An Exploratory Study of Human Performance in Image Geolocation Tasks

Sneha Mehta, Chris North, Kurt Luther
Department of Computer Science and Center for Human–Computer Interaction
Virginia Tech, Blacksburg, VA, USA
{snehamehta, north, kluther}@vt.edu

Abstract

Identifying the precise location where a photo was taken is an important task in domains ranging from journalism to counter-terrorism. Yet, geolocation of arbitrary images is difficult for computer vision techniques and time-consuming for expert analysts. Understanding how humans perform geolocation can suggest rich opportunities for improvement, but little is known about their processes. This paper presents an exploratory study of image geolocation tasks performed by novice and expert humans on a diverse image dataset we developed. Our findings include a model of sensemaking strategies, a taxonomy of image clues, and key challenges and design ideas for image sensemaking and crowdsourcing.

1 Introduction

Image sensemaking, the process of researching and identifying unfamiliar subject matter in an image with little or no context, is an important task in many domains. For example, scientists must recognize animals or plants in remote or satellite imagery, archivists must identify visual materials in their collections, and intelligence analysts must interpret photos of terrorist activity.

In this work, we focus on a key subtask of image sensemaking, *image geolocation*, whose goal is to identify as precisely as possible the geographic location where a photo was taken. Computer vision (CV) has been used extensively for automatic image geolocation (e.g. (Hays and Efros 2008; Weyand, Kostrikov, and Philbin 2016)), but these techniques are constrained by the selection and quantity of geotagged reference images. Humans perform manual image geolocation by recognizing salient clues and synthesizing outside information (e.g. (Crisman 2011; Higgins 2014)). Their techniques could be scaled up with crowds or used to enhance CV algorithms, but they have not been empirically studied.

To address this gap, we present an exploratory study of human image geolocation approaches. We conducted a study of 15 novice and expert participants who performed a series of image geolocation tasks on a diverse dataset we developed. Our contributions include: 1) a model of participants' high level strategy; 2) a rich description and taxonomy of image clues used by participants; 3) an enumeration of key challenges in human image geolocation; and 4) a set of design considerations for supporting image geolocation, sensemaking, and crowdsourcing.

Figure 1: Sample image panoramas for different location factors. Clockwise from top left: Bankok, Thailand (Tropical, high, non-English); Nevada, USA (Desert, low, English); Kütahya, Turkey (Mediterranean, low, non-English); Kempton, Tasmania (Temperate, low, English).

2 Related Work

Researchers have used crowdsourcing to answer questions about images (Noronha et al. 2011; Bigham et al. 2010), but these approaches generally leverage crowds' everyday knowledge. To identify unfamiliar images, e.g. citizen science tasks (Lintott and Reed 2013), some systems use tutorials to help crowds learn what to look for. We build on these efforts by studying how people make sense of images without specialized instructions or familiarity with the contents.

Image geolocation has been researched in computer vision using methods such as IM2GPS (Hays and Efros 2008), which assigns the coordinates of the closest match to the query from similar images retrieved from millions of geotagged Flickr photos. PlaNet (Weyand, Kostrikov, and Philbin 2016) uses a similar approach but treats the problem as a classification task. Another approach (Lin et al. 2015) makes additional use of satellite aerial imagery that provides a more complete coverage, because reference ground-level photos are still elusive for many parts of the world. We aim to complement these pixel-based approaches by understanding which clues humans are uniquely skilled at investigating.

Other researchers have studied how people navigate unfamiliar physical spaces, known as wayfinding (Montello and Sas 2006). We investigate wayfinding practices in virtual environments for the purposes of image geolocation.

3 Study and Methods

3.1 Apparatus and Dataset

To evoke a variety of geolocation challenges and strategies from participants, we systematically generated a diverse image dataset based on 1) geographic biome, 2) population density, and 3) language. Biomes (NASA 2003) included the world's six major biomes: Tundra, Boreal, Temperate, Mediterranean, Desert, and Tropical. Population density (NASA 2010) included two grades: low ($< \frac{100}{m^2}$) and high ($\geq \frac{100}{m^2}$). Language was divided into English and non-English to reflect the fluency of our participants. Allowing for all combinations of these factors, we generated 24 triplets (e.g. Tundra, high density, non-English). We then randomly selected a geographic location (GPS coordinates) satisfying each triplet's conditions and a corresponding image panorama in Google Street View (Figure 1). Locations with obvious landmarks (e.g. Eiffel Tower) were ruled out.

Participants attempted to identify these locations in Geoguessr[1], an online game where players are "dropped into" a Google Street View panorama and challenged to identify their location as precisely as possible on a map. Players can zoom, rotate, or move through the panorama and use a digital compass. Because real-world geolocation is typically performed on static images, we focused on how participants analyzed the image content, not how they moved around.

3.2 Participants and Procedure

Fifteen participants were recruited for this study. We recruited nine experts from an online Geoguessr community[2] to elicit advanced techniques and strategies. These experts lived in four countries with English as an official language (US, UK, New Zealand, and Singapore) and reported playing an average of 125 Geoguessr games each. Experts were compensated 10 USD each. We recruited six novices, mostly students and US residents with no previous geolocation experience, to understand how naïve users such as crowd workers would perform. Participants were aged 18–35 (average=24). Ten were male and five female.

We asked each participant to solve multiple consecutive image geolocation challenges in Geoguessr. Novices were assigned two challenges: a common image (Nevada, USA) and a randomly assigned unique image from our dataset. Experts were assigned three challenges: the common image, a randomly assigned unique image, and a third image that was either unique or also assigned to novices. All participants had 20 minutes for each challenge but were allowed to finish early. Novices visited our research lab to participate, while experts participated remotely over Skype.

We used a think-aloud protocol in which participants externalized their thought processes and justifications by describing them verbally. We recorded these comments and their screen activity and took notes on our observations. After all the challenges, participants completed a post-survey and interview about their strategy, challenges, etc.

[1] http://www.geoguessr.com

[2] https://www.reddit.com/r/geoguessr

4 Results and Discussion

4.1 Sensemaking Strategies

We found that most novice and expert participants followed a strategy of iteratively narrowing down the location from a broad guess of a continent or multi-country region to a specific country, city, and street. They described using image clues to generate a mental model of potential locations that was typically quite diverse at first. Participants then refined this mental model as they explored the environment and researched potential clues.

Participants made sense of image clues using two sources of knowledge. *Internal* knowledge refers to participants' preexisting knowledge about locations, stemming from their cultural background, travel, education, and Geoguessr experiences, if any. *External* knowledge refers to knowledge that participants acquired during the geolocation task by researching image clues using outside information sources, such as search engines.

Internal knowledge tended to come into use early in the session, when participants were considering potential continents or countries. As the geographic space of possibilities was narrowed down, participants increasingly relied on external knowledge to pinpoint specific locations. Experts tended to have larger stores of internal knowledge to draw upon and were more effective at making use of it. Participant P9(E), an expert, summarized her strategy:

> (I) try to get a general idea of where in the world I am based on what road signs look like, what text I can see on signs, which side of the road cars are driving on, etc. If there is a distinctive terrain I might be able to guess that it's in a desert area or beside a large body of water. . . . Otherwise try to find either the name of the city if in a city, or signs indicating the distances to other cities if on a highway. And highway markers or street signs to get the exact location.

In contrast, novices tended to launch almost immediately into using external knowledge to refine their mental models. P3(N), a novice, explained:

> The idea was to find boards or markers that would let me Google and identify the area. For example, street signs, building signs, rail roads, information boards etc. I could make some rough guesses about the economic state of the areas, but that was not very helpful, except to serve as a form of confirmation once I started Googling.

Average identification time per image was 17.3 min (novices) vs. 11.6 min (experts). Average distance from the original location was 285.53 km (novices) vs. 28.2 km (experts). Thus, having a storehouse of internal knowledge and knowing how to map this onto image clues represented a key difference between experts and novices and might be instrumental to achieving accurate and timely results.

In the following sections, we delve into these differences by providing a rich description of the diverse range of image clues novices and experts used in their geolocation practices.

4.2 Clues from Internal Knowledge Sources

These clues were mainly used to identify a continent or a country of the image location.

Architecture Many expert participants, when presented with the image in Nevada, USA identified it as a western country based on the architecture of the houses.

Languages and scripts Knowledge of local languages and scripts and where they are prevalent was helpful. Participant P7(E), presented with an image in Modena, Italy, immediately recognized the language of the signs as Italian.

Some participants recognized the script, but not the precise language. P4(N), viewing an image in Taiwan, first narrowed down the location to somewhere in Asia. She then saw some non-English writing on the side of a truck and recognized the script:

> Okay, that is definitely Chinese writing. I've seen some before. I don't recognize the exact ones but Japan has a writing system Kanji which uses Chinese characters. Looks very similar to that.

This helped her narrow down to countries using Chinese scripts, ruling out many other Asian countries.

Driving rules Knowledge about whether the country has left- or right-side driving helps eliminate possibilities. P9(E), for an image in Nevada thought that the landscape looked somewhere in Australia. But then she saw a road sign on the right side of the road, which ruled out Australia.

Knowledge about the system of units (Imperial or Metric) can provide useful hints for identifying the country. P9(E), noting a speed limit of 25, concluded this was too slow for $\frac{km}{hr}$, so "we're somewhere desert-y in United States."

Sun position P10(E), when presented with an image in Nevada, USA, used the digital compass to note that the sun was shining from the south, so the image must be located in the northern hemisphere.

Animals P2(N), an American, saw a small group of cattle in the Nevada image and noted:

> Cattle. That makes me think I'm not in Utah or in the Southwest for that matter. I don't think people really keep cattle down there, it's more of a Midwest thing.

This participant, a novice, identified a potentially valuable clue, but drew the wrong conclusion because he relied on shaky internal knowledge. In the next section, we describe clues that participants used to find external knowledge and refine their mental models.

4.3 Clues from External Knowledge Sources

These clues typically helped participants narrow down a country to a specific city and street location, but required outside research with search engines or other tools.

Building signs Participant P5(N), for an image in Bankok, Thailand, wasn't familiar with the local language, but found a building with a name in English. Searching for the name online, he figured out that it was a bank based in Bankok, Thailand. He was able to identify both the country and the city with the help of that one crucial clue.

P1(N) searched for a building sign with the text "Ljusdal Tidning" (a Swedish newspaper) to learn that he was in Ljusdal, Sweden.

Road signs For the common Nevada image, many participants used the "Winnemuca Ranches" and "Lovelock" road signs to localize the image to Nevada. Fewer noticed the State Route 401 sign bore the state outline of Nevada.

P4(N) was completely stumped about an image in Taiwan, before she found two road signs in English, "Jingshan Road" and "Yangmingshan National Park," that she could look up.

Telecommunications signs For an image in Cape Town, P3(N) had narrowed down the location to South Africa and spotted a phone number on a board outside a business. She then searched "dialing code 082 south africa" and was able to find the telephone service provider's name, Vodacom. This reinforced her assumption—Vodacom is headquartered in South Africa, but serves 40+ African countries—but did not provide additional information.

P8(E), who had already narrowed down an image in Turkey to the general region of the Middle East, noticed a web address ending in `.tr`. From this top-level domain, he deduced that the country could be Turkey.

Other signage P3(N), for an image in Cape Town, South Africa, saw the phrase "Braai wood" (a kind of wood used for barbecues) painted on a wall. An online search told her that it is sold mainly in the UK and South Africa.

P10(E), who wasn't familiar with the local language for an image in Sri Lanka, found "Polgompola" (a place name), written on a bus. Looking it up helped him narrow down the country. Later, this same participant identified a trash can with the City of Honolulu logo that helped him place the location as Honolulu, Hawaii.

Landmarks on maps P9(E), examining the Nevada image, spotted some natural and man-made landmarks in the distance that she compared to Google Maps to locate her position on a highway:

> It is useful to know we're near a bridge. Once I get it down to what highway we're on, I can look for a river across and maybe a dam and look for something like that on the map.

In the next section, we generalize these clues into several key challenges faced by participants, and the strategies and tools they employed to deal with them.

4.4 Key Challenges

Unreadable signs Google Street View imagery, particularly in rural locations, can be low resolution, making it difficult to read signs and observe other relevant details. Participant P1(N), for the Nevada image, saw a sign that he thought might say, "Bureau of Reclamation", but was unable to read the key phrase that would locate it.

P4(N), encountering similarly blurred signs, used an online image sharpening tool to make a screen capture of the sign sharper and clearer. The phrase "picnic area" came into focus, but other words remained indecipherable.

Unfamiliar scripts Participants struggled to make sense of text in unfamiliar scripts and languages, especially when non-English characters could not be easily retyped and searched or translated. P8(E) commented, "Turkey was the most difficult as no signs were in English which is the only language I understand." Similarly, P2(N) was overwhelmed by signs in an unfamiliar language for an image in Sri Lanka. His guesses for the country ranged from Thailand to India.

P4(N), for an image in Taiwan, could not read the Chinese writing on a sign, so she attempted to use an online image translator to decipher a screen capture of the sign. However, the tool's OCR failed to recognize the Chinese characters.

Fixation Preconceived notions and misleading clues caused participants to go astray at times. For an image in Nevada, P5(N) became fixated on the idea that the location was in Colorado. He later spotted a road sign and misread it as "Route 40" instead of "Route 401" which also happened to pass through Colorado, reinforcing his wrong notion. Consequently, he wasted time and his final guess was far from the actual location.

P2(N) reflected on a similar experience:

> Probably the most difficult part was getting caught with like a red herring . . . you know bad lead type thing. Like I went all the way down that one road trying to read things and I just kept going.

Abundance or lack of information A lack of signage, especially in rural images, posed challenges for participants. P7(E), asked about which image was the most difficult to geolocate, picked an image in the Italian countryside:

> There were no road signs to anywhere major, and no road numbers, so it would have taken me much longer if I had not used Google to search for a place I found.

Other participants were overwhelmed by the large network of streets and abundant signage around their locations. P4(N), referring to an image in Taiwan, said:

> (It) was harder to locate because there were a ton more streets around it on the map, so I had to really narrow down the name of the lane and then reason out where I even was on that lane.

We conclude with design considerations for systems to address these key challenges.

4.5 Design Considerations

We envision a new class of image geolocation systems, informed by our findings, that combine crowdsourcing and computer vision to enable faster and more accurate results than either approach alone. Design considerations include:

Providing tools to extract meaning System designers could provide users with image analysis tools to deal with unreadable signs and unfamiliar scripts and symbols. Few participants—not even experts—were familiar with these tools, and those who were often encountered technical issues that reduced their utility. Streamlined, integrated access to tools for image editing (e.g. sharpening blurred signs), optical character recognition (OCR) and language translation, and reverse image searches (for logos or symbols) could substantially improve users' external information gathering.

Supporting systematic, scalable analysis System designers could encourage systematic consideration of image clues to help users handle an abundance of information and avoid fixation and confirmation bias. Software could automatically extract potential clues (e.g. text, numbers) and present them to users as micro-tasks, encouraging a broad and comprehensive investigation. Some of these tasks could be distributed to crowd workers to leverage the efficiency of parallelized analysis while preventing the user from becoming overwhelmed.

Cultivating insights in sparse areas When few clues are obvious, systems could provide other ways to help users make progress. Computer vision techniques such as PlaNet (Weyand, Kostrikov, and Philbin 2016) can suggest high probability regions as starting points even in rural areas with limited signage. When users lack the internal knowledge to use expert techniques such as sun position or driving rules, the system could suggest them via prompts or perform them automatically with targeted computer vision.

5 Acknowledgements

We thank our participants and anonymous reviewers of this paper. This research is supported by NSF IIS-1527453.

References

Bigham, J. P.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R. C.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; and et al. 2010. *VizWiz: Nearly Real-time Answers to Visual Questions*. UIST 10. ACM. 333342.

Crisman, J. D. 2011. Finder.

Hays, J., and Efros, A. A. 2008. *Im2gps: estimating geographic information from a single image*. 18.

Higgins, E. 2014. Geolocating Tunisian Jihadists in Raqqa.

Lin, T.-Y.; Cui, Y.; Belongie, S.; and Hays, J. 2015. *Learning Deep Representations for Ground-to-Aerial Geolocalization*. 50075015.

Lintott, C., and Reed, J. 2013. Human Computation in Citizen Science. In Michelucci, P., ed., *Handbook of Human Computation*. Springer New York. 153–162.

Montello, D. R., and Sas, C. 2006. *Human Factors of Wayfinding in Navigation*. CRC Press/Taylor Francis, Ltd. 20032008.

NASA. 2003. Major biomes map.

NASA. 2010. Population density.

Noronha, J.; Hysen, E.; Zhang, H.; and Gajos, K. Z. 2011. *Platemate: Crowdsourcing Nutritional Analysis from Food Photographs*. UIST 11. ACM. 112.

Weyand, T.; Kostrikov, I.; and Philbin, J. 2016. Planet - photo geolocation with convolutional neural networks. *arXiv:1602.05314 [cs]*. arXiv: 1602.05314.