Odd Leaf Out

Improving visual recognition with games

Derek Hansen, Darcy Lewis, Dana Rotman, Jennifer Preece
School of Information
University of Maryland
College Park, United States

David Jacobs, Arijit Biswas
Department of Computer Science
University of Maryland
College Park, United States

***Abstract*—A growing number of projects are solving complex computational and scientific tasks by soliciting human feedback through games. Many games with a purpose focus on generating textual tags for images. In contrast, we introduce a new game, Odd Leaf Out, which provides players with an enjoyable and educational game that serves the purpose of identifying misclassification errors in a large database of labeled leaf images. The game uses a novel mechanism to solicit useful information from players' incorrect answers. A study of 165 players showed that game data can be used to identify mislabeled leaves much more quickly than would have been possible using a computer vision algorithm alone. Domain novices and experts were equally good at identifying mislabeled images, although domain experts enjoyed the game more. We discuss the successes and challenges of this new game, which can be applied to other domains with labeled image datasets.**

Keywords-games with a purpose, computer vision, error detection, leaf identification

I. INTRODUCTION

A growing number of scientific projects use images that are created and curated through crowdsourcing. Flickr users submit images of rare species to the Encyclopedia of Life (EOL); space enthusiasts classify Hubble images of galaxies at Galaxy Zoo; and citizen scientists use mobile apps to submit species' photos to online conservation projects such as Project Noah and iNaturalist. Having volunteer enthusiasts collect and classify images helps to tap into enormous reserves of potential human power [1]. It also inevitably introduces classification errors into the underlying datasets, a problem evident in even expertly curated image datasets. Catching a handful of misclassified images in a large corpus of data can be a tedious, time-intensive, and costly process. Automated image-classification algorithms can help catch some of the most egregious errors, but fail to capture more subtle errors that human expertise can uncover. Unfortunately, image classification by humans introduces its own set of problems. In this paper we propose a novel game, Odd Leaf Out, which melds human expertise with computer vision algorithms to help identify misclassified images.

Odd Leaf Out was inspired by other "games with a purpose" (GWAP) [2] that make laborious tasks enjoyable by recasting them as games. In this case, the laborious task is identifying misclassified images. We have constructed an initial dataset of leaf images tagged with their associated plant species, and intend to create a much larger leaf dataset using experts and citizen scientists. The dataset will be used by [blank for blind review], an open-access and mobile image leaf recognition system used by lay-people for species discovery. Assuring the accuracy of the dataset is essential to the project's success.

While the Odd Leaf Out game is specifically focused on identifying leaf image classification errors, its novel game design can be used to find errors in other image datasets used in various scientific endeavors. For example, it could help find errors in image corpora of other biological species (butterflies, bacteria), astronomical phenomena (galaxies, stars), human faces and emotions, and even abstract visual representation of other scientific phenomena (protein structures).

The major contributions of this paper include the novel game mechanics of Odd Leaf Out, an evaluation of its enjoyability, and an assessment of its effectiveness for identifying errors in a dataset. We begin by reviewing the literature on games with a purpose and visual recognition algorithms. Next, we explain the Odd Leaf Out game mechanics and goals, followed by a description of our evaluation methods and results. Finally, we discuss the implications and future possibilities that our work inspires.

II. PREVIOUS WORK

A. Computer Vision

The availability of Internet resources has fueled the rapid accumulation of large datasets of labeled images for use in computer vision. For example, the LabelMe [3] project provides web-based tools for labeling objects and regions in images, along with over 10,000 labeled images. Researchers have used Amazon's Mechanical Turk (MTurk) to acquire over 10,000 labels that describe attributes of face images, such as gender, ethnicity, or facial expression [4]. These are used to build attribute detectors for a face search engine. As [5] note, annotation errors are inevitable in large-scale annotation projects. They suggest some strategies for dealing with these problems, such as having MTurk workers annotate some images with known, ground truth annotations, to assess the accuracy of individual workers. However this does not completely remove mislabeling due to random mistakes (clicking on the wrong button by accident), lapses in judgment, or uncertainty in difficult cases. They also suggest

This work is sponsored by a grant from the National Science Foundation

obtaining redundant annotations, for example, producing annotations based on the two out of three annotators that are most consistent, as a potential solution. However, this can triple the effort required for annotation. Games that can motivate volunteer efforts via intrinsic motivations, such as fun and gratification, rather than extrinsic motivations, such as payment, could enable much larger datasets to be generated and validated.

In its current implementation, the Odd Leaf Out game improves an image dataset of leaves that is used by computer vision algorithms to automatically identify a leaf's associated plant species. Mokhtarian and Abbasi [6] were among the first to explore this domain, applying a general shape-matching algorithm – an approach that continues to be used today. They worked on the problem of classifying images of chrysanthemum leaves, focusing on features based on the curvature of the silhouette, extracted at multiple scales. More recently, [7] and [8] have obtained high accuracy in plant species identification by using general shape matching algorithms. We make use of the algorithms and labeled images used in the LeafSnap system for tree species identification [9], which builds on the work of [10]. The algorithms we adopt from the LeafSnap system compute shape similarity by comparing a multiscale histogram of the curvature of the boundary of the leaf [11].

B. *Games with a Purpose*

The last ten years have seen a proliferation of games such as ESP [12] (the inspiration for Google Image Labeler), Peekaboom [13], TagATune [14] and KissKissBan [15] that provide an enjoyable platform for attracting human input to collectively solve computational problems. von Ahn calls these online activities "games with a purpose (GWAP)" [16].

Many GWAP have the goal of labeling multimedia content such as images [12] [15], audio clips [14], and videos to improve search retrieval [17]. Similarly, citizen science projects like Galaxy Zoo recruit users to classify images of galaxies into preexisting types. Although it is best understood as a volunteer microcontribution site, Galaxy Zoo has introduced social networking features and competitive game-like elements such as leaderboards to increase users' enjoyment while labeling images [18]. Another game, Peekaboom [13] motivates participants to trace the outline of objects in an image, in order to tag sections of the image and create image corpora useful for computer vision algorithms. The novel purpose of the Odd Leaf Out game is finding misclassification errors in an image corpus, a goal that has widespread application, yet no existing games that directly address it. Additionally, Odd Leaf Out could be used to combine initial computer vision identifications with human input, thereby creating the potential for a system that allows for crowdsourcing image identification errors that are best detected by human eyes.

Despite the excitement around games with a purpose, they face several challenges. Perhaps most importantly, such games will fail if players can devise game strategies that improve their game score while producing low quality data. Another challenge is assuring the games are enjoyable enough to elicit play. von Ahn and Dabbish suggest that this can be done by keeping score (specifically, by harshly penalizing incorrect answers through a scoring system that gives increasing points to players when they provide a series of correct answers), showing leaderboards, introducing elements of randomness, advancing through increasingly difficult stages, and including social elements [19]. While these elements have worked for some types of games, particularly those aimed at labeling images, there are many other game mechanics that may be more appropriate for meeting other computational goals or soliciting play in different contexts. For example, most GWAP are based on having agreement between 2 synchronous players. However, many individual games such as Solitaire and Tetris are extremely popular and engaging, suggesting that other models may be fruitfully explored. We describe some of the unique game mechanics of Odd Leaf Out, such as being a single player game, in the following section.

III. ODD LEAF OUT

A. *Game Mechanics*

The purpose of Odd Leaf Out is to provide a platform for non-experts to identify potential errors in datasets consisting of images. Though the game could be played with any set of classified images or visualizations, we will describe our implementation, which uses leaf images and their corresponding plant species.

Odd Leaf Out players are presented with a set of six images in a single round, as seen in Fig. 1. Their task is to select the image that does not belong (i.e., is from a different plant species), which we call the "Odd Leaf".

Once the user clicks on an image they believe is the "Odd Leaf," the game indicates if they are correct (by highlighting the chosen image in green) or incorrect (by outlining the chosen image in red, with the correct image in green), as illustrated in Fig. 2. The player is presented with the next set of images after a two second delay, during which the common name of the plant species is displayed in English. Game play continues until a certain number of incorrect answers are made. In our case, we give each player 3 Lives (i.e., misses) that are represented as leaves in the upper-left-hand corner (Figs. 1 and

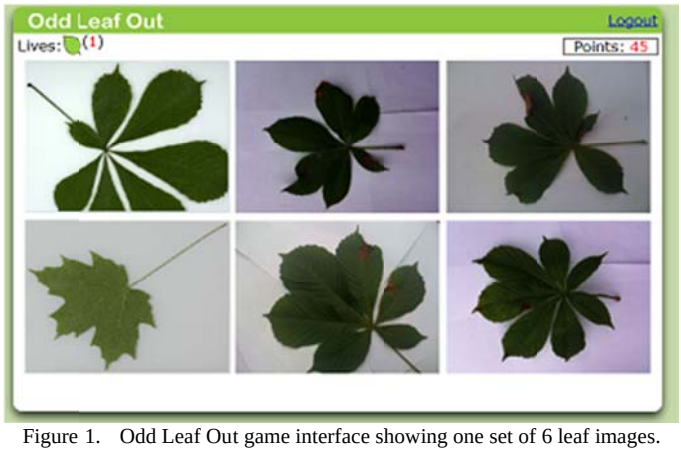

Figure 1. Odd Leaf Out game interface showing one set of 6 leaf images. Players must select the image that is from a different species than the other five images (i.e., the Odd Leaf).

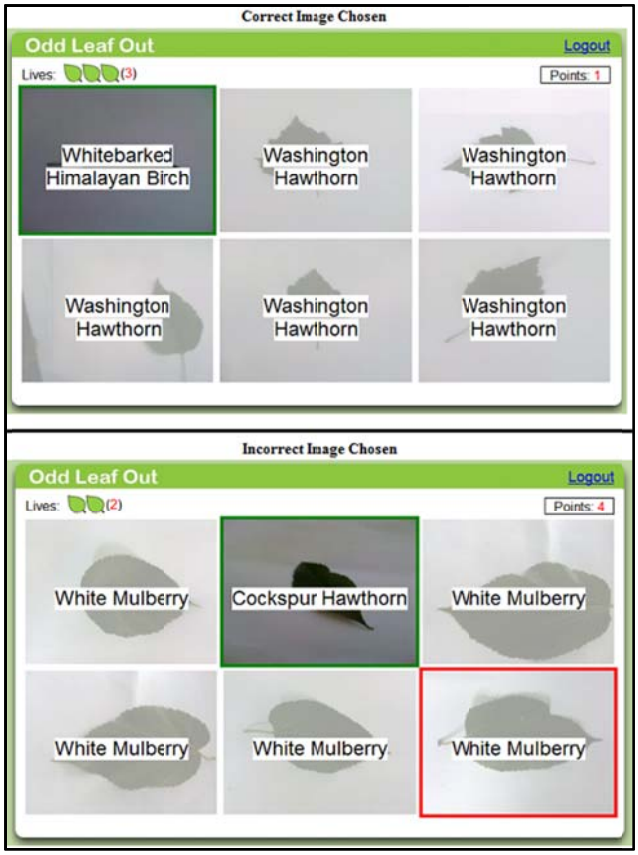

Figure 2. Odd Leaf Out game after correctly selecting the Odd Leaf (top set of images) and incorrectly selecting the Odd Leaf (bottom set of images

2). Points are awarded when correct answers are made, with increasing points for consecutive correct answers, similar to [19]. The object is to gain as many points as possible.

A leaderboard of high scores is available to motivate players. In addition, players are able to revisit each round after the game ends to see which types of leaves they misidentified. Links to the Encyclopedia of Life species pages for each tree are provided to support further self-directed learning. Providing these educational hooks is not common in other GWAP, although they are important to motivating participation in citizen science projects such as Galaxy Zoo [18]. Although the game is presented in English, the basic Odd Leaf Out game mechanics are language independent, making it possible to create similar games for large international audiences.

The decision to make this a single-player game solves some problems while raising others. Single-player games lose some of the thrill of the dynamic interaction that games like ESP provide, where players enjoy guessing what their partner is thinking. However, it avoids problems that arise from the need to match up players in real-time (or the play-against-a-computer alternative). It also avoids collusion strategies that help players score highly, while producing unhelpful data. The success of single-player games like Minesweeper and Solitaire, and lately Angry Birds and Fruit Ninja, suggest that they can be highly engaging despite the lack of a social component.

Finally, Odd Leaf Out offers a new technique for soliciting useful data. Rather than learning from the agreement of multiple parties or from the accurate classification of items, it learns from people's mistakes. We begin with images that have accompanying metadata that identifies them as being a particular image type (i.e., tree species). The initial identifications can be from citizen scientists or from best-matches recommended by a computer vision algorithm. The game's potential lies in its ability to help efficiently find cases where the preliminary identifications are likely to be incorrect. As players make systematic mistakes, they in turn help identify errors in the data. For example, if there is a misidentified leaf among the same-species leaves, players are likely to select it as the Odd Leaf more often than expected. Alternatively, if the Odd Leaf is erroneously the same species as the other five leaves, players are not likely to choose it often.

The challenge of allowing a game to be incorrect is that players can get frustrated when they are penalized for encountering an obvious error. There are two primary strategies to limit the potential frustration of players. First, allow players to provide input on mistakes or avoid them when they occur. For example, we introduce a game variation that allows players to skip a difficult or erroneous set, which we discuss in more detail in the following section. Second, game images can be introduced in such a way that not too many errors will be encountered during play. For example, sets that have already been validated can be used more often if needed.

B. Game Variations

We are interested in creating a game that is enjoyable for players and also elicits the most useful user-generated data, as determined by the efficiency with which players are able to identify mislabeled leaf images. In our case, useful data is tied directly to a feedback mechanism that may cause user frustration, as described above. To curtail any potential ill effects of the game's imperfect dataset, we have developed a game variation that allows players to skip rounds they find particular confusing or difficult. We call this the Skip Variation as described below and shown in Fig. 3. Thus, the two game variations we will use throughout this paper are:

1) Standard Game

The version of the game described thus far.

2) Skip Variation

In this version, a player can pass, or skip a particular set of leaf images up to two times in the course of the game. This option is presented to the user as a button at the bottom of the screen with some explanatory text describing what it does, as illustrated in Fig. 3. Clicking on the button moves the user on to the next set of leaf images without affecting the score or reducing the number of lives remaining in the game. Skipping allows players to have a "way out" if they notice an erroneous set (e.g., two "Odd Leaves" instead of one) or a particularly difficult set. It also provides useful information, as it is possible that skipped rounds will help flag sets with an erroneous leaf, a situation that we test for later.

C. Creating the Leaf Image Sets

Creating the appropriate leaf sets is essential to the game since it determines both the difficulty of the game and the amount of useful information gleaned from play. To make playing the game more enjoyable, we varied the difficulty of the leaf sets, so that players would feel challenged but not

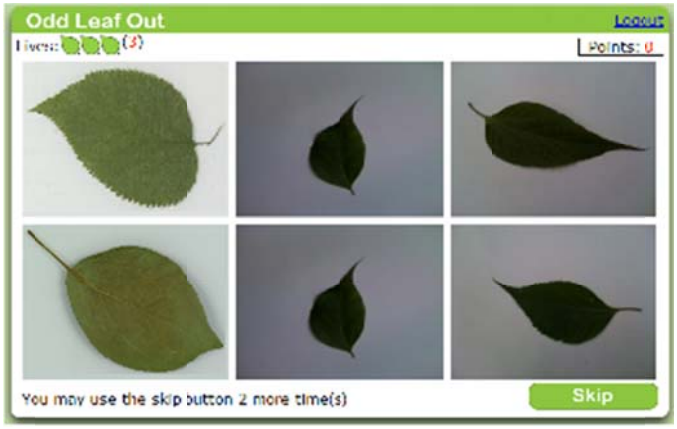

Figure 3. Odd Leaf Out – Skip Variation

frustrated because they were consistently unable to identify the image from a different species (the Odd Leaf).

All leaves used in the game come from an existing database of labeled leaves generated by the LeafSnap [9] team. LeafSnap is a hand-held device that uses computer vision to assist in plant species identification. To use this system, a person photographs a leaf on a plain background. Each leaf image is segmented automatically. The leaf shape is then compared to a collection of leaf shapes from trees of the Northeast US. Using this application, the correct species is identified as the most likely match approximately 75% of the time, and is within the first five possible matches approximately 95% of the time.

To create the game's image sets, one leaf is chosen at random from the image dataset, which we refer to as the initial leaf. Then, four additional leaf images from the same species are chosen, one of which is the *most dissimilar leaf* from the initial leaf (as determined by its pairwise distance from the initial leaf). The sixth leaf is the Odd Leaf because it is always chosen from a different species. However, it has a varying algorithmic similarity to the initial leaf image. To generate a "difficult" set, an Odd Leaf image that has a small distance from the initial leaf, or that more closely resembles the other species, is chosen. Similarly, to generate an "easy" set an Odd Leaf is selected that has a large distance from the initial leaf.

D. Game Development Process & Pretesting

Prototype versions of the Odd Leaf Out were developed and tested by the research team, immediate colleagues, and potential users (casual gamers and a plant sciences professor) in an iterative manner. Several user interface improvements were made based on feedback. Two major changes were made to the game itself based on early feedback. First, we changed the game from a timed game (1 minute of play) to an untimed game based on 3 "lives". The untimed version led to higher accuracy levels since players were not rushed. It also removed the ability to "game the system" by selecting a random answer extremely rapidly. Second, we initially started by choosing the hardest possible Odd Leaf (i.e., minimum pairwise distance between the Odd Leaf and the initial leaf), which proved to be too hard and frustrating to players. We reduced the overall difficulty by using the procedure described in the prior section to generate leaf sets that have a range of difficulty such that players can correctly find the Odd Leaf about 75% of the time.

IV. Methods

A. Player Recruitment and Setup

To evaluate the "Odd Leaf Out" game, we made it available online (at [blank for review]) to a convenient sample of players we recruited via email lists, Twitter, Facebook, and personal contacts. We targeted anyone who wanted to play including friends, family, students, alumni, and other researchers. We also made a special effort to recruit a large number of experienced botanists, plant scientists, and ecologists by disseminating the link through contacts in a university plant sciences department, master gardeners groups, and Flickr groups dedicated to biodiversity projects. Players were presented with the basic instructions of the game and were required to register before playing. Registration entailed providing an email address, password, username, and answering a question about age and general experience in identifying leaves ("How many tree species can you identify in the wild?").

Upon registration players were randomly assigned to one of 2 different game versions described earlier: Standard or Skip Variation. Requiring registration allowed us to assure that players would play the same version of the game if they returned at a later time.

After completing their first game, players were asked to rate the difficulty and the enjoyability of the game on a 1-5 point Likert scale with 5 being the highest. They were also asked to provide any recommendations for improving the game. Players could play as many times as they wanted to by clicking on a "Play Again" button. Upon completion of subsequent games, players were instead taken to the typical recap screen that shows each round they played and information about the leaves and correct answers.

All gameplay information and survey results were stored in a database for later analysis. Email addresses and usernames were anonymized to protect the privacy of players.

B. Creating Images and Image Sets with Errors

A total of 120 image sets were generated and used in the games. All players were served these image sets in a random order to remove any learning effects.

Our primary goal was to determine if game results could be used to identify labeling errors more effectively than standard methods. To test this, we included 12 leaf sets that each had one mislabeled leaf (10% of the image sets). As a baseline error-detection method, for each leaf we compute the average distance to all other leaves of the same species. We call this a leaf's *mean species distance*. When the mean species distance is large, it indicates a leaf that is an outlier, and possibly mislabeled. In a real-world effort to find mislabeled leaves, it would be natural to begin by examining these outliers. Because errors among outliers can be easily detected, we focus on determining whether Odd Leaf Out can be used to detect erroneously labeled leaves with a relatively low mean species distance.

We generated labeling errors by randomly selecting 24 leaf images from our original dataset of 1009 leaf images, and changing their original species affiliation to a random one. Of these, we note half of these initial labeling errors would have been easily identified using our existing algorithm. Therefore, we did not incorporate these errors, leaving 12 mislabeled leaves. We then generated image sets for Odd Leaf Out, as described above. We continued to generate image sets until we obtained 108 image sets that contained no labeling errors, and twelve image sets that each contained a labeling error, with each erroneously labeled leaf appearing in one of these image sets.

C. Prevalence of Error Types in Odd Leaf Out Game

Given some number of mislabeled leaves in a dataset, it is worth considering how often they may show up in different image sets presented in Odd Leaf Out. Each image set may contain the following:

1) *No mislabeled leaves.*
2) *Six leaves of the same species, which occurs when a mislabeled Odd Leaf comes from the same species as the initial leaf. We call this a "No Odd Leaf Error Set."*
3) *Two leaves from species that differs from the initial leaf: the Odd Leaf and a mislabeled leaf. We call this a "Two Odd Leaf Error Set."*
4) *A labeling error, but still with five leaves from the same species, because the Odd Leaf is mislabeled as something different than the other five. These image sets cannot be used to detect labeling errors.*
5) *Two or more mislabeled leaves.*

In generating a large number of image sets, we found that image sets with no labeling errors, and Two Odd Leaf Error Sets are by far the most common. With only about 1% of leaves mislabeled, most image sets (about 76%) contain no mislabeled leaves. Two Odd Leaf Error Sets occur in about 20% of all image sets generated by our procedure. These may occur when any of five leaves are mislabeled. In particular, we generate image sets by adding the farthest leaf from the seed that is believed to be of the same species. It often happens that these are actually mislabeled leaves. This makes the Two Odd Leaf Error Sets relatively common compared to the prevalence of actual leaves with an error (e.g., 20% of sets compared to the 1% of images with errors). Together, these two types of image sets account for more than 95% of image sets, so our focus is on using these to find labeling errors.

We are also interested to determine whether we can detect mislabeled leaves in the No Odd Leaf Error Sets. For our test set, we generated 8 Two Odd Leaf Error Sets, and 4 No Odd Leaf Error Sets, with the remaining 108 images sets not containing any mislabeled leaves.

D. Detecting Errors

We use two different techniques to identify the mislabeled leaves, depending on the type of error set we are dealing with as described below.

1) Two Odd Leaf Error Set Identification Procedure

In these sets, both the Odd Leaf and a mislabeled leaf come from different species than the other four leaves. In this case, we expect that game players will frequently select the mislabeled leaf as the Odd Leaf during gameplay. This will appear to our system as an incorrect answer. Therefore, when game players make a mistake, the leaves that they "incorrectly" choose are the ones most likely to be mislabeled.

To identify potentially mislabeled images, we determine the number of times that a player has selected a non-Odd Leaf as an Odd Leaf. We sort all leaves based on this number and start reviewing those that were incorrectly chosen the most often.

2) No Odd Leaf Error Set Identification Procedure

In these sets all leaves are from the same species. We expect players to find these image sets particularly difficult since there is no real Odd Leaf. In this case, the specific leaf chosen when a player makes a mistake is not significant, because the mislabeled leaf is always the Odd Leaf. So we order the Odd Leaves by the number of times they appear in an image set that produces an erroneous choice by the game player and start reviewing those that show up the most often.

V. RESULTS

In this section, we report two major categories of findings from our analysis of Odd Leaf Out game play: the enjoyability of the game and its ability to help detect mislabeled images. We find that analyzing game results allows us to identify mislabeled images more quickly than can be achieved by using algorithm-created measurements of similarity. We also find that the users reported moderate levels of enjoyment, averaging slightly above 3 on a 1-5 scale, with greater enjoyment by those with more knowledge of leaf identification.

A. Player Experience

A total of 165 individuals played the two game versions during a 3-day period in early June 2011. Table 1 presents overall findings on number of games played, number of rounds played (i.e., number of image sets each player responded to during their games), and player accuracy for the two game versions. Skipped rounds in the Skip Variation are not counted in these numbers. The higher % Correct in the Skip Variation is likely due to the fact that difficult sets can be skipped leaving more easy sets.

TABLE I. SUMMARY GAME STATISTICS

Player Statistics	**Game Version**	
	Standard Game	***Skip Variation***
Players	82	83
Games (Rounds) Played	157 (1,990)	158 (2,345)
% Correct Answer Given	76.1%	80.7%
Avg Games (Rounds) Played	1.9 (24.3)	1.9 (26.3)
Median Games (Rounds) Played	1 (17)	1 (22)
Min Games (Rounds) Played	1 (4)	1 (3)
Max Games (Rounds) Played	10 (162)	10 (157)

Table II provides a summary of the number of players and rounds completed based on players' self-reported ability to identify leaf species. The numbers show that we were able to recruit people of varying levels of domain expertise.

TABLE II. NUMBER OF PLAYERS AND IMAGE SETS (ROUNDS) PLAYED BY SELF-REPORTED EXPERTISE

Expertise Level	Game Rounds Played		
	Players	*Rounds Played*	*Avg Rounds per Player*
Can Identify 0 Leaves	22	717	33.6
Can… 1-5 Leaves	90	1,931	21.5
Can…6-10 Leaves	29	893	30.8
Can…More than 10 Leaves	24	794	33.1
Grand Total	165	4,335	26.3

B. Game's Performance in Error Detection

To assess the effectiveness of the game data in helping identify mislabeled leaves, we compare it with two other methods. For each method we consider how many of the errors we can detect if we look at only a certain fraction of the images (e.g., 1%, 5%, 10%, and 50%). The three methods include:

1) Game playing – Images are sorted based on the methods described in Section IV.D. on Detecting Errors.

2) Mean species distance – Images are sorted from hightest to lowest mean species distance, so that leaves that are on average algorithmically "farther" from the other images in the same species are evaluated first.

3) Random – Images are sorted in a random order to find the errors. If we examine 1% of these leaves randomly, we expect to find 1% of the labeling errors. We report the expected errors detected rather than actual errors detected.

First, we consider how efficiently we can identify mislabeled leaves in Two Odd Leaf Error Sets. Table III summarizes the number of leaves that we would need to examine to find mislabeled leaves using each of the methods outlined above. Note that there are 600 (5x120) total leaves we will consider for possible labeling errors, since we do not consider the Odd Leaves in this type of error set.

TABLE III. THE NUMBER OF LABELING ERRORS DETECTED IN TWO ODD LEAF ERROR SETS USING THREE METHODS.

Errors Detected by Mechanism	Percent of Leaves Examined			
	1%	*5%*	*10%*	*50%*
Game playing	4	5	6	8
Mean species distance	0	0	4	8
Random	0.08	0.4	0.8	4

Using game playing results, the first 4 leaves examined contained labeling errors. Thus, the game data allowed us to find half the labeling errors immediately. Using the mean species distance, none of the labeling errors turn up until we examine 10% of the leaves.

Next we consider how efficiently we can identify mislabeled leaves in No Odd Leaf Error Sets. Table IV summarizes the findings. Again we see that the game data is helpful in improving the error identification.

TABLE IV. THE NUMBER OF LABELING ERRORS DETECTED IN NO ODD LEAF ERROR SETS USING THREE METHODS.

Errors Detected by Mechanism	Percent of Leaves Examined					
	1%	*2%*	*3%*	*5%*	*10%*	*50%*
Game playing	0	1	2	2	2	4
Mean species distance	0	0	0	2	2	4
Random	0.04	0.08	0.12	0.20	0.40	2

In a real application, we will not know which image sets contain labeling errors, and what types of errors they contain. Our results suggest that the most fruitful way of finding errors will be to examine leaves that are incorrectly chosen as the Odd Leaf (since they may be from Two Odd Leaf Error Sets). We might also choose to examine the Odd Leaf in image sets that produce a large number of errors by players (since they may be from No Odd Leaf Error Sets).

We have also examined image sets that were frequently skipped by players in the Skip Variation. We anticipated that players may skip either Two Odd Leaf Error Sets (since they wouldn't know which of the 2 leaves to choose) or No Odd Leaf Error Sets (since they would be very difficult). However, these skips provided less useful information about labeling errors than we had hoped. There were 11 image sets that players skipped 3 times or more. One image set was skipped seven times. This did indeed contain a labeling error in the Odd Leaf (i.e., it was a No Odd Leaf Error Set). However, the other 10 image sets that were skipped between 3 and 6 times did not contain any labeling errors.

We also wanted to know if we could identify errors better using game data from domain experts or novices. We did not find that the results of experts with a knowledge of more trees leads us to more rapidly detect labeling errors than when we use the results of players with knowledge of fewer trees. This may be because the accuracy rates between the two groups are not all that different, as shown in Table V. This is encouraging, because it suggests that we do not need expert input, which is comparatively more costly, to detect labeling errors.

TABLE V. PLAYER PERFORMANCE BY GAME TYPE AND EXPERTISE

Expertise Level	Percentage of Correct Answers	
	Standard Game	*Skip Game Variation*
0	78.24%	75.65%
1-5	75.51%	77.84%
6-10	76.73%	82.52%
More than 10	72.54%	82.52%

Although Odd Leaf Out data can help efficiently identify mislabeled images, it only works because there are multiple people playing each round (i.e. image set). Table VI shows the

expected number of errors identified in the 8 Two Odd Leaf Out Error Sets based on the number of times a given round is played. For example, in the first row, each round is played by five players. The fifth row shows the results with all rounds used; each image set was played a median of 37 times (min = 27 times; max = 50 times). In the first 4 rows, we randomly selected the specified subset of times each round was played (i.e., 5 times for row 1) from among our game data, and show the average over 100 random draws.

TABLE VI. THE NUMBER OF LABELLING ERRORS DETECTED IN TWO ODD LEAF ERROR SETS AS A FUNCTION OF TIMES EACH SET WAS PLAYED

Errors Detected	**Percent of Leaves Examined**					
	1%	***2%***	***3%***	***5%***	***10%***	***50%***
Set Played 5 times	2.7	4	4.2	4.7	5.6	8
Set Played 10 times	3.5	4	4.3	4.9	5.7	8
Set Played 15 times	3.6	4	4.2	4.9	5.9	8
Set Played 20 times	3.8	4	4.1	4.9	5.9	8
Set Played all times	4	4	4	5	6	8

The results suggest that even when each leaf set is played only 5-10 times, half of the misclassified leaves are found in the top 2% of images. Although error detection improves continually, the benefits are diminishing. A similar pattern occurs for identifying errors from the No Odd Leaf Out Error sets.

C. Player Reported Game Enjoyment

A total of 165 unique individuals played the Odd Leaf Out game during our 3 day trial. Of these 165 players, 107 (65%) answered the short questionnaire presented after they played one full game. We examined the data to see if there was any relationship between reported enjoyment and game difficulty, but found there was none. This suggests that people's enjoyment was not due to the fact that they found the game too easy or too hard.

We also looked to see if players enjoyed one of the game variations more than the other. Table VII shows the percentage of players that answered the questions and the average enjoyment score in the Standard Game and Skip Variation. We conducted a Mann-Whitney U Test and found that there was no statistically significant difference between the average scores reported (p=0.76), indicating that neither of the two game variations was particularly more enjoyable than the other.

TABLE VII. PLAYER'S ENJOYMENT BY GAME VERSION

Player Statistics	**Game Variation**	
	Standard	***Skip***
Total Players	82	83
Respondents to Question	52 (63%)	55 (66%)
Average Enjoyment Score	3.19	3.11

We were also interested in the relationship between game enjoyment and domain expertise, as measure by the tree identification question asked at registration. To measure this, we compared players who reported being able to identify six or more leaves to those that were able to identify fewer species. While there was an option for being able to identify more than 10 leaf species, we did not base expertise at that level, as there were too few players in that category to statistically evaluate.

TABLE VIII. PLAYER'S ENJOYMENT BY EXPERTISE

Player Statistics	**Number of Species Player can ID**	
	Less than 6	6 or More
Total Players	112	53
Respondents to Question	72 (64%)	35 (66%)
Average Enjoyment Score	3	3.46

We found that players who reported being able to identify less than 6 leaf species had a mean enjoyment score of 3.00, while those able to identify six or more exhibited a mean enjoyment score of 3.46. The Mann-Whitney U Test found this difference to be significant at the .05 level (p=0.04). Table VIII presents a summary of these findings.

Upon completion of the first game, players had the opportunity to provide any written comments about how to improve the game. As discussed above, 65% of all players (107) provided a response to at least one of the questions in the short questionnaire at the end of the game. Of those, 107 individuals, 47 (44% percent) submitted a comment about how the game could be improved. The most commonly repeated response reflected the sentiment of one player who stated: "the images are not good quality…if they looked more professional, it would be more appealing." This suggests the need to explain why low-quality images from mobile phones are used in the game, something that we did not do in this version.

Several players indicated that they would like more feedback about the leaves, either by giving a "description of the odd leaf out" or text giving some expert knowledge of the difference between the leaves. This emphasizes the interest in educational aspects of the game. One player indicated they believed they encountered an error by telling us we should review them again with statements like "check that last Sassafras, will ya?" Two individuals, both leaf identification experts, went out of their way to contact team members to also express their opinion that there were errors in the dataset not realizing that we had planted them there. They both correctly identified errors. This suggests that having another mechanism for experts to flag errors as they play (even independent of gameplay) may solicit useful information about errors.

Finally, we received more indications of people liking the game than reports of finding it to be unenjoyable. However, the comments did range from "This is not a game. It is not fun" to "It's great! I love this game!".

VI. DISCUSSION

We have created a new game genre for detecting mislabeled images, which we call Odd Leaf Out because of our implementation of the game using leaf images. The game is unique because it uses people's mistakes (rather than their mutual agreement) to help solve a challenging computational problem. Furthermore, the substance of what gameplay contributes is different, as it focuses on identifying mislabeled

images rather than adding metadata to images. Our use of gameplay data from novices and experts enabled us to find mislabeled images far more efficiently than procedures that only use computer vision algorithms. The game was also found to be moderately enjoyable, particularly by those with some expertise in leaf identification.

Though successful, Odd Leaf Out and our study suffer from some limitations. As discussed earlier, about ten players must play each image set to derive most of the benefits. This is far less efficient than games that only require two players. However, because data from novices and experts are equally good at finding errors, the potential player pool is large. One limitation of our study was that it is based on data from a three-day study, which fails to capture long-term trends. Future studies should consider learning effects of players, willingness of players to play repeatedly, and motivators for long-time players. We anticipate offering Odd Leaf Out on an ongoing basis to enable such studies, but feel that our short-term study provided sufficiently strong evidence for the value of the game.

Comments from Odd Leaf Out players suggest the importance of education as a motivator, something not yet well explored in other games with a purpose. Several players wanted more feedback when they missed a round, wanted to see more text about the leaves, and wanted tips from experts about how to identify leaves. The close coupling with a citizen science project suggests the need to bolster these educational elements as a motivation for play that complements the traditional "fun" motivation. Future game iterations could introduce bonus rounds where players can identify a species based on a leaf image and/or images of its other properties (i.e., fruit, bark) or "unlock" expert advice on leaf identification.

Making the game more fun by adding new social elements is also promising. For example, players could "challenge" each other to a live or asynchronous match where they receive the same image sets. Alternatively, players could devise their own image sets from the database of labeled images that others would play, receiving points for the number of misses. This adds an element of strategy and also another source of potentially useful data in identifying errors.

Perhaps the most natural next step is to try the game with other image datasets such as human faces, galaxies, or butterflies. Would errors be more easily detected? Would other domains be more "fun"? Our finding that experts enjoyed the game more suggests that some subject-matter, such as human faces, may be more enjoyable to a broader swath of people because of humans' innate expertise at recognizing faces.

Finally, we may consider ways to better generate and present the image sets to players so as to more efficiently identify mislabeled images. For example, we described how Two Odd Leaf Out Error sets were common due to our method for generating leaf sets, which helps assure that errors are detected more readily. However, not all of the 12 errors were detected easily. In particular, the image set that was hardest to detect was a Two Odd Leaf Error set where the Odd Leaf was a pine and the other 5 leaves, one of which was mislabeled, were not pines. The mislabeled leaf may have been detected as different had the actual Odd Leaf not been so easy to find. Making image sets progressively challenging to those who are doing well would minimize this problem without making the game too hard to be enjoyable. Alternatively, algorithms could be derived that would dynamically create image sets based on the likelihood of identifying errors given prior gameplay results and image similarity metrics.

ACKNOWLEDGMENT

We would like to acknowledge *** and *** for their contribution to the project and extensive feedback on the Odd Leaf Out game.

REFERENCES

[1] C. Shirky, Here Comes Everybody: The Power of Organizing Without Organizations, 1st ed. Penguin Press HC, The, Feb. 2008. [Online]. Available: http://www.worldcat.org/isbn/1594201536

[2] L. von Ahn, "Games with a purpose," in Computer, vol. 39, no. 6, pp. 92-94, Jun. 2006. [Online]. Available: http://dx.doi.org/10.1109/MC.2006.196

[3] B. Russell, A. Torralba, K. Murphy, and W.T. Freeman. "Labelme: a database and web-based tool for image annotation," in International Journal of Computer Vision, 77:157–173, 2008.

[4] N. Kumar, P. Belhumeur, and S.K. Nayar. "FaceTracer: A search engine for large collections of images with faces," in European Conference on Computer Vision, 2008.

[5] A. Sorokin and D. Forsyth. "Utility data annotation with Amazon Mechanical Turk," in IEEE Workshop on Internet Vision, at CVPR, 2008.

[6] F. Mokhtarian and S. Abbasi,. "Matching shapes with self-intersections: Application to leaf classication," in IEEE Trans. on Image Process, 13(5), 2004, pp. 653-661.

[7] H. Ling and D. Jacobs, "Shape classificatino using the inner-distance," in IEEE transactions on Image Processing, 13(5), 2007, pp.653-661.

[8] P. Felzenszwalb and J. Schwartz,, "Hierarchical matching of deformable shapes," in IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[9] LeafSnap. http://leafsnap.com/

[10] P.Belhumeur et al., "Searching the world's herbaria: A system of visual identificatino of plant species," in European Conference on Computer Vision, 2008.

[11] P. Belhumeur, Personal communication with author, 2011.

[12] L. von Ahn and L. Dabbish, "Labeling images with a computer game,"in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 2004.

[13] L. vonAhn, R. Liu and M. Blum, "Peekaboom: A game for locating objects in images," in Proceedings of the SIGCHI conference on Human Factors in computing systems, April 2006.

[14] E. Law and L. vonAhn, "Input-agreement: A new mechanism for collecting data using human computation games," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 2009.

[15] 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 Proceedings of the ACM SIGKDD Workshop on Human Computation, June 2009.

[16] L. vonAhn, "Games with a purpose," in IEEE Computer Magazine, pp. 96-98, June 2006.

[17] S. Thaler, K. Siorpaes, E. Simperl and C. Hofer, "A survey on games for knowledge acquisition," Innsbruck, Australia: STI Innsbrook, 2011.

[18] M. J. Raddick, G. Bracey, P.L. Gay, C.J. Lintott, P. Murray, K. Schawinski, K., A.S. Szalay, and J. Vandenberg, "Galaxy zoo: Exploring the motivations of citizen science volunteers," Astronomy Education Rev., 9, 2010.

[19] L. vonAhn and L. Dabbish,, "Designing games with a purpose," Communications of the ACM, vol. 51, issue 9, pp.58-67, August 2008.