IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 28, 2020, accepted July 2, 2020, date of publication July 9, 2020, date of current version July 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3008270

Cultural Heritage Design Element Labeling System With Gamification

JIEUN LEE, JI HYUN YI, AND SEUNGJUN KIM, (Member, IEEE)
School of Integrated Technology, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, South Korea

Corresponding author: Seungjun Kim (seungjun@gist.ac.kr)

This research was supported in part by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2020, and in part by the GIST Research Institute (GRI) GIAI grant funded by the GIST in 2020.

ABSTRACT Cultural heritage (CH) artifacts, such as ceramics and clothes, reflect the unique characteristics of ancient cultures and have the potential to be sustainably employed in modern design and entertainment. In particular, the shape of ceramics reflects regional and historical characteristics, so datafication is a promising avenue to preserve these assets for future generations. However, design is a specialized domain that requires significant human (expert and novice) labor. This often tedious process decreases the labeler's motivation to complete the task, and data consistency varies with the experience and motivation of the labeler. To increase engagement, we developed an image labeling platform with graphical icon-based labeling methods and introduced gamification. The robust labeling methods with gamification increased novices' engagement and decreased the workload of expert and novice labelers, but decreased data agreement between experts and novices, so we consider opportunities for gamification within the specialized cultural heritage domain.

INDEX TERMS Cultural heritage artifact, design element, gamification, human-based computation, image labelling platform, workload.

I. INTRODUCTION

Cultural heritage (CH) artifacts are an important link between cultures, as well as between the past and the future [1]. These objects have significant historical and educational value; shapes, decorations, and materials give us insight into the beliefs, economic trends, and lifestyles of the people in a particular region or time period [2]–[4].

Unfortunately, almost 90% of the world's CH artifacts cannot be exhibited due to damage or a lack of pre-classified data for classification and reconstruction [5]. To establish a classification system, researchers must rely on labeling to create a quantitative database. Such systematic analysis will save time in analyzing newly excavated artifacts and provide a blueprint for the reconstruction of lost artifacts, thus helping researchers better understand their historical and educational value [6]. Analysis and classification of CH artifacts can bring together diverse fields such as archaeology, historiography art history, fashion and design, and provide a common point of reference for cross-disciplinary collaboration [7], [8]. Further, in classifying specific properties or design features of a given time, the labeling process provides historical evidence and documentation that could inform hypotheses related to historical migrations and cultural propagation. Aside from advancing historical reconstructions, CH artifacts advance tourism and merchandizing and thus contribute significantly to local and global economies.

Among all possible cultural heritage artifacts, ceramic ware is an ideal labeling subject. Ceramic artifacts have been dated to roughly 20,000 years and are plentiful yet diverse in their design features and formative elements [9]. They are closely related to the lives of their users and have been shaped according to their use, such as for cooking, religious events or storing food. In addition to representing users' lifestyles, ceramics reflect the geographic characteristics of a given time, and often incorporate special styles, patterns, or decorative forms linked to a civilization, nationality, dynasty or ancient ethnic identity [10], [11]. For example, the characteristics and patterns of ceramic wares from the Qing and Song Dynasties of 17th-century China differ, yet are similar to European ceramics of the same period.

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

A. LABELING SYSTEMS AND HUMAN LABOR

Image labeling, or image annotation, is a process of identifying and tagging objects and specific features in an image for further analysis and data processing [12]. First, human labelers annotate the image; previous studies have incorporated objects such as handbags and biological images [13]. The generated metadata is then used in human-computer collaborations or to train computer-based automated labelers [14], [15]. With computer vision technology, such as datafication, big data processing, and digital curation are thus promising strategies for determining the value of cultural heritage assets through preservation and reconstruction [16], [17]. However, the historical and artistic value of ceramic pieces are non-linguistic and non-quantified data, which auto-labeling or feature extraction technology cannot classify or annotate as well as humans labelers [18]. Human labor and expertise are needed to extract specialized information, such as specific shapes and symbolic patterns, from cultural heritage artifacts. But, one or a few human experts cannot label a large number of assets, and more volunteers are needed. Moreover, while there are many image labeling methods, such as crowdsourcing and educational tasks, metadata from novice labelers is typically insufficient, and the usability and accessibility of annotation systems for labelers are limited [19], [20]. In other words, data agreement among volunteers is as important as data quantity to ensure consistency [21]. To increase the available labor and data agreement of a labeling system, volunteers of many backgrounds must be able to easily access and learn about the system.

B. LABELING DATA CONSISTENCY

The purpose of design element labeling is converged label data for a given object [22]. Because images can affect emotions, labelers can create rich data. However, in addition to affective experience with an image, many other factors such as gender, age, and style of writing affect the agreement of responses from human labelers, which may result in subjective and uncertain data [23].

Under the best circumstances, labelers most commonly follow the text coding method, describing an object by typing its label into the program or system, which is still a subjective process [24], [12]. We reviewed 33 image labeling programs (e.g. LabelImg (Git code), Lablebox (Git code), Images annotation programme (Git code)), most of which offer the text coding method, where labelers select or draw what they want to label in the image and create text-based labels without guidelines to prevent labeling data divergence. When objects labels do not match, data consistency is likely insufficient for future applications [6], [25]. Thus, the traditional distributed textual labeling method is not ideal for the classification of design elements of cultural heritage artifacts. A more robust method is needed to ensure consistent labeling data for specialized description and comparison.

C. ENGAGEMENT OF LABELER

Motivation is key to performance in labeling tasks, which are repetitive and time-consuming. Consequently, researchers and data engineers should consider the labeler's workload and engagement. Gamification is a promising strategy for decreasing perceived workload and increasing labeling motivation and throughput [26]–[28]. Playing a game reframes the task as enjoyable and goal-oriented, and reduces pressure and tension, as seen in the successful human-based computation game ESP used by Google Image Labeler. Well-known game elements, such as points, badges, and leaderboards (PBL) have likewise been applied to educational assignments, with positive results [29]. By satisfying users' innate psychosocial needs of autonomy, competence, and relatedness, gamification provides intrinsic motivation which leads to increased work throughput without affecting task performance [26], [30]. Despite its success, gamification has not been applied to specialized labeling tasks, like CH design element annotation.

D. CONTRIBUTIONS

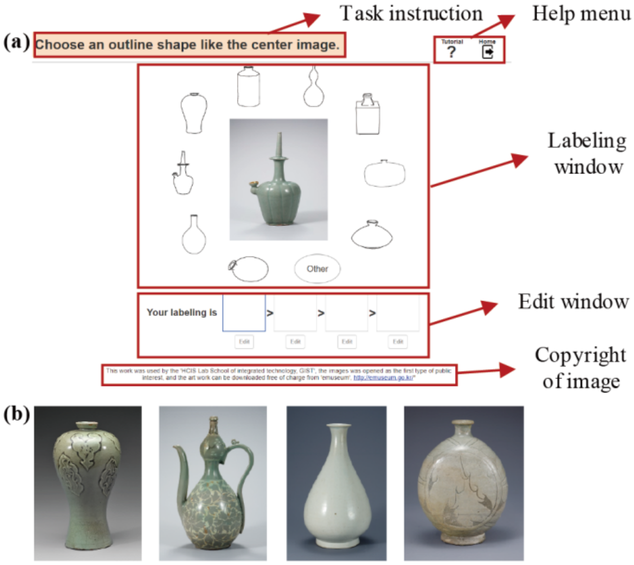
In this paper, we develop an image labeling system for the specialized purpose of annotating cultural heritage design elements in ceramic artifacts. We develop a robust annotation method to increase labeling data agreement between novice and expert labelers, and we evaluate the experimental data agreement. We also add gamification elements. We hypothesize that gamification will increase our labelers' motivation and decrease their workload, and we observe gamification's effect on labeling data agreement. The research questions (RQs) we address are: (RQ1) What factors of the labeling system for the design element of ceramic cultural heritage artifacts encourage labelers who do not have expertise in CH artifacts and design? (RQ2) How can the method for specialized design labeling increase data convergence of novice labelers and decrease divergence in data agreement between expert and novice labelers? (RQ3) How does the gamified labeling task affect motivation, workload, expected continuous working time/work throughput, and labeling data agreement?

Our primary contributions are:

- The graphical icon-based labeling method supports accessible and complete labeling work for participants without CH design expertise and promotes high data agreement between expert and novice labelers.
- Gamification applied to the CH design element labeling system leads to increased time on task and generates more data.
- Gamification increases the intrinsic motivation and reduces the workload of labelers but decreases data consistency of experts.

II. SYSTEM DESIGN

The design elements of ceramics are an exposed visual index reflecting the ancestor's lives, thought, cultural trends and art, making ceramics a good candidate for labeling and further datafication. We chose Korean ceramics as the labeling objects because of the sufficient number and variation of ceramic CH artifacts in the region. To guide the labeler,



Choose an outline shape like the center image.

This work was used by the 'HCIS Lab School of integrated technology, GIST', the images was opened as the first type of public interest, and the art work can be downloaded free of charge from 'emuseum'. http://emuseum.go.kr/"

FIGURE 1. **(a) The labeling stage of the graphical icon-based image labeling system prototype and (b) the different types of ceramic ware for labeling work.**

we analyzed the design elements of 922 ceramics images in museums and identified classes of shapes: nine outline shapes, four body shapes, six mouth shapes, three handle shapes, and each 'other' class. These pre-identified formation elements are graphical icons that guide the work on the labeling system and provide examples. Participants could also classify designs as 'other' (indeterminate), labels that could be later analyzed by experts. We developed the labeling system on the web using the prototyping tool Axure and programming tool Scratch. We added gamification to measure workload, engagement, and labeling data convergence [31], [32].

A. CH ARTIFACTS DESIGN ELEMENT ANNOTATION SYSTEM WITH NEW LABELING METHOD

We developed the graphical icon-based labeling system prototype to encourage engagement in novice labelers. We then analyzed their performance and survey responses to identify system development factors that support them in specialized labeling tasks.

As shown in Fig. 1, the labeling system consists of task instructions for labeling, a help menu, a labeling window with center-positioned ceramic ware images, an edit window, and copyright information. The labeler follows the task instructions for each stage to label the design elements (shape of outline, body, mouth, and handle) and edit the originally-labeled design elements of ceramics image. For example, in the outline shape labeling stage, the labeler selects the outline shape icon that is most consistent with the given object from nine possible outline shape icons. If the labeler cannot find the same or a similar outline shape icon, the labeler selects the 'other' icon. The labeler repeats this process for each of the four design elements for each object.

B. NON-GAMIFICATION VS. GAMIFIED LABELING SYSTEM

We modified the labeling system prototype to test non-gamified and gamified versions of our labeling system. An additional function of both systems was to record the selections of each labeler for further analysis.

To initiate labeling work with the non-gamified labeling system, labelers typed a username, trained how to label by watching a video tutorial, then followed the instructions to annotate design elements in the labeling window by selecting icons that matched the identified shape elements. Labelers could check the number of remaining images and their overall progress. In the non-gamified labeling system, we focused on the influence of the pre-identified icon labeling method to the data convergence of expert and novice labelers, and between individual novice labelers.

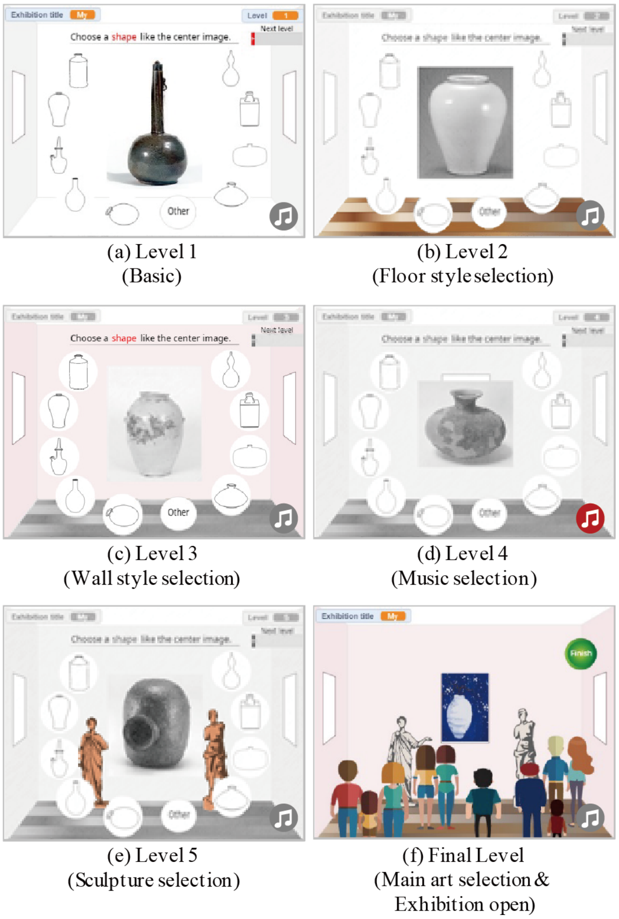
In our gamified labeling system, the game components are the goal setting, the level, the reward, and the customization. These game elements have been shown to increase intrinsic motivation and the amount of data gathered in crowdsourcing [19], [33], [34], [35]. The goal-setting and level elements motivate users to participate in the task, and the level, especially, encourages participants to accomplish the task to reach the next level; it is a promising tool for gathering a significant amount of data. The rewards element encourages novice labelers to continue to participate in labeling. However, once a goal is achieved, rewards are no longer effective. Therefore, the customizing element is used to offer variety and maintain an interest in labeling.

In this study, the goal of the game is for the participant to develop an 'exhibition' of cultural heritage artifacts. In Level 1, participants begin by labeling 10 images. For Levels 2-5, they label an additional 5 images per level (Level 2 = 15, Level 3 = 20, ...), for a total of 100 labeled images. Upon completing each level, participants choose a reward with which to customize their final exhibition space (Level 6). Rewards include interior items for the exhibit space: floor (Level 2), wallpaper (Level 3), music (Level 4), sculpture (Level 5), and a painting (Level 6). When participants reach the final level, the exhibition is held and the labeling task is complete, as shown in Fig. 2. Our non-gamified and gamified labeling systems are available online at https://scratch.mit.edu/projects/324495038 and https://scratch.mit.edu/projects/320155545, respectively.

III. EXPERIMENT

A. STUDY 1

The goal of Study 1 is to find the usability factor to increase novice labelers' motivation for design element labeling in an expert CH domain. We used the snowball sampling method to require volunteers to test and evaluate our labeling system. The 11 volunteer labelers (age: ten 20-29 and one 30-39; gender: nine men, two women) participated in the labeling task. Among them, two labelers had a background in cultural heritage objects and design elements. The labelers were asked to use an image-based design element labeling system



(a) Level 1 (Basic)

(b) Level 2 (Floor style selection)

(c) Level 3 (Wall style selection)

(d) Level 4 (Music selection)

(e) Level 5 (Sculpture selection)

(f) Final Level (Main art selection & Exhibition open)

FIGURE 2. The gamified labeling system includes (a) Level 1-Basic stage, (b) Level 2-Floor style selection rewards, (c) Level 3-Wall style selection rewards, (d) Level 4-Music selection rewards, (e) Level 5-Sculpture selection rewards, and (f) Final Level-Main art pottery painting selection rewards and exhibition opening.

and then responded to a 5-point Likert scale-based modified system usability scale (SUS) questionnaire (1 = Not at all, 5 = Strongly agree). We also added questions about what prior knowledge of cultural heritage objects and of design they believed is required to successfully perform labeling work [36].

B. STUDY 2

In Study 2, we aimed to measure data agreement to analyze the effectiveness of the graphical icon-based labeling method. We observed the game effect on labelers' engagement, workload, and data agreement. Also, we surveyed participants about their expected work throughput and duration using the suggested labeling system. Participants were asked to label four design elements (shape of outline, body, mouth, and handle) for 100 images total, resulting in 400 labels per system. This process was repeated for both the non-gamified and the gamified system. The order of the systems was random.

We considered expertise as CH and design-related knowledge in this domain, but the need for CH knowledge (NCHK) and the need for design knowledge (NDK) were non-essential components for labeling tasks. Therefore, expertise was considered as design-related work experience of more than three years. Using the random sampling method, we recruited participants via SNS. Two participants reported working in the design area for more than three years; the other responders were invited as novices. This proportion of experts to novices is consistent with other data agreement research [21], [37]. Participants' ($n = 26$; 16 men and 10 women) average age was 22.3 ($SD = 2.5$). Participants were compensated $10 (USD) per hour, following an approved Institutional Review Board (IRB) for protection of human subjects in research.

After labeling all 100 images for each system, participants answered surveys to assess their intrinsic motivation. The Intrinsic Motivation Inventory (IMI) is a self-reporting questionnaire that consists of multi-dimensional components related to intrinsic motivation to explain behavior and engagement [30], [38], [39]. The IMI measures participants' intrinsic motivation by completing the given task with non-gamified and gamified labeling systems. Interest/enjoyment (I/E) is interpreted as self-reported intrinsic motivation. Perceived competence (Pcom) and perceived choice (Pch) are interpreted as positive indicators of internal and external measures of intrinsic motivation. Pressure/tension (P/T) is interpreted as a negative indicator of intrinsic motivation.

In this study, we used a version of the IMI questionnaire with 22 items and four subscales. These surveys were conducted using the task evaluation IMI with a 6-point Likert scale (1 = Not at all, 6 = Strongly agree) to measure labelers' interest/enjoyment, perceived competence, perceived choice, and pressure/tension. The 6-point scaling was used to remove the neutral response of odd scaling and eliminate the possibility of misinterpretation [40].

We also assessed labeling workload using the NASA Task Load Index (NASA-TLX) with a 10-point scale [41]. The NASA-TLX rates 1) mental, 2) physical, and 3) temporal demands, as well as 4) performance, 5) effort, and 6) frustration experienced by the user, and then compares in pairs of six factors to find how each factor contributes to the workload [42]. Overall workload was calculated by summing the rating of six indices. We also asked, 'How many images can you label with this system?' and 'How long can you label with this system?' to determine throughput and time on task.

We analyzed data agreement between experts and novices, and between each novice [43], [44], using Cohen's kappa and Fleiss' generalized kappa. Cohen's kappa is used for two raters, and Fleiss's kappa is an adapted version of Cohen's kappa for three or more raters [45], [46]. Because we compared data between experts and novices, and between individual novices, different kappa coefficients were used.

IV. RESULTS

A. STUDY 1

The system usability score was 75.2, which is "good" for an image-based labeling system and suggests our system can be utilized as a real-world labeling system and potentially be

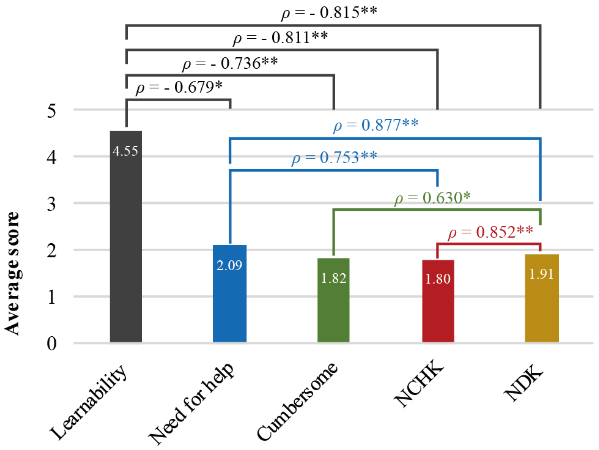


FIGURE 3. Spearman correlations (ρ) of system usability subscales.

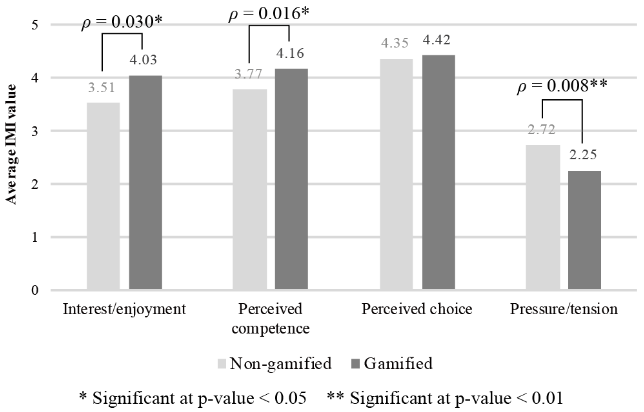


FIGURE 5. Average IMI scores of non-gamified and gamified labeling systems.

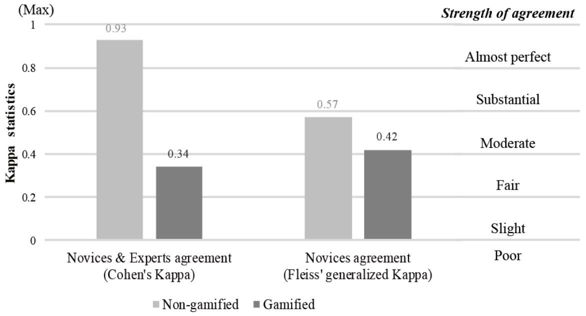


FIGURE 4. Labeling data agreement between novice labelers and expert labelers and individual novice labelers with corresponding strength of agreement levels.

commercialized [47]. We analyzed the correlations (ρ) of system usability factors from SUS and additional questionnaires as NCHK and NDK for the labeling task.

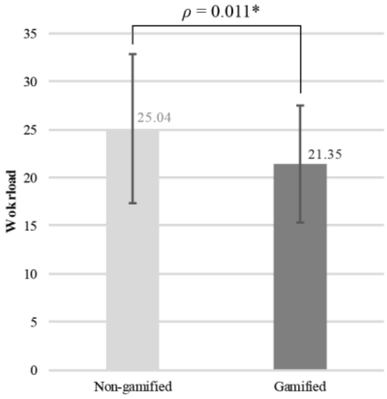
In Fig. 3, Learnability is correlated with several system usability factors and needs of NCHK and NDK. Learnability ($M = 4.55$, $SD = 0.69$) has an inverse relation with Need for help ($M = 2.09$, $SD = 1.04$; $\rho = -0.679$, $p = 0.022$) and Cumbersome ($M = 1.82$, $SD = 0.75$; $\rho = -0.736$, $p = 0.010$), but is proportional to NCHK ($M = 1.82$, $SD = 1.25$; $\rho = -0.811$, $p = 0.002$) and NDK ($M = 1.91$, $SD = 1.04$; $\rho = -0.815$, $p = 0.002$). Need for help is inversely correlated with NCHK ($\rho = 0.753$, $p = 0.007$) and NDK ($\rho = 0.887$, $p = 0.000$). Cumbersome and NDK ($\rho = 0.630$, $p = 0.038$) are inversely proportional, and NCHK and NDK ($\rho = 0.852$, $p = 0.001$) are proportional. These results suggest the novice labeler can easily use our system without help and the system is not cumbersome.

B. STUDY 2

In Fig. 4, Cohen's kappa statistics between novice labelers and expert labelers with the non-gamified labeling system is 0.93, with $p = 0.000$, and Fleiss' generalized kappa statistics of individual novice labelers is 0.57, with $p = 0.000$. The calculated range of the kappa statistic is $<$ 0.00 to 1.00, with six levels at 0.2 intervals. The corresponding strength of agreement includes 'poor,' 'slight,' 'fair,' 'moderate,' 'substantial,' and 'almost perfect' [48]. The kappa statistic is interpreted as the strength of an 'almost perfect' level and a 'moderate' level of agreement. Our graphical icon-based labeling method shows high labeling data agreement. The labeling data agreement with the gamified system between experts and novices is 0.34 with $p = 0.000$ (fair) and data agreement of individual novice labelers is 0.42 with $p = 0.000$ (moderate). The data agreement for the gamified system was much lower; the kappa statistics of novices and experts decreased by 0.59 and those of novices decreased by 0.15 with the gamified condition.

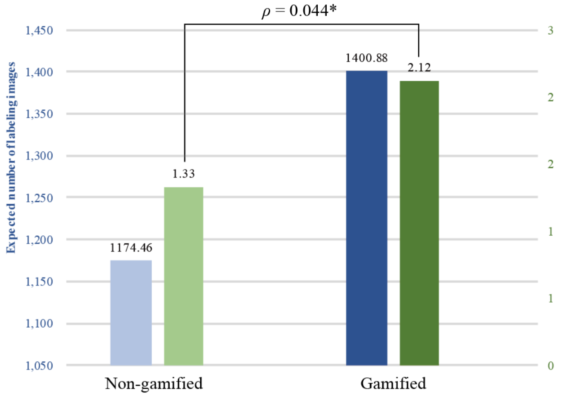
In Fig. 5, the average IMI scores for interest/enjoyment ($M = 4.03$, $SD = 0.89$), perceived competence ($M = 4.16$, $SD = 0.62$) and perceived choice ($M = 4.42$, $SD = 0.79$) of gamified system are higher than average IMI scores for interest/enjoyment ($M = 3.51$, $SD = 1.01$), perceived competence ($M = 3.77$, $SD = 0.89$), and perceived choice ($M = 4.35$, $SD = 0.65$) for the non-gamified system. The pressure/tension responses show the opposite: gamification scores ($M = 2.25$, $SD = 0.88$) lower than without gamification ($M = 2.72$, $SD = 1.19$). We analyzed these values using Wilcoxon signed-rank *post-hoc* test to compare each subscale value that user's experience between the non-gamified and gamified systems. For interest/enjoyment and perceived competence, the gamified system scores significantly higher ($Z = 2.17$, $p = 0.030$) than the non-gamified system ($Z = 2.40$, $p = 0.016$). But the score of the non-gamified system is higher than the gamified system ($Z = -2.66$, $p = 0.008$) for pressure/tension. With the gamified system, labelers were more likely to enjoy and feel engaged with the task.

Workload was measured by NASA-TLX with different labeling systems and analyzed using a paired t-test to compare the non-gamified and gamified systems. The workloads with/without gamification are 25.04 ($SD = 7.75$) and 21.35 ($SD = 6.07$), respectively, and the workload of the gamified system is statistically lower than the workload of the non-gamified system with $Z = -2.54$ with $p = 0.011$, as shown in Fig. 6. The workload with the gamified system decreases



* Significant at p-value < 0.05

FIGURE 6. Workload of non-gamified and gamified systems.



* Significant at p-value < 0.05

FIGURE 7. Expected number of labeling work throughput and work time between non-gamified and gamified systems.

14.74% from that of the non-gamified system. In other words, labelers using the non-gamified system reported a higher workload and lower motivation than labelers who used the gamified system.

To approximate the time for labeling work with and without gamification, the number of images and task duration were added to the survey; we asked how long participants can do labeling work and how many images they can label. To analyze continuous value data like time and number of images, we conducted Pearson correlation and a paired t-test analysis. The expected average number of images labeled and work duration without gamification were 1174.46 ($SD =$ 2264.84) and 1.33 hours ($SD = 1.30$), and they are correlated with the Pearson coefficient $\rho = 0.63$ with $p = 0.001$. The expected number of images labeled and work duration with gamification were 1400.88 ($SD = 2506.19$) and 2.12 hours ($SD = 2.58$), and they have a correlation of $\rho = 0.98$ with $p = 0.000$. The result of the paired t-test between expected labeling work times of non-gamified and gamified systems is t-value $= -2.121$ with $p = 0.044$, as shown in Fig. 7. Additionally, the total labeling task completion time and time to label one image with two different systems were measured. When participants labeled 100 images with and without gamification elements, the average task completion time was 799.81 s (SD = 289.85) and 835.66 s (SD = 144.3), and the average time to complete one label was 2.09 s (SD = 1.2) and 2.00 s (SD = 0.95). The total task completion time difference averaged 35.85 s, and the time difference to complete one label averaged 0.90 s. Gamification saves labelers time and boosts labeler productivity.

As shown in Table 1, without gamification, I/E is correlated with Pcom ($\rho = 0.43$, $p = 0.030$), Pch ($\rho = 0.70$, $p =$ 0.000) and P/T ($\rho = -0.64$, $p = 0.000$). Pcom is inversely correlated with P/T ($\rho = -0.61$, $p = 0.001$) and Pch is inversely correlated with P/T ($\rho = -0.63$, $p = 0.001$). All workload (WL) correlates with the intrinsic motivation factors; I/E ($\rho = -0.52$, $p = 0.006$), Pcom ($\rho = -0.65$, $p = 0.000$), Pch ($\rho = -0.58$, $p = 0.002$), P/T ($\rho = -0.82$, $p = 0.000$). Increased motivation encourages the labeler to experience a lower feeling of workload, but both motivation and workload were not related to the expected work throughput and duration. With gamification, I/E is correlated with Pcom ($\rho = 0.47$, $p = 0.016$). Pcom inversely correlates with P/T ($\rho = -0.44$, $p = 0.026$) and Pch is inversely correlated with P/T ($\rho = -0.63$, $p = 0.003$). WL inversely correlates with I/E ($\rho = -0.49$, $p = 0.011$) and Pcom ($\rho = -0.56$, $p =$ 0.003), but proportionally correlates with P/T ($\rho = 0.76$, $p =$ 0.000). I/E also correlates to Et ($\rho = 0.49$, $p = 0.010$) and Ew ($\rho = 0.45$, $p = 0.021$). With gamification, the I/E, Pcom and P/T affected workload, but perceived choice does not correlate to other factors. One thing to note is interest/enjoyment affects expected work throughput and duration. Gamification increases the labelers' motivation, expected work throughput, and duration, and decreases workload.

TABLE 1. Spearman correlation of intrinsic motivation, workload, expected work throughput and time with/without gamification.

		I/E	Pcom	Pch	P/T	WL	Ew	Et
Non-gamified	I/E	1.00						
	Pcom	0.43*	1.00					
	Pch	0.70**	0.28	1.00				
	P/T	-0.64**	-0.61**	-0.63**	1.00			
	WL	-0.52**	-0.65**	-0.58**	0.82**	1.00		
	Ew	0.25	0.30	0.19	-0.09	-0.23	1.00	
	Et	0.22	0.24	0.07	-0.08	-0.20	0.71**	1.00
Gamified	I/E	1.00						
	Pcom	0.47*	1.00					
	Pch	0.02	-0.14	1.00				
	P/T	-0.33	-0.44*	-0.31	1.00			
	WL	-0.49*	-0.56**	-0.30	0.76**	1.00		
	Ew	0.49*	0.25	-0.18	-0.22	-0.28	1.00	
	Et	0.45*	0.20	-0.18	-0.22	-0.17	0.76**	1.00

* Significant at p-value < 0.05 ** Significant at p-value < 0.01

V. DISCUSSION

The results show that our proposed annotation system for the labeling of CH artifact design elements encourages novice labelers to engage in the labeling task. High learnability, an important HCI principle, is an essential factor for a common labeling system, so users can easily use and be satisfied with the system [49]. However, as current research points out, system usability differs according to the purpose of system development and the user, so key factors of system usability and the intended testing should be adapted accordingly [50], [51]. Cultural heritage artifact design, of ceramics in particular, is special for labeling and the target user is a novice, so learnability is an important requirement. Our results showed that the novice labeler who uses the system and needs help from a developer or system manager may feel that CH and design expertise are needed to complete the task. As the system became less cumbersome, users reported that they could complete the task without design expertise. The low need for help, perception of the system as not cumbersome, and high learnability of the system increased the engagement of participants with no experience in cultural heritage and design.

Specialized annotation tasks are critical in medicine, autonomous driving, and cultural heritage preservation. Data consistency is important and expert participation is required [52], [53]. Despite the computer vision technologies applied, critical heritage and design are artistic domains, so human experts are still needed as evaluators or collaborators for high data consistency [54], [55]. Moreover, the interpretive nature of labeling produces a wide data range, so data convergence is reduced. The text-based labeling method enriches descriptive labels of the image with labelers' sight, emotion, and writing habits. However, this method does not benefit data collection for the classification system, like the patterns of each part shape, and quantitatively measures the convergence of labeling data. To overcome this limitation, in the study we proposed the pre-identified graphical icon-based labeling method to ensure considerable agreement with data from a small number of experts. Also, we quantitatively measured data convergence of label data as kappa coefficients for further data processing and analysis. The graphical icon-based labeling method lowered the entry barrier for the novice labeler. The results showed almost perfect data agreement with experts' data and moderate agreement with other novices' data, which suggests that novice labelers do not experience difficulty in participating in labeling work, and the annotation data consistency is enough to further data processing or traditional design related data prevention. High data agreement means our new labeling method is robust and easy for novice labelers, so the novice labeler can effectively supplement the expert labeler in CH artifact design labeling tasks. The graphical icon-based labeling methods used in this study could improve image recognition and image processing algorithms to approximate human performance.

One of the key means for encouraging participant motivation is gamification. Game elements have been applied to diverse fields, from the Nike running app to healthcare education [27], [56], though gamification is more likely to be used as educational content and museum entertainment in the cultural heritage domain rather than in the critical-need tasks of design labeling or artistic data extraction [57], [58]. The goal-setting, levels, rewards, and customization elements used to develop our gamified system increased labelers' engagement and eased their workload. Labelers also said they might be willing to make more labels and spend more time with the gamified tasks. In this study, participants labeled only 100 images, which is insufficient for data collection and further processing. However, the labeling task completion time and time to label one image with gamification were shorter than those without gamification. Thus if the number of images increases significantly, labelers can save more time using a gamified system. The game affected labelers' internal and external enjoyment and motivation, they felt the gamified task was easier than the non-gamified task, and they reported reduced workload.

The gamification results showed higher intrinsic motivation, expected labeling time and work throughput, and lower workload than non-gamification. However, data convergence of novices with experts and other novices was lower than with non-gamified labeling. Similarly, gamification significantly reduced data agreement and consistency among expert labelers. Both were unexpected results. Compared to the data agreement in the non-gamified task, both data convergence between experts and novices and between each novice were lower in the gamified task. However, the specialized purpose of ceramic design element labeling is to make more consistent labels through data convergence, rather than simply generating proper labels [26]. As a result of labeling data agreement, data consistency among experts decreased with gamification more than data consistency among novices in the same condition. The finding in this study is that gamification interfered with the experts' labeling work and thus is not an effective method to increase data agreement among experts. To obtain consistent labels, gamification could be used for novice labelers. To increase data agreement among novices in a specific domain like CH design, we suggest using a weighted system that additionally rewards labels generated identical to an expert's pre-identified label, similar to the benchmark ESP [59]. To promote sustained enjoyment as well as the data convergence, we suggest another participant role in the gamified system: a label inspector whose goal is to judge the consistency of other participants' labels. The answer of the label inspector could be used to double-check the label data and ensure correctness.

Based on their open feedback, the 11 participants in our study wanted to make more accurate labels and recommended a feedback and editing function. These additional functions could create a more enjoyable and interactive task and thus achieve the same purpose as the system developer. With or without the gamification, the important function of a labeling system is to transfer the desired purpose of the labeling task to the labelers.

VI. CONCLUSION

To digitize the design assets of cultural heritage artifacts, significant expert and novice human labor is required. In particular, participation of novices should be increased to supplement more limited expert resources. However, this must be balanced with the need for highly consistent data in this specialized domain. We suggested a cultural heritage artifact design-related annotation system using a new robust labeling method to learn what factors of a labeling system might affect the novice labeler's engagement with the system and their data agreement with expert data. We added game elements to examine the game effect, and our results indicate that sufficient instruction and the simple design interface of our system positively impacted the novice labelers' levels of engagement as well as their perception of learnability. The suggested graphic icon-based labeling method as the robust labeling method showed significant data agreement between the participants. Our system also demonstrated the positive effects of gamification, including increased motivation to complete the task and decreased perception of workload. Participants' expected work throughput and duration of labeling work increased with gamification. Despite the improvements of gamification, the data agreement of the gamified task was lower than that of the non-gamified task. Gamification may encourage novice labelers to be engaged in labeling work and generate consistent labels. The game elements should have applied to more increase the data agreement of novice labelers, rather than the expert labelers of the specialized labeling work. We plan to explore gamification that better matches the needs of the data collectors.

REFERENCES

[1] A. Belhi, A. Bouras, and S. Foufou, "Leveraging known data for missing label prediction in cultural heritage context," *Appl. Sci.*, vol. 8, no. 10, pp. 1768–1786, Sep. 2018, doi: 10.3390/app8101768.

[2] C. Cintas, M. Lucena, J. M. Fuertes, C. Delrieux, P. Navarro, R. González-José, and M. Molinos, "Automatic feature extraction and classification of iberian ceramics based on deep convolutional networks," *J. Cultural Heritage*, vol. 41, pp. 106–112, Jan. 2020, doi: 10.1016/j.culher.2019.06.005.

[3] I. Saragusti, A. Karasik, I. Sharon, and U. Smilansky, "Quantitative analysis of shape attributes based on contours and section profiles in artifact analysis," *J. Archaeol. Sci.*, vol. 32, no. 6, pp. 841–853, Jun. 2005, doi: 10.1016/j.jas.2005.01.002.

[4] M. K. Buckland, "Cultural heritage (patrimony): An introduction," in *Records, Archives and Memory*, M. Willer and A. J. Gilliland, Eds. Zadar, Croatia: Univ. of Zadar, 2013, pp. 11–25.

[5] A. Belhi, A. Bouras, and S. Foufou, "Towards a hierarchical multitask classification framework for cultural heritage," in *Proc. IEEE/ACS 15th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Oct. 2018, pp. 1–7.

[6] A. Karasik and U. Smilansky, "Computerized morphological classification of ceramics," *J. Archaeol. Sci.*, vol. 38, no. 10, pp. 2644–2657, Oct. 2011.

[7] K. E. Lee, *Korean Traditional Fashion Inspires the Global Runway, in Ethnic Fashion*. Berlin, Germany: Springer, 2016, pp. 47–81.

[8] T. B. Petersen, M. Mackinney-Valentin, and M. R. Melchior, "Fashion thinking," *J. Des., Creative Process Fashion Ind.*, vol. 44, no. 1, pp. 179–195, 2016, doi: 10.1080/17569370.2016.1147699.

[9] X. Wu, C. Zhang, P. Goldberg, D. Cohen, Y. Pan, T. Arpin, and O. Bar-Yosef, "Early pottery at 20,000 years ago in Xianrendong Cave, China," *Science*, vol. 336, no. 6089, pp. 1696–1700, Jun. 2012.

[10] L. Lei and Z. Shouli, "Analysis of the relationship between modeling and decoration in ceramic design," in *Proc. Int. Conf. Econ. Manage. Cultural Ind. (ICEMCI)*, 2019, pp. 868–871.

[11] P. D. Lyons and J. J. Clark, "Interaction, enculturation, social distance, and ancient ethnic identities," in *Archaeology Without Borders: Contact, Commerce, and Change in the U.S. Southwest and Northwestern Mexico*, L. D. Webster, M. E. McBrinn, and E. G. Carrera, Eds. Boulder, CO, USA: Univ. Press of Colorado, 2008, pp. 185–207.

[12] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and Web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, May 2008, doi: 10.1007/s11263-007-0090-8.

[13] A. J. Hughes, J. D. Mornin, S. K. Biswas, L. E. Beck, D. P. Bauer, A. Raj, S. Bianco, and Z. J. Gartner, "Quanti. Us: A tool for rapid, flexible, crowd-based annotation of images," *Nature methods*, vol. 15, no. 8, pp. 587–590, 2018, doi: 10.1038/s41592-018-0069-0.

[14] A. Chortaras, A. Christaki, N. Drosopoulos, E. Kaldeli, M. Ralli, A. Sofou, A. Stabenau, G. Stamou, and V. Tzouvaras, "WITH: Human-computer collaboration for data annotation and enrichment," in *Proc. Companion Web Conf.*, Lyon, France, 2018, pp. 1117–1125.

[15] J. Foley, P. Kwan, and M. Welch, "A web-based infrastructure for the assisted annotation of heritage collections," *J. Comput. Cultural Heritage*, vol. 10, no. 3, p. 14, 2017, doi: 10.1145/3012287.

[16] A. Belhi, A. Bouras, and S. Foufou, "Digitization and preservation of cultural heritage: The CEPROQHA approach," in *Proc. 11th Int. Conf. Softw., Knowl., Inf. Manage. Appl. (SKIMA)*, Colombo, Sri Lanka, Dec. 2017, pp. 1–7.

[17] I. Pateraki and S. Scimeca, *Learning From the Past, Designing Our Future: Europe's Cultural Heritage Through eTwinning*. Brussels, Belgium: Central Support Service eTwinning, 2018. [Online]. Available: https://www.etwinning.net/eun-files/Online%20%E2%80%93%20eTwinning%20publication%2024.09.2018.pdf

[18] C. Harris, "ClueMeIn: Enhancing the ESP game to obtain more specific image labels," in *Proc. Annu. Symp. Comput.-Hum. Interact. Play Companion Extended Abstr.-CHI PLAY Extended Abstr.*, Melbourne, VIC, Australia, 2018, pp. 447–452.

[19] D. G. de Gómez Pérez, M. Suokas, and R. Bednarik, "Crowdsourcing pupil annotation datasets: Boundary vs. center, what performs better?" in *Proc. 7th Workshop Pervasive Eye Tracking Mobile Eye-Based Interact. (PETMEI)*, Warsaw, Poland, 2018, p. 3.

[20] P. Viana and J. P. Pinto, "A collaborative approach for semantic time-based video annotation using gamification," *Hum.-Centric Comput. Inf. Sci.*, vol. 7, no. 1, p. 13, Dec. 2017, doi: 10.1186/s13673-017-0094-5.

[21] A. Dumitrache, L. Aroyo, and C. Welty, "Achieving expert-level annotation quality with crowdtruth," in *Proc. BDM2I Workshop ISWC*, Bethlehem, PA, USA, 2015, pp. 1–13.

[22] G. Castellano, A. M. Fanelli, G. Sforza, and M. A. Torsello, "Shape annotation for intelligent image retrieval," *Appl. Intell.*, vol. 44, no. 1, pp. 179–195, 2016, doi: 10.1007/s10489-015-0693-7.

[23] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. Int. Conf. Multimedia (MM)*, Firenze, Italy, 2010, pp. 83–92.

[24] D. K. Iakovidis, T. Goudas, C. Smailis, and I. Maglogiannis, "Ratsnake: A versatile image annotation tool with application to computer-aided diagnosis," *Sci. World J.*, vol. 2014, pp. 1–12, Jan. 2014, doi: 10.1155/2014/286856.

[25] V. V. Vydiswaran, Q. Mei, D. A. Hanauer, and K. Zheng, "Mining consumer health vocabulary from community-generated text," in *Proc. AMIA Annu. Symp.*, Washington, DC, USA, vol. 2014, 2014, p. 1150.

[26] P. Lessel, M. Altmeyer, L. V. Schmeer, and A. Krüger, "'Enable or disable gamification?' Analyzing the impact of choice in a gamified image tagging task," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Glasgow, U.K., 2019, p. 150.

[27] M. Sailer, J. U. Hense, S. K. Mayr, and H. Mandl, "How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction," *Comput. Hum. Behav.*, vol. 69, pp. 371–380, Apr. 2017, doi: 10.1016/j.chb.2016.12.033.

[28] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. Conf. Hum. Factors Comput. Syst. (CHI)*, Vienna, Austria, 2004, pp. 319–326.

[29] M. Sanmugam, N. M. Zaid, Z. Abdullah, B. Aris, H. Mohamed, and H. van der Meijden, "The impacts of infusing game elements and gamification in learning," in *Proc. IEEE 8th Int. Conf. Eng. Edu. (ICEED)*, Kuala Lumpur, Malaysia, Dec. 2016, pp. 131–136.

[30] E. D. Mekler, F. Brühlmann, A. N. Tuch, and K. Opwis, "Towards understanding the effects of individual gamification elements on intrinsic motivation and performance," *Comput. Hum. Behav.*, vol. 71, pp. 525–534, Jun. 2017, doi: 10.1016/j.chb.2015.08.048.

[31] (2002). *Axure Software Solutions*. Accessed: Jul. 10, 2020. [Online]. Available: http://www.axure.com

[32] M. Resnick, J. Maloney, A. Monroy-Hernández, N. Rusk, E. Eastmond, K. Brennan, A. Millner, E. Rosenbaum, J. Silver, B. Silverman, and Y. Kafai, "Scratch: Programming for all," *Commun. ACM*, vol. 52, no. 11, pp. 60–67, Nov. 2009, doi: 10.1145/1592761.1592779.

[33] R. S. Alsawaier, "The effect of gamification on motivation and engagement," *Int. J. Inf. Learn. Technol.*, vol. 35, no. 1, pp. 56–79, Jan. 2018, doi: 10.1108/IJILT-02-2017-0009.

[34] Y. Jia, B. Xu, Y. Karanam, and S. Voida, "Personality-targeted gamification: A survey study on personality traits and motivational affordances," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2016, pp. 2001–2013.

[35] M. Altmeyer, P. Lessel, K. Dernbecher, V. Hnatovskiy, M. Schubhan, and A. Krüger, "Eating ads with a monster: Introducing a gamified ad blocker," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–6.

[36] J. Brooke, "SUS-A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, no. 194, pp. 4–7, 1996.

[37] L. Aroyo and C. Welty, "Measuring crowd truth for medical relation extraction," in *Proc. AAAI Fall Symp. Ser.*, Arlington, VA, USA, 2013, pp. 1–8.

[38] R. M. Ryan, "Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory," *J. Personality Social Psychol.*, vol. 43, no. 3, pp. 450–461, 1982.

[39] E. McAuley, T. Duncan, and V. V. Tammen, "Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: A confirmatory factor analysis," *Res. Quart. for Exercise Sport*, vol. 60, no. 1, pp. 48–58, Mar. 1989, doi: 10.1080/02701367.1989.10607413.

[40] M. A. Robinson, "Using multi-item psychometric scales for research and practice in human resource management," *Hum. Resource Manage.*, vol. 57, no. 3, pp. 739–750, May 2018, doi: 10.1002/hrm.21852.

[41] S. G. Hart. (1986). *NASA Task load Index (TLX). Volume 1.0; Paper and Pencil Package*. [Online]. Available: http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20000021488_200%200015069.pdf

[42] E. A. Bustamante and R. D. Spain, "Measurement invariance of the NASA TLX," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2008, vol. 52, no. 19, pp. 1522–1526.

[43] F. K. Khattak and A. Salleb-Aouissi, "Improving crowd labeling through expert evaluation," in *Proc. AAAI Spring Symp. Ser.*, Palo Alto, CA, USA, 2012, pp. 1–5.

[44] A. Kulkarni, N. R. Uppalapati, P. Singh, and G. Ramakrishnan, "An interactive multi-label consensus labeling model for multiple labeler judgments," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, pp. 1–8.

[45] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychol. Bull.*, vol. 70, no. 4, p. 213, 1968, doi: 10.1037/h0026256.

[46] B. Dates and J. King, "SPSS algorithms for bootstrapping and jackknifing generalized measures of agreement," in *Proc. Annu. Meeting Southwest Educ. Res. Assoc.*, New Orleans, LA, USA, 2008, pp. 1–11.

[47] A. Bangor, P. Kortum, and J. Miller, "Determining what individual SUS scores mean: Adding an adjective rating scale," *J. Usability Stud.*, vol. 4, no. 3, pp. 114–123, 2009.

[48] L. J. Richard and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[49] A. Chistyakov, M. T. Soto-Sanfiel, E. Martí, T. Igarashi, and J. Carrabina, "Objective learnability estimation of software systems," in *Proc. Int. Conf. Ubiquitous Comput. Ambient Intell.*, Las Palmas, Spain, 2016, pp. 503–513.

[50] X. Ferre, N. Juristo, H. Windl, and L. Constantine, "Usability basics for software developers," *IEEE Softw.*, vol. 18, no. 1, pp. 22–29, Jan. 2001, doi: 10.1109/52.903160.

[51] N. Harrati, I. Bouchrika, A. Tari, and A. Ladjailia, "Exploring user satisfaction for e-learning systems via usage-based metrics and system usability scale analysis," *Comput. Hum. Behav.*, vol. 61, pp. 463–471, Aug. 2016, doi: 10.1016/j.chb.2016.03.051.

[52] C. Ahlstrom, T. Victor, C. Wege, and E. Steinmetz, "Processing of eye/head-tracking data in large-scale naturalistic driving data sets," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 553–564, Jun. 2012.

[53] L. Derczynski, D. Maynard, N. Aswani, and K. Bontcheva, "Microblog-genre noise and impact on semantic annotation accuracy," in *Proc. 24th ACM Conf. Hypertext Social Media (HT)*, Paris, France, 2013, pp. 21–30.

[54] A. Tanyavutti, P. Anuntavoranich, and K. Nuttavuthisit, "An idea generation tool harnessing cultural heritage for design-driven entrepreneurs," *Acad. Entrepreneurship J.*, vol. 24, no. 4, pp. 1–15, 2018.

[55] J. Zhou, Y. Lu, K. Zheng, K. Smith, C. Wilder, and S. Wang, "Design identification of curve patterns on cultural heritage objects: Combining template matching and CNN-based re-ranking," 2018, *arXiv:1805.06862*. [Online]. Available: http://arxiv.org/abs/1805.06862

[56] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: Defining 'gamification,'" in *Proc. 15th Int. Academic MindTrek Conf. Envisioning Future Media Environments (MindTrek)*, Tampere, Finland, 2011, pp. 9–15.

[57] A. Bujari, M. Ciman, O. Gaggi, and C. E. Palazzi, "Using gamification to discover cultural heritage locations from geo-tagged photos," *Pers. Ubiquitous Comput.*, vol. 21, no. 2, pp. 235–252, Apr. 2017, doi: 10.1007/s00779-016-0989-6.

[58] M. A. Ramly and B. B. Neupane, "ExplorAR: A collaborative artifact-based mixed reality game," in *Proc. Asian HCI Symp. Emerg. Res. Collection*, Montreal, QC, Canada, 2018, pp. 1–4.

[59] L. von Ahn and L. Dabbish, "Designing games with a purpose," *Commun. ACM*, vol. 51, no. 8, pp. 58–67, Aug. 2008, doi: 10.1145/1378704.1378719.

JIEUN LEE received the B.S. degree in biomedical engineering from Eulji University, South Korea, in 2013, and the M.S. degree in robotics engineering from the Daegu Gyeongbuk Institute of Science and Technology, South Korea, in 2015. She is currently pursuing the Ph.D. degree with the School of Integrated Technology, Gwangju Institute of Science and Technology (GIST). She is developing a human–computer interaction system for increasing engagement and motivation of people to participate the social community. Her research interests include engagement, gamification, and HCI.

JI HYUN YI received the master's degrees in industrial design and fine arts from the Graduate Program of Industrial Design, University of the Arts, and the Graduate School of Fine Arts, University of Pennsylvania, respectively, and the Ph.D. degree from the Graduate School of Culture and Technology, Korea Advanced Institute of Science and Technology. She is currently an Assistant Professor with the School of Integrated Technology, Gwangju Institute of Science and Technology (GIST). Her research interests include interactive content design, UI/UX design, AR/VR interaction, and creative image creation method.

SEUNGJUN KIM (Member, IEEE) received the B.S. degree in electrical and electronics engineering from the Korea Advanced Institute of Science and Technology, and the M.S. and Ph.D. degrees in mechatronics from the Gwangju Institute of Science and Technology (GIST), South Korea, in 2006. He is currently an Assistant Professor with the Institute of Integrated Technology, GIST, and an Adjunct Faculty Member with the Human-Computer Interaction Institute, Carnegie Mellon University. He currently leads research and development projects concerning human–vehicle interaction, wearable UI/UX technologies, human–robot interaction, sensory augmentation with haptics and augmented reality, and cyber-learning with a sensor support. His research interests include the intersection of human–computer interaction (HCI) and sensor data mining to create intelligent systems that improve the quality of HCI experience based on human attention and cognition.