Fight or Unite: Investigating Game Genres for Image Tagging

Dion Hoe-Lian Goh*
Nanyang Technological University
Wee Kim Wee School of Communication & Information
31 Nanyang Link, SCI Building,
Singapore 637718
Phone: (65) 67906290
Fax: (65) 67915214
Email: ashlgoh@ntu.edu.sg

Rebecca P. Ang
Nanyang Technological University
School of Humanities and Social Sciences
14 Nanyang Drive
Singapore 637332
Phone: (65) 6316-8733
Fax: (65) 6794-6303
Email: rpang@ntu.edu.sg

Chei Sian Lee
Nanyang Technological University
Wee Kim Wee School of Communication & Information
31 Nanyang Link, SCI Building,
Singapore 637718
Phone: (65) 67906636
Fax: (65) 67915214
Email: leecs@ntu.edu.sg

Alton Y. K. Chua
Nanyang Technological University
Wee Kim Wee School of Communication & Information
31 Nanyang Link, SCI Building,
Singapore 637718
Phone: (65) 67905810
Fax: (65) 67915214
Email: altonchua@ntu.edu.sg

**Correspondence to: Dion Goh, Division of Information Studies, Wee Kim Wee School of Communication and Information, Nanyang Technological University, 31 Nanyang Link, Singapore 637718, Telephone: 65-67906290. Email: ashlgoh@ntu.edu.sg.*

Fight or Unite: Investigating Game Genres for Image Tagging

Abstract

Applications that use games to harness human intelligence to perform various computational tasks are increasing in popularity and may be termed as human computation games (HCGs). Most HCGs are collaborative in nature, requiring players to cooperate within a game to score points. Competitive versions, where players work against each other, are a more recent entrant, and have been argued to address shortcomings of collaborative HCGs such as quality of computations. To date however, there is little work conducted in understanding how different HCG genres influence computational performance and players' perceptions of them. In this paper, we study these issues using image tagging HCGs in which users play games to generate keywords for images. Three versions were created: collaborative HCG, competitive HCG and a control application for manual tagging. The applications were evaluated to uncover the quality of the image tags generated as well as users' perceptions. Results suggest that there is a tension between entertainment and tag quality. While participants reported liking the collaborative and competitive image tagging HCGs over the control application, those using the latter seemed to generate better quality tags. Implications of the work are discussed.

Keywords

Image Tagging, Human Computation Games, Collaborative Game, Competitive Game, User Study, Computainment

Introduction

The popularity of social computing coupled with the widespread availability of affordable digital cameras and mobile phone cameras has made the online sharing of digital media, such as photos, considerably easier. Popular sites such as Flickr and Facebook attest to this phenomenon, allowing photos to be shared to one's social network or the public with little difficulty. Parallel with these developments however, is that the resulting proliferation of such media online has made it necessary for techniques to manage them to facilitate their effective and timely retrieval.

In the case of images, techniques are needed to analyze and understand their content so that relevant files are returned in response to a query. For example, the Content-Based Image Retrieval (CBIR) technique supports the process of retrieving desired images from a large collection based on features such as colors, textures and shapes automatically extracted from the images in the collection (Eakins & Graham, 1999). However, the main drawback of CBIR is its inability to satisfy queries represented at higher levels of abstraction, including the identities, meanings and purposes of the objects and scenes depicted in an image. Another technique is known as ALIPR which relies on categorized images to train a dictionary of hundreds of statistical models each representing a concept (Li & Wang, 2003). The process is then reversed so that a user can use freely-assigned keywords to search for a relevant image. Nonetheless, imaging retrieval is not without challenges. For one, pixel-based automatic tagging techniques which work well within a controlled set of images are unable to scale to real world data (Pavlidis, 2009). Another challenge is the semantic gap between human perception and image content. Two images which may have identical meaning to a human could have entirely different pixel values. The problem is further

compounded by the fact that images have multi-faceted attributes including syntactic information such as colors and shape, non-visual information such as metadata, and semantic information such as place and time (Westman, 2009).

In the absence of techniques which could rival humans in terms of understanding image content (Datta, Joshi, Li & Wang, 2008), one approach is to manually annotate images in the form of tags or keywords (Li & Wang, 2008). For example, an image with different types of fruit could be distinguished by human-generated keywords describing their constituent components such as "apple", "grape", and even abstract concepts such as "cornucopia", and so on. Such a task is typically difficult to achieve for an automated algorithm. Image retrieval algorithms can then harness these generated keywords to make sense of the media to meet users' needs. For example, a query for images with the terms "grapes" and "cornucopia" would return the aforementioned image, whereas automatically indexed images may not.

Here however, a conundrum exists. One the one hand, humans can help improve image understanding and retrieval through manual tagging, but on the other, such a process is tedious, labor intensive and potentially costly, involving people sitting in front of computers sifting through huge collections of images, and generating descriptive tags for each image. Stated differently, as long as automated algorithms cannot perform to the level of humans in image understanding, and that human intervention is desired to boost this performance, research is needed to investigate how incentives can incorporated into manual image tagging tasks.

One promising development that could harness human intelligence to perform tasks such as image tagging is the use of computer games. In recent years, gaming has seen an increase in popularity, in part due to better machine performance, high quality graphics, intuitive user interfaces, widespread availability of broadband networking, and engaging gameplay. The popularity of games has spawned new genres that extend beyond pure entertainment. These games offer entertainment, but are meant to accomplish tasks or solve problems, and have been used in a number of areas such as education, advertising, defense, emergency management, and training (e.g. Kankaanranta & Neittaanmäki, 2009; Zyda 2005). Such games could serve as motivators for users to contribute their intellect or creativity to a given endeavor. That is, while users are entertained by playing a game, they are also performing computations as a byproduct of gameplay. In the case of image tagging, the resulting byproducts will be the keywords that describe images.

Human computation games (HCGs) may be termed as "computainment", a portmanteau of the words "computation" and "entertainment". They have been employed relatively successfully in a number of areas including image tagging (von Ahn & Dabbish, 2004), ontology creation (Siorpaes & Hepp, 2008), and location-based annotation authoring (Lee, Goh, Chua, & Ang, 2010). Recently, these games have also been given the name Games With A Purpose (von Ahn & Dabbish, 2008). The ESP Game (von Ahn & Dabbish, 2004) is one of the earlier examples of image tagging HCGs in which two unrelated players are tasked to create matching keywords to randomly presented images within a given time limit. Players not only derive entertainment from the game, but the resulting keywords can be used as tags for the images, and therefore harnessed by image retrieval algorithms.

To date, most HCGs are collaborative in nature, requiring users to work together to fulfill the games' objectives. However, critics of collaborative HCGs have identified a number of problems. In image tagging games, these include the formation of cheating coalitions in which groups of users agree upon a predetermined set of tags to use whatever the image presented, and the tendency for players to generate tags with more generic descriptions, among others (Robertson, Vojnovic, & Weber, 2009). For these reasons, some researchers have proposed competitive variants. In competitive HCGs, players have to work against each other in order to fulfill the games' objectives. Importantly, competition in image tagging games has been argued to address the quality issue as players have to outdo each other to generate tags for images to score points. There is thus no opportunity for collusion and the resulting diversity of tags should help in better descriptions of images to facilitate their future retrieval (Ho, Chang, Lee, Hsu, & Chen, 2009).

Despite increasing interest in the use of HCGs, to the best of our knowledge, there is still a lack of understanding of how different game genres affect players' perceptions and performance. Specifically, in the case of image tagging, there are as yet no studies that compare the collaborative and competitive genres in terms of the quality of tags generated. Instead, work on tag quality typically focuses on the performance of individual applications (e.g. Ho et al., 2010), and the same is typically true for studies of user perceptions (e.g. Lee et al., 2010). Our present work is therefore timely as we seek to investigate the effectiveness of collaborative and competitive HCGs for generating better quality tags, as well as users' perceptions of these genres in terms of their playability.

The remainder of this paper is organized as follows. In the next section, we provide an overview of the work related to HCGs, focusing especially on image tagging games and their associated issues that warrant the present research. Given this background, we introduce our research questions, describe the applications implemented and the methodology of our study to address the research questions. A description of our results follows, covering both quality of tags generated as well as users' perceptions of the applications. We then discuss our findings and implications for image tagging HCGs, as well as identify opportunities for future work.

Literature Review

Human Computation Games

Human computation games blend human problem solving and gaming. They capitalize on people's desire to be entertained and the fact that humans are better at solving certain problems than computers. These games are essentially a class of social computing application as they rely on the participation of users. However, because they blend both gaming and human computation, these objectives could be in competition, hence possibly influencing quality of computations.

Perhaps one of the earlier and more successful examples of HCGs is the ESP Game (von Ahn and Dabbish, 2004). Two unrelated players are tasked to create matching keywords to randomly presented images within a given time limit. Points are earned based on specificity of the keywords, and coupled with a countdown timer, these elements add excitement and hence motivation for players. As described earlier, the objective of the game is for randomly paired players to work together to create keywords to images. These keywords can then be

harvested as metadata for their respective images, facilitating image retrieval. The success of the ESP Game paved the way for subsequent tagging-oriented games. For example, Peekaboom (von Ahn, Liu & Blum, 2006) is a two-player collaborative game for identifying objects in images. One player is presented with the entire image and a word associated with it. This player has to progressively reveal portions of the image to his/her partner in such a way that the partner guesses the word in the least amount of time. In doing so, the revealed portion of the image and the associated keyword may be used by machine vision algorithms for object identification. Another game is Herd It, a Facebook game for tagging music (Barrington, Turnbull, O'Malley, & Lanckriet, 2009). Like the ESP Game, it is collaborative in genre but is played by larger groups of at least 10 simultaneously. Players listen to a music clip and are quizzed on their opinions of the piece (e.g. music sub-genre, prominent musical instrument) through multiple-choice questions. Points are awarded based on the percentage of the other players that agree with a player's choice. As in the other HCGs reviewed above, the selected answers to the quizzes may be used as tags to their respective songs.

HCGs have also extended beyond media tagging and the confines of the desktop computer to other types of computational tasks. OntoGame is a platform that offering games for creating knowledge structures associated with the Semantic Web (Siorpaes & Hepp 2008). Such games range from OntoPronto for creating an ontology from Wikipedia entries, to OntoTube for annotating YouTube videos with ontological elements. Next, Curator (Walsh & Golbeck, 2010) is a collection matching game in which pairs of players are asked to group items into collections. Points are awarded for matches between players, and matches may then be used for recommender systems. Similar ideas that blend human computation and gaming can also be found in mobile applications. One such example is the Gopher Game (Casey et al. 2007).

Gophers are agents that represent missions to be completed, and are carriers of information between players. As players move about their physical surroundings, they pick up gophers and help them complete their missions by supplying them with camera phone images and textual content. By helping gophers complete their missions, content describing specific locations are created, and this can be shared since other users may pick up these gophers and view the images and text associated with them. Like the Gopher Game, Eyespy (Bell et al., 2009) is a mobile HCG that generates photographic and textual descriptions of locations. Players take photos of places and share them with others who have to find out where the photos were taken. This latter process of confirmation validates the photos, resulting in increased confidence that the images are good representations of their associated locations.

The HCGs reviewed above are primarily collaborative in nature. In contrast, competitive HCGs are rarer although they are beginning to emerge. Proponents of such games contend that competition heightens the emotional impact of players who have to respond and react to opponents' abilities. This adds an additional dimension of challenge essential in sustaining interest, and the ability to work against others enhances and amplifies gameplay by increasing player satisfaction (Vorderer, Hartmann, & Klimmt, 2003). More importantly, competition has been argued to address the problem of cheating as there is no opportunity for collusion when winning conditions dictate that players must outdo each other. KissKissBan (Ho et al., 2009) is one example that supports image tagging. This is a three-player game with a "blocker" and a "couple". In each round, a blocker has seven seconds to create a list of blocked words to an image. Thereafter, the couple is presented with the same image and tasked to create tags for it in 30 seconds. If the couple-generated tags are found in the unseen blocked list, points and time are deducted, while the blocker scores points. Otherwise, if

matching tags are made, the couple scores points. An evaluation of the game suggests a more varied set of generated keywords when compared to data extracted from the ESP Game, with good precision and recall. This seems to suggest that competitive image tagging HCGs may also result in tags that are of better quality than collaborative variants, but no follow-up work has yet to be conducted to confirm this.

On the mobile platform, Indagator (Lee et al., 2010) is a competitive variation of a location-based content sharing HCG. While its goals are similar to the Gopher Game and Eyespy, part of the game mechanics require players to plant damage-inflicting traps and other obstacles at various locations. Players who stumble upon these objects lose points while the obstacle-setter gains points. Finally, PhotoCity (Tuite et al., 2010) invites players to contribute photos from different perspectives to places of interest, with the goal of facilitating 3D model construction. In the game, a flag represents a place of interest, and to capture it, players have to earn to points by taking as many as high-quality photos as possible. The player that scores the highest number of points captures the flag. Table 1 summarizes the systems reviewed in this section, showing their genre and the types of computation supported.

[Insert Table 1 here]

Evaluation

In terms of evaluating users' perceptions of games, there are two important research streams requiring further elaboration. The first focuses on identifying frameworks that explain users' reactions and motivations for playing games. Flow theory has often been applied to understand the environmental and individual variables that influence users' intrinsic

motivations for playing games (Csikszentmihalyi, 1990). The theory describes users' experiences as those that are able to sustain long-term focus (O'Brien & Toms, 2008). Another important theoretical framework that is often used is play theory (Stephenson, 1967) which examines the activities that encourage learning and creativity, and the development and satisfaction of psychological and social needs (Rieber, 1996). Studies have found that software interactions that were designed with a "play" focus were often associated with higher user satisfaction (Woszczynski et al., 2002). The second research stream centers on identifying attributes that influence users' experiences with the games they play. Some of these attributes mentioned in the literature include players' engagement (O'Brien & Toms, 2008), sense of social presence (Champion, 2003), and physiological arousal (Rayaia et al., 2004). In the context of HCGs, researchers have suggested that incorporating elements of fun into user interfaces could create positive experience for the users (Sneiderman, 2004; von ahn & Dabbish, 2008). Suggestions from the literature to characterize fun include the ability to create a sense of competence for each player, creating a pleasant and interesting sensory experience and allowing the development of social connections with their partners (Law et al., 2007).

In terms of quality of computations, evaluation results are mixed. In the case of image tagging games, this would refer to how descriptive and useful the tags are for facilitating future retrieval. For example, image tags generated from the ESP Game were found to yield high search precision values (von Ahn & Dabbish, 2004), while those from KissKissBan seemed to perform even better when compared to the former (Ho et al., 2009). In both cases, these values were computed by manually examining a set of images and their assigned tags, and for each, calculating the percentage of tags that were descriptive of their respective

images. Likewise, data from Peekaboom gameplay (von Ahn et al., 2006) showed high accuracy between the user-revealed portions of images and their associated keywords. Put differently, this meant that the bounding box drawn by participants around an image actually revealed an object that was associated with the given keyword in the game.

However, the very same studies plus others also suggest that quality concerns may be valid. Several reasons have been identified. In a collaborative HCG, players could conspire to cheat in order to score points (von Ahn & Dabbish, 2004). For example, entering meaningless or irrelevant terms such as "a" to any image presented. Further, in an analysis of the ESP Game, problems such as tag redundancy/synonymy (e.g. "guy" and "man"), excessively co-occurring tags (e.g. "water" and "blue"), tendency to match on colors, and the inclination to use generic rather than specific tags were identified (Robertson et al., 2009). Similarly, an examination of the music tags generated in Tagatune (Law & von Ahn, 2009) showed evidence of gaming the system in which the tags were used to communicate between players (e.g. "same", "diff", "yes") rather than for the purposes of description. Evaluations of HCGs in domains other than media tagging yield similar patterns of findings (e.g. Bell et al. 2009; Casey et al., 2007).

Despite concerns over the quality of the output of HCGs, they seem to be growing in number, and there are indications that users enjoy using them. For example, evaluations of various games on GWAP.com (e.g. ESP Game, Tagatune, etc.) demonstrate repeated play over time (Law & von Ahn, 2009). Similarly, a survey of Herd It players showed that the game was received positively and that player enjoyment showed an increasing trend during a one-year development and evaluation period (Barrington et al., 2009). Qualitative feedback obtained in

many of these evaluations also attests to user satisfaction. Players of Eyespy (Bell et al., 2009) reported enjoying the game, finding it fun, and that competition among players was a major motivation to continue playing it. In Peekaboom, comments from players suggested it was addictive and that the scoring of points and achieving high scores fostered continued use (von Ahn et al., 2006).

In summary, a common theme among research on HCGs to date is that they focus primarily on the design, implementation and evaluation of a single application. In contrast, comparative studies across different applications in terms of quality of computations and user perceptions are lacking, making it difficult to ascertain whether HCGs fulfill their computainment role effectively. Thus, a key contribution of our present work is that we investigate quality and user perceptions across three different representative applications: two main genres of HCGs, the collaborative and competitive games, and a non-game, manual tagging application. Specifically focusing on image tagging, our study is guided by two research questions: (1) Which image tagging application (collaborative HCG, competitive HCG, manual tagging) generates more quality tags; and (2) Which image tagging application is perceived to be more appealing by users?

Evaluating Game Genres

Experimental Systems Developed

To investigate the research questions, we first developed a set of image tagging applications: a control manual tagging application (serving as a baseline for comparing our results), collaborative HCG and competitive HCG. There were a number of reasons for developing our own applications, as opposed to using existing ones. First, we wanted have better control

over the look-and-feel of the applications to ensure a more consistent user experience during the user study. This was not possible with existing applications such as the ESP Game as customizations could not be performed. Further, there was a scarcity of competitive games to choose from, making it necessary to develop our own version. Finally, using our versions of HCGs made it easier to access the data generated (i.e. tags) for our analyses. This would not have been possible for existing games.

In our study, we constructed a database of 300 images and manually assigned at least 30 tags per image on average. Tags for each image were created by three researchers who held degrees in the field of library and information science, and therefore familiar with indexing concepts. Each researcher first independently assigned tags to a given image for the expressed purpose of facilitating future retrieval. The researchers then gathered to discuss their assignments and collectively decided on the final set of tags for each image. Further, each image was also associated with one or more tags that were designated as "off-limits", that is, keywords, that players could not enter and which no points will be received. These were also determined by the researchers. The purpose of off-limit terms is to prevent players from entering common or obvious keywords, thereby encouraging diversity of tags (von Ahn & Dabbish, 2008). This formed our ground truth data for which to compare the results of the user study. Figure 1 shows an example of an image. Tags that were assigned to it included "grape", "shop", and "tomato". Off-limit terms included "fruit", "red" and "water" as these were deemed either too common or generic.

[Insert Figure 1 here]

Three applications were then developed over this database. First, a control version essentially presented a series of images to a user. The user had to enter between one to five tags before the next image was shown. The user had five minutes to tag as many images as possible. Put differently, the control served as a representative manual image tagging application for which to compare the performance of the HCGs. Figure 2 shows the application. The time left to complete the tagging task is displayed on the top of the user interface. The image to be tagged is displayed prominently in the middle, while players enter up to five tags in the text fields above the image. Off-limit tags are shown next to the image. As mentioned previously, these are tags that cannot be entered for the current image. Points were not awarded for entering tags that matched our database as we wanted to make the control application as close to a standard manual tagging application as possible. Note that while most manual image tagging applications do not have count-down timers, we decided to include it as a form of feedback to users so that they knew when the application would stop.

[Insert Figure 2 here]

Next, the collaborative HCG had a similar design to the ESP Game, in that pairs of players had to enter the same tag, but not necessarily in the same order, to score points. For example, referring to the image in Figure 3, suppose Player A entered the tags “clinton” and “person” in that order, while the partner, Player B, entered the tags “hillary” and “clinton”. Here, “clinton" was entered by both players and was therefore recorded as a matched tag and eligible to receive points. In our version of the game, more points were awarded based on whether the tags were found in the database. If a tag was not in the database, a much smaller number of points was awarded instead. Note that the number of points scored was not used in

our study (to be explained subsequently). Rather, scores were used primarily to motivate players and provide feedback on their performance. Figure 3 shows the collaborative game, which is similar in appearance to the control application. The time left in the game and current score is displayed on the top of the user interface. Again, the image to be tagged is displayed in the middle while players enter tags in the text field above the image. To add an element of urgency, the number of tags entered by the partner is also shown. Beside the image, off-limit tags ("man", "flag", "black") are displayed. The list of tags generated by the user for the current image is displayed below it. The player and his/her partner will see the same interface.

[Insert Figure 3 here]

The competitive HCG also required paired players to enter tags to a presented image. However, the first player in the pair to enter a tag that described the image was awarded points while the other player received none. Points were awarded based on matches against our ground truth data. Players were then presented with the next image, and so on until the time limit of five minutes was reached. This initial design was modified as pilot testing suggested that users perceived it was easier to play than the collaborative game since the game was essentially a guessing game between two independent competing players against the database of tags. To raise the level of challenge so that it was perceived to be comparable with the collaborative version, off-limit tags were not displayed unlike the collaborative version. Further, players who entered any of these tags had points deducted, similar to Ho et al. (2009). The interface for the competitive HCG is similar to the collaborative version in Figure 3 except that the off-limit tags are not shown. As an example of gameplay, suppose in

response to the image in Figure 3, Player A entered the tags "hillary", and "clinton" in that order, while the partner, Player B, entered the tags "man" and "person". Assuming that among all these submitted tags, "clinton" was in the database, Player A would receive points while Player B would have points deducted as "man" was an off-limit tag (which was not shown to the players).

Participants

A total of 103 participants were recruited for the study. They were undergraduate and graduate students from a local university. Of the 103 participants, 58 were males and 45 were females, with ages ranging from 19 to 37, and an average age of 26.8 years. The majority of the participants (85) had a background in computer science, information technology, engineering or related disciplines, while the other 18 came from disciplines such as arts, social sciences and business.

All the participants reported understanding the concept of tagging, with about 82 participants (80%) using tags to access images on a regular basis. In addition, 68 participants (66%) had experience in image tagging, contributing tags to images in various Web sites (e.g. Flickr) on a regular basis. However, most participants (84%) were not familiar with the concept of HCGs, and had not played such games prior to the study. Nevertheless, the majority of the participants (about 64%) were frequent players of online games.

Measures

As discussed in the review of related literature, tag quality is an issue of concern in HCGs as players have been demonstrated to enter generic tags, synonyms and colors, among other

problematic characteristics (Robertson et al., 2009). To overcome this, Google Image Labeler (http://images.google.com/imagelabeler/), a close relative of the ESP Game, introduced a tiered scoring system of between 50 to 150 points depending on how specific the tags being entered were. Previously, all matched tags were given the same scores. Thus, a generic tag such as "man" might earn 50 points, while a more specific "clinton" would earn much more. Note however that the precise scoring system employed by Google Image Labeler is not known.

In our study, our analysis was based on the mean number of generated tags matching our ground truth data, which is a measure of accuracy. Spelling variations and different word forms were allowed, for example, "color" and "colour", or "car" and "cars". In addition, we also classified matched tags into three categories based on the discussion of quality tags in the literature (e.g. Ho et al., 2009; Robertson et al., 2009). This classification was performed by the three researchers described previously. Specifically, we defined Level 1 tags as those that were generic in nature, referred to minor components of the image, or colors. Level 2 tags were more specific, describing the major components of the image, while Level 3 tags were the most specific, describing detailed attributes or characteristics of the image. For example, in Figure 1, the tag "green" to describe the grapes would be classified as Level 1 (color), while "fruit" would be Level 2 (major component of image), and "peach" would be classified as Level 3 (specific component of image). For each image, the number of tags assigned to each level depended on the image content itself, but in general, the distribution of tags in the various levels across images was approximately similar.

Next, a questionnaire (see Appendix for details) elicited participants' perceptions of the applications they used, and questions were rated on a scale of 1 (strongly disagree) to 5 (strongly agree). In particular, perceptions were measured along the following aspects:

- Appeal. Participants were asked to rate the appeal of the application (control, collaborative game or competitive game) they used in the study. In essence, this measured the degree to which they liked the application, serving as an indication of sustained use.
- Challenge. This has been argued by many researchers to be the most important aspect of game design (e.g. Lazzaro & Keeker, 2004; Qin, Rau, & Salvendy, 2010). A game should be challenging enough to capture the interest of the player, providing an enjoyable, interactive experience that engages him/her to continue playing till the final goal has been reached. Yet at the same time, it should not be too challenging such that the player feels frustrated and eventually gives up playing (Goh, Ang & Tan, 2008), nor should it be too easy such the player knows that success is inevitable and it becomes boring (Malone, 1981).
- Usefulness. This referred to how useful the application was for image tagging. Past research has demonstrated that perceived usefulness of a technology has a strong influence on its subsequent adoption (Saeed & Abdinnour-Helm, 2008; Sun & Zhang, 2008). The argument is that if people expect a technology to increase task performance or accomplishment, then it would be natural that their intentions to use it would be greater (Davis, 1989).
- Absorption. To maintain interest, a game must be able to capture the player's attention so that he/she is totally focused on the game (Brockmyer et al., 2009). Put differently, all of the player's skills are needed to deal with the challenges associated with the game, and consequently, there is little time or energy left to handle anything other than the game

itself (Csikszentmihalyi 1990). Research has shown that because of the deep engagement needed, players have a high emotional investment in the game, and this encourages sustained usage (e.g. Jennett et al., 2008; Johnson & Wiles, 2003).

- Control. The ability of a player to feel that he/she has control over his/her actions has been demonstrated to be important for sustained usage of a game (e.g. Brockmyer et al., 2009; Pagulayan et al., 2002). In other words, players should perceive that they are able to translate their intentions into in-game behavior, and that their actions and strategies undertaken will impact the outcome of the game. In doing so, the game becomes highly replayable as players are motivated to improve their skills and explore different strategies to influence the gaming environment (Goh et al., 2008; Sweetser & Wyeth, 2005)
- Learnability. The ability to learn how to play a game quickly contributes to its enjoyment and consequently, continued use (Pagulayan et al., 2002, Pinelle, Wong, & Stach, 2008). This is especially important for casual games such as our collaborative and competitive HCGs which are targeted for a mass audience with varying skill levels, and who may not play games on a regular basis as hardcore gamers (Desurvire & Wiberg, 2008). Consequently, games should provide enough information for players to start playing quickly, and their interfaces should be consistent, intuitive and easy to learn.
- Social interaction. Broadly, social interaction refers to the support for competition, cooperation and communication (Sweetser & Wyeth, 2005). It has been found to be a feature requested by players and also an important factor for a game's success (e.g. Ducheneaut, Yee, Nickell & Moore, 2006; Lee et al., 2010). The premise is that people enjoy interacting with others, whether it is through cooperating with one another to accomplish goals, creating communities, or to compete and be better ranked than other players.

Procedure

Participants were randomly divided into three groups with each playing a specific game. Here, 33 participants were assigned to the control application, 34 to the collaborative HCG and 36 to the competitive HCG. The study was conducted in a lab across separate sessions, with each group (control application, collaborative HCG and competitive HCG) segregated from the others. For the latter two groups, participants were further randomly paired to play their assigned games. Although they were co-located in the same lab, participants did not know who their partners were, and communication among participants was prohibited.

Within each session, the study began with a researcher briefing the participants on image tagging for the purposes of facilitating future retrieval by others. In addition, for the collaborative and competitive game groups, the purpose of HCGs and their potentially useful role in supporting image tagging was explained. Next, participants were also briefed on the usage and gameplay rules of their respective application. They were also instructed to generate tags that were descriptive of the images, being as specific as possible, and anticipate the types of keywords that others may employ to retrieve them. This was followed by a demonstration and participants were then asked to try their assigned application, playing between one to two rounds, as a means of familiarization. Following this, the study commenced and the participants played their assigned applications for one round, after which they completed a questionnaire that captured demographic data (reported earlier) and perceptions of the application they played. Participants were also requested to provide qualitative comments about what they liked and disliked about their assigned application. In addition, the tags generated by the participants were captured for further analysis.

Results

Tag Quality

Table 2 shows the means and standard deviations of the number of matched tags and the breakdown by the three levels of quality. These formed the dependent variables of this portion of our study. The pattern of results suggests that the number of matched tags appears to be influenced by type of application.

[Insert Table 2 here]

To verify this, four one-way ANOVAs were conducted on the dependent variables. Results indicate that there were significant differences with respect to overall matched tags, $F(2, 100) = 14.00$, $p < .001$; Level 1 tags, $F(2, 100) = 26.00$, $p < 0.001$ and Level 2 tags, $F(2, 100) = 9.21$, $p < 0.001$. However, there was no significant difference with respect to the number of Level 3 tags generated across the three applications, $F(2, 100) = 1.09$, $p = .34$. Next, post-hoc comparisons using Tukey's test was conducted (see Table 3). Note that a positive value in the Mean Difference column signifies that the first application type (Type 1) achieved a higher mean value than the second application type (Type 2). These values were derived a subtraction of the respective means in Table 2. Our results revealed the following:

- Overall matched tags. Participants using the control application (M = 19.24) generated significantly more tags that matched the ground truth data than the collaborative (M = 12.29) and competitive (M = 13.67) games. There was however no significant difference in matches between the collaborative and competitive games.

- Level 1 tags. The control application yielded significantly more Level 1 tags (M = 8.65) than the collaborative (M = 3.35) and competitive (M = 6.78) games. In turn, the competitive game performed significantly better than the collaborative game. Therefore in order of performance in generating matching Level 1 tags, the control application ranked first, followed by the competitive and collaborative games respectively.
- Level 2 tags. Again, the control application performed best (M = 8.67), generating significantly more matching Level 2 tags than the collaborative (M = 6.59) and competitive (M = 5.06) games. This time, there was no statistically discernible difference in matches between the collaborative and competitive games.
- Level 3 tags. There were no statistical differences between pairwise comparisons among the three applications. Put differently, the performance in terms of generating matching Level 3 tags was comparable across the control application, collaborative game and competitive game.

[Insert Table 3 here]

In addition, Table 2 suggests that the number of matched tags in each level varies within each application. This was confirmed by running three one-way repeated measures ANOVAs for each application type (control, collaborative, competitive) with the three tag levels (Level 1, Level 2 and Level 3) as the dependent variables. Results indicate that there were significant differences across tag levels for each application type (control application, $F(2, 64) = 48.62$, $p < 0.001$; collaborative game, $F(2, 66) = 31.17$; competitive game, $F(2, 70) = 50.47$, $p < 0.001$). Post-hoc tests revealed the following:

- Control application. The difference between the number of matching Level 1 and Level 2 tags was statistically non-significant. However, there were significantly more matching Level 1 tags than Level 3 tags, and more Level 2 tags than Level 3 tags.
- Collaborative game. All pairwise differences were statistically significant. Interestingly, matching Level 2 tags were generated most, followed by Level 1 and Level 3 tags.
- Competitive game. Again, all pairwise differences were statistically significant. Here, matching Level 1 tags were generated most, followed by Level 2, and Level 3 tags.

Application Perceptions

Table 4 shows the means and standard deviations of participants' perceptions of the applications they used. As in the previous section, these become the dependent variables that were used for further analysis. In addition, qualitative feedback was obtained from the participants. Themes centered around the impressions of the applications they used, focusing especially on the perceived strengths and weaknesses of each application.

[Insert Table 4 here]

One overarching observation as is that participants rated the control application relatively poorly in comparison to the collaborative and competitive games. For example, they found the control application least challenging in terms of usage, least useful for generating tags for images, and found it least appealing, among the three applications evaluated in our study. Qualitative feedback lends support for this. When asked about what their impressions about

the control application, many participants reported there was "*nothing*" they liked about it, and also that it was "*boring to use*".

In contrast, perceptions about the two HCGs were more positive. For the collaborative game, a major reason for enjoyability that emerged was that a player had to put himself/herself in the partner's shoes. One participant remarked that the game was fun because "*you have to guess what the other player is thinking about*". In addition, participants liked the idea of working together to score points. Here, commonly used words included "*liked the team work idea*", "*forming a common understanding is interesting*", and "*two people working together is fun*". The main appeal for the competitive game was the need to challenge one's partner to score points. When participants were asked what they liked about the game, some of the often used words included "*liked the competition*", "*challenging to play*" and "*satisfaction of winning*". This is summed up neatly by one participant who said that "*if I can type faster, I can win a lot of points*". Interestingly, among the two HCG types, the competitive game attracted higher participant ratings than the collaborative game, with the exception of Control, which had the same mean score. In particular, participants seemed to like the competitive game better than the collaborative game.

One-way ANOVAs were performed on the dependent variables in Table 4 to verify whether the differences in participants' ratings were statistically significant. Our analysis indicates that there were significant differences with respect to five of the seven variables: Challenge, $F(2, 100) = 12.62$, $p < .012$; Usefulness, $F(2, 100) = 13.24$, $p < 0.001$; Absorption, $F(2, 100) = 10.27$, $p < 0.001$; Social Interaction, $F(2, 100) = 15.32$, $p < 0.001$; and Appeal, $F(2, 100) =$

15.96, $p < 0.001$. There were however not statistically significant differences among the three applications for Control, $F(2, 100) = .75$, $p = .475$; and Learnability, $F(2, 100) = .64$, $p = .529$.

Post-hoc comparisons using Tukey's test was then conducted (see Table 5) which uncovered the following results:

- Challenge. Participants felt that the collaborative (M = 3.43) and competitive (M = 3.53) games were more challenging than the control application (M = 2.78), and this difference was statistically significant. However, there was no significant difference in ratings between the collaborative and the competitive games.
- Usefulness. In terms of usefulness for generating tags, ratings for both the collaborative (M = 3.82) and competitive (M = 3.98) games were significantly higher than the control application (M = 3.10). The difference in ratings between the two games however, was not statistically significant, suggesting that participants felt that both were just as useful.
- Absorption. Like Usefulness, participants felt that both the collaborative (M = 3.62) and competitive (M = 3.68) games could better capture their attention than the control application (M = 3.03), and the differences in mean ratings were found to be statistically significant. Again, the difference in ratings between the two games were non-significant.
- Social interaction. Unsurprisingly, participants reported that the collaborative (M = 3.38) and competitive (M = 3.94) games could better foster social interaction than the control application (M = 3.03). Here, pairwise comparisons between each game and the control application were found to be statistically significant, but between games, this difference in ratings was not significant.

- Appeal. The collaborative (M = 3.94) and competitive (M = 4.31) games were better liked by their respective participants when compared with the control application (M = 3.06), and these differences were significantly different. However, there was no significant difference between the degree of appeal for the collaborative and competitive games.
- Control. There were no statistical differences between pairwise comparisons among the three applications. Put differently, this suggests that participants felt that they could control the outcome of the applications to a similar degree.
- Learnability. Likewise, pairwise mean differences in ratings between the control application, collaborative game and competitive game was non-significant. This indicates that the learning curve of all three applications were similar.

[Insert Table 5 here]

Discussion

Our study found that participants using the control application generated more tags that matched our ground truth data when compared to the collaborative and competitive games. This was applicable to Level 1 and Level 2 tags, as well as the overall number of tags. The number of Level 3 tags was similar across all three applications, and found to be low in quantity. At the same time, the number of matching tags was also similar between game genres except for one instance in which there were more Level 1 tags produced by the competitive game. Taken together, our study suggests that the manual image tagging application performs better than its game-based counterparts, and that there are no differences

in tag between collaborative and competitive game genres. At first glance, this appears counter-intuitive as existing research seems to suggest that games motivate tag creation (e.g. von Ahn & Dabbish, 2008). However, a closer examination of the literature shows that much of this evidence is anecdotal, and, to the best of our knowledge, there has yet to be empirically-based studies to confirm this notion. For example, most related research has tended to focus on a single game design and implementation, and comparative evaluations against other non-game alternatives are typically not conducted (e.g. Ho et al., 2009; von Ahn & Dabbish, 2004).

We contend that while gaming environments do motivate usage, the mechanisms for gameplay may at times impede the generation of quality computations. Further, an individual's motivations for playing games may not be consistent with the objectives of a particular game for human computation. In our study for example, the pressure to perform, that is, to score points in a fixed amount of time, introduces stresses that may cause participants to submit poorer quality tags. In the collaborative game, there is an implicit obligation to not slow down one's partner by mulling excessively over a presented image. A participant remarked that he did not like the game because of the "… *time factor. We are forced to match the tags within the given time, and so must be as fast as the partner*". In the competitive game, prolonged dwelling over a presented image may cause one to lose points if the opponent is faster at tag generation. One participant lamented, "*if the typing speed is slow, I can't get a score*".

For both cases, the result is that participants likely spent less time thinking about tags, preferring a "shoot first, scatter shot" strategy of tag submission. This led to fewer matching

tags against our ground truth data. Some evidence for this can be gleaned by our finding that the competitive game produced significantly more matching Level 1 tags than the collaborative game. Here, we hypothesize that because a competitive game pits one player against his/her partner, it makes sense to generate as many tags as possible in the least amount of time in the hope of achieving a match. In this case, Level 1 (generic) tags seem a reasonable choice as they are obvious and easier to generate (Robertson et al., 2009; Chung & Yoon, 2009). A comment from a participant lends support for this view, "*I do not like competition because the other player got many points when I was thinking of more specific tags*".

In contrast, the control (manual) tagging application did not have such performance pressures apart from the same countdown timer of five minutes as the two games. This meant that participants using the application could afford to spend relatively more time to think about tags for each image, which consequently meant more matches against our ground truth data. This same observation was made by a participant who said that "*I have time to think about the tags to use*" when asked about her impressions of the control application. This finding therefore appears to concur with the literature on social tagging in which users have been observed to generate quality tags to content even in the absence of gaming mechanisms (Ding et al., 2009; Goh, Chua, Lee, & Razikin, 2009). For example, in the area of images, studies of Flickr content (Stvilia & Jörgensen, 2008; Yoon, 2009) show that the assigned tags were mostly descriptive of the image content and the terms used were comparable to those of existing controlled vocabulary systems and metadata frameworks.

Nevertheless, our results also show that all three applications performed equally poorly in generating the most specific Level 3 tags, suggesting that the ability to produce such tags transcends gaming mechanisms. This is also supported by our observations that the collaborative and competitive games were comparable in terms of the overall number of matching tags generated, as well as for Level 2 and Level 3 tags. Here, we argue the capacity to produce high quality, descriptive tags is inherent in an individual, his/her background, motivation and experiences, expectations of how the content will be used in the future, as well as his/her interaction with the community of taggers, among other factors (Golder & Huberman, 2006; Lee, Goh, Razikin, & Chua, 2009). Consequently, games alone may not be a panacea for tag creation and other mechanisms will have to be investigated such as recommendation systems and automatic keyword extraction algorithms (e.g. Jaschke, Eisterlehner, Hotho, & Stumme, 2009; Melenhorst, Grootveld, van Setten, & Veenstra, 2008). However, whether this finding is generalizable to other human computation domains requires further investigation.

Unsurprisingly, our results showed that participants liked the game-based (both collaborative and competitive) approach to tagging images as opposed to manual tagging, represented by our control application. This concurs with related research (e.g. von Ahn & Dabbish, 2008) demonstrating that games indeed serve as motivators for harnessing human intelligence. Here, the enjoyment one derives from gameplay, together with the challenges to be overcome and incentives received, are just some of the benefits that players obtain (Sweetser & Wyeth, 2005). However, participants were more undecided in terms of liking the collaborative or competitive genres better as borne by our study. It appears that either genre would be acceptable although the competitive game seemed to have a slight, albeit statistically non-

significant, edge over the collaborative game. Perhaps the gratifications associated with challenging and beating an opponent (Vorderer et al., 2003) account for this preference. However, more investigations are required to confirm this assertion.

This pattern of findings was also found for Usefulness, Challenge, Absorption, and Social Interaction, and our results therefore suggest that designers of HCGs should pay heed to these attributes in the development of such applications. However, this is a complex issue because of the need to balance the twin goals of computainment – good game design and effective human computation (Bell et al., 2009; Goh, Lee, & Chua, 2010). On the one hand, a game should harness human intelligence so that quality computations (tags in the case of image tagging applications) can be generated, but at the same ensuring that entertainment and enjoyment is not sacrificed. On the other hand, developers cannot afford to focus primarily on creating an engaging game without thinking about how it can encourage the generation of useful computations. Put succinctly, an entertaining game may not be effective in generating useful outputs, while one that can do so may not be entertaining enough. However, whether these are on opposing ends of a continuum or orthogonal facets of HCG design is an open question that requires further investigation.

The significance of Usefulness in our findings is a case in point. This attribute may be viewed from the two perspectives inherent in HCGs: how effective a game is in entertaining a player, and how effective it is in supporting human computation. In other words, users need to be convinced that they can derive enjoyment through playing the game thereby fostering sustained usage (Hsu & Lu, 2007), and also persuaded that the outputs of the game can have useful applications (e.g. generating tags for improving image retrieval) thereby appealing to

their sense of altruism and/or other motivating drives (Lin, 2007). In promoting the perception of Usefulness, our results suggest that achieving Challenge, Absorption and Social Interaction is critical.

First, Challenge has repeatedly been identified as crucial to game design as it heightens enjoyment (e.g. Qin et al., 2010). Although seemingly simple in design, our collaborative and competitive games support this attribute in the form of time limits, a scoring system commensurate with type of tags generated, and the need to work with or against another player to accomplish the goals of the game. A benefit of a challenging game is that it encourages sustained usage and therefore the potential for generating larger quantities of useful computations. A participant using the competitive game commented that "*I like challenges and I will play it again in the future*". In terms of Absorption, the literature is replete with various principles, some of which include the need to capture the player's attention and focus quickly, provide adequate stimuli, maintain sufficient player workload, and create engaging game scenarios and storylines (Sweetser & Wyeth, 2005). The aforementioned features of our games fulfill many of these principles, with one participant mentioning that "*you must read very fast and must type very fast to ensure that you can score points*". Like Challenge, Absorption fosters sustained usage and the concomitant potential for harvesting more useful computations.

Finally, Social Interaction is becoming increasingly viewed as an important component of game design (Lee et al., 2010) especially with the popularity of social games such as those found in Facebook, and multiplayer role playing games. As people largely enjoy interacting with others either through collaboration or competition, the support for such facilities

promotes replayability. Both our collaborative and competitive games support elements of Social Interaction by their very game mechanics. This is corroborated by participants' comments of both games when asked what they liked about their respective applications. Recurring words include "*competition*", "*challenging*" (to play against one's opponent), "*team work*", and "*working together*". A benefit of Social Interaction is that it can be used to verify the accuracy of the outputs generated. For example, if multiple pairs of players agree on a common tag for an image in a collaborative game, then the confidence that the tag is a descriptive term of the image is increased (von Ahn & Dabbish, 2008).

Interestingly, the differences in Control and Learnability were non-significant across the three applications, suggesting that their ease of use and ease of learning were respectively comparable. Two viewpoints may be adopted here. On a more encouraging note, the usability of the collaborative and competitive games, measured along Control and Learnability, appears to be no different from the manual tagging application. This suggests that the games are likely to be playable by users with different levels of gaming expertise. Simply put, if one knows how to tag images, then one can play our games as well. This perspective therefore bodes well for HCGs in terms of their appeal to users. However, an alternative interpretation is that ease of use and ease of learning does not necessarily translate into effective games for human computation. Returning to our results, the manual tagging application produced more matching tags than the two games even though all three were comparable in terms of Control and Learnability. Again, this finding supports our earlier assertion that game designs need to account for both entertainment and effective human computation.

Conclusion

To summarize, our present study has uncovered a possible tension between entertainment and quality when using games for human computation such as image tagging. Arising from our findings, the following implications may be derived. One, while games may be entertaining and encourage sustained use over nongame-oriented applications, the outputs produced by these games may not always be of high quality and therefore should be treated with caution. There is a possibility that the very game mechanics introduced to heighten excitement may inadvertently hinder the computational process, which in our case refers to image tagging. Two, on a related note, people play games for many reasons, and their motivations may not necessarily align with the objectives of a given game for human computation. Further, in the case of image tagging (and social tagging in general), research has shown that people tag due to a variety of motivations, of which facilitating future content retrieval is just but one of them. Consequently, the generated tags may not be effective content descriptors. Three, if games are deployed, it appears that the collaborative and competitive genres have similar appeal and therefore either may be used. Instead, the challenge is to design games that engage players and motivate repeated play. The significant attributes uncovered in our study provide a starting point for the design of such games.

Although our study has yielded insights, there are a few limitations that should be addressed in future work. First, our results were obtained through a single experiment in a lab setting. A study that involves longer-term repeated use of the applications would be helpful in validating our findings. Second, the study evaluated one instance of each game genre in a specific domain of image tagging, and also did not consider different retrieval scenarios. For better generalizability, it would be instructive to carry out investigations using different game designs, different game mechanics, and different domains of human computation. Next,

participants were primarily undergraduate and graduate students. Replication of this study in other contexts (e.g. more diverse age groups, different educational backgrounds) would be useful to better uncover and understand performance and perceptions of games for human computation. In terms of future work, it should be noted that most HCGs are collaborative in nature, and there is a need to create engaging competitive games to harness the power of the human intellect. In addition, comparing the descriptive/retrieval value of image tags generated by HCGs against professional indexers would be a worthwhile area of study as this has implications for cost versus quality. Finally, while our study has demonstrated that participants' perceptions of games are relatively positive, a possible discrepancy between performance and preference exists, and further research needs to be conducted to better understanding users' motivations for playing such games, as well as designing games that realize the potential for computainment applications.

Acknowledgements. This work was supported by the Singapore National Research Foundation Interactive Digital Media R&D Program, under research Grant NRF NRF2008IDM-IDM004-012.

References

Barrington, L., Turnbull, D., O'Malley, D., & Lanckriet, G. (2009). User-centered design of a social game to tag music. In Proceedings of the 2009 ACM SIGKDD Workshop on Human Computation, (Paris, France, June 28), 7-10. New York: ACM Press.

Bell, M., Reeves, S., Brown, B., Sherwood, S., McMillan, D., Ferguson, J., & Chalmers, M. (2009). Eyespy: Supporting navigation through play. In Proceedings of the 2009

Annual SIGCHI Conference on Human Factors in Computing Systems, (Boston, MA, April 4-8), 123-132. New York: ACM Press.

Brockmyer, J.H., Fox, C.M., Curtiss, K.A., McBroom, E., Burkhart, K.M., & Pidruzny, J.N. (2009). The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. Journal of Experimental Social Psychology, 45(4), 624-634.

Casey, S., Kirman, B., & Rowland, D. (2007). The gopher game: A social, mobile, locative game with user generated content and peer review. In Proceedings of the 2007 International Conference on Advances in Computer Entertainment Technology, (Salzburg, Austria, June 13-15), 9-16. New York: ACM Press.

Champion, E. (2003). Applying game design theory to virtual heritage environments. In Proceedings of the First International Conference on Computer Graphics and Interactive Techniques, (Melbourne, Australia, February 11-14, 2003), 273-274. New York: ACM Press.

Chung, E.K. & Yoon, J.W. (2009). Categorical and specificity differences between user-supplied tags and search query terms for images. An analysis of Flickr tags and Web image search queries. Information Research, 14(3), paper 408. Available at http://InformationR.net/ir/14-3/paper408.html.

Csikszentmihalyi, M. (1990). Flow: The Psychology of Optimal Experience. New York: Harper Perennial.

Datta, R., Joshi, D., Li, J., & Wang, J.Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys, 40(2), article 5. Available at http://doi.acm.org/10.1145/1348246.1348248.

Davis, F.D. (1989). Perceived usefulness, PEOU and user acceptance of information technology. MIS Quarterly, 13(4), 319-340.

Desurvire, H. & Wiberg, C. (2008). Master of the game: Assessing approachability in future game design. In Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, (Florence, Italy, April 05-10, 2008), 3177-3182. New York: ACM Press.

Ding, Y., Jacob, E.K., Zhang, Z., Foo, S, Yan, E., George, N.L., & Guo, L. (2009). Perspectives on social tagging. Journal of the American Society for Information Science and Technology, 60(12), 2388-2401.

Ducheneaut, N., Yee, N., Nickell, E., & Moore, R.J., (2006). "Alone Together?" Exploring the Social Dynamics of Massively Multiplayer Online Games. In Proceedings of the 2006 Annual SIGCHI Conference on Human Factors in Computing Systems, (Montréal, Québec, Canada, April 22–27), 407-416. New York: ACM Press.

Eakins J. P. & Graham, M.E. (1999): Content-based Image Retrieval - A report to the JISC Technology Applications Programme. Institute for Image Data Research, University of Northumbria at Newcastle. Available at: http://www.unn.ac.uk/iidr/research/cbir/report.html.

Goh, D.H., Lee, C.S., & Chua, A.Y.K. (2010). Do games motivate mobile content sharing? In Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries, (Gold Coast, Australia, June 21-25), Lecture Notes in Computer Science 6102, 61-70. Berlin, Germany: Springer.

Goh, D.H., Ang, R.P., & Tan, H.C. (2008). Strategies for designing effective psychotherapeutic gaming interventions for children and adolescents. Computers in Human Behavior, 24(5), 2217-2235.

Goh, D.H., Chua, A., Lee, C.S., & Razikin, K. (2009). Resource discovery through social tagging: A classification and content analytic approach. Online Information Review, 33(3), 568-583.

Golder, S.A. & Huberman, B.A. (2006). Usage patterns of collaborative tagging systems. Journal of Information Science, 32(2), 198-208.

Ho, C.J., Chang, T.H., Lee, J.C., Hsu, J.Y.J., & Chen, K.T. (2009). KissKissBan: A competitive human computation game for image annotation. In Proceedings of the 2009 ACM SIGKDD Workshop on Human Computation, (Paris, France, June 28), 11-14. New York: ACM Press.

Hsu, C.L. & Lu, H.P. (2007). Consumer behavior in online game communities: A motivational factor perspective. Computers in Human Behavior, 23(3), 1642-1659.

Jaschke, R., Eisterlehner, F., Hotho, A., & Stumme, G. (2009). Testing and evaluating tag recommenders in a live system. In Proceedings of the Third ACM Conference on Recommender Systems, (New York, NY, October 23-25), 369-372. New York: ACM Press.

Jennett, C., Cox, A.L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (2008). Measuring and defining the experience of immersion in games. International Journal of Human-Computer Studies, 66(9), 641-661.

Johnson, D. & Wiles, J. (2003). Effective affective user interface design in games. Ergonomics 46(13/14), 1332-1345.

Kankaanranta, M.H. & Neittaanmäki, P. (2009). Design and use of serious games. Netherlands: Springer.

Lazzaro, N. & Keeker, K. (2004). What's my method? A game show on games. In Proceedings of the 2004 Conference on Human Factors in Computing Systems

Extended Abstracts, (Vienna, Austria, April 24-29), 1093-1094. New York: ACM Press.

Law, E. & von Ahn, L. (2009). Input-agreement: A new mechanism for collecting data using human computation games. In Proceedings of the 2009 Annual SIGCHI Conference on Human Factors in Computing Systems, (Boston, MA, April 4-8), 1197-1206. New York: ACM Press.

Lee, C.S., Goh, D.H., Chua, A.Y.K., & Ang, R.P. (2010). Indagator: Investigating perceived gratifications of an application that blends mobile content sharing with gameplay. Journal of the American Society for Information Science and Technology, 61(6), 1244-1257.

Lee, C.S., Goh, D.H., Razikin, K., & Chua, A. (2009). Tagging, sharing and the influence of personal experience. Journal of Digital Information, 10(1). Available at http://journals.tdl.org/jodi/article/view/275/275.

Li, J. & Wang, J.Z. (2003). Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(9), 1075-1088.

Li, J. & Wang, J.Z. (2008). Real-time computerized annotations of pictures. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(6), 985-1002.

Lin, H.F. (2007). Effects of extrinsic and intrinsic motivation on employee knowledge sharing intentions. Journal of Information Science, 33(2), 135-149.

Malone, T.W. (1981). Toward a theory of intrinsically motivating instruction. Cognitive Science, 4(5), 333-369.

Melenhorst, M., Grootveld, M., van Setten, M., & Veenstra, M. (2008). Tag-based information retrieval of video content. In Proceedings of the 1st International

Conference on Designing Interactive User Experiences for TV and Video, (Silicon Valley, CA, October 22-24), 31-40. New York: ACM Press.

O'Brien, H.L. & Toms, E.G. (2008): What is user engagement? A conceptual framework for defining user engagement with technology. Journal of the American Society of Information Science and Technology, 59(6), 938-955.

Pagulayan, R. J., Keeker, K., Wixon, D., Romero, R., & Fuller, T. (2002). User-centered design in games. In J. Jacko and A. Sears (Eds.), Handbook for Human-Computer Interaction in Interactive Systems. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Pavlidis, T. (2009). Why meaningful automatic tagging of images is very hard. In Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, (New York, June 28 - July 3), 1432 – 1435.

Pinelle D., Wong N., & Stach T. (2008). Heuristic evaluation for games: Usability principles for video game design. In Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, (Florence, Italy, April 05-10, 2008), 1453-1462. New York: ACM Press.

Qin, H., Rau, P.P., & Salvendy, G. (2010). Effects of different scenarios of game difficulty on player immersion. Interacting with Computers, 22(3), 230-239.

Ravaja, N., Salminen, M., Saari, T., Laarni, J., Holopainen, J., & Jarvinen, A. (2004). Emotional response patterns and sense of presence during video games: Potential criterion variables for game design. In Proceedings of the Third Nordic Conference on Human-Computer Interaction, (Tampere, Finland, October 23-27, 2004), 339-347. New York: ACM Press.

Rieber, L.P. (1996). Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. Educational Technology Research and Development, 44(2), 45-58.

Robertson, S., Vojnovic, M., & Weber, I. (2009). Rethinking the ESP Game. In Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems, (Boston, MA, April 4-9), 3937-3942. New York: ACM Press.

Saeed, K. & Abdinnour-Helm, S. (2008). Examining the effects of information system characteristics and perceived usefulness on post adoption usage of information systems. Information & Management, 45(6), 376-386.

Seloff, G.A. (1990). Automated access to NASA-JSC image archives. Library Trends, 38(4), 682-696.

Siorpaes, K. & Hepp, M. (2008). Games with a purpose for the Semantic Web. IEEE Intelligent Systems, 23(3), 50-60.

Stephenson, W. (1967). The play theory of mass communication. Chicago, IL: University of Chicago Press.

Stvilia, B. & Jörgensen, C. (2008). User-generated collection-level metadata in an online photo-sharing system. Library & Information Science Research, 31(1), 54-65.

Sun, H. & Zhang, P. (2008). An exploration of affect factors and their role in user technology acceptance: Mediation and causality. Journal of the American Society for Information Science and Technology, 59(8), 1252-1263.

Sweetser, P., & Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. Computers in Entertainment, 3(3). Available at http://doi.acm.org /10.1145/1077246.1077253.

Tuite, K., Snavely, N., Hsiao, D.Y., Smith, A.M., & Popovic, Z. (2010). Reconstructing the world in 3D: Bringing games with a purpose outdoors. In Proceedings of the Fifth International Conference on the Foundations of Digital Games, (Monterey, California June 19-21), 232-239. New York: ACM Press.

von Ahn, L, and Dabbish,.L. (2004). Labeling images with a computer game. In Proceedings of the 2004 Annual SIGCHI Conference on Human Factors in Computing Systems, (Vienna, Austria, April 24-29), 319–326. New York: ACM Press.

von Ahn, L. & Dabbish,.L. (2008). Designing games with a purpose. Communications of the ACM, 51(8), 58-67.

von Ahn, L., Liu, R., & Blum, M. (2006). Peekaboom: A game for locating objects in images. In Proceedings of the 2006 Annual SIGCHI Conference on Human Factors in Computing Systems, (Montréal, Québec, Canada, April 22-27), 55-64. New York: ACM Press.

Vorderer, P., Hartmann, T., & Klimmt, C. (2003). Explaining the enjoyment of playing video games: The role of competition. In Proceedings of the Second International Conference on Computer Games, (Pittsburgh, PA, May 8-10), 1-9. New York: ACM Press.

Walsh, G. & Golbeck, J. (2010). Curator: A game with a purpose for collection recommendation. In Proceedings of the 2010 Annual SIGCHI Conference on Human Factors in Computing Systems, (Atlanta, GA, April 10-15), 2079-2082. New York: ACM Press.

Woszczynski, A.B., Roth, P.L., & Segars, A.H. (2002). Exploring the theoretical foundations of playfulness in computer interactions. Computers in Human Behavior, 18(4), 369-388.

Westman, S. (2009). Image users' information needs and search behaviour. In A. Goker and J. Davies (Eds.), Information Retrieval: Searching in the 21st Century (pp. 63-83), Chippenham, Wiltshire: Wiley.

Yoon, J.W. (2009). Towards a user-oriented thesaurus for non-domain-specific image collections. Information Processing & Management, 45(4), 452-468.

Zyda, M. (2005). From visual simulation to virtual reality to games. IEEE Computer, 38(9), 25-32.

Tables

Table 1. Summary of human computation games.

Game	Purpose	Genre	Reference
ESP Game	Creating keywords for images	Collaborative	von Ahn & Dabbish (2004)
Peekaboom	Identifying objects in images	Collaborative	von Ahn et al. (2006)
Herd It	Creating keywords for music clips	Collaborative	Barrington et al. (2009)
OntoPronto	Creating ontologies from Wikipedia entries	Collaborative	Siorpaes & Hepp (2008)
OntoTube	Annotating YouTube videos with ontological elements	Collaborative	Siorpaes & Hepp (2008)
Curator	Classifying items into collections	Collaborative	Walsh & Golbeck (2010)
Gopher Game	Creating location-based annotations and images	Collaborative	Casey et al. (2007)
Eyespy	Generating photographic and textual descriptions of locations	Collaborative	Bell et al. (2009)
KissKissBan	Creating keywords for images	Competitive	Ho et al. (2009)
Indagator	Creating location-based annotations	Competitive	Lee et al. (2010)
PhotoCity	Creating high-quality location-based photos	Competitive	Tuite et al. (2010)

Table 2. Means and standard deviations of matched tags in the study.

Tags	Game Type					
	Control (N = 33)		Collaborative (N = 34)		Competitive (N = 36)	
	N (%)	Mean (SD)	N (%)	Mean (SD)	N (%)	Mean (SD)
Overall*	633 (100)	19.24 (6.99)	417 (100)	12.29 (5.91)	491 (100)	13.67 (3.87)
Level 1*	285 (45.02)	8.64 (4.04)	114 (27.34)	3.35 (2.30)	224 (45.62)	6.78 (2.59)
Level 2*	286 (45.18)	8.67 (4.20)	224 (53.72)	6.59 (3.69)	182 (37.07)	5.06 (2.45)
Level 3	62 (9.80)	1.88 (1.65)	79 (18.94)	2.32 (1.84)	65 (13.24)	1.81 (1.19)

Note: * Statistically significant differences between the three applications at $p < .05$.

Table 3. Comparison between means of the dependent variables.

Tags	**Type (1)**	**Type (2)**	**Mean Difference (1) - (2)**
Overall	Control	Collaborative	6.95*
	Control	Competitive	5.58*
	Collaborative	Competitive	-1.37
Level 1	Control	Collaborative	5.28*
	Control	Competitive	1.86*
	Collaborative	Competitive	-3.43*
Level 2	Control	Collaborative	2.08*
	Control	Competitive	3.61*
	Collaborative	Competitive	1.53
Level 3	Control	Collaborative	-.44
	Control	Competitive	.07
	Collaborative	Competitive	.52

Notes: * $p < .05$. Type (1) and Type (2) refer to the application types being compared.

Table 4. Means and standard deviations for participants' perceptions.

Variable	Game Type					
	Control (N = 33)		Collaborative (N = 34)		Competitive (N = 36)	
	Mean	SD	Mean	SD	Mean	SD
Challenge*	2.81	.66	3.43	.61	3.53	.63
Usefulness*	3.10	.97	3.82	.63	3.98	.61
Absorption*	3.03	.79	3.62	.57	3.68	.59
Control	3.40	.76	3.57	.61	3.57	.52
Learnability	3.62	.81	3.74	.78	3.83	.75
Social interaction*	3.03	1.01	3.38	.56	3.94	.57
Appeal*	3.06	1.06	3.94	.85	4.31	.89

Notes: * Statistically significant differences between the three applications at $p < .05$. Each variable is measure along a range of 1 (strongly disagree) to 5 (strongly agree).

Table 5. Comparison between means of participants' perception variables.

Variable	**Type (1)**	**Type (2)**	**Mean Difference (1) - (2)**
Challenge	Control	Collaborative	-.61*
	Control	Competitive	-.72*
	Collaborative	Competitive	-.10
Usefulness	Control	Collaborative	-.72*
	Control	Competitive	-.88*
	Collaborative	Competitive	-.16
Absorption	Control	Collaborative	-.60*
	Control	Competitive	-.65*
	Collaborative	Competitive	-.05
Social interaction	Control	Collaborative	-.80*
	Control	Competitive	-.91*
	Collaborative	Competitive	-.11
Preference	Control	Collaborative	-.88*
	Control	Competitive	-1.25*
	Collaborative	Competitive	-.36
Control	Control	Collaborative	-.17
	Control	Competitive	-.16
	Collaborative	Competitive	.00
Learnability	Control	Collaborative	-.12
	Control	Competitive	-.21
	Collaborative	Competitive	-.10

Notes: * $p < .05$. Type (1) and Type (2) refer to the application types being compared.

Images

Figure 1. Example of an image used in the study.

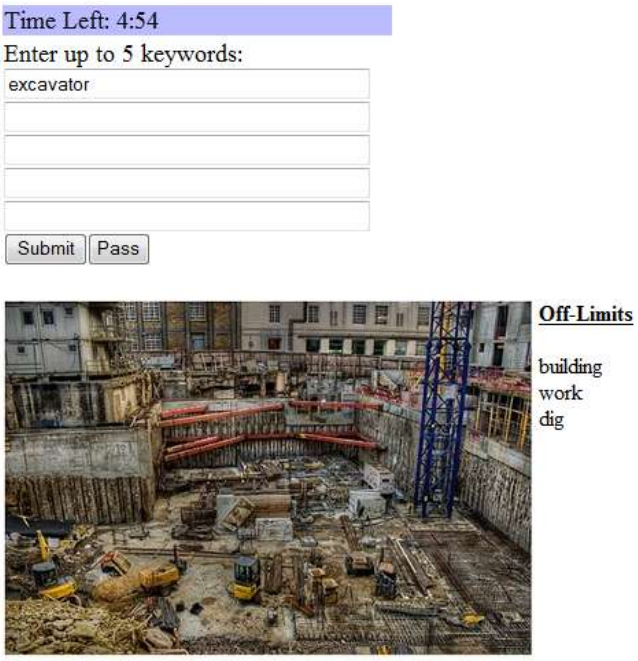

Figure 2. Manual image tagging application developed for the study.

Figure 3. Collaborative HCG developed for the study.

Appendix: User Experience Questions

The following statements elicited participants' perceptions of the applications they used. Questions were rated on a scale of 1 (strongly disagree) to 5 (strongly agree).

1. I felt that the game was sufficiently challenging for me.
2. I felt that the level of challenge increased as the game progressed.
3. The game is able to challenge people with different skill levels.
4. I found the game challenging even after playing many rounds.
5. I found that the game interface is simple and well-designed.
6. I felt bored when I was playing the game.
7. I found that the game was difficult and stressful.
8. I was able to stay focused on the game tasks.
9. The game was intellectually stimulating.
10. I was motivated by the given time-limit and/or scoring system of the game to continue playing.
11. The actions I took in the game could impact my score.
12. The design of the game prevents serious errors from occurring.
13. I was able to recover from errors that I made without affecting the operation of the game.
14. I could learn quickly how to play the game.
15. I could play the game without reading the instructions.
16. I found that learning to play the game was part of the fun.
17. Help was available when I was faced with difficulties in the game.
18. The game supports interaction with other players.

19. The game allowed me to compete/collaborate with other players.
20. The game is a useful means of building social communities with other players.
21. The game is a useful for creating new keywords for images
22. The game encourages me to create new keywords for images.
23. The game is worth playing.
24. I enjoy playing the game.
25. I will continue to play the game if it was available.