ESP: Labeling Images with a Computer Game

Luis von Ahn and Laura Dabbish

School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
{biglou,dabbish}@cs.cmu.edu

Abstract

We present a new interactive system: a game that is fun and can be used to create valuable output. When people play the game they help determine the contents of images by providing meaningful labels for them. If the game is played as much as popular online games, we estimate that most images on the Web can be labeled in a few months. Having proper labels associated with each image on the Web would allow for more accurate image search, would improve the accessibility of sites (by providing descriptions of images to visually impaired individuals), and would help users block inappropriate images. Our system makes a significant contribution because of its valuable output and because of the way it addresses the image-labeling problem. Rather than using computer vision techniques, which don't work well enough, we encourage people to do the work by taking advantage of their desire to be entertained.

Introduction

Images on the Web present a major technological challenge. There are millions of them, there are no guidelines about providing appropriate textual descriptions for them, and computer vision hasn't yet produced a program that can determine their contents in a widely useful way. However, accurate descriptions of images are required by several applications like image search engines and accessibility programs for the visually impaired. Current techniques to categorize images for these applications are insufficient in many ways, mostly because they assume that the contents of images on the Web are related to the text appearing in the page. This is insufficient because the text adjacent to the images is often scarce, and can be misleading or hard to process (Carson and Ogle 1996).

The only method currently available for obtaining precise image descriptions is manual labeling, which is tedious and thus extremely costly. But, what if people labeled images without realizing they were doing so? What if the experience was enjoyable? In this paper we present a new interactive system in the form of a game with a unique property: the people who play the game label images for us. The labels are meaningful even if individual players attempt to disrupt them.

The labels generated by our game can be useful for a variety of applications. For accessibility purposes, visually impaired individuals surfing the Web need textual descriptions of images to be read out loud. For computer vision research, large databases of labeled images are needed as training sets for machine learning algorithms. For image search over the Web and inappropriate (e.g., pornographic) content filtering, proper labels could dramatically increase the accuracy of the current systems.

We believe our system makes a significant contribution, not only because of its valuable output, but also because of the way it addresses the image-labeling problem. Rather than making use of computer vision techniques, we take advantage of people's existing perceptual abilities and desire to be entertained.

Our goal is extremely ambitious: to label the majority of images on the World Wide Web. If our game is deployed at a popular gaming site like Yahoo! Games and if people play it as much as other online games, we estimate that most images on the Web can be properly labeled in a matter of weeks. As we show below, 5,000 people continuously playing the game could assign a label to all images indexed by Google in 31 days.

We stress that our method is not necessarily meant to compete against the other techniques available for handling images on the Web. The labels produced using our game can usually be combined with these techniques to provide a powerful solution.

The Open Mind Initiative

Our work is similar in spirit to the Open Mind Initiative (e.g., Stork and Lam 2000, Stork 1999), which is a worldwide effort to develop "intelligent" software. Open Mind collects information from regular Internet users (referred to as "netizens") and feeds it to machine learning algorithms. Volunteers participate by answering questions and teaching concepts to computer programs. Our method is similar to Open Mind in that we plan to use regular people on the Internet to label images for us. However, we put greater emphasis on our method being fun because of the scale of the problem that we want to solve. We don't expect volunteers to label all images on the Web for us: we expect all images to be labeled because people want to play our game.

General Description of the System

We call our system "the ESP game" for reasons that will become apparent as the description progresses. The game is played by two partners and is meant to be played online by a large number of pairs at once. Partners are randomly assigned from among all the people playing the game. Players are not told whom their partners are, nor are they allowed to communicate with their partners. The only thing partners have in common is an image they can both see.

From the player's perspective, the goal of the ESP game is to guess what their partner is typing on each image. Once both players have typed the same string, they move on to the next image (both player's don't have to type the string *at the same time*, but each must type the same string at some point while the image is on the screen). We call the process of typing the same string "agreeing on an image" (see Figure 1).

Player 1 guesses: purse
Player 1 guesses: bag
Player 1 guesses: brown

Success! Agreement on "purse"

Player 2 guesses: handbag

Player 2 guesses: purse
Success! Agreement on "purse"

Figure 1. Partners agreeing on an image. Neither of them can see the other's guesses.

Partners strive to agree on as many images as they can in 2.5 minutes. Every time two partners agree on an image, they get a certain number of points. If they agree on 15 images they get a large number of bonus points. The thermometer at the bottom of the screen (see Figure 2) indicates the number of images that the partners have agreed on. By providing players with points for each image and bonus points for completing a set of images, we reinforce their incremental success in the game and thus encourage them to continue playing. Players can also choose to pass or opt out on difficult images. If a player clicks the pass button, a message is generated on their partner's screen; a pair cannot pass on an image until both have hit the pass button.

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 they have to "think like each other" and type the same string (thus the name "ESP"). *It turns out that the string on which the two players agree is typically a good label for the image*, as we will discuss in our evaluation section.

Figure 2. The ESP Game. Players try to "agree" on as many images as they can in 2.5 minutes. The thermometer at the bottom measures how many images partners have agreed on.

Taboo Words

A key element of the game is the use of taboo words associated with each image, or words that the players are not allowed to enter as a guess (see Figure 2). These words will usually be related to the image and make the game harder because they can be words that players commonly use as guesses. Imagine if the taboo words for the image in Figure 1 were "purse", "bag", "brown" and "handbag"; how would you then agree on that image?

Taboo words are obtained from the game itself. The first time an image is used in the game, it will have no taboo words. If the image is ever used again, it will have one taboo word: the word that resulted from the previous agreement. The next time the image is used, it will have two taboo words, and so on. (The current implementation of the game displays up to six different taboo words.)

Players are not allowed to type an image's taboo words, nor can they type singulars, plurals or phrases containing the taboo words. The rationale behind taboo words is that often the initial labels agreed upon for an image are the most general ones (like "man" or "picture"), and by ruling those out the players will enter guesses that are more specific. Additionally, taboo words guarantee that each image will get many *different* labels associated with it.

Labels and Good Label Threshold

The words that we use as labels are the ones that players agree on. Although there is additional information that could be utilized (i.e., all other guesses that the players enter), for the purposes of this paper such information will be ignored. We use only words that players agree on to ensure the quality of the labels: agreement by a pair of independent players implies that the label is probably meaningful. In fact, since these labels come from different people, they have the potential of being more robust and descriptive than labels that an individual indexer would have assigned (O'Connor and O'Connor 1999).

To increase the probability that a label for a particular image is meaningful, we can utilize a *good label threshold*. This means that before a label is attached to the image (or used as a taboo word), it must have been agreed upon by at least X number of pairs, where X is the threshold. The threshold can be lenient and extremely low (X=1, one pair agreeing makes a label acceptable) or strict and high (X=40, forty pairs must have agreed on that label before it is attached to the image and made a taboo word). Of course, the lower the good label threshold is, the faster we can assign labels to the entire Web.

When is an Image "Done"?

As a particular image passes through the ESP game multiple times, it will accumulate several labels that people have agreed upon. The question is, at what point is an image considered to have been completely labeled and thus no longer used in the game? Our answer to this question is to remove an image from the game when it is no longer enjoyable to guess its contents with a partner. This will occur when a particular image has acquired an extensive list of taboo words, such that pairs are unable to agree on new labels and consistently ask their partners to pass on the image. Repeated passing notifies the system that an image should no longer be used for the game at that point in time. (Repeated passing might also indicate that the image is too complex to be used in the game; in which case the image should also be removed.)

Fully labeled images should be re-inserted into the game when several months have passed because the meaning of the images may have changed due to maturation effects. The English language changes over time, as do other languages. We want to capture the labels appropriate to an image, and thus if the language referring to that image changes over time, so should our labels. In addition to changes in language, cultural changes may occur since a particular image has last been labeled. Thus a picture of something or someone that was labeled as "cool" or "great" six months prior may no longer be considered to be so. For example, an image of Michael Jackson twenty years ago might have been labeled as "amazing" whereas today it might be labeled as "guilty."

Implementation and Other Details

The current version of the game is implemented as a Java applet and can be played at http://www.espgame.org. The applet connects to a centralized *game server*, which is responsible for the following: pairing up the players, providing each pair with a set of 15 different images with their corresponding taboo words, comparing the players' guesses (currently, guesses can only be 13 characters long), and storing all the information. The game server starts a game every 30 seconds: when a new player logs in, it waits until the next 30-second boundary to pair them with another player and start their game. This is done to make sure that players get paired at random and cannot cheat by logging in at the same time as their friends.

The current implementation is complete except that only 850,000 images are available for playing (rather than all images on the Web). We currently use a good label threshold of X=1.

Pre-Recorded Game Play

Our implementation does not require two people to be playing at the same time: a single person can play with a pre-recorded set of actions as their "partner." This set of actions is recorded at an earlier time when two other people were playing the game simultaneously. For each image, every guess of each partner was recorded, along with timing information. We refer to the set of pre-recorded actions as the "bot." Having pre-recorded game play is especially useful when the game is still gaining popularity. When there are few players, only a single person will usually be playing the game at a time.

Notice that pre-recorded game play does not necessarily stop the labeling process. If the single player and the bot agree on the label that was agreed on when the actions were recorded, we can increase our confidence regarding that label. If the single player and the bot match on another word, we get a brand new label.

Cheating

It is imperative that partners not be able to communicate with each other; otherwise agreeing on an image would be trivial. Similarly, players could cheat by being partnered with themselves or by a large group of them agreeing on a unified strategy (for instance, all the players could agree to type "a" on every image; this could be achieved by posting this strategy on a popular website). The current implementation has several mechanisms in place to counter such cheating.

Notice first that no form of cheating is very likely: the game is meant to be played by hundreds, if not thousands, of people at once, most of which will be in distributed locations. Since players are randomly paired, they will have no information about who their partner is, and they will have no way to previously agree on a strategy. The probability of two cheaters using the same strategy being paired together should be low.

That being said, several additional steps are taken to minimize cheating. First, IP addresses of players are recorded and must be different from that of their partner to further prevent players from being paired with themselves. Second, to counter global agreement of a strategy (e.g., "let's all type 'a' for every image"), we use pre-recorded game-play. Massive global agreement of a strategy can be easily detected by measuring the average time in which players are agreeing on images: a sharp decrease in this average time should indicate agreement on a strategy. If a massive agreement strategy is in place, having a majority

of the "players" be bots acting out pre-recorded sets of actions will make cheating impossible. Once people realize that the massive agreement strategy doesn't work, they will probably stop using it and we can lessen the use of pre-recorded game play.

An alternative mechanism that can be implemented is to enforce taboo words across an entire session. A pair's answer to an image could become a taboo word for the duration of their session together. This, coupled with a good label threshold greater than one ($X > 1$) would also prevent global agreement of a strategy from corrupting our labels. If the strategy was to always type "a" on an image, it would only work for the first image, as "a" would become a taboo word for the duration of the session. If the strategy was something more complicated, like "type 'one' for the first image, 'two' for the second, etc", then the labels couldn't be corrupted because of the good label threshold: in order for "one" to become a label for a certain image, the image would have to occur X times as the first image in games played by cheaters using the same strategy.

We also remark that some amount of cheating is acceptable for certain applications of our labels. In the case of image search, for instance, we expect to see an improvement over the current techniques even if some of the labels are meaningless. The current techniques, which associate most of the text on a website to each image, generate several inappropriate labels.

Selecting the Images

We believe that the choice of images used by the ESP game makes a difference in the player's experience. The game could perhaps be less entertaining if all the images were chosen from a single site and were all extremely similar.

The most basic strategy for picking the images is to select them at random from the Web using a small amount of filtering. This is the strategy employed in the current implementation of the game, except for two minor differences. First, once an image is randomly chosen from the Web, we reintroduce it into the game several times until it is fully labeled. Second, rather than picking the images from the Web in an online fashion, we collected 850,000 images in advance and are waiting until those are fully labeled to start with the whole Web. The images were chosen using http://random.bounceme.net, a website that selects a page at random from the Google database. "Random Bounce Me" was queried repeatedly, each time collecting all JPEG and GIF images in the random page, except for images that did not fit our criteria: blank images, images that consist of a single color, images that are smaller than 20 pixels on either dimension, and images with an aspect ratio greater than 4.5 or smaller than 1 / 4.5. This process was repeated until 850,000 images were collected. The images were then rescaled to fit the game applet. For each session of the game, we choose 15 *different* images from our set of 850,000.

Spelling

The game server is equipped with a 73,000-word English dictionary that alerts players when they have misspelled a word. It does so by displaying the misspelled word in yellow rather than in white in the "Your Guesses" area (Figure 2). This is useful when one of the players doesn't know how to spell a word, or makes a typing mistake.

Extension: Context-Specific Labels

Presenting images randomly selected from the Web to a wide-ranging audience is likely to result in labels that are general. There might be more specific labels for some images, which could be obtained if the correct population of users was doing the labeling. For example, when presented with pictures of faculty members at a certain university, the average ESP game player might enter general words such as man, woman, person, etc. However, if the users playing the ESP game were all students at that university, they might input faculty member names.

In order to generate these kinds of specific labels for certain categories of images, we suggest the usage of "theme rooms" for the ESP game. These more specific theme rooms can be accessed by those who wish to play the ESP game using only certain types of images. Some players might want images from certain domains or with specific types of content (e.g., images of paintings). Images for these theme rooms can be obtained using either Web directories or the labels generated during the "general category" ESP game. The labels generated in such theme rooms are likely to be more specific and thus more appropriate for certain applications. In the "art" theme room, for instance, images of paintings could be labeled with the names of their creators and maybe even the year in which they were made.

Notice, however, that proper general labels will already provide a vast improvement for many applications. For the visually impaired, for example, knowing than an image has a man in it is better than not knowing anything about it. The current version of the game implements the "general category" ESP game.

Inappropriate Content

A small percentage of images on the Web are inappropriate for children (e.g., pornography). This means that the "general category" ESP game may also be inappropriate for children. Our suggested solution to this problem uses theme rooms as described above: children would only be allowed to play the "children's version" of the game. This version would obtain its images from the general category ESP game. Only images that have obtained a certain number of labels can go to the children's version; all of the labels for these images must be "safe." To be more rigorous, we can combine this with text-based filtering. Images coming from web pages containing inappropriate words, etc., would not be allowed. We believe these strategies would prevent inappropriate images from reaching the children's version.

Notice also that the actual percentage of freely accessible images on the Web that are pornographic is small (the exact number of such images is hard to estimate, and varies depending on the source). Our game will only display images that are freely accessible.

Evaluation

We present data supporting our claims that people will want to play the ESP game and that the labels it produces are useful. In general it is difficult to predict if a game will become popular. One approach, which we followed early on, is to ask participants a series of questions regarding how much they enjoyed playing the game. Our data were extremely positive, but we follow a different approach in this paper: we present usage statistics from arbitrary people playing our game online.

We also present evidence that the labels produced using the game are indeed useful descriptions of the images. It's not the case that players *must* input words describing the images: players are never asked to describe anything. We show, however, that players do input words describing the images. To do so, we present the results of searching for randomly chosen keywords and show that the proportion of appropriate images when searching using the labels generated by the game is extremely high. In addition, we present the results of a study that compares the labels generated using the game to labels generated by participants that were asked to describe the images.

Usage Statistics

For the past few months we have been running the ESP game over the Web, allowing independent users to sign up for accounts and play the game. The statistics here presented are for the time period starting October 5, 2003 and ending on September 5, 2004. A total of 29,435 people played the game during this time, generating 4,015,231 labels for 850,000 different images. Over 80% of the people played in more than one occasion (by this we mean that more than 80% of the people played on different dates). Furthermore, some people play an average of over 20 hours a week.

We believe these numbers provide evidence of how much fun the game is: over 4 million labels were collected with only 29,435 players, some of which spent over 20 hours a week playing the game! We emphasize that there is no monetary compensation from playing the game and players are there simply for the fact that they find the game entertaining.

Labeling Rate

The usage statistics also allowed us to determine the rate at which images are labeled. The average number of images labeled per minute was 3.89 (std. dev. = 0.69). At this rate, 5,000 people playing the ESP game 24 hours a day would label all images on Google in 31 days. This 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: on a recent weekday afternoon, the authors found 107,000 people playing in Yahoo! Games, 115,000 in MSN's The Zone and 121,000 in Pogo.com. A typical game on these sites averages over 5,000 people playing at any one time.

The time it takes players to agree on an image depends on the number of taboo words associated with the image. Our calculation of the labeling rate, however, is independent of the number of taboo words: every session of the game has roughly the same number of images with 0 taboo words, the same number of images with 1 taboo word, etc.

Quality of the Labels

We give evidence that players do input appropriate labels for the images, even though they are only trying to maximize their score in the game. We give the results of three distinct evaluations. The first is a measure of precision when using the labels as search queries. The second compares the labels generated using the game to labels generated by experimental participants asked to describe the images. The third consists of asking experimental participants whether the labels generated using the game were appropriate with respect to the images. (In addition to the results of our evaluations, the reader is also encouraged to see the prototype search engine based on the labels generated by the game that can be found at http://www.espgame.org.)

Search Precision. We performed an evaluation similar to that of (Lempel and Soffer 2000): we examined the results of searching for all images associated to particular labels. To do so, we chose 10 labels at random from the set of all labels collected using the game. We chose from labels that occurred in more than 8 images.
Figure 3 shows the first 14 images having the label "car" associated with them: all of them contain cars or parts of cars. Similar results were obtained for the other 9 randomly chosen labels: dog, man, woman, stamp, Witherspoon (as in "Reese Witherspoon"), smiling, Alias (the TV show), cartoon, and green.

100% of the images retrieved indeed made sense with respect to the test labels. In more technical terms, the *precision* of searching for images using our labels is extremely high. This should be surprising, given that the labels were collected not by asking players to enter search terms, but by recording their answers when trying to maximize their score in the ESP game.

Comparison to Labels Generated by Participants Asked to Describe the Images. To further determine whether the words that players agreed on were actually describing the image, we asked 15 participants to input word descriptions of images and we compared their descriptions to the labels generated by the game. The participants were between 20

and 25 years of age and had not played the game during the trial period.

Method. Twenty images were chosen at random out of the first 1,023 images that had more than 5 labels associated to them by the game (1,023 is the number of images that had more than 5 labels associated to them at the time this experiment was performed). All 15 participants were presented with each of the 20 images in randomized order. For each image, the participant was asked to do the following:

> Please type the six individual words that you feel best describe the contents of this image. Type one word per line below; words should be less than 12 characters.

Results. The results indicate that indeed players of the ESP game were generating descriptions of the images. For all (100%) of the 20 images, at least 5 (83%) of the 6 labels produced by the game were covered by the participants (i.e., each of these labels was entered by at least one participant). Moreover, for all (100%) of the images, the three most common words entered by participants were contained among the labels produced by the game.

Figure 3. First 14 images that had the label "car" associated to them by the ESP game (some of them have been slightly cropped to fit the page better).

Manual Assessment of the Labels. In addition to the previous evaluations, we had 15 participants complete a questionnaire about the quality of the labels generated using the game. The participants were chosen as independent raters because they had not played the ESP game. None of the participants of this evaluation took part in the previous one and vice-versa. All participants were 20 to 25 years of age.

Method. Twenty images were chosen at random out of the first 1,023 images that had more than 5 labels associated to them by the game. All 15 participants were presented with each of the 20 images in randomized order. For each image the participant was shown the first six words that pairs agreed on for that image during the game, as shown in Figure 4. For each of the 20 image-word sets they were asked to answer the following questions:

1. How many of the words above would you use in describing this image to someone who couldn't see it.
2. How many of the words have **nothing** to do with the image (i.e., you don't understand why they are listed with this image)?

Dog
Leash
German
Shepard
Standing
Canine

Figure 4. An image with all its labels.

Results. For question 1, the mean was 5.105 words (std. dev. 1.0387), indicating that a majority (or 85%) of the words for each image would be useful in describing it. The mean for question 2 was 0.105 words (std. dev. 0.2529), indicating that for the most part subjects felt there were few (1.7%) if any labels that did not belong with each image.

Previous Techniques for Processing Images

To this point, we have presented a method for labeling images on the Web and we have presented evidence that it does indeed produce high-quality labels. There are a variety of other techniques for processing images on the Web, all of which are different in nature from ours. We now survey the different techniques and contrast them with our method.

Computer Vision

There has been considerable work in computer vision related to automatically labeling images. The most successful approaches *learn* from large databases of annotated images. Annotations typically refer to the contents of the image, and are fairly specific and comprehensive (and building these kinds of databases is expensive). Methods such as (Barnard, Duygulu and Forsyth 2001 and Barnard and Forsyth 2001) cluster image representations and annotations to produce a joint distribution linking images and words. These methods can

predict words for a given image by computing the words that have a high posterior probability given the image. Other methods attempt to combine large semantic text models with annotated image structures (Duygulu et al. 2002). 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 (Duygulu et al. 2002) only gave reasonable results for 80 out of their 371 vocabulary words (their evaluation consisted of searching for images using the vocabulary words, and only 80 of the words resulted in reasonable images).

A different line of research attempts to find specific objects in images. (Schneiderman and Kanade 2002), for instance, introduced a method to locate human faces in still photographs. These algorithms are typically accurate, but have not been developed for a wide range of objects. Additionally, combining algorithms for detecting specific objects into a single general-purpose classifier is a non-trivial task.

The ESP game provides a possible solution to the image-labeling problem, but having a computer program that can label images remains a more important goal. One application of the ESP game is in the creation of such a program: the limitations of the current computer vision techniques partly arise from the lack of large databases of annotated images, which could be constructed using methods similar to our game.

Text and Content-Based Image Retrieval

The World Wide Web contains millions of images and finding effective methods to search and retrieve from among them has been a prevalent line of research, both academically and in industry. Text-based image retrieval systems such as Altavista's annotate images with text derived from the HTML documents that display them. The text can include the caption of the image, text surrounding the image, the entire text of the containing page, the filename of the containing HTML document, and the filename of the image itself. More recent proposals such as (Lempel and Soffer 2000) also make use of the link structure of the Web to assign "authority" values to the images. Images that come from more authoritative web pages (e.g., pages with higher PageRank) are displayed before images coming from less authoritative pages. This improves the quality of the results by typically showing more relevant images first. Another possibility that has been explored involves combining text-based systems with computer vision techniques as in webSeek (http://www.ctr.columbia.edu/webseek). This approach allows different types of queries to be processed (e.g., similarity queries).

Most image search engines rely on the images being related to the text in the web page. This heuristic works in part because almost any simple query matches thousands of images on the Web — the results appear satisfactory even if the search engine finds only half of all the relevant images. However, not all queries are handled properly, and finding a particular image on the Web can be impossible. We claim that our game can improve the quality of image retrieval systems by providing meaningful labels that are independent of the content of the websites.

Inappropriate Content Filters

Inappropriate content filters such as N2H2 (http://www.n2h2.com) attempt to block certain images from being displayed. Typically these filters try to block pornographic sites from reaching children at home or employees in the workplace. Since computer vision techniques for this purpose are not highly accurate (Fleck, Forsyth and Bregler 1996), content filters usually analyze the text inside web pages to determine whether they should be blocked.

Most filters are reasonably accurate, but have several flaws. First, they only work for a few languages and in most cases only work for pages in English. Second, they work poorly when the pages don't contain any "incriminating" text: a page with a nude image and nothing else in it would not be correctly identified. For this reason, in order to ensure that inappropriate content does not get posted, dating services and websites that allow users to post images have to hire people to look over every single picture to be posted. Third, content filters have to be constantly updated: imagine what happens when a new porn star named Thumbelina comes out; suddenly every search for "Thumbelina" will return pornography. Google Image Search offers a content filter (called SafeSearch), which attempts to block all inappropriate images from being displayed in their search results. At the time of writing this paper, a query for "interracial" returns several inappropriate images (and a more direct query like "wet tshirt" returns even more inappropriate results). We argue that having proper labels associated to each freely available image on the Web would improve content filtering technology.

Using our Labels

This paper is mostly concerned with obtaining appropriate labels for images, and not with how these labels should be used. In the case of image search, building the labels into the current systems is not difficult, since they can be thought of as HTML captions or text appearing right next to the image. This naïve strategy would already signify an improvement over the current techniques, as there would be more useful data to work with. More intelligent techniques could be conceived, such as assigning a higher weight to labels coming from the ESP game as opposed to regular HTML captions, or a numerical weight based on the "good label threshold". However, arriving at an optimal strategy for using the labels is outside the scope of this paper and is left as future work.

In the case of providing textual descriptions for the visually impaired, using the labels is slightly less trivial. Our game produces labels, not explanatory sentences. While keyword labels are perfect for certain applications such as image search, other applications such as accessibility would benefit more from explanatory sentences. Nevertheless, having meaningful labels associated to images for accessibility purposes is certainly better than having nothing. Today's screen-reading programs for the visually impaired use only image filenames and HTML captions when attempting to describe images on the Web — the majority of images on the Web, however, have no captions or extremely descriptive filenames (Milliman). We propose that all the labels be available for use with screen readers and that users determine themselves how many labels they want to hear for every image. Again, extensive tests are required to determine the optimal strategy.

Conclusion

The ESP game is a novel interactive system that allows people to label images while enjoying themselves. We have presented evidence that people will play our game and that the labels it produces are meaningful. Our data also suggest that 5,000 people playing the game for 24 hours a day would enable us to label all images indexed by Google in a matter of weeks. This is striking because 5,000 is not a large number: most popular games on the Web have more than 5,000 players at any one time. Having proper labels associated to each image on the Web could allow for more accurate image retrieval, could improve the accessibility of sites, and could help users block inappropriate images.

Although the main application of the ESP game is to label images, our main contribution stems from the way in which we attack the labeling problem. Rather than developing a complicated algorithm, we have shown that it's conceivable that a large-scale problem can be solved with a method that uses people playing on the Web. We've turned tedious work into something people want to do.

Perhaps other problems can be attacked in a similar fashion. For instance, the ESP game can be used, with only minor modifications, to attack the problem of labeling sound or video clips (i.e., there is nothing inherent about images). Of course, the success of these variations of the ESP game depends on whether people will enjoy playing them. The same mechanism can also be used to attach labels to images in other languages. Other problems that could be solved by having people play games include categorizing web pages into topics and monitoring security cameras. One of the main stumbling blocks for installing more security cameras around the world is that it's extremely expensive to pay humans to watch the cameras 24 hours a day. What if people played a game that could alert somebody when illegal activity was going on? One could imagine many other applications. We hope that others may be inspired to develop systems similar in approach to the ESP game or the Open Mind Initiative.

Acknowledgements

We thank Lenore and Manuel Blum for their unconditional support and advice. We also thank Aditya Akella, Sonya Allin, Scott Hudson, Steven Katz, Lenore Ramm, Chuck Rosenberg, David Stork and the anonymous CHI'04 and KCVC'05 reviewers for insightful comments. We thank Shingo Uchihashi for allowing us to use his image retrieval tool. This work was partially supported by the National Science Foundation (NSF) grants CCR-0122581 and CCR-0085982 (The Aladdin Center). Luis von Ahn is partially supported by a Microsoft Research Graduate Fellowship, and Laura Dabbish is partially supported by a National Defense Science and Engineering Graduate Fellowship.

References

Barnard, K., Duygulu, P., and Forsyth, D. A. Clustering Art. *IEEE conference on Computer Vision and Pattern Recognition*, 2001, pages 434-441.

Barnard, K., and Forsyth, D. A. Learning the Semantics of Words and Pictures. *International Conference of Computer Vision*, 2001, pages 408-415.

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*, 1996, Vol. 19 No. 4.

Duygulu, P., Barnard, K., de Freitas, N., and Forsyth, D. A. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *7th European Conference on Computer Vision*, 2002, pages 97-112.

Fleck, M. M., Forsyth, D. A., and Bregler, C. Finding Naked People. *ECCV*, 1996, pages 593-602.

Lempel, R. and Soffer, A. PicASHOW: Pictorial Authority Search by Hyperlinks on the Web. *WWW10*.

Milliman, R. E. *Website Accessibility And The Private Sector*. www.rit.edu/~easi/itd/itdv08n2/milliman.html

O'Connor, B. and O'Connor, M. Categories, Photographs & Predicaments: Exploratory Research on Representing Pictures for Access. *Bulletin of the American Society for Information Science* 25.6, 1999, page 17-20.

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

Stork, D. G. and Lam C. P. Open Mind Animals: Ensuring the quality of data openly contributed over the World Wide Web. *AAAI Workshop on Learning with Imbalanced Data Sets*, 2000, pages 4-9.

Stork, D. G. The Open Mind Initiative. *IEEE Intelligent Systems & Their Applications, 14-3*, 1999, pages 19-20.