Human Computation (2019) 6:1:1-11

ISSN: 2330-8001, DOI: 10.15346/hc.v6i1.1

Local Crowdsourcing for Audio Annotation: the Elevator Annotator Platform

THEMISTOKLIS KARAVELLAS, Netherlands Institute for Sound and Vision

ANGGARDA PRAMESWARI, Netherlands Institute for Sound and Vision

OANA INEL, Department of Computer Science, Vrije Universiteit Amsterdam

VICTOR DE BOER, Department of Computer Science, Vrije Universiteit Amsterdam

ABSTRACT

Crowdsourcing is useful in collecting large numbers of annotations for various datasets. Local crowdsourcing is a variant where annotations are done at specific physical locations. This paper describes a local crowdsourcing concept, platform and experiment to gather annotations for an audio archive. For the experiment, we developed a hardware platform and its supporting software functionality, designed to be deployed in building elevators. To evaluate the effectiveness of the platform, test the impact of location, and the interaction interface on the annotation results, we set up an experiment in two locations. In each location we used two different user interaction modalities. Our results show that our simple local crowdsourcing setup achieves significant accuracy levels and generates up to 4 annotations per hour, depicting as well the correlation between location and accuracy.

1. GOALS

Crowdsourcing is a proven method for outsourcing a variety of tasks to a large network of people (Howe, 2006). This includes annotation and enrichment tasks, where for specific datasets, crowd annotators are asked to provide (additional) metadata. The domain of Cultural Heritage is where many such initiatives have been explored to enrich digital archives for various retrieval and research tasks (Oomen and Aroyo, 2011). Crowdsourcing campaigns are typically executed online, with the help of a variety of online platforms such as Amazon Mechanical Turk[1] (AMT) or

[1] https://www.mturk.com/mturk/welcome

FigureEight[2]. In many cases, the contributors of such platforms have no relation to the source material, which could help in motivating them.

Rather than engaging a crowd using online platforms, in this paper, we explore offline crowdsourcing. More specifically, we explore *local crowdsourcing* which mimics much of the features of online crowdsourcing with the addition of exploiting the characteristics of the physical location and the relation that contributors can have with this location. Local crowdsourcing serves the same purpose as online crowdsourcing but exploits the physical environment as a source and exposes the crowd to the possible influences from the location (Agapie et al., 2015). We study the feasibility of using local crowdsourcing for content enrichment and metadata gathering. Considering the Spatial Content Production Model (SCPM) classification (Hecht and Gegle, 2010) our experiment follows a hybrid approach by introducing a task for the creation of User Generated Content (UGC). The approach follows the "flat-earth" SCPM but explores the "you have to be there" SCPM effect on the results.

We present a case study which concerns annotation of audio heritage, archived at the Netherlands Institute for Sound and Vision (NISV[3]). Specifically, we describe an experiment in the elevator of this institute and compare results to those of a deployment in the elevator of an educational institution. Employees and visitors using this elevator are asked to perform an audio-based micro task concerning musical instrument identification. The annotations gathered through our experiment are preserved in the archives of NISV. We used elevators as an opportunity to gather annotations by taking advantage of the idle time that people spent in these elevators, with the intention of yielding the cognitive surplus, which is the basis of the success of crowdsourcing (Shirky, 2010).

In this paper, we therefore investigate the effectiveness of local crowdsourcing in the generation of accurate metadata enrichments for archival content collections. We furthermore present the implementation of this method, called the *ElevatorAnnotator* platform, based on small and affordable hardware.

2. Related work

Local crowdsourcing extends human computation into the physical environment through the use of on-site contributors to perform a given collaborative or distributive task (Agapie et al., 2015). Many local crowdsourcing approaches emerged through the use of mobile phones (Väätäjä et al., 2011; Gupta et al., 2012). People were advised to gather news-related information such as images or short updates by means of mobile phones (Väätäjä et al., 2011). Furthermore, people located in the vicinity of an event were asked to provide local information related to the development of the event and the information gathered was curated remotely by domain experts to generate event reports (Agapie et al., 2015). The user-generated event reports received 50% of additional media

[2] https://www.figure-eight.com/

[3] http://www.beeldengeluid.nl/en/netherlands-institute-sound-and-vision

content compared to conventional newspapers. Such approaches are clearly based on the "you have to be there" SCPM (Hecht and Gegle, 2010) following the incentive of recognition for the given contribution, as no other form of reward was given to its creators.

To take advantage of the short time frames when people check their phones, Vaish et al. introduced the Twitch mobile phone application (Vaish et al., 2014), which asks people to solve short tasks, for one or two seconds, to unlock their mobile phone. The authors replaced the standard slide-to-unlock mechanism with low time consuming and low cognitive load tasks such as image ranking or relevant information extraction. This is an example of the "flat earth" SCPM with operational incentives, as users are rewarded by gaining access to the desired operation. Community sourcing (Heimerl et al., 2012) is another example that uses vending machines, placed in college campuses near lecture halls, to crowdsource work from targeted expert groups, such as students. Students were asked to grade field-related exams and they were rewarded with snacks upon completion. This is another "flat earth" SCPM example where the incentive, though, is the receipt of the reward after the task completion. Research findings show that local crowdsourcing is better suited than traditional single-expert grading and that, comparatively, the AMT crowd lacks in performance on grading exams type of task.

Given the above mentioned studies we identify a number of dimensions to be considered in the design of crowdsourcing tasks; namely: a) the SCPM b) the incentive-reward combination c) the profile of the expected task participants.

Literature shows little focus on local crowdsourcing in the cultural heritage domain and this paper attempts to fill that gap. The most common crowdsourcing approach in the cultural heritage domain is through online platforms. Within this domain, crowdsourcing has several distinctive categories, depending on the use of the results. These categories are: correction and transcription, contextualization, complementation of the collection, classification, co-curation, and crowdfunding (Oomen and Aroyo, 2011).

Research conducted in the context of a video labeling game called "Waisda"[4] (Hildebrand et al, 2013) examined how gaming can be used as a method to enrich television heritage. The broader scope of the research was to determine the suitability of such an approach for integrating UGC tags with professional annotations (Oomen et al., 2010), (Gligorov et al., 2010). The results showed that the crowd sourced time-based metadata, i.e. tags, can be successfully used by media professionals to access specific media fragments. Such results framed the possibilities of this study to explore other methods for annotating archival collections.

[4] http://waisda.beeldengeluid.nl/

3. Experimental Methodology

This section describes the methodology followed in the experiment design and the design decisions regarding the micro-task to be performed by the participants.

The inspiration for the creation of the *ElevatorAnnotator* is the observation that there is sufficient time within an elevator ride to allow for the completion of a micro-task. In our case we examined the elevators of two buildings (NISV and the VU University building) of 7 and 6 floors respectively, and we found that the time needed for an elevator to "travel" between 3 floors is on average 30 seconds.

Having the intention to provide a crowdsourcing task that would yield UGC for the benefit of Cultural heritage we decided, following the framework of Oomen and Aroyo (2011), to implement a classification task, i.e. gathering of descriptive metadata for cultural heritage objects. Moreover, we chose the micro-task to address only audio collection annotations and not include text or video annotations to avoid further need of sensorial stimulation (other than the hearing of the participants). Regarding the type of audio enrichment, we examined multiple options such as singer identification, genre identification, etc. Considering the need for a simple and short task we decided to address musical instrument identification. We believe that the accuracy of the identification results would aid in investigating the effectiveness of the *ElevatorAnnotator* platform.

Furthermore, we also explore the effect of the location on the results accuracy: a cultural heritage institution where annotators could potentially have domain knowledge (i.e., nichesourcing) and an educational institution where we have no background information about the annotators. Research on nichesourcing (de Boer et al., 2012) showed that combining the knowledge of professionals with the knowledge of the crowd can optimize the results of human-based computation tasks in the cultural heritage domain. This dimension directly relates to the design variables mentioned in Section 2.

To summarize, we developed a platform that takes advantage of the cognitive surplus of workers during an elevator ride to perform a classification task (Oomen and Aroyo, 2011). The content to be classified is audio, the type of classification is musical instrument identification and the classification dimension to be examined is accuracy. This is done in order to require minimal sensorial stimulation of the participants and offer a simple micro-task that can indicate the effectiveness of the platform. The audio content chosen for the experiment is licensed under PD access and features minimal vocals to allow focus on the instruments. The task is a "flat-earth" SCPM (Hecht and Gegle, 2010), follows an intrinsic motivation model (Zheng, Li & Hou, 2011), and relates to the location through the profile of the expected elevator visitors. Last, we address two UI modalities, speech and buttons, and measure the difference in results between them.

The goal of the *ElevatorAnnotator* design is to provide a reusable platform for collecting accurate UGC, in this case audio annotations. *ElevatorAnnotator* is implemented as a portable, self-powered, audio annotation platform, based on the concept of pervasive computing and ubiquitous

computing technology[5]. The design, source code, audio content and hardware for the platform can be found at the project's homepage and GitHub repository[6].

The main design guidelines for the platform stem from the need for the platform to be (re)deployable in various locations. Also, since users are only briefly exposed to the platform, annotations must be added quickly and with limited user interaction. Following, we enumerate the main design decisions:

- *Location*: the platform's primary deployment locations is in elevators. This location was expected to get people's attention as well as offer them a short-time occupation while travelling between floors.
- *Stand-alone device*: Since in such locations power sockets, internet connections etc. might be unavailable, the device needs to be stand-alone. This includes power, processing and storage of the audio content as well as the annotations.
- *Process Control*: All processes had to be running on a small form factor machine without Internet connection and start, run and terminate automatically inside the elevator. Thus, we use a motion sensor to trigger a run cycle.
- *User Interaction:* To optimize the short contact time, we need a simple and natural user interaction. We implement two options for user interaction: speech recognition and simple large buttons. The user interaction is limited to binary operations ("yes/no") for both speech and buttons.

3.1 Hardware

A list of the hardware components used to assemble the platform is provided at the project's homepage. For the main processing unit of the device, we use a Raspberry Pi. We selected it because of its size, community support, as well as compatibility with many different hardware and software components. An external power bank supplies power to the platform, to fulfil the stand-alone requirement.

To initiate the user interaction, we use a passive infrared motion sensor that is able to detect participants entering the annotation space. A USB microphone and two 3-inch speakers are used to capture audio and play the audio file, respectively. Last, two labeled push buttons also allow for binary input (Y green for "yes", N red for "no").

The hardware was assembled and bundled in a small and friendly-looking box (Figure 1).

[5] https://en.wikipedia.org/wiki/Ubiquitous_computing

[6] https://ajprameswari.github.io/ElevatorAnnotator/

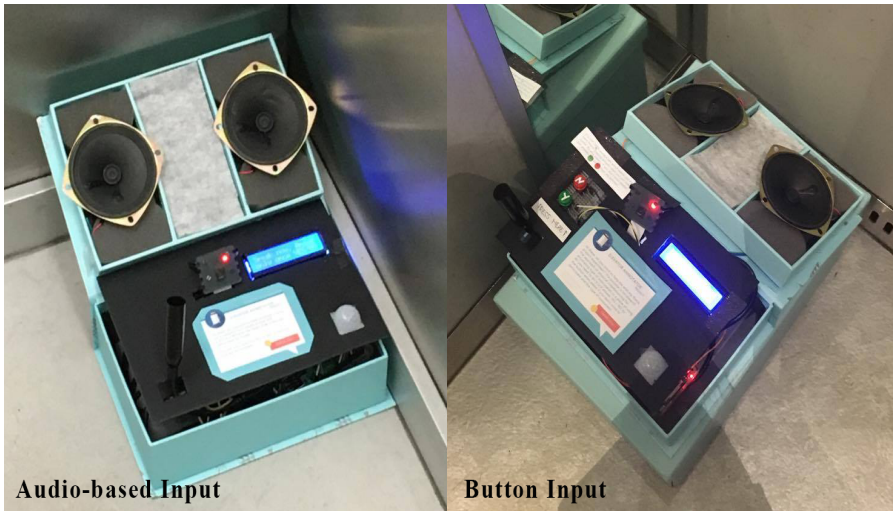

***Fig. 1* - The ElevatorAnnotator hardware**

3.2 The Annotation Process

The main behaviour of the platform is orchestrated by a Python script. We illustrate this behaviour through the following user interaction scenario:

1. When in *idle* mode, the motion sensor is enabled
2. When motion is detected the *annotation process* micro-task is initiated.
3. A greeting message ("Hello! Would you like to play a game?") is played by the speaker followed by a question to the user regarding his intention for participation.
4. Depending on the UI modality (speech/button) the user's choice is registered. In case the user does not intend to participate the process ends with a "Thank you" message. In case the user is willing to participate we proceed to step 5.
5. A random audio track is selected from the audio collection database. A sample of 7 seconds of the audio track is played out.
6. An annotation question is generated and played, e.g. "Did you hear a piano?"
7. The user answers through the UI at hand. In the case of speech input, the user is expected to answer verbally with a simple "Yes" or "No".
8. The elevator annotator stores the captured audio as well as the processed answer. The additional metadata (timestamp, experiment id, length of audio annotation) are also stored.
9. A "thank you" audio message is played and the system returns to *idle* mode.

In the case of audio mis-recognition, repetitions were limited to 3 attempts and 10 seconds of timeout was selected for buttons input. These timeframes are low to avoid trailing wait for user input. For the speech recognition functionality we used the Pocketsphinx[7] open-source library, mainly because of the lightweight processing requirements as well as its suitability for offline operation.

3.3 Experiment Design

In this experiment, we focus on the annotation of musical instruments in audio fragments, part of the NISV audio-visual archives. Musical instrument information is an important metadata attribute for audio files facilitating Music Information Retrieval (MIR). Furthermore, the identification of musical instrument can also be beneficial to classify musical genre, in which instruments were used as its distinctive features (McKay and Fujinaga, 2005).

The specific dataset used in this experiment was acquired from the Europeana Collection[8] of which we selected 10 random public domain licensed audio tracks. For each track we selected three time based audio samples at the 25%, 50% and 75% time marks of the original track. To evaluate our annotations, we created a gold standard by manually annotating the audio samples in advance. The dataset is available at the project's homepage and Table 1 shows the pre-annotated answers.

[7] https://github.com/cmusphinx/pocketsphinx

[8] http://www.europeana.eu/portal/

***Table 1* - Experiment audio sample list**

No	Title	Classification	Pre-annotated instrument answers			Time
			25%	50%	75%	
1	Oktāvu etīde	Piano music	Piano	Piano	Piano	1m34s
2	Tautas polka	Polkas, Folk dancing	Violin, Trumpet	Violin, Trumpet	Violin, Trumpet	2m40s
3	Dienā jaukā	Popular music	Violin, Trumpet	Trumpet	Violin, Trumpet	2m59s
4	Florentine	Popular music, Foxtrots	Trumpet	Trumpet	Violin, Trumpet	2m59s
5	Mana dzimtene	Foxtrots	Violin, Trumpet	Trumpet, Violin	Trumpet, Violin	2m57s
6	Meitenes sirsniņa	Operas	Piano, Violin	Violin, Piano	Violin, Piano	2m08s
7	Kādēļ tik ilgi vilcinies tu?	Foxtrots, Jazz	Violin, Trumpet, Piano	Trumpet Violin	Trumpet, Piano, Violin	2m46s
8	Dziedu tev	Popular music	Violin, Piano	Violin, Piano, Trumpet	Violin, Piano	2m53s
9	Serenade iz operas	Operas, Arranged	Piano	Piano	Piano	1m57s
10	Serenade	Violin w/ orchestra	Violin, Piano	Violin, Piano, Trumpet	Violin, Piano, Trumpet	2m31s

The goal of the experiment is two-fold: (1) to compare results based on the location of the platform, and therefore on the profile of the participants, and (2) to compared the results between the two available UI modalities. The platform was deployed in two locations, in the employee elevator of NISV and the elevator of a local university (VU Amsterdam). The number of floors in both locations is similar, NISV has 7 floors counted from the ground level, whereas VU has 6 floors. In each location, experiments were carried out for a total period of 16 hours. Each UI modality was deployed for 4 hours in each locations.

4. Results

Through the deployment of the *ElevatorAnnotator,* as described above, we collected a significant number of annotations. A complete view of the experimental results is available as supplemental material to this paper (Anggarda, 2017). The four conditions of the experiment yielded 264 recorded responses in 930 minutes, out of which 141 responses were correctly identified. Of these 141, 65 participants agreed to join the experiment. Their answers were compared to the gold standard. Table 2 shows the resulting annotation statistics, where the platform's operation accuracy in the designed task is depicted.

Given that in total the experiment lasted for 930 minutes from which 60 identified and valid instrument annotations were collected, we calculate an average of 3.9 successful annotations per hour. This number depicts a positive effectiveness of the *ElevatorAnnotator* platform as a method of UGC collection for audio content annotation.

***Table 2* - ElevatorAnnotator experiment results**

		Duration (mins)	Total recorded answers	Unidentified answers	Identified answers	Instrument annotation answers			Accuracy[9] rate
						Correct	Incorrect	Invalid	
NISV	Audio	240	55	35	20	5	2	2	0.71
	Button	210	81	32	49	19	5	0	0.79
VU	Audio	240	67	38	29	3	6	2	0.33
	Button	240	61	18	43	11	9	1	0.55
Location Total	NISV	450	136	67	69	24	7	2	0.77
	VU	480	128	56	72	14	15	3	0.48
Modality Total	Audio	480	122	73	49	8	8	4	0.50
	Button	450	142	50	92	30	14	1	0.68
Total		930	264	123	141	38	22	5	0.61

Chi-squared test shows that the difference in annotation accuracy between the two locations (0.77 for NISV vs 0.48 for VU) is statistically significant ($p=0.019$). We therefore see a positive correlation between the location, and more specifically the profile of the people at that location, and the accuracy in annotating audio content, compared to participants in a content irrelevant location, i.e. VU. We also observed that the difference between the two UI variants (0.68 vs 0.50) is not significant.

5. Conclusion

Our local crowdsourcing approach, designed and implemented in a reusable platform, the *ElevatorAnnotator*, offers an effective solution for eliciting annotations from on-site participants. For this specific annotation task, we are able to achieve an accuracy of 61%. Executing the same experiment in online crowdsourcing or with automatic tools and comparing the results to the currently presented results is left as future work.

In our experiment we present a binary annotation case, limiting the extrapolation to additional annotation options, or even arbitrary textual annotations. Moreover, we follow a hybrid combination of "flat-earth" SCPM and "you have to be there" SCPM, where the location is irrelevant to the task but relevant to the users' profile and therefore, to the results. More complex

[9] Accuracy = Correct answers / (Correct answers + Incorrect answers)

types of UGC collection or annotation are left as open research for the future (e.g. translation of language fragments, elevator quality feedback collection, video/image annotation, etc.).

The described crowdsourcing approach, which combines pervasive computing components, is a promising method for metadata gathering for audio collections. Our experiments show that there is a significant association between the different locations and the annotation results accuracy as well as a significant level of platform effectiveness (3.9 annotations per hour). This indicates that local crowdsourcing can be a valuable way of eliciting high accuracy annotations.

6. Acknowledgements

This research was partly funded through the Indonesia Endowment Fund for Education.

7. References

Agapie, E., Teevan, J., & Monroy-Hernández, A. (2015, September). Crowdsourcing in the field: A case study using local crowds for event reporting. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

Anggarda, P. (2017). Experiment Results. figshare. Retrieved 1 July 2017, from https://doi.org/10.6084/m9.figshare.5106844.v1

De Boer, V., Hildebrand, M., Aroyo, L., De Leenheer, P., Dijkshoorn, C., Tesfa, B., & Schreiber, G. (2012, October). Nichesourcing: harnessing the power of crowds of experts. In *International Conference on Knowledge Engineering and Knowledge Management* (pp. 16-20). Springer, Berlin, Heidelberg.

Gligorov, R., Baltussen, L. B., van Ossenbruggen, J., Aroyo, L., Brinkerink, M., Oomen, J., & van Ees, A. (2010). Towards integration of end-user tags with professional annotations.

Gupta, A., Thies, W., Cutrell, E., & Balakrishnan, R. (2012, May). mClerk: enabling mobile crowdsourcing in developing regions. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1843-1852). ACM.

Hecht, B. J., & Gergle, D. (2010, February). On the localness of user-generated content. In Proceedings of the 2010 ACM conference on Computer supported cooperative work (pp. 229-232). ACM.

Heimerl, K., Gawalt, B., Chen, K., Parikh, T., & Hartmann, B. (2012, May). CommunitySourcing: engaging local crowds to perform expert work via physical kiosks. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1539-1548). ACM.

Hildebrand, M., Brinkerink, M., Gligorov, R., Van Steenbergen, M., Huijkman, J., & Oomen, J. (2013, October). Waisda?: video labeling game. In Proceedings of the 21st ACM international conference on Multimedia (pp. 823-826). ACM.

Howe, J. (2006). The rise of crowdsourcing. Wired magazine, 14(6), 1-4.

McKay, C., & Fujinaga, I. (2005, March). Automatic music classification and the importance of instrument identification. In Proceedings of the Conference on Interdisciplinary Musicology.

Oomen, J., Belice Baltussen, L., Limonard, S., van Ees, A., Brinkerink, M., Aroyo, L., Vervaart, J., Asaf, K. & Gligorov, R. (2010). Emerging practices in the cultural heritage domain-social tagging of audiovisual heritage.

Oomen, J., & Aroyo, L. (2011, June). Crowdsourcing in the cultural heritage domain: opportunities and challenges.In Proceedings of the 5th International Conference on Communities and Technologies (pp. 138-149). ACM.

Shirky, C. (2010). Cognitive surplus: Creativity and generosity in a connected age. Penguin UK

Väätäjä, H., Vainio, T., Sirkkunen, E., & Salo, K. (2011, August). Crowdsourced news reporting: supporting news content creation with mobile phones. In Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (pp. 435-444). ACM.

Vaish, R., Wyngarden, K., Chen, J., Cheung, B., & Bernstein, M. S. (2014, April). Twitch crowdsourcing: crowd contributions in short bursts of time. In Proceedings of the 32nd annual ACM conference on Human factors in computing systems (pp. 3645-3654). ACM.

Zheng, H., Li, D., & Hou, W. (2011). Task design, motivation, and participation in crowdsourcing contests. International Journal of Electronic Commerce, 15(4), 57-88.