IMPROVING IMAGE SEARCH WITH PHETCH

Luis von Ahn, Shiry Ginosar, Mihir Kedia, and Manuel Blum

Computer Science Department, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15213
biglou@cs.cmu.edu, {shiry,majin,mblum}@cmu.edu

ABSTRACT

Keyword-based image search engines are hindered by the lack of proper labeling for images in their indices. In many cases the labels do not agree with the contents of the image itself, since images are generally indexed by their filename and the surrounding text in the webpage. Another popular approach to image search, content based image retrieval, suffers from a gap between the available low level data and the semantic needs of user searches. To overcome these problems we suggest human annotation of images with natural language descriptions. To this end we present Phetch, an engaging multiplayer game that allows people to attach accurate explanatory text captions to arbitrary images on the Web. People play the game because it is fun, and as a side effect we collect valuable information that can be applied towards improving image retrieval. Furthermore, the game can also be used for other novel applications.

Index Terms— Distributed knowledge acquisition, Accessibility, Web-based games, Image retrieval.

1. INTRODUCTION

Current image search engines employ one of two approaches: keyword based or content based image retrieval. Neither approach performs optimally. Keyword based search engines index images by using textual data such as filenames, image captions and/or adjacent text on the Web page. Unfortunately, such data can be insufficient or even deceptive, making it hard to return accurate search results [10]. Moreover, Hollink *et al.* find that while people describe images that they are searching for in abstract, perceptual terms, they translate these into much more specific terms when performing a keyword search. This may be due to the fact that when using traditional search engines which rely on a small number of tags for each image, specific queries lead to higher precision results than abstract queries [7]. These findings suggest that by labeling images with natural language text we may be able to capture some of the semantic meaning of images and allow users to search using more abstract perceptual terminology.

Content based image retrieval (CBIR) relies on image content extracted by computer vision methods instead of textual labels. Unfortunately, the performance of these systems does not match the needs of the average user [5]. While users typically search for images that show a certain object or have a certain meaning, CBIR metadata relies on low level image features. Smeulders *et al* argue that the aim of content based methods must be to bridge this mismatch between the richness of user semantics and the simplicity of technology, which they term the *semantic gap* [8]. Natural language captions for images could provide the missing semantics needed to close this gap.

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. On the one hand, humans have little difficulty describing the contents of images, although they typically would not find this task engaging. On the other hand, people tend to spend considerable amounts of time involved in activities they consider "fun". Thus, we address the image captioning problem by creating a fun game that produces the data we aim to collect. This method is similar to the one taken by the ESP Game [1] (a.k.a., Google Image Labeler).

To this end, we implemented Phetch, an online multiplayer game which provides natural language descriptions of images as a side effect of game play. Phetch was originally presented in [2] as a system which improves Web accessibility. This paper reviews previous work on image captioning, describes Phetch in detail and continues to demonstrate how Phetch can be used to improve image search in many ways.

2. OTHER APPROACHES TO IMAGE CAPTIONING

Most attempts to attach textual metadata to images may be divided into two main heuristics. The first heuristic focuses on automatically captioning images. Srihari *et al.*, for example, exploit the fact that images on the web are often accompanied by surrounding text. They improve image retrieval by combining image and text processing [10]. Unfortunately, while automatic methods do not require human intervention, the quality of annotations they provide cannot be compared to human-generated captions [5]. The second heuristic provides interfaces for manual annotation by humans. Here, some researchers attempt to collect data

that can be directly used to build ontology. Volkmer *et al.* present a collaborative image annotation tool in which users are allowed to attach any of 39 distinct concepts to each image [11]. However, there is no incentive given to annotators other than the benefit of participating in a scientific experiment. Since requesting humans to annotate data without proper incentive is often a long and expensive effort [8], this method has an inherent scaling problem. Other researchers use data collected by popular tagging sites (e.g. www.flickr.com) in order to infer a faceted ontology in hindsight [9]. While social networked tagging sites provide some incentive for tagging, the correctness of data produced is often unverified and overwhelmingly, users only tag their own generated data rather than publicly available images. By approaching image captioning from a different angle, we can collect verifiable human captions for arbitrary images in a fast and inexpensive way, as was previously introduced by the ESP Game [1]. In contrast to ESP however, Phetch collects natural language descriptions instead of individual keyword labels. We show below that this type of annotation is far more descriptive.

3. PHETCH GAME MECHANICS

Phetch is designed as an online game played by three to five players. Initially, one of the players is chosen at random as the "Describer" while the others are "Seekers." The Describer is given an image and helps the Seekers find it by textually describing 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 text, the Seekers can find the image using an image search engine. However, they are not cued as to how to extract a search query from the given text. The first Seeker to find the image obtains points and becomes the Describer for the next round. The Describer also gains points if the image is found. Intuitively, by observing the Describer's text, we can collect natural language descriptions of images (see Figure 1).

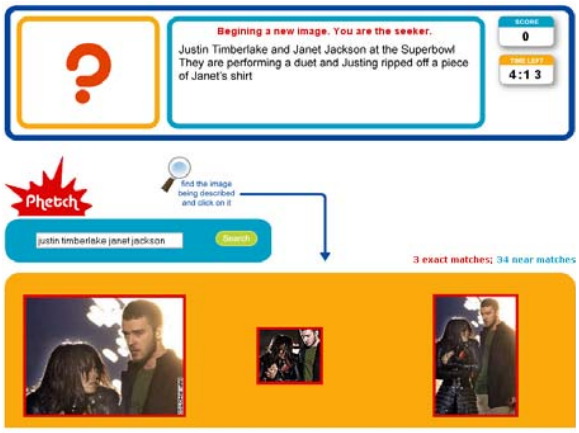

Figure 1. The Seeker's interface. The Describer's text is a natural language description of the correct image.

Each session of the game lasts five minutes during which the players go through as many images as they can. The Describer can pass, or opt out, on an image if they believe it is too difficult. In this case, the Describer gets a new image and is penalized by losing a small amount of points. To prevent Seekers from clicking on too many images, each wrong guess results in a strong point penalty. This penalty also ensures that the Describer's text is a reasonably sufficient description of the image, since Seekers will tend to not guess until they are certain.

4. IMPLEMENTATION DETAILS

We implemented a Phetch game server as a stand alone Java application that communicates with applet clients and a Perl-based image search engine, which we discuss below.

4.1. The Image Search Engine

The search engine given to the Seekers is a crucial component of the game for several reasons. First, the search space cannot be so large that it requires Seekers to filter through thousands of query results. Second, we must guarantee that the correct image is usually returned given a good query. The original design of Phetch [2] achieves both these properties by using a restricted search engine based on keywords collected with the ESP Game [1], which are associated with some hundreds of thousands of images.

In general, any reasonable image search engine can be used provided that two modifications are in place. First, the search space should be of the right magnitude. This does not mean that the entire index should only contain a small number of images, but that each session of the game can only be played on a subset of the indexed images. Second, in order to ensure proper game flow, the results should artificially contain the Describer image whenever a query is "accurate enough". An "accurate" query may be defined, for instance, by the percentage of the query words that also appear in the Describer's text so far. Note that a Seeker still needs to be provided with an accurate description in order to locate the correct image. This method of inserting images into the search results enables us to use previously unlabeled images as input to Phetch. Moreover, we can easily expand the engine's index to include these new images. If a Seeker is able to locate the desired image, we assign it the natural language description as a set of labels.

4.2. Use of Randomly Chosen Images

Phetch uses a collection of randomly chosen images obtained by a Web crawl. We regard these images as a subset of all images in the Web and assume that their aggregate bears no special context or meaning beyond the individual instances. Therefore seekers have no additional context knowledge to aid them in their search process.

5. EMULATING PLAYERS

As stated, Phetch requires three to five players: one Describer and two to four Seekers. In order to account for a single player, we can also pair up users with computerized players, or "bots," when needed. Bots use previously collected game data to emulate players. When simulating a Describer, a bot merely needs to replay an old description (in fact, we show below how this is also useful to further ensure accurate descriptions). To simulate a Seeker, a bot only guesses the correct image if the Describer's text contains a sufficient number of known keywords associated with the image (from previous sessions of the game).

6. DESCRIPTION ACCURACY

6.1. Ensuring Description Accuracy

In the following, we describe some of the strategies used to ensure the accuracy of entered descriptions.

Description testing: When simulating a Describer, the bot plays back a previously entered description to the Seekers. If a Seeker can still find the correct image, this is a significant indicator that the description is of high quality: two different people chosen at random were able to select the correct 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") to guarantee description accuracy within a given percentage.

Random pairing of the players: We randomly assign players to sessions. This helps prevent players from colluding with each other.

Success of the Seekers: We use the amount of time it takes the Seekers to find the correct image as an indicator of the description's quality. Furthermore, if the Seekers cannot find the image, we discard the Describer's text.

6.2. Experimental Results

In [2], experimental data showed that the descriptions collected using Phetch are precise and complete, and that they are an improvement over a list of accurate keywords collected from the ESP Game [1]. To determine this, a study was conducted in which participants were assigned to one of two conditions: PHETCH or ESP. Participants in each condition were asked to single out one image among other similar ones, based on either a natural language description collected using Phetch or a set of word labels from ESP. The experimental data showed that while 98.5% of the descriptions collected using Phetch were sufficient for finding the correct image, only 73% of ESP Game keyword cues led participants to successfully single out the correct image. This shows that natural language descriptions are far more descriptive than keyword labels.

7. USAGE STATISTICS

Phetch may theoretically produce useful data, but if it is not engaging people will not play and output will not be produced. Hence, it is crucial to test our claim that Phetch is entertaining. To do so, we enlisted test players to play the game by offering random players from another gaming site (www.peekaboom.org) the opportunity to play for a two-month period. A total of 7,120 people played Phetch during this trial, generating 81,950 captions for images. Each session of the game lasted five minutes and, on average, produced captions for 6.8 images. Roughly 75% of the players returned to play more than once, and some played for over 110 hours. We believe this data shows that the game is indeed enjoyable.

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.

8. PHETCH OUTPUT USAGE

8.1. Improving Web Accessibility

One of the major accessibility problems of the Web 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 web page into synthesized speech. Although screen readers are helpful, they cannot determine the contents of images that do not have a descriptive html alt caption. Unfortunately, many images online are not accompanied by proper captions and therefore are inaccessible to the blind [3].

Phetch was originally designed to aid accessibility by use of a centralized database to store all collected captions that can later be used as alt tags [2]. Alternatively, Bigham *et al.* propose sending all images that cannot be automatically captioned to be human-annotated by Phetch players. Additionally, they present a method to automatically assess the quality of Phetch descriptions for alt tag usage, relying on the content of the text as well as on the context of the site containing the image [3,4].

8.2. Improving Image Retrieval

As stated above, the natural language annotations collected by Phetch could be used to improve upon both keyword and CBIR methods by providing the missing high level semantic data. In addition, Yee *et al.* show that such sentinel descriptions can be semi-automatically converted to hierarchical metadata categories in order to be used in faceted metadata image retrieval. In their implementation they convert similar natural language descriptions of images

using an algorithm that compares the words in the descriptions to higher-level category labels in WordNet [12, 6]. Using such heuristics, images described by Phetch can be form a collection with attributed hierarchical metadata which allows for faceted metadata image search.

8.3. Human-Powered Descriptive Search

Phetch can also be used to provide free, real time assistance for people unable to find the images which they are looking for, by recruiting others to search in their place. When an image search engine user types a complete description of a desired image, his description can be fed to a currently active game of Phetch as a "bonus round." In this round, all players act as Seekers and aim to find an image based on the description given. Whenever they find an image that matches the description it can be sent back to the user. If the describing individual accepts one of these as the target of his search, the Seeker who found it wins the round.

In essence, this would provide a service similar to Google Answers, but with an almost immediate turnaround time. In Google Answers, users submit questions for which they cannot find an answer (e.g., "How can I get to Antarctica?"), and for a monetary fee, other people attempt to find answers for them over a period of several hours or days. Phetch could provide a similar service for images, except that it would be free and immediate: the players would perform the searches as a part of the game.

9. CONCLUSION

This paper describes a novel way to improve image captioning and image retrieval methods. In order to allow the use of semantically rich metadata, we suggest annotating images with natural language descriptions. Instead of developing computer vision algorithms or CBIR methods, we make use of a game which allows people to correctly describe images. Even if we could pay people to annotate images, it would incur a great cost and require manual accuracy verification for each and every caption. Phetch generates these captions for free. Moreover, the results from this game are guaranteed to be statistically accurate. Phetch has generated thousands of image captions, and if it were placed on a popular gaming site, it could create captions for every image on the Internet within months. It is our intention to integrate Phetch with a large-scale search engine in order to improve the state of the art in image retrieval.

10. ACKNOWLEDGEMENTS

We thank Laura Dabbish, Susan Hrishenko, and Roy Liu for their insightful comments. This work was partially supported by the National Science Foundation (NSF) grants CCR-0122581 and CCR-0085982 and by a generous gift from Google, Inc. Luis von Ahn was also partially supported by a Microsoft Research Fellowship and a MacArthur Foundation Fellowship.

10. 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] von Ahn, L., Ginosar, S., Kedia, M., Liu, R., and Blum, M. Improving Accessibility of the Web with a Computer Game. In *ACM Conference on Human Factors in Computing Systems (CHI),* 2006, pp. 79-82.

[3] Bigham, J., Kaminsky, R., Ladner, R., Danielsson, O. and Hempton, G., Webinsight: Making Web Images Accessible. In *ACM SIGACCESS Conference on Computers and Accessibility (ASSETS),* 2006, pp. 181-188.

[4] Bigham, J., Increasing Web Accessibility by Automatically Judging Alternative Text Quality. In *Conference on Intelligent User Interfaces (IUI),* 2007.

[5] Eakins, J. Towards intelligent image retrieval. In *Pattern Recognition*, 35(1):3–14, 2002.

[6] Fellbaum, C. editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[7] Hollink, L., Schreiber, A.Th., Wielinga, B., Worring, M. Classification of User Image Descriptions. In *International Journal of Human Computer Studies*, (61/5):601-626, 2004.

[8] Smeulders, A., Worring, M., Santini, S., Gupta, A., and Jain, R. Content-based image retrieval at the end of the early years. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12), 2000.

[9] Schmitz, P. Inducing ontology from Flickr tags. In *Workshop on Collaborative Web Tagging, World Wide Web Conference*, 2006.

[10] Srihari, R., Zhang, Z., and Rao, A. Intelligent Indexing and Semantic Retrieval of Multimodal Documents. In *Information Retrieval*, 2(2/3):245-275, 2000.

[11] Volkmer, T., Smith, J., Natsev, A., Campbell, M., Naphade, M., A Web-based System for Collaborative Annotation of Large Image and Video Collections. In *ACM Conference on Multimedia*, 2005, pp. 892-901.

[12] Yee, K.-P., Swearingen, K., Li, K., and Hearst, M. Faceted metadata for image search and browsing. In *ACM Conference on Human Factors in Computing Systems*, 2003, pp. 401–408.