Solving Hard AI Problems Using Computer Games

Luis von Ahn M. Ian Graham Laura Dabbish
David Kitchin Lenore Blum

November 2002
CMU-CS-02-191

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Automatically determining the contents of an image is a problem far from being solved by Artificial Intelligence (AI) techniques. In this paper we introduce a simple game that is fun to play and has the following unique property: when humans play the game, they help computers determine the contents of images. We present a study providing evidence that the game is fun. If played by a large number of people, this game could provide a resource capable of classifying all images on the World Wide Web. This represents a novel form of interaction between humans and computers, a symbiosis in which humans playing the game are entertained and computers running the game obtain answers to problems they can't solve. Thus, we introduce a much more general idea: computer games can be used to solve hard AI problems.

This material is based upon work partially supported by a National Science Foundation Aladdin Grant.

Any opinions, findings, conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

Keywords: Hard AI Problems, Games, Image Classification

INTRODUCTION

Writing a program that can fully determine the contents of an image is still impossible. Even reading slightly distorted text, a much simpler sub-problem, is hard for current computer programs [16]. To get around this, image search engines on the World Wide Web (such as Google [13] or Altavista [1]) classify images according to file names: an image named "car.jpg", for instance, is classified as an image of a car. This method, though somewhat successful, is clearly not optimal. First, there is no reason for anybody to accurately name the image file, and second, a single file name is not enough to describe the contents of an image. Text appearing adjacent to the images in web pages can also be used as an aid in the classification, but most images have little or no associated text, and even when it's there, such text can be hard to process and is oftentimes unstructured and misleading [5]. Thus, a significant percentage of all images on the World Wide Web are incorrectly classified and can't be found through reasonable search queries.

A possible solution to this problem is manual classification. Manually classified image databases such as The Corbis Collection [6] and The Getty Images [12] allow for very accurate search results. However, manually classifying all images on the World Wide Web could be extremely expensive (there are 390,000,000 images on the World Wide Web, according to Google Image Search [13]).

On the other hand, millions of people around the world play computer games. On a recent weekday afternoon, for instance, the authors found 106,000 people playing on Yahoo Games [19], 120,000 on Pogo.com [15], and 115,000 on MSN's Gaming Zone [14]. In fact, according to a survey recently conducted by the Interactive Digital Software Association, 145 million Americans play computer or video games on a regular basis [4].

In this paper, we explore the possibility of having humans solve hard Artificial Intelligence (AI) problems while playing computer games. We introduce a game that is fun to play (even addictive), and which can be used to determine the contents of an image. We stress that the game should be shown to be fun, so humans will play it like any other game. If our game is deployed at a popular gaming site like Yahoo Games, we estimate that all images on the World Wide Web can be reasonably classified in less than a month.

This introduces a novel approach to solving hard AI problems, as well as a new way of looking at computer users. Even if our game doesn't eventually classify all images on the World Wide Web, our main goal is to provide a proof of concept: games *can* be used to solve hard AI problems. We hope that this small seed will encourage research along these lines.

Related Work in Computer Vision

Over the years, there has been considerable AI work on trying to automatically determine the contents of images. The most successful attempts learn from large databases of annotated images. (Annotations typically refer to the contents of the image, and are fairly specific and comprehensive.) Some of these methods cluster image representations and annotations to produce a joint distribution linking images and words [2,3]. Such methods can predict words for a given image by computing the words that have a high posterior probability given the image. Other algorithms attempt to combine large semantic text models with annotated image structures [8].

Though impressive, such algorithms based on learning don't work very well in general settings and work only marginally well in restricted settings. For example, the work described in [8] only gave reasonable results for 80 out of 371 vocabulary words (their evaluation procedure consisted of searching for images using the vocabulary words, and only 80 queries resulted in reasonable images).

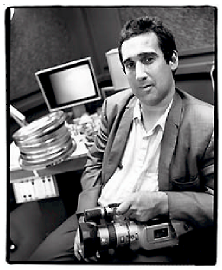
Another line of work that is relevant is one that attempts to find specific objects on images. Schneiderman and Kande [17], for instance, introduced a method to locate human faces in still photographs. Such algorithms are typically accurate, but have not been developed for a wide range of objects. Additionally, combining algorithms for detecting different objects into a single general-purpose classifier is a non-trivial task.

Thus, even a method that can produce reasonable classifications (not necessarily good classifications) for images in general would be a considered a breakthrough.

DESCRIPTION OF THE GAME

We call our game *the ESP game*. It is a two-player game, but it's meant to be online and played by a large number of people at once. We will assume that the two players don't know each other's identity and can't communicate with each other (partners will be randomly assigned among all the people playing the game). Players can't see what their partners are typing, and the only thing two partners have in common is an image that they can both see.

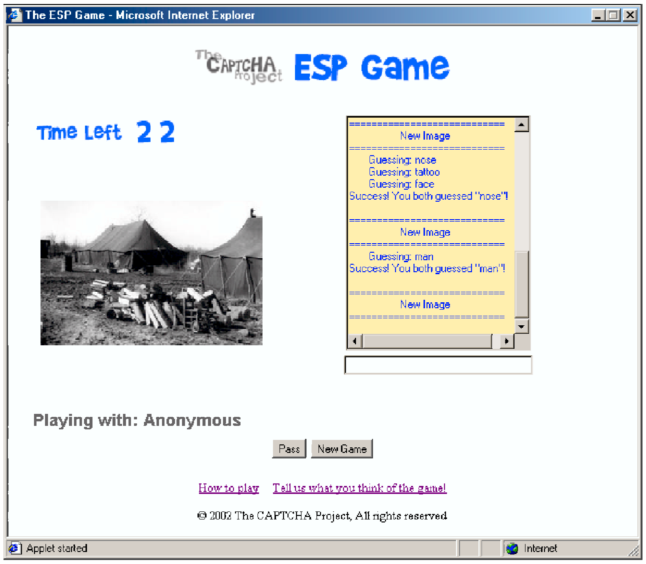
The idea of the ESP game is to guess what your partner is typing. Once the players have produced a matching string, they get a new image. Players don't have to type the string at the same time, but rather must have one matching guess among several made for the same image. Partners have to "agree" on as many images as they can in 90 seconds (see Figures 1 and 2).



Player 1 guesses: man
Player 1 guesses: camera
Player 1 guesses: film
Success! You agree on "camera"

Player 2 guesses: guy
Player 2 guesses: camera
Success! You agree on "camera"

Figure 1: Two players "agreeing" on an image.



```
============================
        New Image
============================
     Guessing: nose
     Guessing: tattoo
     Guessing: face
Success! You both guessed "nose"!

============================
        New Image
============================
     Guessing: man
Success! You both guessed "man"!

============================
        New Image
============================
```

Playing with: Anonymous

© 2002 The CAPTCHA Project, All rights reserved

Figure 2: The ESP Game. Players try to "agree" on as many images as they can in 90 seconds. "Agreeing" on an image means that both players type the same string of characters (not necessarily at the same time).

Since the players can't communicate and don't know anything about each other, the easiest way for both to type the same string is by typing something related to the common image. Notice, however, that the game doesn't ask the players to describe the image: all they know is that to agree on an image, they have to "think like each other" and type the same string. As our experimental results will suggest, the string on which the two players agree is a very good description of the image.

A final element of the game is the use of taboo words. Some images will have taboo words associated with them. Players shouldn't type an image's taboo words, nor should they type singulars, plurals or phrases containing the taboo words. This makes the game harder. Imagine if the taboo words of the image in Figure 1 were "man, camera, film"; how would you then agree on the image?

Taboo words will be obtained from the game itself. The first time an image is used in the game, it will have no taboo words. When the image is used again, it will have one taboo word: the word that resulted from the previous classification. Notice that taboo words might not necessarily be actual words: some images might be classified with a number or a phrase. Taboo words allow the game to find a variety of descriptions for each image.

If deployed at a popular gaming site, we expect the ESP game to be played by a large number of people and thus allow the classification of all images on the World Wide Web. In later sections we will report data from two experiments that provides evidence in support of this belief. Before that, however, we discuss several issues related to the design of the game.

Design Issues

1. Why do we expect the ESP game to be played by a large number of people?

First we use an argument of similarity: the ESP game is similar to other games that are known to be fun. We briefly discuss similarities to two games: "Family Feud" and Taboo™.

"Family Feud" [9] is a TV game show that was extremely popular in the 80s. The game featured two teams each composed of five members from the same family. The teams competed against each other to match answers to the results of a survey of one hundred people. For instance, the teams could be asked: "name an item of clothing worn by the three musketeers." The team's answers to the question were judged according to the survey that was performed prior to the game: one hundred people were asked the same question, and a team's answer was worth more if more people in the survey answered in the same way. If 80 of the 100 respondents answered "a hat" and 20 answered "pants", then the answer "a hat" would be worth more than the answer "pants." Thus, the game was not about giving correct answers but about giving popular answers. The ESP game is similar in this respect: it's best for players to type popular descriptions. We believe there is something fun about trying to guess what other people will say.

The ESP game is also similar to the game Taboo™ [18], which is played between two teams. The teams take turns guessing words. On a given team's turn, one of the players tries to get their team members to say a word, like "apple." The player can talk to her team members, but she cannot say the word "apple" nor any of the words in the taboo set for apple: red, fruit pie, cider, and core. She has to get them to say "apple" by saying phrases like "New York is the big X" that don't contain the taboo words. The point of the game is to come up with creative ways to make your team members say a word without saying any of the taboo words. The taboo words in the ESP game share this flavor: how to describe an image without using the taboo words?

The second reason we expect our game to be played by a large number of people is that, in its final implementation, it will possess key qualities that are known to motivate people to play games. Particularly, it will posses the motivations of *proving oneself*, *need for acknowledgement*, and *exercise* outlined by Crawford [7].

The motivation of *proving oneself* involves the function of games as a means of demonstrating prowess. The ESP game supports this function in that players receive a score based on how many images they classify. Scores of all players could be displayed in a high score list, recording the screen names of the top-scoring players.

The *need for acknowledgement* relates to our need to be recognized by other people. By showing players their partner's unique screen name, the ESP game can fulfill this need, allowing players to get a sense of the personality of whomever they are playing with.

The ESP game also provides the motivation of mental *exercise*. It allows players to exercise their skills of cognition and intuition in reasoning or intuiting how the other player might describe a particular image, and in attempting to avoid taboo words when making their own guesses.

According to Crawford [7], these motivations draw people to play and keep playing particular games. We believe they will motivate users to play the ESP game.

Notice that games do not need to be complex in order to draw people to play them. Wordwhomp™ [15], for instance, is an extremely popular game that is based on the simple idea of finding all the possible words that can be formed with a specific set of letters. Though the ESP game is a simple game, it should not be disregarded because of its simplicity.

The above arguments are, of course, only pieces of evidence and are by far not conclusive.

2. Why should the classifications given by the ESP game be any good?

As we said before, the only thing that the players know about each other is that they are seeing the same image, so the easiest way to "agree" is to type a word related to the image. Moreover, much like in "Family Feud," players are encouraged to type popular descriptions of the images (rather than descriptions that are significant to themselves only).

The use of taboo words allows for quite specific classifications. The first time around, without any taboos, we expect the classification of an image to be a very general one, like "man." Subsequent classifications should become progressively more specific: unable to use "man," players might start giving answers about how the man is dressed or what he is doing.

Our goal is to associate a set of words to each image. This set of words should only contain words that are in fact related to the image. The bigger the set, the better. We do not believe that the ESP game will give "the best classification" of an image (whatever that means), nor do we want it to. We want to classify images so as to improve the search process. The first step towards a good image search engine is to be able to associate all relevant words with each image. This is the goal of the ESP game.

3. What about cheating?

Cheating is an issue that should be seriously considered when using games to solve hard AI problems. In the ESP game, for instance, what prevents two players from cheating by using a previously decided strategy? Two players, for example, could agree to always type the letter "a" for each image. This would allow them to agree much faster and get higher scores. What about other forms of cheating, like playing against yourself?

Our first argument against cheating is one of large numbers: the game is meant to be played by hundreds, if not thousands, of people at once. Most of these people will be in vastly different locations. Since players are randomly paired, they will usually have no information about whom their partner is, and they will have no way to previously agree on a strategy. Additionally, it takes two players to cheat in the ESP game, and the probability of two cheaters being paired together should be low.

Our next argument has to do with fun. It's actually very easy to cheat in several of the most popular online games. In Wordwhomp™ [15], for instance, it's fairly straightforward to write a computer program that will help you cheat. In fact, such programs can be found online. The reason not everybody does this is because it stops being fun after the first time.

Several steps can also be taken to guard against cheating in a full-scale implementation of the ESP game:

- IP addresses may be checked in order to prevent cheating in the form of a single person posing as two players and having the (unlikely) chance of being paired with themselves.
- An answer to an image might become a taboo word for the duration of the game. This would prevent cheating that results by agreeing to a simple strategy such as typing "a" for every image.
- Since it's easy to check whether a particular player is cheating, a warning about canceling cheaters' accounts could be posted. Such a measure is taken by Pogo.com, for instance.
- Rather than having people play an entire game with the same person, partners might be changed after each image is classified. If enough people are playing at the same time, this would present no logistical problem, and cheating would be significantly harder. Of course, two people should be paired only if they have similar remaining times.
- The game might also require that all answers be words from the dictionary.

4. Selecting Images

We believe that the choice of images used by the ESP game makes a difference in players' experiences. The game would perhaps be less entertaining if all the images were chosen from a single website containing second by second shots of an almost still scene. Similarly, different people might enjoy playing with different sets of images, and several images found online might not be appropriate for all audiences. Such issues should be considered when writing a full-scale implementation.

In the most basic setting, the images are chosen at random from the World Wide Web with a small amount of filtering: no blank images, no images that consist of a single color, no images that are smaller than 4x4 pixels, no nudity, *et cetera*. Such filtering can be easily done automatically (filtering for nudity is possible using the techniques of [10]).

More specific "theme rooms" might be created for those who wish to play the ESP game using only certain types of images. Some players might want images coming from certain domains or with specific types of content. Images for these "theme rooms" can be obtained either using web directories or the classifications given by the "general category" ESP game.

EXPERIMENTAL EVIDENCE

It is important to show that the ESP game is enjoyable and that the classifications it produces are reasonable. We performed two simple experiments to evaluate the game. The experiments also allowed us to estimate the rate at which images are classified using the EPS game.

In the first experiment, participants were asked to play the ESP game and answer a questionnaire about their experience. Our purpose was to learn about the game's design and about how likely people are to play it. In the second experiment, participants rated the classifications that resulted from the first experiment.

These experiments are only meant to provide further evidence in conjunction with the rest of the paper. They should be taken for what they are: one element of many demonstrating the feasibility of this approach. The only way to truly determine whether the ESP game can classify a large number of images is to release it on a popular gaming site.

Experiment 1: Game play

We arranged two experimental sessions, during which groups of distributed participants played the game at a scheduled time. By using large groups (10-12 people) of distributed participants at scheduled times, we ensured that no player would know their partner for a particular round of the game. Each participant was instructed to play the game 10 times, and then complete a questionnaire about the experience. Partners were randomly assigned for each individual 90-second game.

The experimental version of the game did not allow for a list of high scores, a component that we believe would make the game more fun in practice. Interpretations of our results should take this fact into account.

Three hundred images were selected at random from the set of results that Google returned on the query "jpg" when moderate SafeSearch™ was on. Our implementation of the game selected one of these 300 images at random every time a new image was required. We believe our set of images is representative of a random set of images from the World Wide Web.

Participants were recruited from CMU and the surrounding area, but were not necessarily CMU students. Overall 22 people participated in the experiment: 9 males and 13 females. The mean age for the participants was 23.4 (std. dev. = 1.29).

As Fulton [11] suggests, feedback on gaming experiences should accurately represent the opinions of target gamers. In our case, the target gamers are those who frequently play online games such as those in www.pogo.com or Yahoo Games. *We were not able to recruit subjects based on their gaming preferences*. This said, perhaps more accurate data could be gathered by recruiting the target gamers.

Results of Experiment 1

To get feedback on a question of game design, we asked participants whether they would have liked to see their partner's guesses (on a scale of 1 to 5, where 1 is strongly agree and 5 is strongly disagree). All but one of the participants responded "strongly agree" to the statement "You would like to see what your partner's guesses are after each round is over." (The other participant responded with "agree.") In our experimental implementation of the ESP game, we chose not to show the partner's guesses to prevent the possibility of cheating by communicating through the unmatched guesses. Perhaps in a full-scale implementation of the game these guesses can be shown while still ensuring that participants can't communicate.

Participants were asked to rate their agreement with the following statements on a Likert scale of 1 to 5 (where 1 was "strongly agree", and 5 was "strongly disagree"):

- You enjoy playing the ESP game.
- You could see yourself playing the ESP game regularly.

In response to the statement "You enjoy playing the ESP game," participants responded as follows: 18% strongly agree, 64% agree, 14% neutral, 4% disagree and 0% strongly disagree. We believe this provides evidence towards how enjoyable the ESP game is.

In response to the statement "You could see yourself playing the ESP game regularly," participants were distributed as follows: 14% strongly agree, 32% agree, 18% neutral, 22% Disagree and 14% Strongly Disagree. This supports our belief that many people will play the game if put on a popular gaming site. (By the way, it would be ridiculous to expect a large majority of our participants to claim that they would play any similar game on a regular basis: they were not selected from among the target gamers.)

Rate of Classification

Experiment 1 also allowed us to determine the rate at which images are classified. The mean number of images agreed upon by our subjects in 90 seconds was 6.64 (std. dev. = 3.07). This translates to a little over 4 images per minute. At this rate, 4000 people playing the ESP game 24 hours a day would classify all images on the web (390,000,000) in 31 days. This, of course, would only associate one word to each image. In 6 months, 6 words could be associated to every image. Notice that this is a perfectly reasonable estimate: 4000 is a typical number of people that one might find playing any single game in a popular gaming site like Yahoo Games. Additionally, 6.64 images every 90 seconds is somewhat conservative, as we believe that players get faster with practice.

Experiment 2: Quality of the classifications

To show that the ESP game is a reasonable method for classifying images, it is important to assess the quality of the classifications. To do so, we had 5 participants perform a simple evaluation of the classifications produced during Experiment 1 (described above). The participants were between 21 and 26 years of age. Two of them were female and three were male.

Each participant was shown all images classified in Experiment 1, along with all classifications produced for each image. Participants were asked to mark any classification that did not match the corresponding image. All 300 images had at least one classification, and some had as many as 9.

Notice that we did not test whether the classifications were the best possible for each image. Participants were merely asked to say whether the words made sense with respect to the image. "The best classification of an image" is a very subjective notion. Furthermore, the goal of the ESP game is exactly to associate as many reasonable words as possible to each image (see point 2 in the section labeled "Design Issues").

Results of Experiment 2

None of the classifications were judged to be unreasonable. This strongly supports our belief that the ESP game produces reasonable image classifications.

CONCLUSION

Though the results of modern-day Artificial Intelligence are extremely impressive, there is still a long way to go until many problems are solvable by computers alone. The authors believe there will be a day in which computers will be as good as or even better than humans in all cognitive skills. In the meantime, however, humans are the only resource capable of solving several important problems.

We have provided evidence that one of these important problems can be solved by humans playing games. Perhaps other problems can be attacked in a similar way.

ACKNOWLEDGMENTS

We thank Manuel Blum for his unconditional support and advice. We also thank Chuck Rosenberg, Steven Katz and Lenore Ramm for their insightful comments. This work was partially supported by the National Science Foundation (NSF) grants CCR-0122581 and CCR-0085982 (The Aladdin Center). Laura Dabbish is also partially supported by a National Defense Science and Engineering Graduate Fellowship.

REFERENCES

1. Altavista:
   http://www.altavista.com/
2. Barnard, K., Duygulu, P., and Forsyth, D. A. Clustering Art. In *IEEE conference on Computer Vision and Pattern Recognition*, 2001, II 434-441.
3. Barnard, K., and Forsyth, D. A. Learning the Semantics of Words and Pictures. In *International Conference of Computer Vision*, 2001, 408-15.
4. Ceangal Assorted Gaming Statistics Page: http://www.ceangal.com/gaming/statistics.html
5. Carson, C., and Ogle, V. E. Storage and Retrieval of Feature Data for a Very Large Online Image Collection. *IEEE Computer Society Bulletin of the Technical Committee on Data Engineering*, Dec. 1996, Vol. 19 No. 4.
6. The Corbis Collection:
   http://www.corbis.com/
7. Crawford, C and Crawford, L. L. *The Art of Computer Game Design*. Osborne McGraw-Hill, January 1984.
8. Duygulu, P., Barnard, K., de Freitas, N., and Forsyth, D. A. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *Seventh European Conference on Computer Vision,* 2002, IV 97-112.
9. Family Feud Homepage:
   http://www.familyfeud.tv/
10. Fleck, M. M., Forsyth, D. A., and Bregler, C. Finding Naked People. *ECCV 1996.*
11. Fulton, B. Beyond Psychological Theory: Getting Data that Improve Games. *Game Developer's Conference,* 2002.
12. The Getty Images:
    http://www.gettyimages.com/
13. Google:
    http://www.google.com/

14. MSN Gaming Zone:
http://zone.msn.com/

15. Pogo:
http://www.pogo.com

16. Rice, S., Nagy, G., and Nartker, T. *Optical Character Recognition: An Illustrated Guide to the Frontier*. Kluwer Academic Publishers, Boston, 1999.

17. Scheniderman, H. and Kanade, T. Object Detection Using the Statistics of Parts. *International Journal of Computer Vision,* 2002.

18. Taboo™ Game Instructions: http://www.centralconnector.com/GAMES/taboo.html

19. Yahoo Games:
http://games.yahoo.com/