4.3)
5	Web (5.1) = Mobile (5.1) > VR (3.9)	*11*	VR (3.9) > Web (3.3) > Mobile (3.0)	*17*	Web (3.0) > Mobile (2.8) > VR (2.5)
6	VR (5.0) > Mobile (4.4) > Web (4.2)	*12*	Web (4.2) > VR (4.0) > Mobile (3.1)	*18*	Mobile (5.3) > Web (4.5) > VR (3.5)
Overall	VR (4.8) > Web (4.6) > Mobile (4.3)	***Overall***	Web (4.3) > VR (4.2) > Mobile (3.7)	***Overall***	Mobile (4.9) > Web (4.4) > VR (4.0)

Table 5: Cognitive task load scores (NASA-TLX) given by 18 workers across three platforms and three task types.

Find task		Fix task		Verify task	
Worker	**NASA-TLX score**	***Worker***	**NASA-TLX score**	***Worker***	**NASA-TLX score**
1	Web (41) < Mobile (56) < VR (64)	*7*	Web (15) < Mobile (43) < VR (49)	*13*	Mobile (5) < Web (6) < VR (20)
2	Mobile (11) < Web (30) < VR (41)	*8*	Web (48) < Mobile (50) < VR (57)	*14*	Mobile (28) < VR (35) < Web (40)
3	Web (5) < VR (27) < Mobile (29)	*9*	Web (30) = VR (30) < Mobile (38)	*15*	Mobile (5) < Web (8) < VR (22)
4	Web (9) < Mobile (19) < VR (30)	*10*	Web (31) < Mobile (41) < VR (46)	*16*	Mobile (5) < Web (13) < VR (53)
5	Web (21) < Mobile (29) < VR (49)	*11*	Web (12) < Mobile (30) < VR (55)	*17*	Web (37) < Mobile (38) < VR (55)
6	Web (23) < VR (30) < Mobile (45)	*12*	Web (45) < Mobile (47) < VR (65)	*18*	Mobile (5) < Web (13) = VR (13)
Overall	Web (21) < Mobile (31) < VR (40)	***Overall***	Web (30) < Mobile (41) < VR (50)	***Overall***	Mobile (14) < Web (19) < VR (33)

is reasonable since a VR-based task needs more physical interactions for both drawing bounding boxes and inputting labels. However, it should also be noted that the VR platform did output the most accurate bounding boxes in fix tasks, as seen in Table 1.

User engagement

For each task type and platform, a UES-SF score is calculated based on the responses workers gave for the UES-SF questionnaire. This is a score on a scale of 1 to 7, with a higher score meaning the workers were more engaged while completing the tasks. The user engagement scores across platforms and task types are presented in Table 4.

Since this preliminary study recruited only six workers per task type, we did not conduct any statistic significance tests. In terms of find tasks, we found that 4 (out of 6) workers agreed that the VR platform had the highest user engagement, while 3 of them gave the mobile platform the lowest score. As to fix tasks, the majority (4 out of 6) of the workers reported that the web platform had the highest user engagement, and the mobile platform had the lowest user engagement. Furthermore, most workers found that the VR platform also had high user engagement, being slightly lower than the web. For verify tasks, 4 (out of 6) workers reported that the mobile platform could engage users the best, while none of them gave the mobile platform the least score. Finally, the overall user engagement scores across all task types basically align with our findings of individual scores.

Cognitive load

Based on the responses of NASA-TLX, the average cognitive task load scores could be calculated, where a lower score means a lower cognitive load. These averages are presented in Table 5.

We found that 14 (out of 18) workers reported that they perceived the heaviest task load from the VR platform. Most workers experienced it as a much more intensive experience, as it requires more physical interaction and is more immersive, yielding higher mental demand and physical demand.

For both find task and fix task, the web platform had the lowest TLX score (except only 1 out of 12 worker), as the workers found it intuitive to navigate through the streets using the mouse controls. Particularly for the fix task, all the workers indicated that they perceived the least cognitive load on the web platform, since it required less effort to draw the bounding boxes, and to type labels using keyboard. And finally for the verify task, the mobile platform had the lowest cognitive task load scores (according to 5 out of 6 workers), as the workers found it intuitive and convenient to navigate using the touch controls.

Discussion

Workers feedback

We recorded additional comments from the workers to further analyze the effects of three platforms on their performances.

For find tasks, workers generally reported higher physical and mental demand on VR platform since the operation was not as convenient as web/mobile platforms. Especially, a worker suggested that using physical buttons (instead of gazing) in VR setup would better help users roam and complete objectives in the targeted area:

"We should be able to move by pressing a button."

We found that with the fix task, the frustration factor was high for the mobile workers as they sometimes were not able to be as accurate as they wanted to be, due to the limited size of screen and the inaccurate nature of screen-touching. Furthermore, we found that a worker suggested that a zoom-in function could be added to the mobile platform for fix tasks, to address the disadvantage of small screen size:

"Zooming in would be a nice feature to add."

Similarly, for the fix task, the frustration factor was also high for the VR platform. In addition, the VR platform generally has longer execution time and higher cognitive load, as workers experienced a steeper learning curve:

"When using VR, I spent some time getting familiar with the operations. In addition, the adjustment of images on screen dizzle me when fulfilling the task."

All mobile platform workers responded with the highest user engagement score and the lowest cognitive load score for verify tasks, indicating that the mobile platform was very clear for the workers and was much easier to use compared to other platforms:

"I personally think the mobile platform would be easier to propagate in daily use, since people would like to make a little movement on their fingers and do a little favor for the task."

Implications for design

The experiments showed that the location estimation method using two angles is able to achieve accurate results, especially with the mouse on the web platform. It is much easier to accurately click on a specific location with a mouse (web) than using a finger (mobile). With the VR platform the workers have to select the location using their head movement. Since it is relatively difficult to keep the head completely still, it could affect the estimation accuracy. The data enrichment showed that the VR platform is more accurate in drawing bounding boxes. The high bounding box accuracy of VR could be a result of the fact that bounding boxes (especially for annotating large urban objects) could be drawn without the limit of the field of view in the VR platform.

The mobile platform proved to be the quickest for all tasks in terms of execution time, which can be explained by the advantage of the screen size and touch control, as a worker commented:

"In comparison, the operations on mobile seems easier for me. Using finger pointing is more convenient than using mouse clicking."

In the mobile platform, very element that can be interacted with on the screen is within finger's reach, making the interactions much easier.

User engagement and cognitive load were other factors that were measured during the experiments. The VR platform received higher TLX scores given the more immersive and physically demanding nature of the platform. However, this did not appear to heavily impact the user engagement. Even with high task load, the VR platform was proved to have relatively high user engagement, especially for find and fix tasks. Most workers were surprised while executing VR tasks and either told us or commented: *"VR is cool."*

All results considered, the platforms in their current implementation all appear to have a specific task type at which it performs the best. The web platform seems best suited for the find tasks, with the high accuracy, reasonable execution time, high engagement score and lowest cognitive load score. Similarly the VR platform performs well at the fix tasks, although the execution time and perceived task load have to be considered when utilizing the platform. Finally, the mobile platform seems well suited for verify task execution. Workers using the mobile platform executed the tasks very quickly, whilst still being reasonably accurate, receiving high user engagement scores and low cognitive load.

Limitations and future work

We found that the VR platform has significantly lower geolocation estimation accuracy, compared with the web platform, due to the unstable head movement. A solution could be to enable the VR hand/touch controller. Furthermore, The VR platform performed worse at generating relevant labels, which is possibly caused by errors from the voice recognition. Moreover, the evidence have shown that the VR platform could better engage users, implying the VR platform could be a promising tool for online mapping tasks, as the VR interactive technique advances.

Since our experimental setup needs specific guidance and devices, and our experiment is a preliminary study, another limitation is the limited amount of participants. Previous work has shown the feasibility of using VR in crowdsourcing marketplaces (Ma et al. 2018). Future work could recruit real crowd workers from crowdsourcing marketplaces instead of conduct lab experiments, to enable statistic significant tests. This could also benefit from a more diverse demographic, as in this study all participants were affiliated with our institute.

The results of these experiments come from a single case study with lamp posts. Further research needs to be done with different types of urban objects to determine whether the type of object influences the worker performance.

Conclusion

In this work, we analyzed the worker performance while completing online mapping tasks across web, mobile and VR platforms. To answer **RQ1**, we designed and developed an online crowd mapping application supporting the task execution on multiple platforms. To answer **RQ2**, we conducted a preliminary study to analyze the role of platforms in online mapping tasks. Results show that the web platform outperforms other platforms on find and fix online mapping tasks, while the mobile platform is suitable for verify tasks. The VR platform shows the strong potential to be a useful mapping tool in the future since it provides high-accuracy bounding boxes and high user engagement.

Acknowledgements. We thank all the participants from Delft University of Technology who participated in our experiments.

References

Bader, M. D.; Mooney, S. J.; Bennett, B.; and Rundle, A. G. 2017. The promise, practicalities, and perils of virtually auditing neighborhoods using google street view. *The ANNALS of the American Academy of Political and Social Science* 669(1):18–40.

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 313–322. ACM.

Caminha, C., and Furtado, V. 2012. Modeling user reports in crowdmaps as a complex network. In *Proceedings of 21st International World Wide Web Conference. Citeseer*.

Eagle, N. 2009. txteagle: Mobile crowdsourcing. In *International Conference on Internationalization, Design and Global Development*, 447–456. Springer.

Gonzalez-Franco, M., and Lanier, J. 2017. Model of illusions and virtual reality. *Frontiers in psychology* 8:1125.

Goodchild, M. F. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4):211–221.

Gupta, A.; Thies, W.; Cutrell, E.; and Balakrishnan, R. 2012. mclerk: enabling mobile crowdsourcing in developing regions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1843–1852.

Hara, K.; Le, V.; and Froehlich, J. 2012. A feasibility study of crowdsourcing and google street view to determine sidewalk accessibility. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, 273–274.

Hart, S. G., and Staveland, L. E. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52. Elsevier. 139–183.

Li, X., and Ratti, C. 2018. Mapping the spatial distribution of shade provision of street trees in boston using google street view panoramas. *Urban Forestry & Urban Greening* 31:109–119.

Li, X.; Zhang, C.; Li, W.; Ricard, R.; Meng, Q.; and Zhang, W. 2015. Assessing street-level urban greenery using google street view and a modified green view index. *Urban Forestry & Urban Greening* 14(3):675–685.

Ma, X.; Cackett, M.; Park, L.; Chien, E.; and Naaman, M. 2018. Web-based vr experiments powered by the crowd. In *Proceedings of the 2018 World Wide Web Conference*, 33–43.

Meier, P. 2012. Ushahidi as a liberation technology. *Liberation technology: Social media and the struggle for democracy* 95–109.

Mora, F. 2011. Innovating in the midst of crisis: A case study of ushahidi. *Submitted for publication to SAGE Convergence Journal*.

O'Brien, H. L.; Cairns, P.; and Hall, M. 2018. A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short form. *International Journal of Human-Computer Studies* 112:28–39.

Qiu, S.; Psyllidis, A.; Bozzon, A.; and Houben, G.-J. 2019. Crowd-mapping urban objects from street-level imagery. In *The World Wide Web Conference*, WWW'19, 1521–1531. New York, NY, USA: Association for Computing Machinery.

Quercia, D. 2013. Urban: Crowdsourcing for the good of london. In *Proceedings of the 22nd International Conference on World Wide Web*, 591–592. ACM.

Rogers, K., and Scholz, R. 2011. Crowdmapping the classroom with ushahidi. *Learning Through Digital Media Experiments in Technology and Pedagogy*.

Saha, M.; Saugstad, M.; Maddali, H. T.; Zeng, A.; Holland, R.; Bower, S.; Dash, A.; Chen, S.; Li, A.; Hara, K.; et al. 2019. Project sidewalk: a web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Salesses, P.; Schechtner, K.; and Hidalgo, C. A. 2013. The collaborative image of the city: mapping the inequality of urban perception. *PloS one* 8(7):e68400.

Shahid, A. R., and Elbanna, A. 2015. The impact of crowdsourcing on organisational practices: The case of crowdmapping. In *ECIS 2015 Completed Research Papers*, 1–16.

Slater, M. 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364(1535):3549–3557.

Tobias, E. 2011. Using twitter and other social media platforms to provide situational awareness during an incident. *Journal of business continuity & emergency planning* 5(3):208–223.

Wegner, J. D.; Branson, S.; Hall, D.; Schindler, K.; and Perona, P. 2016. Cataloging public objects using aerial and street-level images-urban trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6014–6023.

Yan, T.; Marzilli, M.; Holmes, R.; Ganesan, D.; and Corner, M. 2009. mcrowd: a platform for mobile crowdsourcing. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, 347–348.