Karido: A GWAP for Telling Artworks Apart

Bartholomäus Steinmayr, Christoph Wieser,
Fabian Kneißl and François Bry
Institute for Informatics, Ludwig-Maximilian University of Munich
Oettingenstr 67, 80796 Munich, Germany
Email: bry@lmu.de

***Abstract*—Creating descriptive labels for pictures is an important task with applications in image retrieval, Web accessibility and computer vision. Automatic creation of such labels is difficult, especially for pictures of artworks. "Games With A Purpose" strive to create image descriptions by harnessing the intelligence of human players. Existing implementations are highly successful in terms of player count and number of collected labels, but hardly create comprehensive tag sets containing both general and specific labels. We propose Karido, a new image labeling game designed to collect more diverse tags. This paper describes the design and implementation of the game, along with an evaluation based on data collected during a trial deployment. The game is compared to an existing image labeling game using the same database of images. Results of this evaluation indicate that Karido collects more diverse labels and is fun to play.**

***Index Terms*—J.8.g Games, H.3.3.d Metadata, J.5.c Fine arts**

I. Introduction

Games With A Purpose (GWAPs), first proposed by Luis von Ahn [1] and based on projects such as the Open Mind Initiative [2], are a kind of *serious games*. GWAPs strive to circumvent problems which are currently difficult to solve computationally, such as optical character recognition, speech transcription, translation and semantic image analysis. The goal of GWAPs is to transform given problems into games, which –by harnessing the cognitive capabilities of their human players– collect solutions to problem instances. Two goals must be met in the design of a GWAP: Firstly, the game must be fun, to ensure sufficient player participation. Secondly, the game must ensure the collected data is correct.

As described above, semantic image analysis (or *image labeling*) is an important application of GWAPs. While a number of successful image labeling games exist (see Section II), research indicates that existing implementations do not yield comprehensive image descriptions. We propose a novel game called Karido, designed to collect diverse tag sets for arbitrary images. The game has been implemented as part of the Artigo platform (see Section IV-A), a joint project of computer science and humanities at the University of Munich.

This paper describes the motivation, design, implementation and evaluation of Karido and presents the following contributions:

- Conception of a novel game design for image labeling
- Implementation of a fully functional version of the game (publicly available at http://www.artigo.org)
- Experimental evaluation of the created game

A. Problem

The goal of image labeling is to create textual descriptions of images. Our primary goal for Karido is to create labels for image retrieval. While generic tags are used more often [3] in image retrieval queries, using highly specific tags for image queries –especially on the Web– commonly leads to inferior results. We therefore propose to design a game to collect both generic and specific tags.

The basic mechanics of existing games are designed to enforce correct labels, but not necessarily comprehensive ones. Probably the most important example for this is the ESP game, created by von Ahn and Dabbish [4]. In this game, two players are randomly paired and have to agree on a description of an image without being able to communicate. Thus, each player does not know anything about her partner and has to assume she is not an expert in a given domain. Therefore, players are more likely to achieve a match if they enter generic terms as opposed to specific ones. This has been proven using a game theoretic approach by Jain and Parkes [5] (see Section II). Thus, it is unlikely that the basic game design of the ESP game yields specific descriptions. Although measures for correcting this behavior have been proposed, research indicates that they are only partially successful (see Section II).

B. Proposed Solution

As one possible solution to the issues described above, we propose a novel game called Karido. It is designed to inherently ensure both tag validity and tag diversity. Karido is a cooperative game in which two players strive to gain as many points as possible by describing the content of images. The two players alternatively assume the roles of Guesser and Describer.

Before each round of Karido, nine similar images are randomly selected from a given database. The selection of these images is crucial to the goal of increasing tag diversity and is discussed in more detail in Section III. The selected images are displayed as regular grids to both players (see Figure 1). In the view of the Describer, a single image is highlighted (this is called the *goal*). The Describer's task is to explain the goal image to the Guesser. To achieve this, the Describer can send short textual messages to the Guesser.

The Guesser's view shows the same images as the De-

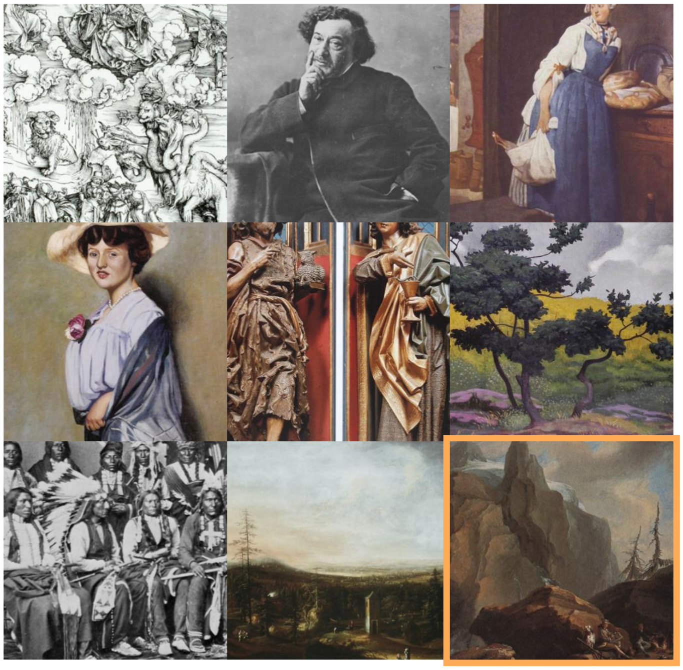


Fig. 1: Image grid for Describer (Hovering over a picture displays the whole image)

scriber's, but in randomized order[1] and without a highlighted image. The Guesser's task is to deduce the current goal image from the given description. At any time during the game, the Guesser can select an image by doubly clicking it. If this guess is correct, the image is removed from the grid of both players and both receive a score bonus. Furthermore, the text entered by the Describer is considered a valid description of the image. The Describer must then select another image, which becomes the next goal. The game proceeds in this fashion until only one image remains. As selecting this image would be trivial, the current game round is ended at this point, the players switch roles and a new round is initiated.

KARIDO falls into the class of inversion-problem games, as defined by Luis von Ahn [1]. In these games, one player (the Describer) transforms a given input (the selected goal image) into an intermediary output (the textual description). The second player (the Guesser), tries to transform the intermediary output back into the original input (by selecting the correct image). Success of the players implies that the intermediary output is a representation of the original input.

II. RELATED WORK

Arguably the most important and successful of all image labeling games is the ESP GAME[2], created by von Ahn and Dabbish [4]. The ESP GAME has been adapted by Google to improve the results of their image search engine. The game has been in productive use since 2006[3] and has been very successful in terms of numbers of collected tags and players. As described in the problem statement, the basic design of the ESP GAME tends to collect generic tags. To solve this issue, von Ahn and Dabbish propose the use of so called *taboo words*. The game keeps track of the number of times a given tag has been applied to an image. Once this number exceeds a given threshold, the tag is added to the list of taboo words for the given image. This list is shown to both players and all tags on it can no longer be used. Taboo words are meant to force players to use different and eventually more specific terms.

A. Game-theoretical analysis of ESP Game

Jain and Parkes present a game-theoretic model of the ESP GAME [5]. The goal of this model is to formalize player strategies in the ESP GAME and prove which ones are most successful. The authors analyze the most basic version of the ESP GAME in which the scores players receive for a match are independent of the matched word. Therefore, the goal of the players is to complete as many rounds as possible within the given time limit. Jain and Parkes call this *match-early preferences*. They furthermore assume no taboo words are used in the game.

The model designed by Jain and Parkes assumes that a describing set of words exists for each image (which the provider of the game is trying to learn). The words in this set possess and are ordered by a *frequency*, which defines the likelihood of a word being assigned to the given image. Jain and Parkes then proceed to prove that playing the describing terms in order of descending frequency leads to a Bayesian-Nash equilibrium for the ESP GAME. In such an equilibrium, no player can gain an advantage by changing their strategy. The authors conclude that the basic version of the ESP GAME tends to lead to common tags. The analysis of taboo words and incentive structures which would lead to more uncommon tags is left as future work.

B. Analysis of Taboo Words

Weber et al. performed an extensive evaluation [3] of the results of a version of the ESP GAME. In the data they extracted, the authors found a number of redundant and generic labels. Furthermore, many tags where highly correlated. Weber et al. therefore argue that this kind of data does not necessarily need to be created by human players. To prove their point, they implemented a software that successfully plays the ESP GAME. This is somewhat paradoxical, since the main objective of the game (i.e., labeling images) cannot yet be achieved reliably by computers.

The software created by Weber et al. disregards the visual content of the images and predicts likely tags by analyzing the taboo words. The software played over 400 games and achieved a match about 80% of the time. This means that the

[1] This randomization ensures that players cannot describe images by their position in the grid. If both players shared the same view, the content of the images would not have to be described.

[2] The name ESP GAME stems from *Extra-Sensory Perception*, a concept that describes the communication of information between humans without relying on senses, but solely on the mind.

[3] http://images.google.com/imagelabeler/help.html

tags entered by human players are highly predictable given the taboo words. Thus, human players add little information to the existing tags even when taboo words are enforced.

C. Further Image Labeling Games

Aside from the works of von Ahn et al., a number of alternative image labeling games with very different approaches have been proposed. For example, *PhotoSlap* by Ho et al. [6] translates an existing board-game into a GWAP and allows four players to cooperate. *Picture This*, by Bennett et al. [7] is designed not to label images directly, but instead improve query results using existing tags.

KissKissBan: To solve the issue of tag diversity in the ESP GAME, Ho et al. created KISSKISSBAN [8]. The fundamental difference between the two games is the introduction of a third player and a competitive element in KISSKISSBAN. The first two players are called *Couple* and try to achieve the same goal as in the ESP GAME. The third player in KISSKISSBAN is called the *Blocker* and is competing with the Couple players. Before each round, the Blocker has seven seconds to enter as many words as possible which the Couple players are not allowed to use. In contrast to the taboo words in the ESP GAME, the Couple players cannot see this list of words. If a player enters a blocked word, five seconds are subtracted from their remaining time. If the timer runs out before the Couple players achieve a match, their scores are decreased and the Blocker's score is increased. If the Couple players succeed, the opposite is the case.

Phetch: Von Ahn et al. argue that the labels created by the ESP GAME are very well suited for image retrieval, but do not allow humans to form an accurate mental model of the described image. To solve this problem, von Ahn et al. propose PHETCH, an image labeling game targeted at collecting entire sentences describing an image [9]. PHETCH is designed to be played by three to five players. One of the players is randomly selected as the *Describer*, whereas the other players form the group of *Seekers*. A picture is shown to the Describer, who can enter arbitrary sentences to describe the image to the Seekers. The Seekers must then use a special search engine to locate the described image. If a Seeker selects the correct image from her list of results, she and the Describer are awarded a score bonus. To discourage random guessing, points are deducted whenever a player ventures a wrong guess. After the correct image has been found, the Seeker who found it becomes the Describer in the next round.

Peekaboom: PEEKABOOM was created by von Ahn et al. to collect tags for images, along with the regions of the respective objects depicted in an image [10]. Like in most other Games With A Purpose, PEEKABOOM matches two random players for each game session. An image is shown to the first player, along with a description of an object in the image. These descriptions have been generated using the ESP GAME. The first player can successively reveal parts of the image to their partner. The second player can only see the areas of the image that were revealed to her. Her goal is to guess the term that was shown to the first player. If this goal is achieved, both players are awarded a score bonus and a new round is initiated.

III. DESIGN OF KARIDO

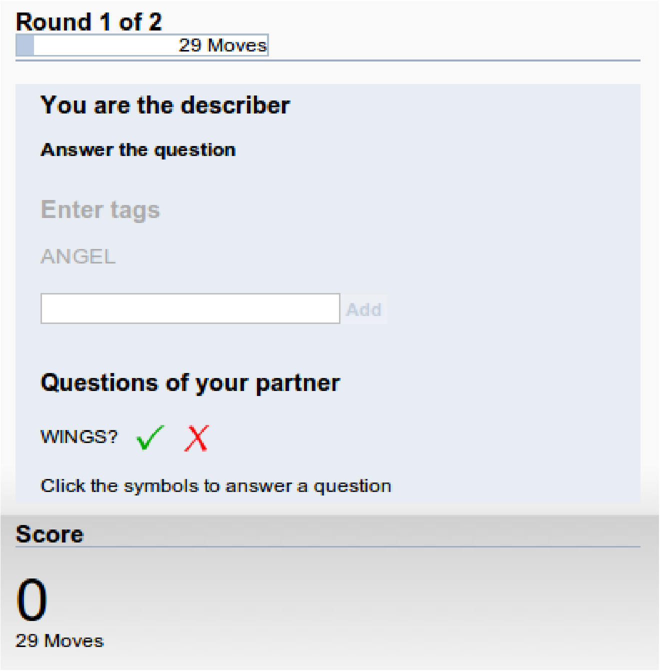
The primary goal of KARIDO is to collect more diverse tags than previous image labeling games, which –as discussed above– struggle to create comprehensive tag sets. The key element for achieving this goal is what we call "input-similarity". By selecting similar images to display, the players of KARIDO are forced to find differentiating properties of the images. Therefore, without the use of explicitly restricting methods such as taboo words, players must contribute new information. For example, consider a grid in which each image contains only a unique single color. In this case, the Describer can use the color to clearly describe the goal. In contrast, given a grid in which all images contain a red car, the Describer can no longer use "red" or "car" and thus has to find characteristic traits of the goal image that differentiate this image from all others.

KARIDO relies on player-created tags as a measure of image similarity. After selecting a random base image, the game calculates the number of tags shared with this base image for all other images in the database. This selection method implies that for a new set of images without any tags, all images are considered equally similar. In this case, the images in the grid are selected randomly and can be expected to be relatively diverse. As a result, players can use general tags to distinguish the images. As a result of these tags, the images now possess different similarity ratings. For example, all photographies in a collection could be tagged "photo", whereas paintings could be tagged "painting". Thus, the players will be given grids which contain either only photographies or paintings. Therefore, the attributes "photo" and "painting" can no longer be used to distinguish the images and the Describers are forced to find new properties of the images.

If any two images have no distinguishing tags, they will necessarily both be selected for a game round at some point. One of the images will become the goal image, while the second image is still in the grid. Thus, the players either have to come up with a label that distinguishes these two images or use random guessing. As the scoring of the game is designed to make random guessing a poor strategy (see below), a player who is able to find a distinguishing feature is likely to use this trait to describe the image. Thus, the process of refinement of the labels continues until all images possess a set of tags that sets them apart from all other images in the game.

A. Player Communication

Most GWAPs rely on player agreement to verify that the entered data is correct. It is therefore essential that players do not have the possibility of artificially creating an agreement. The most common way of creating such an agreement is by using a communication channel outside of the game. However, KARIDO does not produce matches on entered tags, but requires selection of an image. Therefore, to circumvent the verification method, players would need be able to communicate the content of the goal image directly (for example, by using a screen-sharing software). This in turn implies that arbitrary textual communication inside the game could be allowed. However, to ensure proper keywords for the images are collected (as opposed to free-form texts), descriptions in Karido are restricted to a maximum of three words and all punctuation is removed.



Round 1 of 2

29 Moves

You are the describer

Answer the question

Enter tags

ANGEL

Questions of your partner

WINGS? ✓ ✗

Click the symbols to answer a question

Score

0

29 Moves

Fig. 2: User interface of Karido for the Describer

To allow the Guesser to help the Describer, Karido allows posing Yes/No-questions, which the Describer can answer by clicking on an icon instead of typing (see Figure 2). In addition to aiding the Describer, these Yes/No-questions enable the Guesser to actively take part in the labeling process: Whenever a question is answered with "Yes", the question is added as a label to the description of the current goal image. If a question is answered with "No", a negative label could be attached to the current image. However, because of the questionable informative value of negative labels (the list of things *not* depicted in any given image is nearly endless), no tags are added for questions answered with "No".

B. Game modes

To make the scores given to players comparable and to provide a goal for the players to achieve, the duration of a game session must be limited. There are several possibilities for limiting the duration of game sessions and game rounds. Limiting the duration of an entire game session is the most commonly used approach for GWAPs. It is also possible to use the inverse approach of limiting the number and duration of the rounds in a game session. This approach is used in Karido. Each game session consists of two rounds. While this seems very little in comparison to the ESP Game, each round in Karido contains nine images and generally lasts longer than a round in other games.

The most straight-forward metric for "duration" is time. Therefore, the first game mode of Karido limits each round to 90 seconds. A second game mode relies on the number of actions performed by the players to limit the duration of game rounds. Sending tags and questions and performing guesses are the atomic actions of the game. By limiting the number of allowed actions, the players can take as much time as they want for considering, typing and submitting their labels. This means that success in the game only depends on the quality of the labels, making the number of actions a reasonable alternative to the metric of time. However, whereas time is always shared equally between the players of a game, more actions are "consumed" by players who take less time. Thus, a quick player can acquire a disproportionate share of the available actions. This introduces a potential competition between the players, which goes against the cooperative nature of Karido.

To enforce an equal distribution of the actions between the players in the second mode of Karido, players are forced to take turns to describe images, send and answer questions and venture guesses. However, there are scenarios in which a rigid succession of player turns can be a problem. Occasionally, only two images are potential candidates for the Guesser. Thus, if the Guesser takes her turn and guesses wrong, the correct solution is immediately obvious. Using a strict turn-based approach, the players would have to waste another action (the Describer's turn, which has become superfluous) and the Guesser would have to wait her turn until she could perform the now trivial action of selecting the right image. To avoid such situations, the Guesser in Karido can take a guess at any point in the game. While this approach allows the guesser to take a disproportionate share of the available actions, we consider it an acceptable trade-off for an improved game flow.

C. Scoring

In other GWAPs such as the ESP Game, both players enter their answers in form of textual labels. Therefore, the set of possible answers contains all character sequences and thus millions of items. Thus, the probability of a player being successful by chance or guessing is very low. In contrast, the Guesser in Karido can only select from a given set of images. At most, this set comprises nine images (at the beginning of a game round). It would therefore be easy for a player to try all possible answers. This would eliminate the verification of the entered tags, as a player relying solely on guessing can ignore the description given by her partner. It is therefore necessary to take measures that discourage random guessing.

In Karido, the score of both players is reduced as a penalty for selecting a wrong image. This penalty exceeds the bonus for selecting the correct image. Therefore, the expected value of the score for a player relying on guessing is always negative. However, malicious users could still cooperate to introduce wrong data into the system by entering wrong descriptions and using random guessing to verify the entered data.

D. Data Verification

To ensure that the collected descriptions are valid, KARIDO assigns a real-valued relevance score to each pair of an image and a tag. When a tag is first applied to an image, this score is initialized with zero. Once a player correctly guesses a goal image, the score of all tags in the description leading to the correct guess is increased. To avoid gathering wrong data, the amount by which the scores are increased depends on the number of wrong guesses performed before the selection of the goal. If the Guesser tried more than 30% of the images before selecting the right one, all tags are considered irrelevant. For example, at the beginning of a game (with nine images remaining), the Guesser must be right on the third try or the labels will be discarded. The full score increase is only assigned to a tag if the Guesser is right on the first try. Otherwise, the increase is interpolated linearly between 1 and 0 depending on the percentage of wrongly selected images (from 0% to 30%).

In contrast to the ESP GAME, KARIDO can collect several labels for an image in one round of the game. The Describer usually enters several tags before the Guesser correctly selects an image. Thus, the labels entered first were not sufficiently precise to enable the Guesser to select the correct image. It is therefore reasonable to assume that the labels assigned later are more relevant to the image. KARIDO uses a weighting scheme to take this increasing relevance into account. In addition to a base weight of 0.5, an additional weight of 0.5 is distributed over all tags in a linearly increasing fashion. The result is multiplied with the factor calculated above and added to the existing verification score.

IV. IMPLEMENTATION

The implementation of KARIDO is based on the common platform provided by the ARTIGO project (see below). This platform is based on the SEAM[4] Web framework. In KARIDO, players need to be able to communicate with each other and perform actions which need to be propagated to their partner. Especially in the time-limited game mode, actions and sent descriptions and questions often follow in rapid success and need to be relayed with as little delay as possible. SEAM does not provide mechanisms to support such real-time communication between multiple users. Therefore, an additional framework based on SEAM was created during the implementation of KARIDO. It coordinates the queuing and matching of all players and the synchronization between the players participating in a game. The framework is kept modular and can easily be used for many different kinds of multi-player online games.

An additional issue arises from the Web-based nature of KARIDO. The Hypertext Transfer Protocol used on the Web does not allow the server to notify the client of any changes (such as the arrival of a new description). There are experimental approaches to allow Web servers to notify clients, but at the time of writing, there is no standard method for achieving this. Therefore, the KARIDO client sends light-weight requests to the server in regular intervals and is then informed of any changes.

A. Artigo Platform

From the very beginning, KARIDO was planned to be based on the so-called ARTIGO 4.0[5] platform. The name ARTIGO references both the platform as well as a GWAP designed by a team around Hubertus Kohle at the University of Munich. The ARTIGO game uses the same mechanics as the ESP GAME, but specifically aims to label historical artworks. This is an important task, as large databases of artworks exist but often lack sufficient meta-information to allow efficient image retrieval. ARTIGO has been publicly available online for over four years and has collected over three million tags.

ARTIGO 4.0 is a complete reimplementation of the game and also provides a common framework and platform to develop and deploy other GWAPs. KARIDO is the first such game that has been added alongside the ARTIGO image labeling game. In addition to sharing a common technological base (thus increasing maintainability and decreasing code size), ARTIGO and KARIDO are accessible through the same Web site and share a common image database. This has several benefits. Firstly, the users of the existing game are directly made aware of the new game, thus allowing the latter to quickly gain users (for example, for early evaluation). Additionally, the shared database allows the direct comparison of the results of both games. Finally, the data collected in the first game can be used to bootstrap the simulated players of the second game (see below).

B. Simulated Player

Although most GWAPs are advertised as multi-player games, many implementations [4], [9]–[11] include simulated players (or *bots*), which enable single (human) players to participate in the game. There are two reasons for supporting single-player operation. Firstly, the number of players in the game can be uneven, leading to players who cannot be paired with a partner. Additionally, a single person may be the only player of the game at any given time and thus would have to wait for a second player to join the game. The bot of KARIDO must fulfill two different roles in the game, Describer and Guesser.

The Describer's task is to explain the current goal image and answer the Guesser's questions. The first task can be achieved by replaying previous rounds. The tags of the goal image are sent using the same delays and in the same order as in the recorded round. Once the right image has been selected, the simulated Describer chooses a new image. To ensure that all tags are replayed in the proper context, the simulated Describer selects the same sequence of goal images that was played in the previous round. If a human Guesser takes longer than the Guesser in the original round to select the correct image, there are no further tags that can be replayed. In this case,

[4] http://www.seamframework.org

[5] http://www.artigo.org

the simulated Describer selects random tags and sends them with randomized delays. In addition to sending descriptions, a simulated Describer must be able to answer questions posed by the human Guesser. To achieve this, the bot relies on the labels already assigned to the current goal image. If the content of a question has ever been used to describe an image, the question is answered positively. If no label with the content of the question exists, the question is answered negatively.

Like the Describer bot, the Guesser bot in KARIDO must fulfill two tasks: Firstly, it must interpret the descriptions given by the Describer and select an image once it is reasonably certain that it is the goal image. Secondly, if several potential images remain, it should ask questions to reduce the number of candidate images. For both tasks, the bot relies on the tags previously assigned to the images. As described in Section III-D, the behavior of the simulated Guesser does not directly influence the quality of the data collected in the game, as all tags are validated by an independent player. However, if players discover that they can score points as Describers without entering valid descriptions, no more valid information will be entered into the system and the verified information will stagnate. It is therefore necessary to simulate a Guesser that only selects the goal image if it has been described accurately.

As a first step, the percentage of entered tags assigned to the images in the grid is calculated. This value will from now on be called the *match percentage* of an image. For example, if the tags “car” and “dog” have been entered, an image tagged “car” has a match of 50%, an image tagged “dog”, “house”, “car” has a match of 100%. All images with a match percentage larger than zero (and which have not yet been wrongly guessed) are sorted by decreasing match percentage. The resulting list of images is used to decide which image to select. The first image from the list is selected if the list contains only one image, contains only one image with a match of 100% or if its match percentage is at least 1.2 times larger than the one of the next image. Additionally, the first image is selected if a sufficient number of descriptions have already been sent. This ensures that the Guesser bot starts to guess randomly if the Describer fails to refine the description.

If none of these criteria are applicable, the simulated Guesser prepares to ask a question. All images which have at least half the match percentage of the first image are used to ask a question. The simulated Guesser selects the first image from this list of candidates. From all tags assigned to this image, those which have been applied to any of the remaining candidate images are discarded. The remaining tags are unique for the selected image. One tag is randomly selected from the three most common tags of this list and sent to the Describer as a question. This approach yields questions which are well suited to discriminate the candidate images. At the same time, the selected questions tend to be obscure.

Lastly, answered questions must be used to reduce the number of candidate images. All questions which were answered positively are treated like tags created by the Describer. In contrast, questions which were answered negatively are added to a list of blocked tags. For each candidate image, the number of blocked tags assigned to this image is calculated. This value is subtracted from the number of matching tags used for calculating the match percentage of the images.

A problem arises if the current goal image has not yet been tagged a reasonable number of times. In this case, whether a tag entered by a human player is valid cannot be determined and therefore the Guesser bot must make a trade-off. Either unknown tags are accepted with a high probability, at the risk of inciting players to enter wrong tags, or unknown tags are rejected with a high probability, effectively punishing players for entering new tags. As any data collected by the game is independently verified, we have decided to choose the first alternative in KARIDO.

V. CASE STUDY

To assess whether KARIDO fulfills its goal of collecting more diverse tag sets, an empirical evaluation was performed. On April 13th, 2011, the ARTIGO 4.0 platform replaced the previous version of the ARTIGO game. KARIDO was released alongside the new version of the ARTIGO game. For the first three weeks after the release, no public announcement of the update was made. Therefore, the users of the game primarily consist of regular players of the old ARTIGO game, returning to the Web site. On May 3rd, 2011, announcements of the newly released games were made on several channels (among others, the Web site of the University of Munich). Another three weeks later, a snapshot of the database of the games was taken on May 24th, 2011. This snapshot has been used to gather the results described below.

During the evaluation period, 664 players completed at least one round of the KARIDO. The participants played a total of 1939 game sessions, consisting of 3542 rounds. In the same timespan, 3766 sessions of ARTIGO were played. The higher popularity is likely caused by the large number of regular players of ARTIGO participating in the evaluation.

A. Results

An analysis of the number of played rounds for individual players reveals that the majority of players completed only two rounds. This number indicates that players do not enjoy playing the game. However, this contradicts the high satisfaction scores of the game (see below). An alternative explanation can be found in the way players are identified by the game. Players may choose to play the game without registering. In this case, only the number of played rounds during one visit to the Web site can be counted. This number must necessarily be less than or equal to the total number of played rounds of a player. The number of registered players is too small to draw reliable conclusions, however, the distribution of played rounds of registered players appears to be more even than the distribution of played rounds of all players.

To measure the subjective enjoyment players gain from KARIDO, a rating mechanism was added to the user interface. After each completed game session, players are invited to “rate this game session”. To submit their rating, players can select from a scale of five stars. This scheme is commonly used to express ratings from dislike (one star) to approval (five stars). A total of 1051 ratings were created. As described above, 1939 game sessions were played by a human and a simulated player, implying at most one rating per game session. Thus, a rating was submitted for 54% of all sessions.

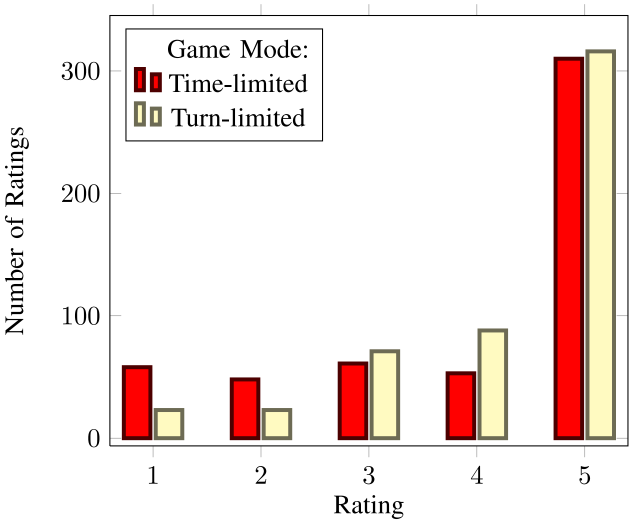


Fig. 3: Histogram of the game session ratings submitted during the evaluation.

Figure 3 contains a histogram of the submitted ratings for both modes of KARIDO. In contrast to the low average number of played rounds (indicating low acceptance of the game), the submitted ratings show a strong peak for the maximum rating of 5. The turn-limited mode is slightly more popular, with an average rating of 4.2 (as opposed to 4.0 for the time-limited mode). It should be noted that the design of the rating interface might introduce a bias: As described before, the rating interface is displayed after each completed game session. Therefore, players who do not complete a session do not get a chance to vote. Additionally, players who dislike the game and stop playing only get to vote once, whereas players who like the game and play many sessions can vote more often. To eliminate the influence of multiple votes, the collected ratings are averaged in a grouped fashion. First, the average rating for each player is calculated. Then, the average of these ratings is calculated, attributing an equal weight to the votes of each player. This results in an average score of 4.0 for the time-limited mode and 4.4 for the turn-limited mode. As returning players cannot always be identified correctly, a bias might remain in the grouped data. A comparison between KARIDO and ARTIGO is not possible, because no game session ratings were collected for the latter.

One problem that became apparent during the evaluation is the player matching mechanism. If a human player has not been matched with another partner after ten seconds, she is automatically matched with a simulated player. Even with a relatively large number of concurrent players, the probability of two players being matched in this ten second window is quite low. As a result, all rounds in the evaluation were played by a human and a simulated player. However, the high satisfaction ratings of the game indicate that the simulated player –while certainly no replacement for an actual human– is adequate at the very least.

The primary goal of KARIDO is to collect more diverse sets of tags. To evaluate whether this goal was reached, we consider the average number of unique tags created per game round. We define a tag to be unique if it has not been previously assigned to any image. During the six week evaluation period the players created 6933 unique tags in KARIDO and 16635 unique tags in ARTIGO. This corresponds to 2.0 (KARIDO) and 1.2 unique tags per round respectively. Thus, KARIDO arguably leads to a higher rate of unique tags per round. However, this rate can be expected to decrease when more tags are submitted, thus favoring KARIDO (because of its lower number of tags). To compensate for this, we have also calculated the average number of unique tags during the first three weeks of the trial. It amounts to 2.1 (an increase of 5%) for KARIDO and 1.6 (an increase of 33%) for ARTIGO. This indicates that KARIDO indeed collects more unique tags and thus more diverse tag sets.

The evaluation has shown an issue concerning the verification of entered tags. Of 11372 tags (unique per image), only 2364 posses a verification score larger than 1.0. This behavior is to be expected: Usually, each player alternates between creating tags and verifying tags. Thus, as long as the creation rate of unique tags remains high, these tags are –on average– only verified once. See the next Section for our proposed solution.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have presented a novel Game With A Purpose to label images. Our evaluation has shown strong indication that this game is both fun to play and succeeds in collecting the desired data. Nevertheless, several tasks remain to be completed in the future.

With its short round duration and relatively simple game mechanics, KARIDO would be well suited for mobile devices such as Smartphones or Tablets. However, the lack of a keyboard makes text entry tedious on these devices. Therefore, the role of the Describer cannot be played properly. To solve this issue, as well as the issue of slow verification of the entered tags (see above), a new game mode could be created specifically for mobile devices. In this mode, the player would always perform the role of the Guesser and verify replayed tags. Additionally, questions should be disabled, thus requiring no text entry whatsoever from the player. Finally, a different user interface should be devised to allocate as much screen space as possible to the pictures.

The verification scheme of KARIDO implies that tags with a score exceeding a given threshold can be considered reliable. However, there is currently no way to distinguish between tags which have a low score because they are wrong and tags which have not yet been validated. The game should not continue to validate tags which seldom or never lead to correct guessing, in order to avoid annoying the Guessers. Similarly, tags which have been verified to be correct should also be verified no longer, to ensure that new tags can be verified quickly. The data model of KARIDO could be enhanced to keep track of the number of times a given tag has been replayed. The verification could then be stopped after a tag has been replayed a certain number of times. However, this would enable malicious players to remove tags from the verification pool, by deliberately entering wrong tags. Alternative schemes could use median values of the number of times all tags have been replayed to balance the verification process.

While KARIDO ensures that its players provide new labels, it does not impose restrictions on the form in which these labels are applied. In a grid in which all images share a common property, both players are likely to recognize this common theme. Therefore, describing the common element of the images does not increase the chance of the Guesser selecting the right image. As a result, the Describer will only send additional information, but the semantic connection between the common tags and the added tag is lost. This loss of information can be a problem, for example in the case of image retrieval. To collect the semantic connections between tags, an additional version of the game could be devised. The basic layout of KARIDO remains unmodified. In contrast to the basic design of the game, the Describer cannot enter tags directly. Instead, the Describer is given a list of all labels assigned to the current goal image. To describe the image, the Describer must select a pair of labels by clicking them. This pair of tags is sent to the Guesser and an association between the tags is stored.

Lastly, while the evaluation of KARIDO indicates that the game collects the desired data, the sample size collected during the six week trial is small. To draw statistically significant conclusions and especially to evaluate the quality of the collected tags, more extensive studies are necessary. It should furthermore be ensured that players in the evaluation are matched with other human players. Finally, controlled lab experiments as performed by social sciences might yield interesting results.

ACKNOWLEDGEMENTS

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG) within the project Play4Science number 578416 (cf. http://www.play4science.org).

REFERENCES

[1] L. von Ahn and L. Dabbish, “Designing games with a purpose,” *Communications of the ACM*, vol. 51, no. 8, pp. 58–67, 2008.

[2] D. G. Stork, “Open data collection for training intelligent software in the Open Mind Initiative,” *IEEE Expert Systems and Their Applications*, vol. 14, pp. 16–20, 2000.

[3] I. Weber, S. Robertson, and M. Vojnović, “Rethinking the ESP game,” in *Proc. of 27th intl. conf. on Human factors in Computing Systems*, ser. CHI ’09. New York, NY, USA: ACM, 2009, pp. 3937–3942.

[4] L. von Ahn and L. Dabbish, “Labeling images with a computer game,” in *CHI ’04: Proc. of SIGCHI conf. on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2004, pp. 319–326.

[5] S. Jain and D. Parkes, “A Game-Theoretic Analysis of Games with a Purpose,” in *Internet and Network Economics*, ser. LNCS. Berlin, Heidelberg: Springer Berlin / Heidelberg, 2008, vol. 5385, ch. 40, pp. 342–350.

[6] C. J. Ho, T. H. Chang, and J. Y. J. Hsu, “PhotoSlap: a multi-player online game for semantic annotation,” in *Proc. of 22nd nat. conf. on Artificial Intelligence*. AAAI Press, 2007, pp. 1359–1364.

[7] P. N. Bennett, D. M. Chickering, and A. Mityagin, “Picture this: preferences for image search,” in *Proc. of ACM SIGKDD Workshop on Human Computation*, ser. HCOMP ’09. New York, NY, USA: ACM, 2009, pp. 25–26.

[8] C. J. Ho, T. H. Chang, J. C. Lee, J. Y. jen Hsu, and K. T. Chen, “KissKissBan: a competitive human computation game for image annotation,” in *Proc. of ACM SIGKDD Workshop on Human Computation*, ser. HCOMP ’09. New York, NY, USA: ACM, 2009, pp. 11–14.

[9] L. von Ahn, S. Ginosar, M. Kedia, R. Liu, and M. Blum, “Improving accessibility of the web with a computer game,” in *CHI ’06: Proc. of SIGCHI conf. on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2006, pp. 79–82.

[10] L. von Ahn, R. Liu, and M. Blum, “Peekaboom: A Game for Locating Objects in Images,” in *Proc. of SIGCHI conf. on Human Factors in Computing Systems*. ACM, 2006, pp. 55–64.

[11] L. von Ahn, M. Kedia, and M. Blum, “Verbosity: a game for collecting common-sense facts,” in *In Proc. of ACM CHI 2006 Conf. on Human Factors in Computing Systems*, 2006, pp. 75–78.

Bartholomäus Steinmayr has recently finished his diploma thesis in computer science at the Ludwig-Maximilian University of Munich. KARIDO has been designed and developed as the subject of this thesis.

Christoph Wieser graduated in computer science at the Ludwig-Maximilian University of Munich in 2006. Afterwards, he worked as a Researcher and Developer for Salzburg Research. Currently, he is teaching assistant at University of Munich. Christoph Wieser is interested in Social Software, Games with a Purpose and Higher-Order SVD.

Fabian Kneißl is working at the Institute for Informatics of the Ludwig-Maximilian University of Munich as a doctoral student. His research is concerned with games with a purpose together with the humanities. In 2010, he graduated with a Dipl.-Inf. in computer science from this institute and furthermore from the elite graduate program “Technology Management” from the Center for Digital Technology and Management in Munich.

François Bry is a full professor at the Institute for Informatics of the Ludwig-Maximilian University of Munich, Germany, heading the research group for programming and modeling languages. He is currently investigating methods and applications related to querying answering and reasoning on the Web and social semantic Software and Media. Before joining University of Munich in 1994, he worked in industry in France and Germany, in particular with the research center ECRC.