# ClueMeIn: Obtaining More Specific Image Labels Through a Game

**Christopher G. Harris**
School of Mathematical Sciences
University of Northern Colorado
Greeley, CO 80639 USA
christopher.harris@unco.edu

**Abstract**

The ESP Game (also known as the Google Image Labeler) demonstrated how the crowd could perform a task that is straightforward for humans but challenging for computers – providing labels for images. The game facilitated the task of basic image labeling; however, the labels generated were non-specific and limited the ability to distinguish similar images from one another, limiting its ability in search tasks, annotating images for the visually impaired, and training computer vision machine algorithms. In this paper, we describe ClueMeIn, an entertaining web-based game with a purpose that generates more detailed image labels than the ESP Game. We conduct experiments to generate specific image labels, show how the results can lead to improvements in the accuracy of image searches over image labels generated by the ESP Game when using the same public dataset.

**Keywords:** ESP Game, Image Labeler, Games With a Purpose, GWAP, Computer Vision

## 1. Introduction

There are numerous benefits to image recognition and labeling (i.e., tagging), such as providing better accuracy in image searches to greater accessibility for visually impaired users. Despite impressive advances in computer vision, many challenges remain in image recognition; it remains an intractable problem. Unlike machines, humans are capable of image recognition and labeling but require incentives. Games With A Purpose (GWAP) are a class of games that developed to bridge the gap between human and machine abilities. One goal of GWAPs is to aid in annotation or labeling of items for training machine learning algorithms.

In 2004, the ESP Game was created to assist in these tasks of image recognition and labeling (Von Ahn and Dabbish, 2004). By integrating image recognition and labeling tasks into an entertaining game, it provides an incentive to human players by aligning game performance with task achievement.

The ESP Game randomly matches two players with no other means of communication. The two players are shown the same image, and they both enter words that can be used to describe the image. The objective is for the two players to enter the same word or phrase, which earns them points and becomes a label, or tag, to describe the image. Labels successfully assigned to that image become "taboo" words, which do not earn points in future games. There is a time limit to increase player engagement: players have 150 seconds to label 15 images.

One clear limitation of the ESP Game is that the tags given to the images are generic and rarely provide enough information to discriminate between *similar* images (see Figure 1). In this paper, we introduce a game, ClueMeIn, to address this problem. We designed ClueMeIn to generate more specific image labels, to improve image search accuracy, to train machine learning algorithms, and to increase accessibility for the visually impaired. In the next section, we discuss the limitations of the ESP Game as an image labeler; in Section 3, we describe other games that have also been designed to label images. In Section 4, we discuss the design and creation of ClueMeIn. We describe experiments in Section 5, followed by analysis in Section 6. Last, we conclude and mention future work in Section 7.

## 2. Limitations of the ESP Game

The ESP Game was adapted in 2008 as the Google Image Labeler. Starting with a collection of 350k images, the game later used randomly selected images from the web to create its image dataset.

One limitation of the ESP Game is that players are given incentives to type the most obvious labels, which maximizes agreement with other players (and consequently points). Due to its reliance on matching, the ESP Game rewards players providing generic terms and punishes players for the use of more informative (but rare) terms. The reward mechanism ensures players are more likely to achieve a match if they enter generic terms as opposed to specific ones. This has been demonstrated through a game-theoretic approach by Jain and Parkes (2009). Moreover, the generic nature of the ESP Game labels defeats the advantages that human computation provides.

Figure 1: A game to distinguish between similar images, such as these boats, can create more meaningful labels.

The use of more generic terms by players also encourages redundancies in the labels; Weber, Robertson, and Vojnovic (2009) indicated that 81% of images labeled with the "guy" were also labeled with "man." Thus, the more general the generated terms, the less informative they are in describing the image.

A second limitation of the ESP Game is that generic labels such as "car" provide little benefit to image collections, except at a superficial level; a search on "car" on any popular image search engine, such as those provided in Bing or Google, will return more than 100M images. Many labels have a strong association with one another and can be predicted through simple word association, like "sky" and "clouds." Also, there is a strong tendency to rely on colors as labels – an aspect of computer vision that machines can already detect with high accuracy. Therefore, the ESP Game favors general labels for an image over specific ones, as this is the best strategy to match other players and to generate the most points. However, this is less useful for generating labels for search tasks.

A third limitation is that labels can be ascertained using language models or other means, limiting the human-added value. For example, Weber, Robertson, and Vognovic developed a program to play the ESP Game without the need to evaluate the actual image. Their program disregards the visual content of the images and predicts likely tags by analyzing the taboo words and then applies a probabilistic language model. It manages to agree with the human partner on a label for 69% of images, growing to 81% of images with at least one assigned taboo term. Thus, human players provide little additional information to the existing tags even when taboo words are used. ClueMeIn overcomes these limitations by providing more informative labels than the ESP Game is able to do; our focus is on having participants identify a single image from a set of similar images by specifying increasingly precise labels.

## 3. Related Work

Since its initial development, the ESP Game has inspired other image labeling games. *Peekaboom*, by the same creators as the ESP Game, looks for pixel boundaries of objects in images (Von Ahn, Liu, and Bloom, 2006). Human annotators enhance image metadata to create better learning algorithms. While the outputs are different from the ESP Game, the methods of collecting data are similar.

*Karido* uses a collaborative framework to tell works of art apart (Steinmayr et al., 2011). In Karido, nine similar images are randomly selected from a given database of artwork with the objective of increasing tag diversity. Players take turns either playing the Guesser or describer of the image selected by the system to be described. To discourage random guessing, the score of both players is reduced as a penalty if a wrong image is selected. This penalty exceeds the bonus for selecting the correct image.

*Phetch* is not designed to collect image labels but to collect entire sentences that described an image (Von Ahn et al., 2007). Three to five players play each round of Phetch, one of which is randomly selected as the describer while the remaining players become seekers. Initially, a picture is shown to the describer, who enters descriptive sentences to guide the seekers. The seekers use a search engine within the game to locate the described image. If a seeker selects the correct image, that seeker and the describer are awarded a score bonus. Once the correct image has been found, the winning seeker becomes the describer in the subsequent round. To discourage random guessing, points are deducted whenever a seeker makes an incorrect guess.

One issue with the ESP Game is the lack of tag diversity. Ho et al. created *KissKissBan* (2009), which introduces a third player and a competitive element in KissKissBan. The first two players (called a couple) try to achieve the same goal as in the ESP Game. The third player in KissKissBan, called the blocker, is competing with the other two players. Before each round begins, the blocker can see the image and has seven seconds to enter as many words as possible, which the couple is not allowed to use. Unlike the taboo words in the ESP Game, the couple cannot see this list of words. If one of the players in the couple enter a blocked word, five seconds are deducted from their allotted time. If the timer runs out before the couple achieves a match, their scores are decreased and the blocker's score is increased; if the couple has a successful match, their score increases while the blocker's decreases.

*PhotoSlap* by Ho et al. (2007) is a web-based variation of Snap, a popular card game. PhotoSlap engages users in an interactive game that capitalizes on the human ability to quickly decipher whether the same person shows up in two consecutive images presented by the computer. The game mechanism encourages rational play; in other words, from a game-theoretic view, the optimal player strategy is not to collude, but balance cooperation with competition.

*Picture This*, by Bennett et al. (2009) is designed not to label images directly, but rather to improve query results using existing tags. Other image labeling tools exist in non-gamified formats as well. *LabelMe* (Russell et al. 2007), a web-based tool for annotating images and sharing those annotations within a community of users, provides an easy-to-use interface for manual labeling of object information, including position, shape, and object label. Likewise, *ImageTagger* (Fiedler, Bestmann, and Hendrich, 2018) is a collaborative labeling tool that allows also includes an automated photo annotation option.

## 4. ClueMeIn: Designing for Informative Labels

ClueMeIn falls into the class of inversion-problem games, as defined by Von Ahn and Dabbish (2004) In these games, one player transforms a given input (the selected goal image) into an intermediary output (i.e., the textual description). The second player tries to transform the intermediary output back into the original input (i.e., by selecting the correct image). Inversion-problem games are designed for player success to be associated with the degree

to which the intermediary output becomes a representation of the original input.

### 4.1 Dataset

For our dataset, we use the IAPR TC-12 image retrieval benchmark, a collection of 20k images created for the CLEF cross-language image retrieval track (ImageCLEF) (Grubinger, 2006) In our initial experiment, we manually selected 473 similar images on several themes (e.g., boats, waterfalls, birds, churches). We assigned these images to 40 image pools based on a single theme (e.g., sailboats, waterfalls, clouds). Image pool sizes ranged between 5 and 18 with a mean size of 11.83. ClueMeIn randomly assigned images for a single image pool in groups of 3 to 9 for each game session. As with the ESP Game, clues provided by players for an image in earlier games became "taboo" words for that image in subsequent games. To test image similarity, we focused on images taken at different angles and of very similar items, such as those seen in Fig. 2.

Figure 2: Some images are challenging to come up with unique labels, such as with these four images.

### 4.2 Game Design

Unlike the ESP Game, which examines a single image, our game, ClueMeIn, presents the pair of players with between three and nine *similar* images. These similar images can be selected using those with identical labels from the ESP Game or other sources. ClueMeIn is designed to develop labels that distinguish similar images from one another. It, therefore, focuses more on providing informative labels without the penalties associated with generic labels.

Players take turns playing two roles- one player serves as the *Guesser* while another serves as the *Cluegiver*. Players are each presented with the same set of images in a randomized order. The game identifies the one image for the Cluegiver to describe to the Guesser (see Fig. 3). Because the order is randomized, providing clues based on the relative position of each image will not help describe a specific image, nor will providing comparative words (most *term+er*, *least term+er*, *term+est*, etc.) as these are not permitted. As in Karido, label inputs are restricted to a maximum of three words and all punctuation is removed. Because there is less of a focus on matching and more of a focus on using the human-provided clues to discriminate


Figure 3: Screenshots from the game indicating the view of Guesser (top) and Cluegiver (bottom). Players take turns in each role and have different incentives to facilitate meaningful clues.

between images, the information contained in the labels themselves is better at describing that image.

In ClueMeIn, each of the two players serves as Guesser and Cluegiver five times on different sets of images either from the same image pool or different image pools. Each player alternates between the two roles, Guesser and Cluegiver, in an attempt to maximize the number of points. ClueMeIn assigns points based on different behaviors.

- *Cluegivers* are given points based on how unique their clues (words or phrases) are – we examine the label frequency, and once a clue has been mentioned three times (across multiple games), it is added to the list of "taboo" words. By dividing the number of labels supplied for that image overall by the number of instances the label has appeared previously, we arrive at a raw score. We apply some normalization and smoothing to arrive at an overall score for that label, rewarding more unique labels more than commonly-used ones. As each image is evaluated more frequently, the label quality increases since the more commonly-used clues are awarded fewer points or become taboo words after they are given for the third time for that image. Taboo and comparative clues are not conveyed to the Guessers; however, an error message is returned to the Cluegiver, indicating the word is off-limits.
- *Guessers* are given points based on how few guesses they use to identify the correct image. They can only make a single guess after a clue has been provided by the Cluegiver. We count the number of guesses, minus the chance they would guess randomly as the raw score. We apply some normalization and some smoothing to arrive at an overall score for a correct selection. Therefore, if five images are presented, Guessers are given more points for guessing the first image correctly than guessing the second time correctly out of the remaining four.
- To prevent the Cluegiver from supplying intentionally useless labels or the Guesser from making intentionally

poor guesses, a portion of points are assigned equally to both players per session based on their mutual performance. Although this reward is the opposite of the penalty assigned for random guessing in Karido, it has a similar effect. The number of points given to each is 25% of the combined number of points the two players achieve in that round (see Fig. 4 for an example). Players were provided this information in advance to persuade them not to be adversarial.

### 4.3 Game Interface

The game interface was designed in Flash to be played through a web browser. Image categories, each pulled from a separate image pool, were randomly selected, as were the images from each pool. Each participant could only evaluate a given group of images once. All players had the option to create and log into an account or remain anonymous (but were tracked by a userID only). ClueMeIn provides a leaderboard for players who logged in to see their overall rank (given as a percentile) for the day and the overall campaign (see Fig. 5).

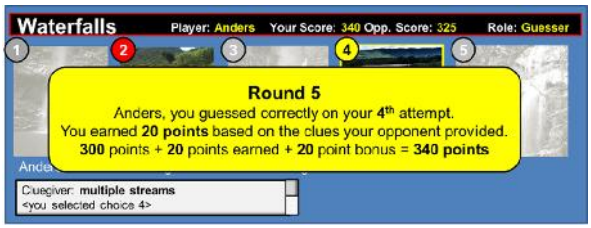


Figure 4: Screenshot of the information given to the players at the end of each round. The bonus given is 25% of the number of points earned by both Guessers and Cluegivers.


Figure 5: A screenshot of the information given to each player who logs in (players can also play anonymously).

### 4.4 Other Game Design Considerations

We also ran some small usability experiments to determine the best method for achieving informative labels. We examined the use of a countdown timer, but the competitive nature of our game often made the two players more adversarial – some players intentionally slowed each other down to achieve a higher score. A survey of player satisfaction suggested we get rid of the timer (which we did). We plan to explore other options to enhance the speed of the games.

Deciding how cooperative versus competitive to make the game was another consideration. There is some emerging research (e.g., Siu, Zook, and Riedl, 2014; Siu and Riedl, 2016) that examines the role of competition vs. cooperation in games. Von Ahn and Dabbish (2008) argue that games like the ESP Game work better because it is cooperative, while most entertainment-related games work best when the environment is competitive. Emmerich, K., & Masuch (2013) found that the desirable game characteristics of immersion and flow were greater in competitive gaming formats, while empathy was greater in cooperative gaming formats. We experimented with cooperation by taking the average of the scores obtained by the Guesser and Cluegiver and giving the same score to each. While this seems equitable, it made the game less enjoyable based on our player satisfaction survey. We found that providing the scoring approach for each player described earlier made it competitive without producing adverse effects, such as misleading clues or bad guesses.

## 5. Experiments

We conducted a series of experiments over six weeks to evaluate the design of our game. These experiments build on the preliminary studies found in Harris (2018). For label generation, we recruited and randomly distributed 40 participants, comprised of students all proficient in the English language from a four-year university, into two groups with a 60-40 split.

### 5.1 Gathering Labels

We then replicated the ESP Game format using the 473 images in our dataset. Our objective was to determine the labels the ESP Game could generate; this became our *baseline*. We had 16 participants (average age = 23.2, males = 13) play 373 five-round games of the ESP Game, generating a total of 2098 labels, or 4.44 labels per image. Of these 2098 labels, 997 (47.5%) were "taboo" at the end of the six-week gaming period (i.e., they had been given as clues three or more times).

Next, we had 24 participants (average age = 22.4, males = 18) play a total of 886 games of ClueMeIn with the same 473 images, averaging 36.9 games per participant. We gathered a total of 4514 unique labels across the 473 images, averaging 9.54 labels per image. Of these 4514, taboo labels totaled 2437 (54%) at the end of the campaign.

### 5.2 Determining Label Quality

We evaluated the quality of the generated labels from the ESP Game and ClueMeIn; high-quality labels should be specific enough to identify an image from a pool of images.

To accomplish this, we provided the generated labels obtained for all images to 10 participants (who did not participate in the labeling tasks). Four were asked to use the 2098 labels generated by the ESP Game and the other six using the 4514 labels from ClueMeIn.

Each participant was asked to identify which image (from the 473 total images) was best represented by the provided label. When the labels were created, participants were only able to see the subset of images from that pool that appeared in that round of the game; however, good quality labels should identify the correct image (even those that

were unseen as choices in the game) when a particular label was created. Although we divided up the labels among the participants, 20% of the image labels were evaluated by more than one participant to examine inter-rater reliability (IRR), a measure of consistency among observational ratings provided by multiple coders. We obtained a Fleiss' $\kappa$ of 0.610 and 0.672 for the ESP Game and ClueMeIn evaluators, respectively, indicating substantial agreement (Landis and Koch, 1977). Participants performed the label matching task independently (i.e., not as a group).

Each of the 10 participants evaluated multiple searches. Some of the searches were provided with search results in three formats:

- *ordered* lists (ordered in decreasing order by term frequency, but no frequency was provided)
- *unordered* lists (a list of search terms listed in random order without the knowing the number of times players generated each term
- *ordered weighted* lists (ordered in decreasing order by term frequency, where the frequency count was provided)

The number of searches participants received with each type of list, whether they received the ClueMeIn generated list of terms vs. the ESP Game list of terms, as well as the assignment of list type to each search was each independently and randomly determined.

When participants were provided with an unordered list, the average accuracy (calculated as the number of correctly assigned labels/total number of labels) was 68.0% for those generated by the ESP Game and 88.7% for those generated by ClueMeIn, a substantial difference. This shows that even when information about the frequency of terms is not given, the quality of labels generated using ClueMeIn is superior to those generated for the same images using the ESP Game.

When an ordered weighted list was provided, the average accuracy increased to 78.1% for those generated by the ESP Game and 96.6% for those generated by ClueMeIn, also a large difference. Since both the ESP Game and ClueMeIn results were provided with the same type of list. Again, these results showing a consistent jump in accuracy indicate that it was not the dataset used, but the game format, that made a difference in label quality.

The improvement of results between the three list types as the information becomes more meaningful (first ordered, then both ordered and weighted) is attributable to a form of bias called *search result bias*. A violation of search neutrality, this bias occurs when people scan a list of terms from top to bottom and perceive the ones towards the top are more important than those further down the list (Kulshrestha et.al., 2019). This has known to have an impact on various aspects of daily life, from searching through a phone directory to find a business to the order candidate names are listed on election ballots. However, we examine these because most labels have an implied order, and the use of these ordered, weighted list of labels provides a more realistic scenario than the unordered list.

We note that while participants selected from all 473 images, the pools were distinct enough that possible labels were, in practice, restricted to a single pool of images (e.g., sailboats). Although the average image pool size to select a given a label from was small (11.83), we believe the method in which labels generated for an image show promise to enhance the accuracy of image searches overall.

## 6. Analysis

Better quality labels help us generate more meaningful annotations for images, more descriptive image tags for the visually impaired, and richer information for training machine learning algorithms. The better accuracy achieved by human evaluators indicates the design of the ClueMeIn game by which labels are generated for an image show promise to enhance the accuracy of image searches overall relative to that used in the ESP Game. We also note the number of labels (4514 vs. 2098) and the diversity (the number of non-taboo tags: 2077 vs. 1101) was more than double using ClueMeIn; this is also a measure which implies the richer language used in creating labels through ClueMeIn.

One may observe that more games were played of ClueMeIn than the ESP Game; however, both game campaigns ended when the rate of new label generation fell below 0.5 (defined as the average number of new non-taboo labels generated for an image per round of the game). This also indicates the ability of ClueMeIn to generate more diverse labels. With a larger pool of images, we believe the diversity of labels would increase with ClueMeIn (due to the need to create specific labels to distinguish between similar images), but not necessarily with the ESP Game (which examines a single image at a time).

We used the 2015 version of the Linguistic Inquiry and Word Count (LIWC) to evaluate aspects of the language used in each label. Our analysis using LIWC is limited because our labels are limited to three words and contained a few of the features normally common in free-form text. Some comparisons on key linguistic features between the ESP Game and ClueMeIn labels were possible and are given (in a normalized form) in Table 1.

| Metric | ESP Game | ClueMeIn |
|---|---|---|
| Words>6 letters | 0.768 | 0.845 |
| Dictionary words | 0.923 | 0.881 |
| Use of Numbers | 0.217 | 0.294 |
| Use of Quantifiers | 0.265 | 0.338 |
| Cognitive Terms | 0.373 | 0.460 |
| Perceptual Terms | 0.318 | 0.377 |

Table 1: Comparison of LIWC metrics between labels obtained from the ESP Game and ClueMeIn

From this, we can see that the language used in the ClueMeIn labels use longer words (>6 letters), more numbers and quantifiers, use words that are more cognitive and more perceptual but use fewer dictionary words than labels generated on the same dataset using the ESP Game.

These are linguistic characteristics often associated with more specific, meaningful terms (e.g., Chuang et.al. 2012, Pitt and Samuel, 2006).

We designed ClueMeIn to be entertaining – that is, participants enjoy playing the game and don't perceive it as a task. To examine this, we asked our 40 participants to evaluate the game they were assigned to play on enjoyment (how much fun it was to play relative to other games) and engagement (how sticky the game was) on a five-point scale, 1 = lowest, 5 = highest. Participants, on average, found the ClueMeIn game more enjoyable (3.58 vs. 3.06) and more engaging (3.71 vs. 3.38) than the ESP Game, indicating a greater potential for participants to enjoy the game and play for longer periods. See Fig. 6 for a box-and-whisker plot of the results for each.

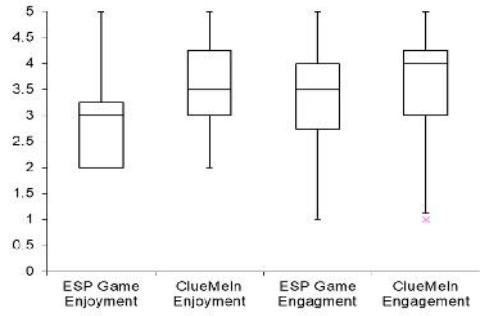


Figure 6. A box and whisker plot comparing enjoyment and engagement for the ESP Game and ClueMeIn.

## 7. Conclusion and Future Work

We have implemented an entertaining web-based game, ClueMeIn, to provide more specific image labels and improve the accuracy of image searches. The design of ClueMeIn addresses some of the weaknesses of the popular ESP Game (Google Image Labeler). While the ESP Game was designed to provide broad labels, advancements in computer vision have propelled past what the ESP Game was intended to accomplish. ClueMeIn can build upon the initial tags generated by the ESP Game to create image pools (i.e., "house") which in turn can provide a game environment to compare similar images of houses with one another, forcing the adoption of more specific labels, (i.e., "Greek Revival architecture"). This increase in granularity can be repeated (all houses with a "Greek Revival architecture" label can then be compared using ClueMeIn once again to get even more specific labels). It can also advance word embeddings that tied to images; people not familiar with architecture may understand columns (see Fig 7), but more specific labels will help build word embeddings that capture the similarities and differences.

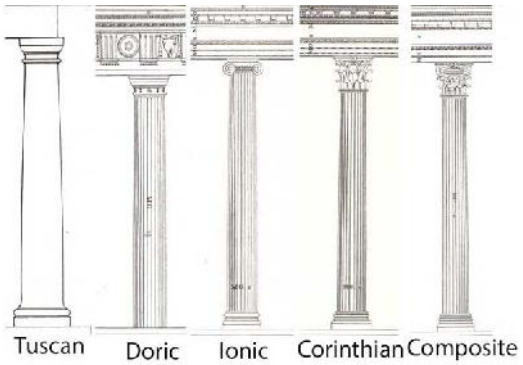


Figure 7. Five classical orders of columns. ClueMeIn can help illustrate the differences in these orders, enhancing word embeddings and leading to more descriptive labels. Image from Mitrović (1999).

As the clues get more specific and the list of "taboo" words grows, the clues that separate images become less and less important to the images. This feature is especially true when identifying the image from a larger pool of similar images. In Fig. 8, we see that the words used to separate these two images, 'grass" and "rocks," will be winning clues in our game but are not very descriptive of the image overall.

Figure 8. Sometimes seemingly irrelevant facts can separate two similar images. "Rocks" was a label given to the image on left, but "grass" was a label given to the image on the right. We resolve this by assigning weights to these labels based on Cluegiver frequency

Some challenges remain. One challenge is how the game should properly weigh the labels. Image labels identified in earlier sessions become "taboo" words in later sessions for other players, but these labels contain more obvious identifiers and need to be weighed higher in the label metadata. We are currently exploring how to properly model and apply term weights to these image labels.

We will continue to apply the game to an expanding pool of similarly themed images. Once the pool of images is sufficiently large (e.g., "cars"), we plan to examine the game's labeling effects on large-scale image searches. Initial results are promising.

We also plan to explore how we can make the game more enjoyable and immersive for players. We are exploring the addition of game elements to improve the flow of the game. We are also looking at other game mechanisms such as scoring, collaborative vs. competitive elements, and how to reward players who devote a significant amount of time (and provide significant value) recognizing and labeling images in ClueMeIn.

One further use for ClueMeIn is that it has the possibility of helping language learners understand and apply terms in a second language to images they already know and

understand, help build a better list of synonyms and possibly help build a stronger, more robust set of word embeddings that can be tied to a specific image.

## 8. Bibliographical References


Bennett, P. N., Chickering, D. M., & Mityagin, A. (2009). Picture this: preferences for image search. *In Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 25-26).

Chuang, J., Manning, C. D., & Heer, J. (2012). "Without the clutter of unimportant words" Descriptive keyphrases for text visualization. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(3), 1-29.

Emmerich, K., & Masuch, M. (2013). Helping friends or fighting foes: The influence of collaboration and competition on player experience. *In FDG* (pp. 150-157).

Fiedler, N., Bestmann, M., & Hendrich, N. (2018). Imagetagger: An open-source online platform for collaborative image labeling. *In Robot World Cup* (pp. 162-169). Springer, Cham.

Grubinger, M., Clough, P., Müller, H., & Deselaers, T. (2006). The iapr tc-12 benchmark: A new evaluation resource for visual information systems. *In International workshop ontoImage* (Vol. 2).

Harris, C. (2018). ClueMeIn: Enhancing the ESP Game to Obtain More Specific Image Labels. *In Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts* (pp. 447-452).

Ho, C. J., Chang, T. H., & Hsu, J. Y. J. (2007). Photoslap: A multi-player online game for semantic annotation. *In Proceedings of the National Conference on Artificial Intelligence (Vol. 22, No. 2, p. 1359).* Menlo Park, CA; AAAI Press; MIT Press.

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 acm sigkdd workshop on human computation* (pp. 11-14).

Jain, S., & Parkes, D. C. (2013). A game-theoretic analysis of the ESP Game. *ACM Transactions on Economics and Computation (TEAC)*, 1(1), 1-35.

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2019). Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal*, 22(1-2), 188-227.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Mitrović, B. (1999). Palladio's Theory of the Classical Orders in the First Book of I Quattro Libri Dell'Architettura 1. *Architectural History*, 42, 110-140.

Pitt, M. A., & Samuel, A. G. (2006). Word length and lexical activation: Longer is better. *Journal of Experimental Psychology: Human Perception and Performance*, 32(5), 1120.

Robertson, S., Vojnovic, M., & Weber, I. (2009). Rethinking the ESP game. *In CHI'09 Extended Abstracts on Human Factors in Computing Systems* (pp. 3937-3942).

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3), 157-173.

Siu, K., Zook, A., & Riedl, M. O. (2014). Collaboration versus competition: Design and evaluation of mechanics for games with a purpose. *In FDG*.

Siu, K., & Riedl, M. O. (2016). Reward systems in human computation games. *In Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play* (pp. 266-275).

Steinmayr, B., Wieser, C., Kneißl, F., & Bry, F. (2011). Karido: A GWAP for telling artworks apart. *In 2011 16th International Conference on Computer Games (CGAMES)* (pp. 193-200). IEEE

Von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. *SIGCHI conference on Human factors in computing systems (CHI '04)*. 319-326. ACM.

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

Von Ahn, L., Ginosar, S., Kedia, M., & Blum, M. (2007). Improving image search with phetch. *In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07 (Vol. 4, pp. IV-1209).* IEEE.

Von Ahn, L., Liu, R., & Blum, M. (2006). Peekaboom: a game for locating objects in images. *In Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 55-64).