Game of Words: Tagging Places through Crowdsourcing on Public Displays

Jorge Goncalves, Simo Hosio, Denzil Ferreira, Vassilis Kostakos
Department of Computer Science and Engineering, University of Oulu
Pentti Kaiteran katu 1, FI-90014 Oulu, Finland
firstname.lastname@ee.oulu.fi

ABSTRACT
In this paper we present *Game of Words*, a crowdsourcing game for public displays that allows the creation of a keyword dictionary to describe locations. It relies on crowdsourcing and gamification to identify, filter, and rank keywords based on their relevance to the location of the public display itself. We demonstrate that crowdsourcing on public displays can leverage users' knowledge of their environment, can work with a generic gaming task, and can be deployed on displays with multiple concurrent services. Our analysis shows that our approach has important benefits, such as the ability to identify undesired input, provide words of high semantic relevance, as well as a broader scope of keywords. Finally, our analysis also demonstrates that the chosen game design coped well with the challenges of this complex setting (i.e. public urban space) by disincentivising incorrect use of the system.

Author Keywords
Crowdsourcing; public displays; gamification; location.

INTRODUCTION
We present *Game of Words*, a gamification and crowdsourcing approach that allows the creation of a keyword dictionary to describe locations on public displays. It relies on these techniques to identify, filter, and rank keywords based on their relevance to the location of the public display itself. The establishment of such a contextual keyword dictionary is valuable for researchers to provide adapted services, for display owners to select relevant content for the screens [28], for urban planners to better understanding citizens' mental map of an area [3], and even for marketing and advertising purposes where the industry is mostly built around the use of keywords [30].

Location *keywords* are convenient to store and process; they can be efficiently searched and retrieved; they can be used with automated tools to model emotion, artefacts, events, and even identify relevant photographs or other media; and they can be used for tagging or labelling.

Yet obtaining keywords to describe a place is challenging. Fully automated ways to characterise a place remain immature, since *place* consists of much more than just the physical surroundings as they can also include actions, patterns and behaviour of people [13]. Recognizing a space is an ongoing machine vision challenge [17], so far with limited success beyond controlled settings. Another approach is to harvest and analyse geo-tagged social media (*e.g.*, Twitter or Facebook tags), but without human "curation" [29] or additional sources like user diaries [22], it is challenging to deal with "noise", *i.e. undesired input from users*, and validate the semantic relevance of the outcome.

Maps, location directories, and encyclopaedias do offer a valuable resource for automatically characterising a location, but these sources may not be granular or can remain static and miss out on dynamic changes to locations and settings. Recent work has shown how urban mobility patterns can be used to derive keywords to describe locations through a fully automated analysis [21], but this approach remains to be validated in a broader context. More straightforward approaches do exist, such as surveys and interviews, but their potentially disembodied nature (*e.g.*, answering an online survey at home about a place you visited last week) when conducted on a large scale can confound the results.

Another potential approach is to rely on *crowdsourcing*, *i.e.*, to ask a crowd of users to perform the task of characterising a location. However, the online nature of traditional crowdsourcing markets (*e.g.*, Amazon Mechanical Turk) lacks the potential to recruit users who are locals. On the other hand, crowdsourcing on smartphones requires infrastructure development and enrolment effort, with potentially additional costs for data transmission and is not easily accessible to everyone. For these reasons, public displays have recently emerged as an alternative platform for crowdsourcing [9].

Interactive public displays allow *localized* crowdsourcing, a geo-fenced and granular crowdsourcing environment. Although with potentially fewer "workers" than its online crowdsourcing counterpart, this approach has been shown to reduce noise and bias in "crowd-data" [9]. An important limitation of public display-based crowdsourcing, however,

DIS '14, June 07 - 11 2014, Vancouver, BC, Canada

ACM 978-1-4503-2902-6/14/06…$15.00.
http://dx.doi.org/10.1145/2598510.2598514

is that they need to rely on simple user interfaces and be effortless to use [24]. They also need to support "walk up and use" so that users can learn from others or start using the display and its services immediately [4]. On the other hand, a benefit of this technology is that people approach public displays when they have free time, and use them without clear motives [24], for example to play games [26]. Thus, public displays provide a crowdsourcing opportunity for people to donate their time.

In this paper we report on a study where a game, *Game of Words*, was deployed on several public displays across multiple locations in a city, for the purpose of building a dictionary of keywords to describe their deployment locations. We compare the keywords obtained through the game with i) keywords obtained manually through interviews, ii) keywords obtained through a fully automated approach [21], and iii) a random set of keywords. Our analysis shows that our approach has important benefits, such as the ability to effectively identify noise, provide words of high semantic relevance, as well as a broad scope of keywords. Our analysis also demonstrates that the chosen game design coped well with the challenges of this complex setting [24] (i.e. public urban space) by disincentivising incorrect use of the system.

RELATED WORK

Our work explores crowdsourcing on public displays that are *accessible to everyone, without restrictions* or constant *supervision* from researchers. Earlier public display studies have shown that this kind of technology can produce "noisy" results [10,15,30] due to unpredictable behaviour from users. However, recent work [9] demonstrated the appeal of public displays for crowdsourcing by reporting the key to overcoming these limitations is designing for short attention spans and fairly effortless tasks.

Crowdsourcing on the Go

Until recently, mobile phones have championed the push towards crowdsourcing on the go. Most of these platforms have been deployed in developing countries targeting low-income workers providing them with simple tasks [e.g., 11]. Recent advances in mobile technologies have also allowed for more intricate and creative tasks. For instance, location-based distribution of crowdsourcing tasks has allowed its workers to perform real-world tasks for others. Some examples of this include giving location-aware recommendations for restaurants [1], providing an instant weather reports [1] or authoring news articles by requesting photographs or videos of certain events from workers [34].

Recently, researchers have explored ways in which mobile phones can enable a new empowering genre of mobile computing usage known as Citizen Science [27]. Citizen Science can be used collectively across neighbourhoods and communities to enable individuals to become active participants and stakeholders, typically through crowdsourcing. Mobile phones have a major appeal for this movement due to their affordances (*e.g.*, majority of people have one, users take them everywhere, etc.). However, there are potential barriers for this type of crowdsourcing like additional configuration effort or even possibly additional financial costs. Contrary to mobile environments, crowdsourcing on public displays does not require workers to make any deployment effort or bear financial costs [9].

Gamification

A growing body of literature within HCI is focused on gamification, *i.e.*, using selected features of "serious gaming", such as rewards and competitiveness, for purposes beyond pure gaming [6]. Perhaps the best known example of gamification today is the social network "*Foursquare*". Its popularity is largely based on the perceived value of "badges" and status rewards, such as "mayorships" and other digital rewards.

Social and collaborative games have been used to improve the social interaction and experience of museum visitors by gamifying the digital museum guide for the visitors [8]. Another application domain where gamification has been applied is citizen sensing (*i.e.*, using humans as sensors) [2]. Perhaps the most beneficial use case for gamification, from a societal perspective, is learning. People are willing to spend hours learning the "pleasantly frustrating" gameplay and features of games. To apply such, at times frustrating, engagement effort to learning is beneficial [7].

In this work, we leverage gamification on public displays. We do so by turning a menial task (i.e. "voting on relevance of words") into a game through the use of game elements (e.g., score, leaderboard) effectively gamifying the crowdsourcing task. Games in general are appealing to public display users and are often cited as "unexpectedly popular" [26], implying that people like to use public displays in a casual way, to spend free time. For instance, the playful design of the Ubinion system masked a "serious" civic engagement application as a service to play and have fun with [15]. Its idea was to bootstrap an online community by using content generated by users of public displays. Ubinion demonstrated that public displays can be used to rapidly gather large numbers of societally relevant input from citizens if the design is fun and the display location fitting for the purpose. Another prototype, FunSquare, was designed as a quiz-game on public displays to enhance sense of community among its users [25]. FunSquare's lure was based on dynamic facts about the deployment environment itself and on the heavily gamified design that appealed both to children and adults. Furthermore, gamification has been shown to hinder privacy concerns as well as help recruiting initial users on public displays [5]. Attracting the first users for an application is crucial because of the *honeypot effect* [4] simply because the presence of users on a display will attract more users.

Game of Words leverages fundamental elements of gamification and design recommendations for public displays: keep it *effortless* and easy to use [4]; it does not

offer badges or status improvements that would require strong association to the game, like user accounts and registration potentially hindering participation, but provides instant gratification by a scoring mechanism and a top-5 leaderboard; the game attempts to leverage the fact that public display users are often already in a willing state to spend free time using its services [24], particularly games [26] and it provides a walk up and use interface [23,25].

STUDY

Our study investigates whether crowdsourcing on public displays can be used to generate a list of keywords that describe five different locations in a city. To assess the potential of such a crowdsourcing approach, we contrast it against alternative ways to obtain such keywords.

One simple but at times troublesome approach is to interview people and collect keywords from them directly. Another alternative way to characterize a location is to use a fully automated method such as Location Archetype Keyword Extraction (LAKE) [21]. This method is used to automatically discover semantically relevant keywords by correlating local mobility patterns with nationwide Google search trends. Given a longitudinal dataset of pedestrian movements at any particular location, LAKE automatically generates keywords that have been shown to be semantically relevant to that location.

In our study, we compared 4 ways of generating keywords that semantically relate to a location: **I)** keywords generated algorithmically through LAKE analysis [21], **II)** keywords obtained by interviewing local citizens, **III)** random keywords, and **IV)** keywords generated with the crowdsourcing *Game of Words* deployed on public displays.

Study Setup

The displays used in this study to deploy *Game of Words* are single touch-enabled, large (57") interactive displays, situated in five different public locations. The game was deployed for one month during the summer of 2013, on all five displays at the same time. The displays are accessible to all users without dedicated supervision and they have been deployed in their respective locations for 3 years. As such, they have become an accepted part of the public city infrastructure itself. This is an important point because many reported field trials often suffer from novelty bias. It should be also mentioned that the displays were not dedicated to this study alone, but about 25 other services were simultaneously offered via a directory of services. The most popular services on these displays are typically games, such as adaptations of the traditional Hangman and Tetris games. Other applications include news, service directories, public transport information, and commercial advertisements [26].

The five locations we selected for the deployment were a popular swimming hall (SH), a lobby of a local sports/event hall (S/EH), the main market square (MS), the main library in our city (L), and a university campus (U). These locations offer a rich and diverse set of different spaces and audiences, as to minimize sampling bias. Figure 1 depicts the five display deployment locations.

Figure 1. Display deployment locations. Top row from left: library, swimming hall lobby and market square. Bottom row from left: university campus and sports/event hall lobby.

Game of Words

Game of Words is based on the idea of players categorizing 10 different words that sequentially appear on the screen as relevant or irrelevant to the current location. The game consists of 5 screens shown in Figure 2. The game was designed for sole users because previous work [9] has shown that users completing crowdsourcing tasks on public displays alone are willing to spend more time and have better performance.

When users launch the game on the public displays, it occupies the right half of the screen. Screen 1 of the game has instructions and a top-5 leaderboard of nicknames that have scored the currently highest scores at that location. These instructions asked for a chronic characterisation of the space and not just what was going on around the participant at the time. So while the location might, for example, be "sunny" or holding an event during a particular game, we asked players to consider the broader context. On Screen 2, before the actual game starts, players are prompted to type a word that they feel is relevant to their current location given the aforementioned instructions. The soft keyboard was designed following literature recommendations [18] to minimize error rates. The typed words are added to the gameplay dictionary. Naturally, each of the five locations in our study has a dedicated dictionary.

Screen 3 consists of the main gameplay, during which words will sequentially float inside a cloud from the bottom of the screen towards the top. A player has five seconds to decide whether the word is *relevant* or *irrelevant* to their current location by pressing the respective large buttons in the game's user interface. A single game consists of 10 words, *i.e.*, a player will "vote" for 10 words in each game. At the start of each game the placement of identically coloured "Relevant" and "Irrelevant" buttons was randomized (relevant on left/irrelevant on right or relevant on right/irrelevant on left) to avoid any bias. Players could skip a word by simply not pressing anything and letting the cloud float away.

After playing all 10 words (or interrupting the gameplay by pressing a "finish" button) the player is shown their final score on Screen 4. If a player achieved a top-5 score, they were instead shown Screen 5 where they were asked to optionally type a nickname to appear in the leaderboard and their email address. In either case, the game subsequently returned to Screen 1.

We spent considerable effort optimising the duration of gameplay. A very short gameplay would not allow for enough variation and might become boring, resulting in lower uptake. A very long gameplay may become tedious and physically tiring causing users to give up and walk away from the display halfway through a game. We followed the findings in [9] which suggest that crowdsourcing tasks on public displays should be kept short otherwise the users are likely to abandon the task and walk away. More importantly, their work reports a cut-off point in terms of task difficulty, beyond which error rates increase and completion rate dramatically drops. Thus we tested a variety of gameplay settings, finally settling on a gameplay of 10 words with up to 5 seconds per word. Our pilot testing showed that these settings did not cause fatigue yet provided players enough time to classify a word as relevant or irrelevant.

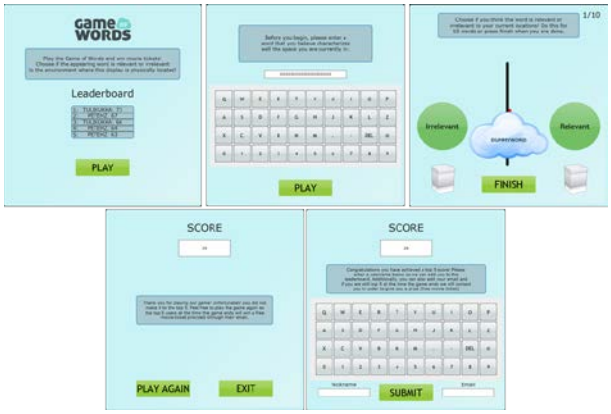

Figure 2. In-game interfaces. Top row from left: start screen with instructions and top-5 leaderboard; a virtual keyboard to type a new word; actual gameplay screen with buttons to choose between relevant and irrelevant and a finish button. Bottom row: score screen with no new highscore; and the same screen if the player gets to the leaderboard and is allowed to type a nickname and an email address.

Word Selection and Scoring

At the start of each game the system selected the 10 words to display from the dictionary. The dictionary per location included words that players added on Screen 2, and words that were added more than once did not bear any additional weight. The 10 chosen words were selected in such a way so that all words in the dictionary are voted equally. Hence, the chosen words were those with the least amount of votes at the time, and random selection was used to choose between words with the same number of votes. We did this in order to ensure that by the end of the study most words had a similar amount of total votes.

The scoring mechanism of the game was based on how previous players categorized the words. The more a player agrees with previous players on whether a given word is relevant or not, the more points are awarded. The choices of previous players are not shown to the current player. We illustrate this mechanism with an example: if the currently displayed word has previously been voted as relevant eight times and irrelevant three times, and the current player chooses "relevant", then we award five points (relevant minus irrelevant). On the other hand, if the player chooses "irrelevant", we award minus five points (irrelevant minus relevant). Finally, the points from all 10 played words in each game are added to derive the final score.

This scoring mechanism has the feature that the same word will yield different points at different times, and over time the scores get higher. We decided to use such an evolving mechanism to keep the game engaging and to ensure that it does not become too static or worst: boring. The players were not made aware of the mechanism, but we expected them to get an intuitive feel for it through gameplay. The final score was immediately shown to players after each game, so feedback on their performance was instant.

Bootstrapping the Game

To bootstrap the game we compiled for each of the five locations a distinct dictionary of 30 keywords, using 3 different sources. First, we algorithmically [21] generated 10 keywords for two locations: the main library and university campus. Because this automated method requires a substantial amount of pre-recorded urban mobility data we were not able to use this method for all locations. Second, we interviewed five volunteers to obtain ranked keywords by their perceived order of relevancy for each of the five locations. The volunteers were not shown each other's keywords. To select the final set keywords for each location we aggregated their ranking across all volunteers. We note that these volunteers had lived in the city for over 10 years, and had visited repeatedly all locations in question. Third, we generated a random set of keywords as a control condition, chosen randomly from the official language dictionary. We summarise the set of keywords we used to bootstrap the game in Table 1.

	SH	S/EH	MS	L	U
Algorithm	-	-	-	10	10
Volunteer	20	20	20	10	10
Random	10	10	10	10	10

Table 1: The number of keywords generated to bootstrap the game using each method for every location (Swimming Hall, Sports/Event Hall), market Square, Library, University).

Data Collection and Interviews

During deployment our moderation included correcting keywords that had minor spelling mistakes, thus avoiding the existence of multiple versions of the same keyword as

well as removing particularly offensive words in a handful of instances. We did not remove any keywords that were clearly irrelevant to the location as we wanted to rely on the crowd-based moderation as a filtering mechanism. However, we note that we did not remove words that were consistently voted as irrelevant as it would adversely affect our analysis as we wanted the keywords to have the same amount of votes.

All user interactions with the game were logged. This includes number of games played, all choices made for each word by each player, the number of words played per player in one session, the duration of games and how long it took to vote on a word, all new words typed by players, and the timestamps of all the above. In our analysis we attribute games played in quick succession without a timeout to the same player.

At the end of the deployment we calculated a metric of every player's session performance as a number between 0 and 1 by comparing their answers to final community consensus. A vote by player P for word W was deemed to agree with the community if that vote was the same as the majority outcome of all votes for that word. Our formula considers how many times a player voted and of those votes how many agree with the community.

$$player\ agreement = \frac{\#\text{ of votes in agreement with community}}{\#\text{ of votes by the player}}$$

Furthermore, we measured the *relevance* of each word in relation to its location as a number between -1 and 1 using the formula:

$$word\ relevance = \frac{(\#\text{ voted relevant} - \#\text{ voted irrelevant})}{\#\text{ of votes the word received}}$$

This metric does not account for frequency of votes as our game did not randomly select keywords from the database but prioritised those with the least votes. We also ensured that a user could not see the same word during a single session. By the end of our study, the majority of keywords had been voted the same number of times – only the recently added words had much fewer votes and were discarded from our analysis. Furthermore, during the study we conducted unobtrusive in-situ observations of players, surveys and semi-structured interviews. The interview subjects (N = 30), six per each location, were recruited on-site. We made sure all interviewees were long-time residents of the city. We then asked them to play the game while "thinking aloud", so researchers could take notes of the playing experience. After gameplay, they completed a short survey that contained a mix of Likert-scale and open-ended questions. These included demographics, prior experience with these public displays, the social context in which they would likely play this game, the ease/difficulty of coming up with a word to characterise the location, the ease/difficulty of voting for words, and overall experience.

RESULTS

The *Game of Words* was deployed for 1 month at 5 public locations. During this deployment it was played 632 times, collecting 6009 votes (M=9.51, SD=1.82) and 362 keywords. Table 2 shows a breakdown of these results for each location separately and in total. The swimming hall was the most popular location, even though the market square and library yielded more keywords. The vast majority of users played all 10 words and did not abandon the game before its completion, resulting in an overall high average number of votes per game (M=9.51, SD=1.82). We also note that the median number of votes per game session was 10 across all locations. Some examples of keywords added for each location are: Swimming Hall (water, locker room, meeting point), Sports/Event Hall (exercise, badminton, artificial turf), Market Square (seagull, ice cream, shopping), Library (librarian, literacy, history) and University (research, lecture hall, study).

	SH	S/EH	MS	L	U
# of votes	1958	787	1263	1131	870
# of games	213	82	128	120	89
Avg. # of games	9.19	9.60	9.87	9.43	9.78
# of keywords collected	66	65	90	73	68

Table 2: Breakdown of results obtained from each location.

Table 3 summarizes what portion of the keywords entered in the game had either i) already been entered by a previous player, ii) been given by a volunteer before the deployment, iii) had been generated automatically through LAKE, or iv) were unique (i.e. not giving by volunteers, not generated by LAKE and only given by one player).

	SH	S/EH	MS	L	U	Total
given by other players	25%	16%	18%	26%	24%	22%
given by volunteers	1%	53%	24%	1%	17%	19%
given by LAKE	-	-	-	0%	10%	5%
Unique	74%	31%	58%	73%	49%	57%

Table 3: Overlap between newly added player words and existing words in the dictionary from other sources.

We also looked at the popularity of the game over time. Figure 3 shows the progression of the total number of games played, and new words added, during the 30 days of deployment. The progression for both measures remained constant, suggesting that the game did not lose its appeal after a few days.

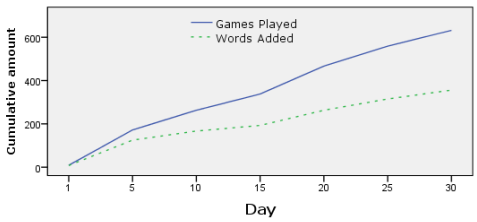

Figure 3. Cumulative progression of numbers of games and words added during the deployment

Manual Coding of Crowdsourced Keywords

All 362 keywords collected by the game were subjected to content analysis as discussed in [16]. This type of analysis is appropriate when existing theory or research literature on a phenomenon is limited. Therefore, we avoid using preconceived categories, instead allowing the categories and names for categories to emerge from the data [19]. We chose to categorize all the user-generated keywords, instead of just the relevant ones, to get a thorough understanding of what kind of terms people like to use for characterizing locations. This process consisted of open and axial coding and was conducted independently by two researchers. The resulting coding scheme was discussed and iterated, and all reports were classified in one of five categories. Interrater reliability was satisfactory (Cohen's K = 0.98).

The five categories that emerged from the analysis were:

- **activity** (N=46): depicts something that is typically done by humans in the space (e.g., swimming, badminton),
- **object** (N=123): tangible objects, often close to the public display and perhaps seen by the player even at the same time when playing (e.g., book, chair),
- **atmosphere** (N=37): adjectives, words that describe the overall feeling of the place (e.g., innovative, sunny),
- **concept** (N=90): more abstract notions that are not physically present but associate with the space, for example ideologies that the space is built for (e.g., meeting point, literacy), and
- **noise** (N=66): nonsensical text such as random character sequences or profanities (e.g., asdlfasdilj, hiiiiiiii).

We also looked into the word relevance for each of these categories. The activity category was voted the most relevant (0.84), followed by object (0.68), concept (0.64), atmosphere (0.57) and finally, as expected, noise (-0.35). We found a significant relationship between the location and the category of the word added at that location (χ^2 (16, N=362) = 176.90, p < .01). Figure 4 shows the popularity of each category in the five different locations.

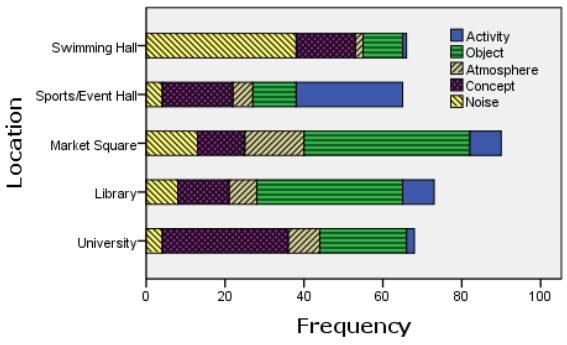

Figure 4. Breakdown of collected keywords by category and location.

Disagreement between Players

We analysed the extent to which players agreed with each other in their voting. For every single vote in our dataset we analysed whether that vote agreed with the majority of votes for the same particular word at the same location. We then count the number of votes that disagreed with the majority. The results are shown in Figure 5 where we see the Swimming hall followed by the Library as the two locations with most disagreement, both just over 20%.

Furthermore, we analysed each player's agreement with the majority. For example, if a player voted on 10 words and the community agreed with him on 6 of those, that player had a 0.6 agreement rating on a scale of 0 to 1. Figure 6 depicts the distribution of this metric for all players across all locations. We note a high level of player agreement throughout with an average of 84.21% (SD=21.55) which highlights the overall consensus regarding voting during our deployment.

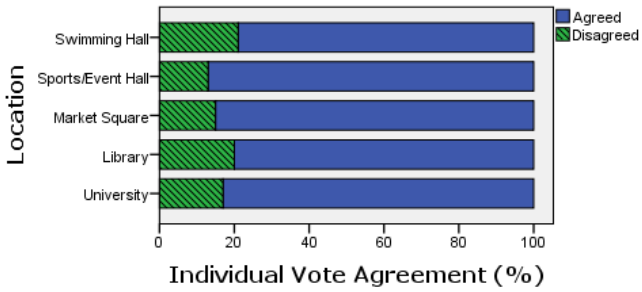

Figure 5. Proportion of agreement for each individual vote.

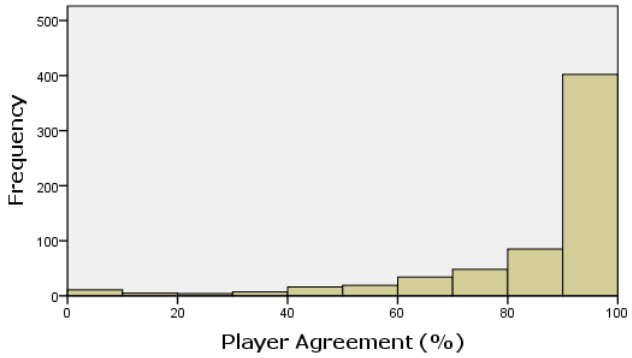

Figure 6. Distribution of player agreement across all locations.

Differences between Sources of Words

We analysed how the keywords produced by each of our 4 sources were voted by players. Specifically, we averaged the relevance of the keywords produced by the players, by the volunteers before this study, by the LAKE algorithm, and the random set we generated. In addition to these, we also calculated the performance of the top-10 *player* words for each location as well as the top-10 *volunteer* keywords (for the locations that originally had 20 keywords). We defined these additional sets to make the comparison between the groups more fair as the LAKE algorithm already chooses the top 10 most relevant words while the

number of crowdsourced words and volunteer words for 3 locations was higher. In Table 4 we show the average relevance, while in Figure 7 we show the distribution of words that resulted in those average values.

	SH	S/EH	MS	L	U	Overall
LAKE	-	-	-	-.67	.78	.06
Volunteer	.54	.85	.81	.45	.74	.68
Volunteer top-10	**.72**	**.99**	.92	.45	.74	.76
Random	-.47	-.82	-.51	-.55	-.90	-.65
Player	-.35	.37	.46	.39	.74	.32
Player top-10	.65	.86	**1.00**	**1.00**	**1.00**	**.90**

Table 4: Average relevance for each source of words in each location and overall. The most successful method is highlighted.

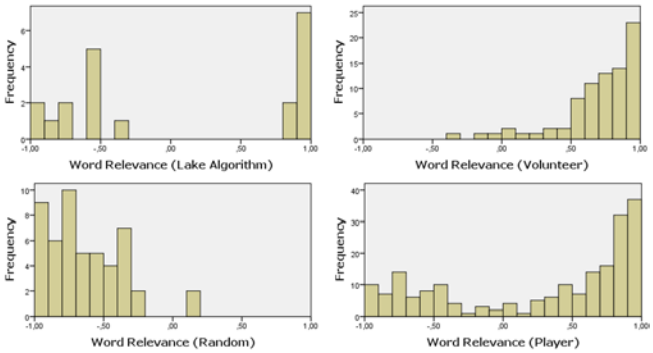

Figure 7. Histogram of word relevance for each of the source of words.

Interview Data

We conducted 30 semi-structured interviews (15 male, 15 female) across the 5 locations of the study (6 per location). The participants' average age was 33. We made sure all participants were long-time residents of the city in order to obtain more reliable feedback. Every participant reported having noticed the displays around the city. Seventeen participants stated that they used the displays before, mostly for games. When asked under what circumstances would they prefer to play *Game of Words*, the majority would rather do it together with friends (N=15, 50%) or while waiting for something (N=14, 47%), or "when alone" (N=9, 30%). One participant claimed she would play the game "when there are no other people around the display" and another "when using the display for some other reason". Finally, 6 participants reported they would play the game for other reasons (13%) citing being bored and to kill time.

Next, we enquired about their perceptions regarding the game itself. Participants reported that the instructions of the game were clear and therefore it was easy for them to understand what to do. Using a 5-point Likert scale the ease (1: Very Easy, 5: Very Difficult) we asked about the ease/difficulty of coming up with a relevant word for their location. While some participants had some difficulties, the overall consensus was that this initial part of the game was easy (M=1.63, SD=0.96, min=1, max=4). Further analysis of the keywords added by the participants during the interviews confirmed that they understood the instructions. As for the main gameplay (voting) the overall consensus was that this was a relatively easy step ((M=1.80, SD=.96, min=1, max=5) with only 2 people considering it either difficult or very difficult. When asked if they would like to play again or not, the majority said yes (N=14, 47%), followed by maybe (N=10, 33%) and no (N=6, 20%).

DISCUSSION

Through gaming we can encourage people to donate their free time for a research purpose or an otherwise valuable task. Our study sought to leverage people's free time, and games are the archetypal tools to spend free time on situated public displays [24]. Game of Words was described as a fun game despite its underlying research purpose. Interview comments revealed that the perception of the game's purpose does not even matter to the players: *"I think this game exists just to kill time, it's obvious that this cannot be used for anything else than just lightweight fun"*, "*This game might be used to develop the mental model of the market square and maybe use that to develop it"*, and *"Maybe this is used for marketing this display"*. These comments were all made by people who indicated their willingness to play the game again in the future, and by doing so further contribute their time and ultimately to the keywords' dictionary.

Public Displays as a Crowdsourcing Platform

While previous work [9] has discussed public displays as a potential crowdsourcing platform, that study was conducted in a rather contrived setting. For instance, the study consisted of a task with a relatively high *intrinsic* motivation: counting blood cells infected with malaria for the purpose of developing better software. It could be argued that such a "worthy" task could in itself act as an incentive to participate appealing to altruism, *i.e.*, perhaps users' innate desire to "do something good". Furthermore, the task was deployed on a bespoke display, whose sole purpose was to crowdsource. Finally, this task was location-agnostic, and did not tap into the unique knowledge of the local community.

The study we present here addresses these limitations and further demonstrates that public displays can be a potentially valuable crowdsourcing platform. First, we show evidence that tasks that are perceived as less "collectively useful" – like the one we presented here – can still attract attention and produce viable results. In our study more than 6000 votes were collected by our game in one month. We argue that providing tasks that are *localised* can attract people's interest. This is something that previous work had also demonstrated [29], and there is evidence to suggest that tasks that are *not* specific enough may suffer from severe appropriation on public displays [13]. In situations that call for strong local and in-situ knowledge, such as our case of characterizing a location, geographically distributed online mechanisms are inadequate: online users are disembodied from the space and thus have restricted

contextual knowledge of the space, and attracting users with sufficient local knowledge is challenging.

Furthermore, our study shows that even with multi-purpose displays (*i.e.*, displays with multiple applications), crowdsourcing can still attract people's attention. Public displays are becoming increasingly embedded with several services "competing" for the users' attention [14,20]. In our particular deployment, *Game of Words* was deployed on displays with two dozen other applications and it accounted for just about 7.5% of all applications launches, being somewhere in the middle of the list in terms of popularity. This suggests that crowdsourcing on public displays can still work reliably even when the displays have multiple services competing for the user's attention.

Finally, and most crucially, we demonstrate that crowdsourcing on public displays can genuinely benefit from local people's knowledge of the environment. One of these displays' key affordance, their serendipitous nature [24], provides a perfect opportunity for this type of knowledge sharing. Our game did not lose popularity over time (Figure 3), counter-reacting the novelty effect. Many interviewees stated they would play the game while waiting for someone or "just for fun". Hence, we argue that given their strong locative nature, public displays have the inherent potential to "harvest" local knowledge. They can enhance knowledge sharing by lowering temporal and spatial barriers between those that wish to share it and those that want to acquire it.

Bootstrapping and Automation

An important challenge in creating and maintaining public deployments is content creation and reproduction [32]. Creating the initial content for prototypes is often nontrivial and requires training or talent. In the case of our game, keywords constitute the content. In fact, the keywords in *Game of Words* define the gameplay experience: if all the words that appear are clearly irrelevant or relevant, the gameplay would require little effort, being perhaps boring and not challenging. This actually means that having a portion of keywords that are not relevant to the location can contribute to the gameplay experience positively by increasing its difficulty. It also highlights the importance of bootstrapping such a game with randomized words that are less likely to be perceived as relevant by the players. In essence, people play not because they are personally interested in solving a particular computational problem but because they wish to be entertained [36] for which challenge is a key aspect [33].

The four techniques we explored for bootstrapping semantically relevant keywords are each very different and have both drawbacks and benefits. We observed that their performance varied greatly across different locations. Our analysis shows that random keywords and algorithm-generated keywords both yield a significant amount of irrelevant votes (81% and 48% of all votes respectively), although the algorithmically generated keywords for the University setting performed well (word relevance = .78). The keywords provided by volunteers before the study also performed relatively well (overall word relevance = .68). As for the crowdsourced keywords they also contained several entries that were deemed to be irrelevant (N=65). However, as the game proceeded, the crowdsourced validation process (the gameplay itself) proved effective and the top-10 keywords from this list outperformed all the other methods (.14 and .84 higher overall word relevance when compared to volunteer top-10 and LAKE respectively). These are encouraging results, as the moderation of unwanted input has always been one of the core problems in deployments on public displays [30].

An important aspect of the keyword generation techniques we evaluated is to consider their potential for automation. Crucial to the development of effective public deployments is the ability of our systems to perceive and understand the context in which they are located. In this sense, an automated self-learning process is rather beneficial. In the case of games, one approach is to not consider input as correct until a certain number of players have entered it [36], or in our case voted for it.

This means that, in the long run, the utility of volunteer-based keyword generation is limited due to the difficulty in automating this reliability mechanism. A purely algorithmic approach such as [21] is promising but offers limited success, and its prerequisite of collecting long-term mobility traces can prove challenging and time-consuming.

A gamified, crowdsourced approach as we demonstrated here is fast, accurate, and produces non-trivial keywords that all other approaches miss. Specifically, more than 50% of the words collected by *Game of Words* were unique and not identified by asking the volunteers or through automated analysis (Table 3).

Location Matters

In our study, players performed differently in different locations. For example, the swimming hall location was much "noisier" than the others (-.35 word relevance for player words) with a significant amount of irrelevant keywords being provided (N=35). Not surprising, as the swimming hall display is used mostly by teenagers and pre-teenagers, who tend to "misbehave" in front of their friends by doing something that is seen as forbidden [31]. In this case the forbidden act was to type nonsensical words to the displays. However, what is remarkable is that even in that location the *voting* of words seemed to work well: the random character strings and inappropriate words were all quickly voted as irrelevant by other players, and in fact player agreement was exceptionally high in that location (~80%).

Of course, our use of the term "noise" has an unnecessarily bad connotation. After all, appropriation and unpredictable usage is to be expected when introducing new technologies to the crowds [12]. Our data was full of examples where in

addition to showing off to friends we observed numerous word entries that are merely nicknames, which we presume to be the players' nicknames. Prior studies have witnessed similar behaviour in public deployments, and the psychological need for "self-presentation and advertising" has been documented before when using public technologies [35]. Location can have an impact on how playful or freed from external pressure and supervision a player feels, and therefore it also impacts how much of such appropriation is to be expected. The key point is that despite "misbehaving" participants, the crowd provided a reliable "noise cancelling" mechanism.

The display in the library also provided erratic results. Our analysis indicates that the algorithmic keywords performed rather badly in this location (word relevance = -0.65). We attribute this to the fact that the library is a place where people go to do research or look up information on many different topics. Thus, the players disagreed about the relevance of the keywords provided by the LAKE algorithm (word relevance = -.67) and to a lesser extent about those generated by volunteers or other players (word relevance = .45 and .35 respectively). Such erratic locations can be troublesome due to being highly heterogeneous or having widely different meanings to people [22], which can further explain the relatively low word relevance throughout. Even so, the player top-10 list emerged as very relevant (word relevance = 1) as it consisted mostly of objects that are less likely to be influenced by subjectivity.

Furthermore, our findings showed that in different locations users provided different types of words, in terms of the categories that emerged in our analysis. Not surprisingly, in the University the most popular word category was concept and for the Sports/Event Hall it was activity, reflecting the reality of both these locations. However, overall the object category was the most used with the majority of library and market square words being in this category. We argue that this is mainly due to the importance of visual stimuli in their decision to select a word. As reported by several interviewees their strategy to choose a word was simply to look around and select something they saw in the space. Given our interviews and data analysis, we believe that in follow-up work it would be interesting to explore giving users more *specific instructions* about providing keywords. In our case the instructions were rather vague (Figure 2), but it should be possible to provide instructions that focus on objects, activities, or concepts. Similarly, we could ask them to indicate something that they like or dislike about the place. In fact, the game could alternate between different instructions, thus nudging users to provide a wide range of keywords by getting them to think about the place from a variety of perspectives.

Finally, we also note the potential for priming: perhaps players submitted keywords that reflected the categories that emerged during our bootstrapping. However, our results indicate otherwise for two reasons. First, an inspection of the words from the random and LAKE bootstrapping reveals that they do not reflect the final categorisation, while the volunteer keywords do to some extent. This suggests that location, rather than bootstrapping, is affecting what keywords people think of. Second, players typed their own word before seeing the words in the game, and in 65% of instances players added only that single word. It is difficult for us to know how many players did return to the game at a different point, but these results suggest that most words were collected before players were exposed to other keywords.

Limitations

We acknowledge certain limitations in the application itself as well as in the conducted field trial. Although we make the case for full automation of characterizing locations in the future, our deployment does not support it quite yet due to moderation. So, naturally this needs to be automated as well, but presents a new research problem as itself.

Furthermore, we acknowledge that it is possible that some participants may have misinterpreted the instructions; however a careful look into the results and interview data confirms that overall participants followed the instructions as intended. Finally, the use of LAKE keywords was technically impossible to generate in some locations (lack of prerequisite data). While introducing this extra independent variable can be rightly questioned, our intent was to establish a comparison between crowdsourcing and an automated approach of generating such keywords.

CONCLUSION

In this paper we demonstrate the feasibility of crowdsourcing on public displays. Unlike previous work, we demonstrate that crowdsourcing on public displays can benefit from users' knowledge of their environment, can work with a generic gaming task, and can be deployed on displays with multiple concurrent services. Specifically, we relied on a gamified task to establish a list of keywords that describe particular locations. Our results show that our approach can produce *more accurate* and *richer results* than algorithmic approaches or interviews.

Our work provides compelling evidence that public displays can be a potentially valuable crowdsourcing platform. By relying on the crowd to both provide input and evaluate the input, we show that despite their public nature public displays provide reliable results. In our ongoing work we are interested in further investigating any potential bias that may arise in crowdsourcing on public displays, particularly in terms of the effects of the surroundings to the crowdsourcing users.

REFERENCES

1. Alt, F., Shirazi, A., Schmidt, A., Kramer, U., Nawaz, Z. Location-based crowdsourcing: extending crowdsourcing to the real world. *Proc. NordiCHI '10*, ACM (2010), 13-22.

2. Bowser, A., Hansen, D., Preece, J. Gamifying citizen science: Lessons and future directions. Position paper presented at the Gamification Workshop (2013).
3. Brewer, J., Dourish, P. Storied spaces: Cultural accounts of mobility, technology, and environmental knowing. *International Journal of Human-Computer Studies 66*, 12 (2008), 963-97.
4. Brignull, H.,Rogers, Y. Enticing people to interact with large public displays in public spaces. *Proc. INTERACT 2003*, IOS Press (2003), 17-24.
5. Cao, X., Massimi, M., Balakrishnan, R. Flashlight jigsaw: an exploratory study of an ad-hoc multi-player game on public displays. *Proc. of CSCW 2008*, ACM Press (2008), 77-86.
6. Deterding, S., Dixon, D., Khaled, R. and Nacke, L. From game design elements to gamefulness: defining gamification. *Proc. MindTrek*, ACM (2011), 9-15.
7. Gee, J.P. Learning by design: Good video games as learning machines. *E-Learning and Digital Media 2*, 1 (2005), 5-16.
8. Ghiani, G., Paternò, F., Spano, L.D. Enhancing Mobile Museum Guides with Public Displays. www.comp.lancs.ac.uk/~corina/CHI08Workshop/Papers/Ghiani.pdf.(2008), Accessed in 19-08-2013.
9. Goncalves, J., Ferreira, D., Hosio, S., Liu, Y., Rogstadius, J., Kukka, H., Kostakos, V. Crowdsourcing on the Spot: Altruistic Use of Public Displays, Feasibility, Performance, and Behaviours. *Proc. of UbiComp 2013*, ACM (2013), 753-762.
10. Goncalves, J., Hosio, S., Liu, Y., Kostakos, V. Eliciting Situated Feedback: A Comparison of Paper, Web Forms and Public Displays. *Displays 35*, 1 (2014), 27-37.
11. Gupta, A., Thies, W., Cutrell, E., Balakrishnan, R. mClerk: enabling mobile crowdsourcing in developing regions. *Proc. CHI '12*, ACM (2012), 1843-1852.
12. Harper, R. Texture: Human expression in the age of communication overload. MIT Press (2011).
13. Harrison, S., Dourish, P. Re-place-ing space: the roles of place and space in collaborative systems. *Proc. CSCW '96*, ACM (1996), 67-76.
14. Hosio, S., Goncalves, J., Kostakos, V. Application Discoverability on Public Displays: Popularity comes at a Price. *Proc. PerDis'13*, ACM (2013), 31-36.
15. Hosio, S., Kostakos, V., Kukka, H., Jurmu, M., Riekki, J., Ojala, T. From school food to skate parks in a few clicks: using public displays to bootstrap civic engagement of the young. *Proc. Pervasive 2012*, Springer-Verlag (2012), 425-442.
16. Hsieh, H.F., Shannon, S.E. Three approaches to qualitative content analysis. *Qualitative Health Research 15*, 9 (2005), 1277-1288.
17. Kai, N., Kannan, A. Criminisi, A., Winn, J. Epitomic location recognition. *Computer Vision and Pattern Recognition* (2008), 1-8.
18. Ko, S., Kim, K., Kulkarni, T., Elmqvist, N. Applying mobile device soft keyboards to collaborative multitouch tabletop displays: design and evaluation. *Proc. ITS '11*, ACM (2011), 130-139.
19. Kondracki, N.L., Wellman, N.S. Content analysis: Review of methods and their applications in nutrition education. *Journal of Nutrition Education and Behavior 34*, 4 (2002), 224-230.
20. Kostakos, V., Kukka, H., Goncalves, J., Tselios, N., Ojala, T. Multipurpose Public Displays: How Shortcut Menus affect Usage. *IEEE Computer Graphics and Applications 33*, 2 (2013), 56-63.
21. Kostakos, V., Juntunen, T., Goncalves, J., Hosio, S., Ojala, T. Where am I? – Location Archetype Keyword Extraction from Urban Mobility Patterns. *PloS ONE 8*, 5: e6398 (2013).
22. Krumm, J., Rouhana, D. Placer: semantic place labels from diary data. *Proc. UbiComp 2013*, ACM (2013), 163-172.
23. Kukka, H., Oja, H., Kostakos, V., Goncalves, J., Ojala, T. What Makes You Click: Exploring Visual Signals to Entice Interaction on Public Displays. *Proc. of CHI'13*, ACM (2013), 1699-1708.
24. Müller, J., Alt, F., Michelis, D., Schmidt, A. Requirements and design space for interactive public displays. *Proc. of Multimedia*, ACM (2010), 1285-1294.
25. Memarovic, N., Elhart, I., Langheinrich, M. (2011). FunSquare: First experiences with autopoiesic content. *Proc. of MUM' 11*, ACM (2011), 175-184.
26. Ojala, T., Kostakos, V., Kukka, H., Heikkinen, T., Linden, T., Jurmu, M., Hosio, S., Kruger, F., Zanni, D. Multipurpose Interactive Public Displays in the Wild: Three Years Later. *IEEE Computer 45*, 5 (2012), 42-49.
27. Paulos, E., Honicky, R.J., Hooker, B. Citizen Science: Enabling participatory urbanism. *Handbook of Research on Urban Informatics*, IGI Global (2009), 414-436.
28. Ribeiro, F.R., José, R. Timely and keyword-based dynamic content selection for public displays. *Proc. of CISIS*, IEEE (2010), 655-660.
29. Rogstadius, J., Teixeira, C., Vukovic, M., Kostakos, V., Karapanos, E., Laredo, J. CrisisTracker: Crowdsourced Social Media Curation for Disaster Awareness. *IBM Journal of Research and Development 57*, 3 (2013).
30. Schroeter, R., Foth, M., Satchell, C. People, content, location: sweet spotting urban screens for situated engagement. *Proc. of DIS '12*, ACM (2012), 146-155.
31. Schwarz, O. Subjectual Visibility and the Negotiated Panopticon: on the Visibility-Economy of Online Digital Photography (2011).
32. Storz, O., Friday, A., Davies, N., Finney, J., Sas, C., Sheridan, J.G. Public ubiquitous computing systems: Lessons from the e-campus display deployments. *Pervasive Computing 5*, 3 (2006), 40-47.
33. Sweetser, P., Wyeth, P. GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment 3*, 3 (2005), 3-3.
34. Väätäjä, H., Vainio, T., Sirkkunen, E., Salo, K. Crowdsourced news reporting: supporting news content creation with mobile phones. *Proc. MobileHCI '11*, ACM (2011), 435-444.
35. Van House, N.A. Collocated photo sharing, story-telling, and the performance of self. *International Journal of Human-Computer Studies 67*, 12 (2009), 1073-1086.
36. Von Ahn, L., Dabbish, L. Designing games with a purpose. *Communic. of the ACM 51*, 8 (2008), 58-67.