Tile-o-Scope AR: An Augmented Reality Tabletop Image Labeling Game Toolkit

Sofia Eleni Spatharioti
spatharioti.s@northeastern.edu
Northeastern University
Boston, MA

Borna Fatehi
fatehi.b@northeastern.edu
Northeastern University
Boston, MA

Melanie Smith
m.smith@northeastern.edu
Northeastern University
Boston, MA

Avery Rosenbloom
rosenbloom.a@husky.neu.edu
Northeastern University
Boston, MA

Josh Aaron Miller
miller.josh@husky.neu.edu
Northeastern University
Boston, MA

Magy Seif El-Nasr
m.seifel-nasr@northeastern.edu
Northeastern University
Boston, MA

Sara Wylie
s.wylie@northeastern.edu
Northeastern University
Boston, MA

Seth Cooper
se.cooper@northeastern.edu
Northeastern University
Boston, MA

ABSTRACT

Crowdsourcing games involving image labeling tasks are commonly digital, played online, and have rules set by designers. In this work we explore the potential of tabletop image labeling games, incorporating physical elements, in-person community-based gameplay, and support for customizable rules. We developed an augmented reality game toolkit called Tile-o-Scope AR and conducted two studies. The first study demonstrates how the toolkit can facilitate in-person discussions through collaborative image labeling, and the toolkit's potential adaptability to other games and applications. The second study, using three different activities designed for the toolkit, demonstrates the toolkit's flexibility for creating customized experiences for audiences of different backgrounds.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

augmented reality; image labeling; tabletop

ACM Reference Format:
Sofia Eleni Spatharioti, Borna Fatehi, Melanie Smith, Avery Rosenbloom, Josh Aaron Miller, Magy Seif El-Nasr, Sara Wylie, and Seth Cooper. 2020. Tile-o-Scope AR: An Augmented Reality Tabletop Image Labeling Game Toolkit. In *International Conference on the Foundations of Digital Games (FDG '20), September 15–18, 2020, Bugibba, Malta.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3402942.3403002

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

FDG '20, September 15–18, 2020, Bugibba, Malta

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-8807-8/20/09…$15.00
https://doi.org/10.1145/3402942.3403002

1 INTRODUCTION

The human eye can be a powerful mechanism for analyzing data, where automated systems fail to provide reliable results. Thus, organizations have turned to crowdsourcing, with popular applications found in image labeling tasks, due to the relatively low complexity and cost [11]. The monotonous nature of the image labeling task, however, makes it challenging to engage participants over time [8].

Gamification in general is used as an approach to improve engagement [2]. Prior work has looked at mitigating disengagement via approaches such as gamification and task variety [1, 9, 10]. Additionally, performing image labeling in a co-located setting can open up in-depth discussions, with potential to enhance the crowdsourcing experience and outcomes, such as achieving longer-term engagement [12]. Moreover, using Augmented Reality (AR) may offer additional benefits, due to its potential effectiveness in education and collaboration [7].

To explore how AR and co-location may enhance image labeling tasks by fostering discussions among participants, we designed Tile-o-Scope AR (TOSAR), a multi-person, co-located image labeling game toolkit. TOSAR utilizes AR technology to read physical tags and show images on a mobile device in their place. TOSAR does not enforce any specific rulesets beyond the core matching mechanics, instead opting for a customizable design where players can determine their own games and rules.

In this work, we describe Tile-o-Scope AR and two studies of the toolkit. A first study showed potential for using TOSAR as a conversation facilitator among participants when asked to label images by playing a simple game. A second study, using three different activities, revealed that each participating group had a different preference, highlighting the toolkit's capacity towards engaging contributors from different backgrounds.

2 RELATED WORK

Prior work has explored applications or AR in activities with physical components. Synflo [6] uses Sifteo Cubes as a physical component for completing actions. Evaluation of the system revealed peer

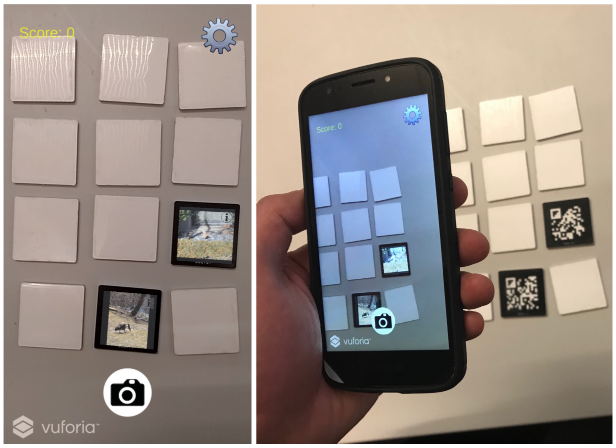

Figure 1: Memory being played using Tile-o-Scope AR. In-game tile images from iNaturalist.

collaboration patterns among participants. BacPack [5] utilizes tokens as the physical component and a multitouch tabletop interface as the digital component, with a goal of increasing collaboration by creating tangible experiences that provoke discussions. Juan et al. [4] present an AR game about endangered animals, using tangible cubes. ARToolKit [3] is another cost effective toolkit used for prototyping in AR. While these tools aim for co-located activities using AR, each requires either a separate platform with predefined game rules, or are focused on more general prototyping, or, in some cases, need a considerable investment (Synflo), or loss in portability (BacPack). TOSAR is a low cost, portable option with potential to combine different AR labeling games under one platform.

3 TILE-O-SCOPE AR

Tile-o-Scope AR[1] is developed using Unity[2] and Vuforia[3]. The physical component (i.e., scannable tags) can vary, from paper print-outs to laser-cut pieces (Figure 1). First, players can choose a set of images and categories. The mapping from tag to image can be randomized to support sets with more images that there are tags. While playing, images are revealed on screen by pointing the camera over the tiles. Players can then make matches by using the camera button to select images they believe belong to the same category. Upon confirmation, text and audio feedback is provided. Making a match can apply a category to images. Along with the images to be categorized (whose categories are unknown), images with ground truth (whose correct category is already known) can be added to the playable set; then every selection falls under four possible cases:

- **Only ground truth images of the same category:** Players receive feedback that the match is *correct*.
- **Ground truth images of the same category, one or more unknown images:** A unique category for all images can be identified. That category can be applied to any unknown images in the match. In this case, players receive feedback that they made a *potential* match.
- **Ground truth images of multiple categories, zero or more unknown images:** A unique category cannot be identified, therefore the match is *incorrect*.
- **All unknown images:** The player is asked to pick a category, which is applied to unknown images in the match. Players receive feedback that they made a *potential* match.

4 STUDIES

4.1 Playing a Game and Designing New Games

The first study was exploratory in nature, and consisted of three parts. First, participants were asked to play Memory[4] competitively, using the toolkit. Next, we held a brainstorming session for proposing other games. Finally, we asked a set of open-ended questions about their experience, such as what they enjoyed most/least, improvements, and whether they saw benefits of using the toolkit based on their research field. The total duration of the study was 1 hour. We used two datasets, one for animal identification (*Animals*), sourced from iNaturalist[5], and one for identifying flooded manure lagoons in North Carolina after Hurricane Florence (*Florence*), sourced from NOAA[6]. We used 12 tiles for time purposes.

We recruited 14 students and faculty (5 male, 9 female; 6 from Computer Science and 8 from Social Sciences) from the university, aiming for varying experience in games, so as to get more diverse feedback. Participants were put into groups of two ($n = 6$) or four ($n = 8$). All sessions were transcribed and then sorted using Affinity Diagrams via three independent raters. IRB approval was received and participants were compensated with a $15 Amazon Gift Card.

4.1.1 Findings. The average total time across all groups was 8.26 minutes ($n = 5$, $sd = 2.33$, $min = 4.63$, $max = 10.55$). The average time per match was 1.06 minutes ($n = 39$, $sd = 0.78$), which can be attributed to both the game, but also to the conversations between selections. The average percentage of correct or potential matches was 64% (Florence = 62%, Animals = 66%). These may also have been affected by the difficulty of the datasets, and searching for images. Our Affinity Diagram findings are summarized below:

Collaborative In-Person Image labeling: All groups started collaborating almost immediately, exchanging ideas and observations, effectively helping their opponents. One group evolved from a competition to a collaboration as a team. When asked about using TOSAR for building community, one participant noted "I think that it would have pretty strong potential for that." Different groups commended TOSAR's ability to initiate conversations by getting people together in a room and helping them understand the problem better. One group also felt they could see neighbors using it as a fun way of monitoring nearby industrial facilities.

Motivation was also linked to meaningful contribution ("I liked that I was helping someone," "The scientific layer to it is super interesting"). The potential for educational purposes was also raised by participants, with one group discussing how they felt they were learning about identifying flooded manure lagoons.

[1] Link to toolkit: https://cartosco.pe/ar
[2] https://unity.com
[3] https://developer.vuforia.com
[4] Link to Memory game rules: https://cartosco.pe/ar_games
[5] https://www.inaturalist.org
[6] https://hddsexplorer.usgs.gov

Customizable Rulesets: Participants offered suggestions along three directions. First, we got suggestions of existing games, such as Go Fish, Candy Crush, Mahjong Solitaire, and CAPTCHA games, which we expect can be readily played with the toolkit. Second, adaptations for Memory, such as flipping coins to chain matches, or making tiles visible for a certain time limit. Finally, novel game ideas suggested were "Convince-a-Match", a game involving subterfuge and negotiation tactics, where players must convince or fool others into making correct or incorrect matches, and "BattleTile", where participants must describe images to other players. Story-driven games were also proposed ("A puzzle or story game where you start with one animal and it wants you to do something for another animal and you have to find that animal"). Although designed primarily as multiplayer, some participants suggested using TOSAR as a single player game, when traveling or relaxing.

Adaptability to Other Projects: Some participants found value in using the toolkit in disaster response and damage assessment, by counting "how many people lost their roofs or not a couple of days after the storm". When discussing neighborhood collaboration on the topic of hog farms, another participant pointed out the potential towards identifying and reporting violations. Another suggestion involved predator-prey relationships and animal conservation. Such discussion indicated that community-based analysis of local aerial imagery may raise privacy issues.

Interactions with Physical Component: The combination of physical and digital components was generally well received ("I liked that it was a virtual game but also something you touch with your hand"). Participants stated they enjoyed being able to interact with physical tiles. Some also stacked their tiles and used it to playfully showcase their skills. The stacked formations were also viewed as motivation from others for improvement. Multiple participants praised the ability to customize the tiles. Some considered mixing different materials (i.e. plastic, wood) in custom decks.

4.2 Engagement across Different Backgrounds

To explore how customizable rulesets may engage participants from varying backgrounds, we conducted a second, mixed study. Groups from differing research fields were asked to perform three activities and answer questions about their experience for each one, as well as rate them at the end. We used the *Florence* dataset, and implemented some improvements based on feedback from the first study, such as adding a more interactive tutorial, and enabling labeling individual images. We used 3 activities: a) A no-game condition (**Sorting**), where the goal is to sort images into available categories. Participants could choose to either sort as a team, in subgroups, or individually; b) a competitive game (**Memory**); and c) a collaborative game (**TrekStack**[7], which involves working together and using a "hand tile" to push matching tiles, moving a game piece to a desired location). We used the Intrinsic Motivation Inventory's (IMI) enjoyment subscale after every activity, and a ranking question for enjoyment at the end. The duration was 1 hour. To observe how different groups would approach the toolkit, and increase feedback diversity, we recruited the following groups (4 participants each, 1 device per participant, randomized activity order):

- *Game Designers* (GD): Graduate students with a Game Design background (*avg. age* = 27; 2F, 1M, one chose not to disclose gender). Some participants knew each other prior to the study. Experience in games may yield feedback on the design.
- *Sociology* (SC): Faculty from Social Sciences (*avg. age* = 53; 3F, 1M). All participants knew each other as colleagues. This group could provide constructive criticism on potential dangers, such as regarding free labor.
- *Environmental Health & Justice* (EH): Researchers from an environmental health fellowship program attending a conference (*avg. age* = 38; 2F, 2M). Due to expertise in real-life applications, this group may offer insights on adapting TOSAR in the wild. This group played during dinner at a restaurant, allowing exploration of how the toolkit could potentially be used in a less formal setting.

4.2.1 *Findings.*

The average percentage of correct and potential matches was 88% (GD: 90% , SC: 79 %, EH: 100%; Sorting: 77%, Memory: 83%, TrekStack: 100%). The increase may be attributed to the interactive tutorial. The EH group failed to verify matches via the app for Sorting, leading to some missing accuracy information. 42% of participants had no prior experience with AR. IMI and ranking data (Figure 2) indicate different preferences for each group. The GD group enjoyed TrekStack, the collaborative game, the most, while EH and SC preferred Memory and Sorting respectively. We found no apparent relationship between order of activities and reported enjoyment levels. Our observations are summarized as follows:

Sorting: Two groups chose to *collaborate as a team* (GD, EH), and one group (SC) chose to sort *individually*. In the first two, we saw a *division of labor*, where the deck was split among players to label. These groups also opted to *create piles* for the sorted categories, and tended to discuss images among each other, asking for second opinions on images they were unsure of. On the contrary, the SC group did not communicate much overall. In some cases, tiles that had been already categorized were picked up by others to categorize as well, resulting in an *overlap of labor*.

Memory: The GD group opted for a *competitive approach*, limiting discussions to a minimum. The SC group also played competitively, but had more discussions on the images, thus effectively helping their opponents, as was the case in our first study. This group was unable to finish the game in the time allotted. Finally, the EH group chose to play *as a team*, and had the highest level of discussions, with players offering arguments for opposing views.

TrekStack: This was viewed as most challenging. The GD group was the most efficient, and the only ones to finish in time. They grasped the game quickly, and formed a good strategy for winning. Group votes were called before making any moves. Even though they played as a team, the SC group looked at images individually. They were unable to formulate a strategy, sometimes making counterproductive moves. Overall, they found the game complicated and were often disappointed by lack of progress. The EH group was more engaged and really wanted to win. They were quicker to identify a strategy, but sometimes got stuck, as there were no possible matches. In these cases, a new hand tile was drawn.

Group Interactions: The SC and EH groups were friendlier with each other during gameplay, with an increased sense of community and mutual support observed in SC. In Memory for example, a

[7] Link to game rules: https://cartosco.pe/ar_games

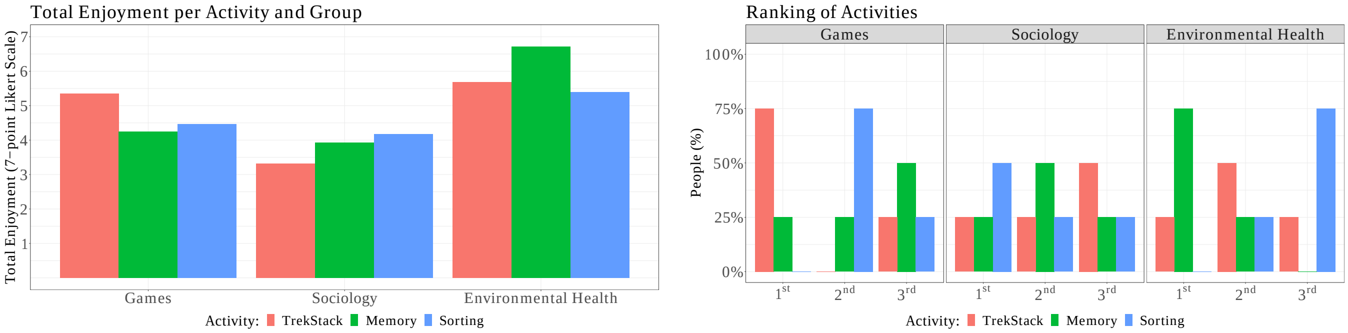

Figure 2: Results from (left) Intrinsic Motivation Inventory (IMI) enjoyment subscale and (right) ranking of activities. Ranking is organized by group, and shows percentage of participants that ranked an activity as first, second or third most enjoyable.

participant offered some of their tiles to another with none. The EH group appeared to be having the most fun overall, regardless of their performance. They were often engaged in strong discussions about the images, although this may be in part due to the environment they were in (restaurant). Finally, in the GD group, players were mostly focused on winning the games than discussing.

Open-Ended Discussion: All groups commended the *group aspect* of the game. Interestingly, one participant (SC) disliked the competition in Memory. The biggest issue for SC was image size and resolution, and the presence of a score, even though they were informed it was not applicable for the chosen activities. The GD group liked Memory the least, as the rules led to the game “snowballing fast”. They also suggested locking images on the screen. When asked to suggest games, participants mentioned games like Catan, Ticket to Ride, and Photo Hunt. Applications suggested included labeling geological formations, traffic, and tabletop exercises for housing management after disasters. The GD group felt that having to physically move tiles in TrekStack made the game much more engaging, compared to a potential purely digital version. Two participants claimed this approach could make them more interested in longer term contributions, and one thought it promoted the multiplayer component. However, most agreed that Memory was at times cumbersome.

5 CONCLUSION

We presented Tile-o-Scope AR, a co-located, multiplayer AR game toolkit for image labeling. TOSAR could be a new avenue for raising awareness during publicly organized events, beyond methods like presentations and static tutorials. The variety of activities suggested may indicate its applicability to broader applications. By supporting multiple activities, TOSAR may be an option for simultaneously engaging groups with varying game and AR experience, as witnessed in our second study. Finally, by bringing people physically together, TOSAR may potentially create more social cohesion, and less invisibility about who is doing the crowd work, as raised by some participants.

An upcoming version will allow toggling elements like dice and timers. One group (EH) did not use the app to verify the validity of their matches during Sorting, opening future directions for mitigating this behavior. Moreover, we are also interested in evaluating repeated use over longer periods of time, and comparing against a purely digital implementation. Finally, it would be interesting to generalize this set of design ideas to other crowdsourcing tasks.

ACKNOWLEDGMENTS

We thank all participants. This material is based upon work supported by the National Science Foundation under grant no. 1816426 and a Northeastern University TIER 1 grant.

REFERENCES

[1] Carrie J. Cai, Shamsi T. Iqbal, and Jaime Teevan. 2016. Chain Reactions: The Impact of Order on Microtask Chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3143–3154.

[2] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From Game Design Elements to Gamefulness: Defining "Gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (MindTrek '11)*. 9–15.

[3] Eva Hornecker and Thomas Psik. 2005. Using ARToolKit Markers to Build Tangible Prototypes and Simulate Other Technologies. In *Proceedings of the 2005 IFIP TC13 International Conference on Human-Computer Interaction*. 30–42.

[4] Carmen M. Juan, Giacomo Toffetti, Francisco Abad, and Juan Cano. 2010. Tangible Cubes Used as the User Interface in an Augmented Reality Game for Edutainment. In *Proceedings of the 2010 10th IEEE International Conference on Advanced Learning Technologies*. 599–603.

[5] Anna Loparev, Lauren Westendorf, Margaret Flemings, Jennifer Cho, Romie Littrell, Anja Scholze, and Orit Shaer. 2017. BacPack: Exploring the Role of Tangibles in a Museum Exhibit for Bio-Design. In *Proceedings of the Eleventh International Conference on Tangible, Embedded, and Embodied Interaction*. 111–120.

[6] Johanna Okerlund, Evan Segreto, Casey Grote, Lauren Westendorf, Anja Scholze, Romie Littrell, and Orit Shaer. 2016. SynFlo: A Tangible Museum Exhibit for Exploring Bio-Design. In *Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*. 141–149.

[7] Iulian Radu. 2014. Augmented Reality in Education: A Meta-Review and Cross-Media Analysis. *Personal and Ubiquitous Computing* 18, 6 (Aug. 2014), 1533–1543.

[8] Henry Sauermann and Chiara Franzoni. 2015. Crowd Science User Contribution Patterns and Their Implications. *Proceedings of the National Academy of Sciences* 112, 3 (Jan. 2015), 679–684.

[9] Sofia Eleni Spatharioti and Seth Cooper. 2017. On Variety, Complexity, and Engagement in Crowdsourced Disaster Response Tasks. In *Proceedings of the 14th International Conference on Information Systems for Crisis Response And Management*. 489–498.

[10] Sofia Eleni Spatharioti, Sara Wylie, and Seth Cooper. 2019. Using Q-learning for Sequencing Level Difficulties in a Citizen Science Matching Game. In *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. 679–686.

[11] Luis von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 319–326.

[12] Sara Wylie and Len Albright. 2014. WellWatch: Reflections on Designing Digital Media for Multi-Sited Para-Ethnography. *Journal of Political Ecology* 21, 1 (2014), 321–348.