Improving Accessibility of the Web with a Computer Game

Luis von Ahn, Shiry Ginosar, Mihir Kedia, Ruoran Liu and Manuel Blum
Computer Science Department, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15213
biglou@cs.cmu.edu, {shiry,majin,ruoranl,mblum}@cmu.edu

ABSTRACT
Images on the Web present a major accessibility issue for the visually impaired, mainly because the majority of them do not have proper captions. This paper addresses the problem of attaching proper explanatory text descriptions to arbitrary images on the Web. To this end, we introduce Phetch, an enjoyable computer game that collects explanatory descriptions of images. People play the game because it is fun, and as a side effect of game play we collect valuable information. Given any image from the World Wide Web, Phetch can output a correct annotation for it. The collected data can be applied towards significantly improving Web accessibility. In addition to improving accessibility, Phetch is an example of a new class of games that provide entertainment in exchange for human processing power. In essence, we solve a typical computer vision problem with HCI tools alone.

Author Keywords
Distributed knowledge acquisition, Accessibility, Web-based games.

ACM Classification Keywords
H5.3 [**HCI**]: Web-based interaction.

INTRODUCTION

The Web is not built for the blind. Only a small fraction of major corporate websites are fully accessible to the disabled, let alone those of smaller organizations or individuals [5]. However, millions of blind people surf the Web every day, and Internet use by those with disabilities grows at twice the rate of the non-disabled [3].

One of the major accessibility problems is the lack of descriptive captions for images. Visually impaired individuals commonly surf the Web using "screen readers," programs that convert the text of a webpage into synthesized speech. Although screen readers are helpful, they cannot determine the contents of images on the Web that do not have descriptive captions. Unfortunately the vast majority of images are not accompanied by proper captions and therefore are inaccessible to the blind (as we show below, less than 25% of the images on the Web have an HTML ALT caption). Today, it is the responsibility of Web designers to caption images. We want to take this responsibility off their hands.

We set our goal to assign proper descriptions to arbitrary images. A "proper" description is *correct* if it makes sense with respect to the image, and *sufficient* if it gives enough information about its contents. Rather than designing a computer vision algorithm that generates natural language descriptions for arbitrary images (a feat still far from attainable), we opt for harnessing humans. It is common knowledge that humans have little difficulty in describing the contents of images, although typically they do not find this task particularly engaging. On the other hand, many people would spend a considerable amount of time involved in an activity they consider "fun." Thus, like the ESP Game [1], we achieve our goal by working around the problem, and creating a fun game that produces the data we aim to collect.

We therefore introduce Phetch, a game which, as a side effect, generates explanatory sentences for randomly chosen images. As with the ESP Game, we show that if our game is played as much as other popular online games, we can assign captions to all images on the Web in a matter of months. Using the output of the game, we mention how to build a system to improve the accessibility of the Web.

Design of a Useful Game

A traditional algorithm is a series of steps that may be taken to solve a problem. We consider Phetch as a kind of algorithm. Analogous to one, Phetch has well-defined input and output: an arbitrary image from the Web and its proper description, respectively.

Because it is designed as a game, Phetch needs to be proven *enjoyable*. We do so by showing usage statistics of a one-week trial period. Because it is designed to collect a specific kind of data, Phetch's output needs to be proven both correct and sufficient. We prove this through a specifically designed experiment.

CHI 2006, April 22–28, 2006, Montréal, Québec, Canada.

Related Work: The ESP Game

One possible approach to assigning descriptions to images in order to improve accessibility is the ESP Game, which allows people to label images while enjoying themselves [1]. The game collects random images from the Web and outputs keyword labels that describe the contents of the images. For example, for an image of a cat and a dog, the ESP Game outputs "cat," "dog," "animals," etc. If all images on the Web were labeled by the ESP Game, screen readers could use these keywords to point out the main features of an image. This, however, can be vastly improved. While keyword labels are perfect for certain applications such as image search, they may be insufficient for describing the contents of a picture. This is demonstrated in Figure 1. We therefore design a different game that collects sentences instead of keyword descriptions. We further show below that the descriptions collected using our game are significantly more expressive than the keywords collected using the ESP Game.

Figure 1. Two inherently different images that share the same ESP labels: "man" and "woman." The Phetch descriptions, however, are different: "half-man half-woman with black hair" and "an abstract line drawing of a man with a violin and a woman with a flute."

GAME MECHANICS

Phetch is designed as an online game played by 3 to 5 players. Initially, one of the players is chosen at random as the "Describer" while the others are the "Seekers." The Describer is given an image and helps the Seekers find it by giving a textual description of it. Only the describer can see the image, and communication is one-sided: the Describer can broadcast a description to the Seekers, but they cannot communicate back. Given the Describer's paragraph, the Seekers must find the image using an image search engine. The first Seeker to find the image obtains points and becomes the Describer for the next round. The Describer also wins points if the image is found. (See a screenshot of the Seeker's interface in Figure 2.) Intuitively, by observing the Describer's text, we can collect natural language descriptions of arbitrary images.

Each session of the game lasts five minutes, and the players go through as many images as they can in that amount of time. The Describer can pass, or opt out, on an image if they believe it is too difficult for the game. When deciding to pass, the Describer gets a brand new image and is penalized by losing a small amount of points.

As mentioned, the Seekers have access to an image search engine, which, given a text query (either a word or a set of words), returns all images related to the query. Once a Seeker believes she has found the right image, she clicks on it. The server then tells them whether their guess was correct. The first Seeker to click on the correct image wins and becomes the next describer. To prevent Seekers from clicking on too many images, each wrong guess carries a strong penalty in points. This penalty also ensures that the text given by the Describer is a reasonably sufficient description of the image, since Seekers will tend to not guess until they are certain.

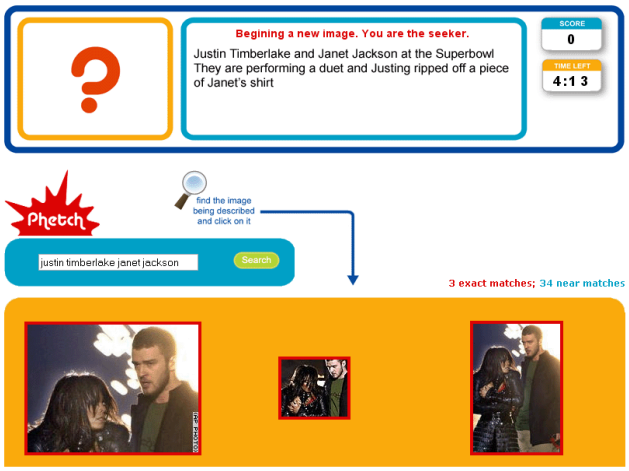

Figure 2. A screenshot of the Seeker's interface.

The ESP Game-Based Search Engine

The image search engine given to the Seekers is a crucial component of the game. In particular, it must be accurate enough to find a single image given a sufficient description of it. Unfortunately, the majority of popular image search engines (Google Images, Yahoo Image Search, etc.) are not nearly accurate enough. Although the results they return appear convincing (e.g., many images returned on the query "car" are really cars), attempting to find a specific image (even one that is guaranteed to be indexed) is nearly impossible for two reasons. First, the search space is very large: Google Images, for instance, searches over a billion images. Second, and more importantly, each image indexed by standard image search engines on the Web has a small number of accurate keywords associated with it. Current search engines on the Web index images by using their filenames, their captions and the text around them on the website [1]. Unfortunately, this method attaches very few useful keywords to most images.

For these reasons, we choose to use a significantly more accurate search engine based on keywords collected by the ESP Game. ESP has already collected millions of accurate keyword labels for arbitrary images from the Web [1].

To remedy the problem of a large search space, 100,000 labeled images were chosen from the ESP dataset as a basis

for our search engine. Each of these images has an average of nine keyword labels associated to it.

We note that the dependence on the ESP Game can be eliminated or mitigated by using a standard image search engine along with some modifications to allow successful game play. Here, however, we do not concern ourselves with this issue.

EMULATING PLAYERS FOR SINGLE-PLAYER GAMES

As stated, Phetch requires three to five players: One Describer and 2-4 Seekers. Given that the total number of players on the website may not always be split perfectly into games of 3-5 players (for instance, when there is only one player), we can also pair up players with computerized players, or "bots."

Essentially, a "bot" emulates a person by playing pre-recorded game actions. Whenever a real 3-5 person game takes place, we record every piece of text the Describer enters, how long it takes the Seekers to find the image, etc.

Emulating a Describer is easy: simply display text that was previously recorded on a successful session of the game. Emulating Seekers in a convincing manner is more difficult. If a real player enters useless descriptions of the image, we do not want the emulated Seekers to find it. Although this is not a significant problem since most Describers enter accurate descriptions of images, we nevertheless address it to protect the illusion that a real game is being played. The solution to the problem relies on the fact that we are using images from a familiar database, so the keywords associated with the images in the search engine are already known. In this implementation we utilize the ESP Game database. We therefore emulate Seekers by not guessing the right image until a number of the ESP keywords have been entered by the Describer.

In addition to allowing every player to quickly start a session of the game, player emulation further acts as a mechanism that helps ensure the accuracy of the descriptions obtained (as we show below).

ENSURING DESCRIPTION ACCURACY

The following strategies ensure description accuracy:

- **Success of the Seekers.** We use the amount of time taken by the Seekers to find the proper image as an indicator of the quality of the description. If the Seekers do not find the image, we discard the Describer's text.
- **Random pairing of the players.** Players could collude to poison our data. For example, two officemates could play at the same time; one as a Describer entering bogus descriptions and the other as a Seeker that magically finds the images. However, this is unlikely. Phetch is meant to be played online by many players at once. By randomly assigning players to sessions, those who wish to collude have low probability of playing together (and even if they do play together, the mechanism described below ensures they are unable to poison our data).
- **Description testing.** Most importantly, to determine whether a description is accurate, we use the single-player version of the game. We play back previously entered descriptions to a single-player Seeker. If they can still find the correct image, we get a significant indicator that the description is of high quality: two different people chosen at random were able to single out the image given just this description. Indeed, we can use this strategy more than once: "*N* people chosen at random were able to find the image."

Note that since Phetch game play is time-constrained, the collected descriptions are not guaranteed to be in proper English, and may well be in a chat-like language. This, however, is acceptable for our purposes since we are mainly concerned with improving the current situation.

VERIFYING DESCRIPTION ACCURACY

We now show that captions collected by Phetch are indeed correct and sufficient with respect to the images. Moreover, we compare the natural language descriptions collected to keyword labels produced by the ESP Game. To this extent, we conducted a study in which participants were assigned to one of two conditions: ESP or PHETCH. Participants in each condition were asked to single out one image among other similar ones, based on either a natural language description, or a set of word labels from the ESP Game.

Eight participants, between 25 and 38 years old who had never played the game were asked to perform our task. Fifty images were chosen at random from the images already processed by Phetch (these images were annotated by players in a one-week trial period of the game described below). For each of these images, we compiled a list of 29 ESP images that were most closely related to the original one (two images are closely related if they share as many ESP labels as possible).

Participants in the PHETCH condition were given a description of an image taken from Phetch. Participants in the ESP condition were given a list of all the ESP labels related to the image. All the participants were shown a randomly ordered 5x6 grid of 30 images containing the target image plus the set of 29 similar images. They were then asked to choose one image based only on the description or the list of keywords, depending on which condition they were in. This process was repeated for the 50 images in the test set. For each image, we tested the participant's ability to correctly single it out.

For the PHETCH condition, participants were able to single out the correct image 98.5% of the time (with standard error of 1.25%), whereas for the ESP condition, participants were able to select the correct image only 73.5% of the time (standard error = 1.25%). The difference is statistically significant ($t = -14.048$, $p = 0.0002$). This result is important for two reasons. First, it shows that the captions produced by Phetch are indeed correct and sufficient — had they not been, participants in the PHETCH condition would not have been able to pick the correct images such a large

percentage of the time. Second, it shows that captions obtained by Phetch are significantly more descriptive than the keywords obtained from ESP (Figure 1).

ENJOYABILITY

Our game may theoretically output data that is useful, but if it is not engaging, people would not play, and output would not be produced. Hence it is of major importance to test our claim that Phetch is entertaining.

Although Phetch has not been formally released to the public, we present the results of having test players interact with the game. These people were partly obtained by offering select random players from another gaming site (http://www.peekaboom.org) the opportunity to play.

A total of 129 people played the game in this one-week trial, generating 1,436 captions for images. Each session of the game lasted five minutes and, on average, produced captions for 6.8 images. Therefore, on average, each player spent 24.55 minutes playing the game in this one-week period, and some people played for over two hours straight. These numbers show how enjoyable the game is.

Given the average number of captions produced in a single game of Phetch, 5,000 people playing the game simultaneously could associate captions to all images indexed by Google in just ten months. This is striking, since 5,000 is not a large number compared to the number of players of individual games in popular gaming sites [1].

IMPROVING ACCESSIBILITY

Although it is common knowledge that image captions are a major accessibility problem [3], we were unable to find a reference for the percentage of images throughout the Web that lack a caption. We therefore conducted a study to determine the percentage of images that have an ALT tag (ALT is the caption mechanism in HTML). 2,700 websites were collected at random (using a large crawl of the Web) in which only 28,029 (24.9%) of the 112,343 total images had an ALT tag at all. (Although this number is indicative of the percentage of captions throughout the Web, other factors such as whether the captions were actually related to the image were not taken into account.)

We showed that Phetch produces accurate descriptions for images on the Web, but how can such captions be used to improve accessibility? One way is as follows. All captions could be stored in a centralized location. Whenever a visually impaired individual using a screen reader visits a website, the screen reader could contact the centralized location to ask for all captions associated to the images in the site. The screen reader would then read the caption out loud based on user preferences. Logistically, this mechanism is similar to browser toolbars (e.g. The Google Toolbar) that contact a centralized server to obtain information about the websites that are being visited.

TURNING IMAGE DESCRIPTIONS INTO QUERIES

Thus far we have described Phetch as a tool to obtain captions for arbitrary images. Although this is the main application of our game, other pieces of useful data can be obtained. The process of Seekers finding the correct image involves turning a plain English description into a sequence of appropriate search queries. By recording the search queries the Seekers enter versus the original description, we can obtain training data for an algorithm that converts natural language descriptions into successful keyword-based queries. Such an algorithm would have important applications in information retrieval (e.g., [2]). We do not investigate this application here, but remark that the problem of using such data to train an NLP system has been studied before [2].

CONCLUSION

We have introduced a novel game designed to attach descriptive paragraphs to arbitrary images on the Web. Our game uses the idea of harnessing human computation power in order to solve a problem that computers cannot yet solve People engage in our game not because they want to do a good deed but because they enjoy it.

We have shown that the descriptions obtained using Phetch are correct and sufficient. Although Phetch has not been formally released to the public, we hope that in the near future we can collect millions of annotations for arbitrary images on the Web and thus build a system that can improve accessibility. In today's advanced society, driven forward by the information exchange throughout the Internet, it can no longer be acceptable that a large portion of the population cannot make full use of this resource.

ACKNOWLEDGEMENTS

We thank Jaime Arguello, Laura Dabbish, Susan Hrishenko and the anonymous CHI 2006 reviewers for their insightful comments. This work was partially supported by the National Science Foundation (NSF) grants CCR-0122581 and CCR-0085982 (The Aladdin Center) and by a generous gift from Google, Inc. Luis von Ahn was also partially supported by a Microsoft Research Graduate Fellowship.

REFERENCES

1. von Ahn, L., and Dabbish, L. Labeling Images with a Computer Game. In *ACM Conference on Human Factors in Computing Systems (CHI),* 2004, pp 319-326.
2. Brill, E., Dumais, S., and Banko, M. An Analysis of the AskMSR Question-Answering System. *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 2002.
3. National Organization on Disability Website. http://www.nod.org/
4. Stork, D. G. The Open Mind Initiative. *IEEE Intelligent Systems & Their Applications, 14-3*, 1999, pages 19-20.
5. Watchfire Corporation Website. http://watchfire.com.