h_da
hochschule darmstadt
institute of communication and media

member of EUT+ EUROPEAN UNIVERSITY OF TECHNOLOGY

esa

→ PREPARATION

Generating an open AI training data set for Moon craters with crowd science via a videogame

Executive Summary
Study

Open Channel
Affiliation(s): Hochschule Darmstadt Univeristy of Applied Sciences (Prime)

Activity summary:

This study investigates the integration of citizen science tasks into a fully realized gaming environment as an innovative solution to challenges in participant retention and data generation. We developed a moon-base building game that incorporates the annotation of small lunar craters (10–100 m) using high-resolution Lunar Reconnaissance Orbiter Camera (LROC) images. Unlike traditional gamification approaches, our design embeds the annotation tasks directly within the core gameplay mechanics, creating a more immersive and engaging user experience.

The results demonstrate that while annotation precision and recall were comparable to those achieved in other citizen science projects, participants in the game-based environment marked significantly more craters. Additionally, the game fostered considerably higher long-term engagement, with users remaining active for extended periods compared to those in conventional, non-gamified setups.

Publishing Date: 2024-11-28
Contract Number: 4000142966
Implemented as ESA Standard Procurement

→ THE EUROPEAN SPACE AGENCY
ESA Discovery & Preparation
From breakthrough ideas to mission feasibility. Discovery & Preparation is laying the groundwork for the future of space in Europe
Learn more on https://www.esa.int/preparation
Activity Page: https://activities.esa.int/4000142966
Deliverables published on https://nebula.esa.int

Generating an open AI training data set for Moon craters with crowd science via a videogame

1. Introduction

The increasing availability of high-resolution planetary imagery has amplified the need for annotated datasets to support artificial intelligence (AI)-driven analysis. For lunar missions, accurate identification of surface hazards, particularly craters, is critical to ensure safe landing and operational planning. Although automated algorithms have shown promise, they often fall short of the accuracy achieved by human annotations due to limited training datasets [1][2]. To address this gap, crowd-sourcing through citizen science has emerged as a viable method for generating annotated data. However, conventional approaches face significant challenges in sustaining long-term participant engagement, particularly for repetitive and labor-intensive tasks such as crater marking [3][4][5]

Hypothesis 1 (H1). *We hypothesize that integrating annotation tasks into a fully realized game, rather than relying on conventional gamification techniques, can significantly enhance user motivation and retention.*

While prior studies have leveraged elements such as leaderboards and reward systems to maintain interest [6], embedding tasks into the core mechanics of an engaging game promises a more robust solution. This approach aligns with research indicating that serious games can effectively sustain participation and improve task precision through immersive gameplay and intrinsic motivation [7][8][9].

To test this hypothesis, we developed a game that incorporates the task of identifying and marking small lunar craters (10–100 m diameter) using high-resolution images from the Lunar Reconnaissance Orbiter Camera (LROC) Narrow Angle Camera (NAC). Recognizing that such tasks can become monotonous, we integrated these activities into the context of a moon-base construction game. Players receive in-game rewards, such as resources for base expansion, based on their annotation performance. This novel approach aims to not only maximize data quantity and quality but also sustain participant engagement over extended periods.

Furthermore, we conducted a comparative analysis to evaluate the effectiveness of this game-based approach. This includes examining marking performance across different platforms (Windows and Android) and testing an optimized marking workflow tailored for mobile devices. By streamlining the annotation and tutorial processes, we seek to assess whether reducing cognitive overhead can further enhance task efficiency and user satisfaction.

Generating an open AI training data set for Moon craters with crowd science via a videogame

2. Project Background

Project Overview

The final game

The game was released on Android and PC. To maintain player engagement, we opted for a fun science fiction design rather than a scientific simulation. Players build a lunar base and send rovers to conduct research and collect resources. The game follows a tree-structured progression, where players spend their collected resources to unlock new technologies and buildings for their base. (Fig. 1)

The citizen science task is integrated by providing a special resource to the player. Similar to a premium resource in free-to-play games, Helium-3 (He-3) can be used to speed up every process in the game, reducing wait times from hours and days to seconds. The only way to collect Helium-3 is by engaging with the citizen science part of the app. Here, players are presented with small tiles of LROC NAC images. During the onboarding tutorial, players are informed about the purpose of the citizen crater marking task and trained to recognize craters in the presented images. They then continue to mark images, being rewarded with He-3 for their efforts. We collect the data in a database, gathering multiple markings per NAC tile to cross-reference player markings. After cross-referencing, the database contains positional and radius information alongside a confidence score based on the cumulative score of all players who marked a crater.

It is important to note that the game can be fully played without engaging with the citizen science part. Likewise, players are given the choice whether they want to engage with the game or simply mark craters. This also ensured that we obtained a group of non-players that represent the null hypothesis.

Generating an open AI training data set for Moon craters with crowd science via a videogame

3. Methodology

Marking Process

At the start of the app, users are given the choice wether they want to play the game or just do the marking process. Every user regardless of playing group was given the possibility to mark NAC images.

To test the effects of simplified marking methods on accuracy and marking quantity we randomly assigned users to one of two Groups: Fixed Radius Marking and Dynamic Radius Marking. While the Dynamic Radius Marking Group marked the images with the traditional Center + Radius method, the Fixed Radius Marking Group were asked to mark only craters of a specific radius. To get markings for a wide range of radii, the Fixed Radius Marking Group was shown the same pictures at multiple zoom levels, varying the effective crater size while keeping the radius of the annotation tool the same. The Fixed Radius Marking was implemented under the assumption that mobile users might have a much easier time marking craters about the size of their finger and the fixed radius was chosen accordingly.

Users have to first complete a short tutorial, explaining what the task is about and how to use the annotation tool. We make sure to communicate clearly that the results will be used for real scientific analysis and are part of a study. Users then have to mark a sequence of pre-annotated images correctly, before being presented with the actual marking dataset. The presented data consists of unknown 128x128 px NAC Tiles mixed with known pre-annotated data to continuously reevaluate the users performance. After marking one tile, the user is presented with either an evaluation of their marking results in case of a known image or a predetermined constant reward for an unknown image, making sure to provide feedback while hiding the known-state from the user until they submit the data.

After submission the marking data is stored in the database. If the marked section was part of the preannotated dataset we compare the users crater submissions to the known craters using the Intersect over Union (IoU) as a way to determine annotation similarity. [1]

$$IoU = \frac{A_{\text{intersection}}}{A_{\text{union}}}$$

where:

$$\begin{aligned} (A_{\text{intersection}}) &= \text{area of the intersection of the two circles.} \\ (A_{\text{union}}) &= \text{area of the union of the two circles.} \end{aligned}$$

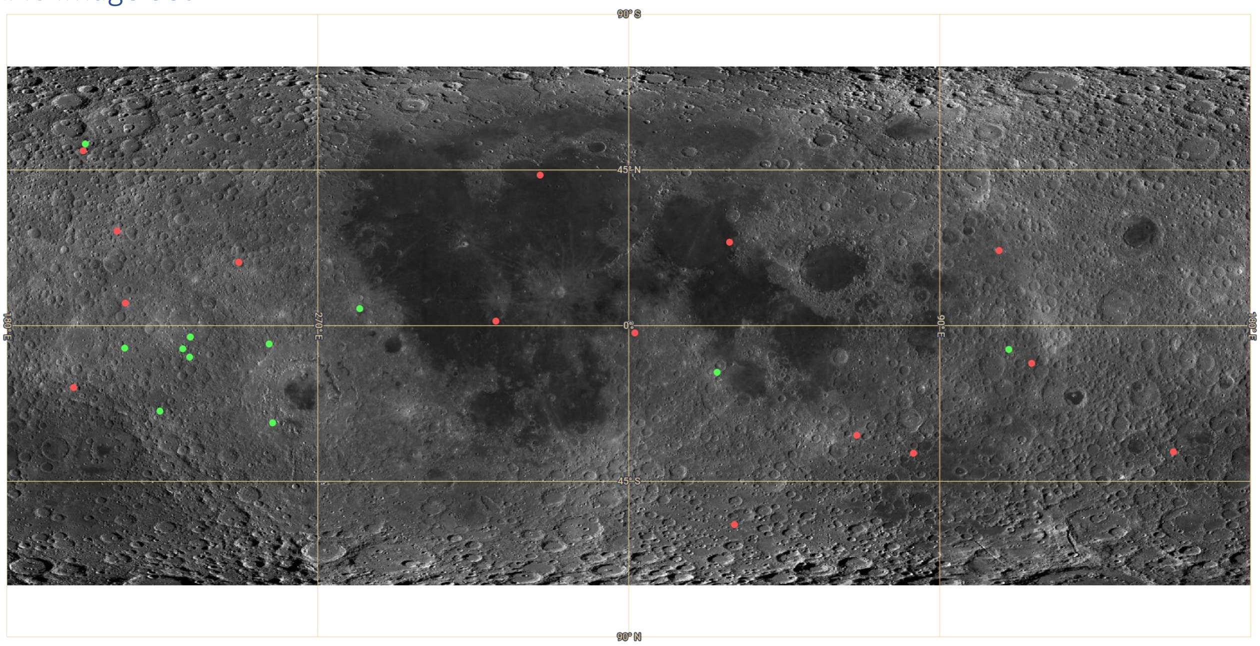
The IoU similarity score is then compared against a threshold. (we chose 0.4) Dynamic Radius Marking users get direct feedback for their markings in the app.

For Fixed Radius Marking users an additional preprocessing step is employed. Craters that are too big or too small should not be marked by these users. To determine whether a crater should be marked every known crater is compared against an artificial marking of the fixed radius size to figure out if they should be required. They are then put into two categories based on a threshold value. A high threshold for craters with a very similar radius to the fixed radius (IoU >= 0.85) creates the must have set. A lower threshold (IoU >= 0.4) defines a can have set. Craters in the can have set don’t show up as missed craters when they are not marked, but also don’t show up as false if marked. This way we give the users a bit of leeway when marking the dataset.

Every user is assigned an accuracy score based on a exponential moving averageof their current marking performance on known image tiles.

This score is used to determine the amount of preannotated images served to the user ($p_{\text{preannotated}} = 1 - S_t$), as well as the annotation progress of unknown tiles. Every unknown tile keeps track of the accuracy score sum of all marking users, setting a 'finished' flag once a threshold (in our case 300%) is reached. This way we ensure that high performing player's contributions have a higher impact on progress while least 3 people have marked every given slice before further analysis can take place.

NAC Image Set



The chosen NAC locations (green regions have been marked at the time of publishing)

For the marking set 26 NAC Sites were picked from various parts of the moon. The main selection criteria were a wide variety of terrain types and good lighting conditions making marking as easy as possible for the participants. (Fig. 2 As the LROC NAC covers the same area multiple times at different lighting conditions, marked craters can later be remapped to other NAC Images under worse lighting conditions.

Target Group and Participant Acquisition

We launched the App at the 2024 gamescom event, targeting space interested gamers. The game was presented and playable on the show floor for 5 days, giving the players the option to download the app to their own phone or pc via the Google PlayStore and the Steam Storefront.

Additionally, the game was covered by multiple press outlets on various media sites.

- ARD
- RTL
- FAZ
- Darmstädter Echo
- GameStar
- ESA Blog
- TikTok
- Golem

As a citizen science project, we allowed anyone to play the game without restraining the sample group to any particular demographic or background.

Data Collection

We classify motivation in two categories: Time from first login to last login (User Retention) and total craters marked per user. We do not include the marking accuracy at this point, as it is highly dependent on the marking method and target device. For this analysis we compared the null-hypothesis group of users who did not play the game against the playing group in a t-test.

Crater Merging

To combine the individual user markings, crater submissions were grouped by image and marking mode. Each group was then processed individually.

Markings were iterated from largest radius to smallest and compared by IoU similarity with a very low threshold (0.1). Each marking was either merged into an existing annotation group or created a new one if no plausible candidates were found. Annotation groups were then combined into one meta crater by calculating the weighted average of all contributing craters by the users accuracy score, adding the cumulative score of all craters as an additional data point to the meta crater.

As a second refinement step the users scores were updated with the newly generated crater dataset now taking the euclidean distance and radius difference into account for each crater. The score was generated as the weighted reciprocal.

The merging algorithm was run a second time using the new marking score to ensure we eliminated users who were particularly good at marking the preannotated images while failing on others and put a bigger emphasis on users who had a high agreement with the average marking.

4. Key Findings

Influence of the Game on Motivation and Accuracy

Participants were given the option to solely mark craters without engaging in the gameplay. Twelve percent of participants chose this option ($n_{\text{Marked Only}} = 487$), while the remaining participants engaged with the game at least once ($n_{\text{Played Game}} = 3726$). On average, players marked 403 craters, whereas non-players marked 106 craters. A t-test on the data resulted in a p-value of ($p=9.45 \times 10^{-10}$), indicating statistical significance. Players engaged with the game for an average of 10.5 days, compared to 4 days for non-players. This suggests that players not only participated for a longer duration but also marked more craters within the same timeframe.

The average accuracy for non-players was $a_{\text{Marked Only}} = 0.63$, and for players, it was $a_{\text{Played Game}} = 0.64$. Accuracy was not significantly influenced by gameplay, as indicated by a t-test (p=0.6).

Influence of the Platform on Motivation and Accuracy

A total of $n_{\text{Android}} = 1311$ users played on Android, while $n_{\text{Windows}} = 2902$ users played on Windows. There was no significant difference in play duration (p=0.13) or the number of craters marked (p=0.27).

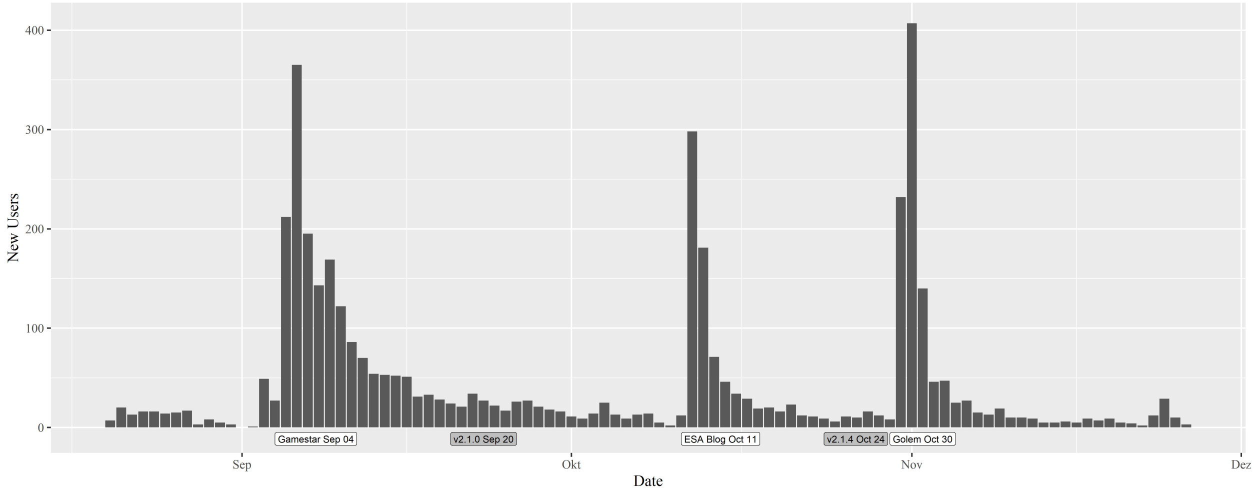
It was hypothesized that users on touch devices would have less accurate markings compared to users using a mouse. Under this assumption, a significant difference was found between the two groups ($p=3.44 \times 10^{-6}$), indicating that Android users marked slightly less accurately than users with access to a mouse.

Generating an open AI training data set for Moon craters with crowd science via a videogame

Crater Database

Following the merging process, a comprehensive database was generated, comprising 40 croppings of NAC slices, each with a resolution of 2048x2048 pixels. The dataset includes a total of 1,420,000 user-submitted crater markings, which, after merging, resulted in 127,000 distinct craters. The craters in the dataset have pixel radii ranging from 4 to 100 pixels, corresponding to a pixel density of approximately 0.5-2 meters per pixel. While the precision of the markings appears to be high, further evaluation is necessary to confirm this. The recall rate, although not at 100%, is notably high for a citizen science project. Ongoing evaluations aim to further assess these metrics. It is worth noting that some users marked rocks despite instructions in the tutorial advising against this behavior.

Press Impact



User registrations per day

Press coverage significantly influenced user registrations, as illustrated in Fig. 3. In contrast, game updates did not have a meaningful impact on the number of user registrations. This underscores the importance of raising awareness about the project to achieve a large sample size. The potential for games with a purpose to generate engaging press stories is substantial, highlighting the necessity of informing news outlets about such initiatives.

Limitations

Participants experienced difficulties in distinguishing between craters and rocks, underscoring the continued importance of effective tutorialization. The primary motivation for users to engage with the game remained the Citizen Science aspect. However, once the game content was completed, most players ceased their participation. This indicates that while Citizen Science may initially attract users, maintaining high levels of engagement requires a sufficient quantity of game content.

Developing a game is a substantial endeavor, and creating one that remains engaging over long periods necessitates a dedicated team working over extended durations. This project was developed by one full-time employee and an intern over the course of one year, successfully creating a small-scale game. However, the project has much greater potential, which could be realized with a larger team.

5. Conclusion

In this study, we successfully launched a game incorporating crater marking within its design. Our findings indicate that the gameplay aspect significantly influenced participant motivation and crater count, while maintaining accuracy at a stable level.

1,420,000 user markings were merged into an open database consisting of 127,000 distinct craters distributed across 40 NAC croppings, each with a resolution of 2048x2048 pixels.

Participants engaged with the app for the entire duration of the available gameplay content but ceased marking craters once the content was exhausted. This demonstrates the potential for large-scale "games with a purpose" projects to incorporate citizen science tasks effectively, thereby keeping users engaged over extended periods.

6. Future Work

Future research should focus on several key areas to enhance the effectiveness and engagement of the crater marking game. One important area is the influence of tutorialization on marking performance. This includes evaluating different tutorial methods to determine the most effective approach for training users to distinguish between various lunar hazards, such as rocks and craters, or adding more detailed labeling to the annotations, like crater age.

Another area of focus is the introduction of more collaborative features. Allowing users to see the markings of other players and provide feedback could foster a sense of community and improve the overall quality of the crater markings. Additionally, developing an expert mode to enable high-performing players to refine the dataset further could offer advanced tools and challenges to keep experienced users engaged and contribute to higher data quality.

Due to administrative reasons, it was not possible to launch on iOS devices. Expanding the game to these platforms could yield similar crater counts as the Android version and provide insights into the differences in annotation behavior between Android and iOS users.

Creating a larger game with a more extensive development team and a longer timeframe could help in understanding how sustained engagement can be achieved in large-scale citizen science projects and gather substantial amounts of data for the open dataset. Furthermore, exploring the inclusion of other marking tasks, such as identifying galaxies or other celestial objects, could attract a broader audience and maintain user interest.

Additional specific moon regions could be added to the database. The polar regions might be of special interest and have thus far been excluded from the NAC set due to their difficult to mark lighting conditions. A special game mode for well performing players could be included to have them mark polar regions as a challenge.

References

[1] J. Fairweather, A. Lagain, K. Servis, G. Benedix, S. Kumar, and P. Bland, “Automatic mapping of small lunar impact craters using LRO NAC images,” *Earth and Space Science*, vol. 9, Jul. 2022, doi: 10.1029/2021EA002177.

[2] S. J. Robbins *et al.*, “The variability of crater identification among expert and community crater analysts,” vol. 234, pp. 109–131, May 2014, doi: 10.1016/j.icarus.2014.02.022.

[3] R. Bugiolacchi *et al.*, “The moon zoo citizen science project: Preliminary results for the apollo 17 landing site,” *Icarus*, vol. 271, pp. 30–48, Jun. 2016, doi: 10.1016/j.icarus.2016.01.021.

[4] J. Sprinks, J. Wardlaw, R. Houghton, S. Bamford, and J. Morley, “Task workflow design and its impact on performance and volunteers’ subjective preference in virtual citizen science,” *International Journal of Human-Computer Studies*, vol. 104, pp. 50–63, 2017, doi: https://doi.org/10.1016/j.ijhcs.2017.03.003.

[5] S. P. Schwenzer, M. Woods, S. Karachalios, N. Phan, and L. Joudrier, “LabelMars: Creating an extremely large martian image dataset through machine learning,” in *50th lunar and planetary science conference*, 2019, p. 1970. Available: https://www.hou.usra.edu/meetings/lpsc2019/pdf/1970.pdf

[6] L. von Ahn, “Games with a purpose,” *Computer*, vol. 39, no. 6, pp. 92–94, 2006, doi: 10.1109/MC.2006.196.

[7] J. S. van ’t Woud, J. A. Sandberg, and B. J. Wielinga, “The mars crowdsourcing experiment: Is crowdsourcing in the form of a serious game applicable for annotation in a semantically-rich research domain?” *2011 16th International Conference on Computer Games (CGAMES)*, pp. 201–208, Jul. 2011, doi: 10.1109/cgames.2011.6000339.

[8] O. Troyer, “Towards effective serious games,” in *2017 9th international conference on virtual worlds and games for serious*, Sep. 2017, pp. 284–289. doi: 10.1109/VS-GAMES.2017.8056615.

[9] J. S. S. Van’t Woud, J. A. C. Sandberg, and B. J. Wielinga, “Cerberus: The Mars Crowdsourcing Experiment,” *Communicating Astronomy with the Public Journal*, vol. 12, p. 28, May 2012.