GroundTruth: Augmenting Expert Image Geolocation with Crowdsourcing and Shared Representations

SUKRIT VENKATAGIRI, Virginia Tech, USA
JACOB THEBAULT-SPIEKER, Virginia Tech, USA
RACHEL KOHLER, Virginia Tech, USA and BNSF Railway Company, USA
JOHN PURVIANCE, Virginia Tech, USA and Bloomberg L.P., USA
RIFAT SABBIR MANSUR, Virginia Tech, USA
KURT LUTHER, Virginia Tech, USA

Expert investigators bring advanced skills and deep experience to analyze visual evidence, but they face limits on their time and attention. In contrast, crowds of novices can be highly scalable and parallelizable, but lack expertise. In this paper, we introduce the concept of shared representations for crowd–augmented expert work, focusing on the complex sensemaking task of image geolocation performed by professional journalists and human rights investigators. We built GroundTruth, an online system that uses three shared representations—a diagram, grid, and heatmap—to allow experts to work with crowds in real time to geolocate images. Our mixed-methods evaluation with 11 experts and 567 crowd workers found that GroundTruth helped experts geolocate images, and revealed challenges and success strategies for expert–crowd interaction. We also discuss designing shared representations for visual search, sensemaking, and beyond.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**; **Collaborative and social computing design and evaluation methods**;

Additional Key Words and Phrases: crowdsourcing; crowd-augmented expert work; geolocation; investigation; verification; journalism; misinformation; shared representations; real-time crowdsourcing; expert; crowd; sensemaking; visual search

ACM Reference Format:
Sukrit Venkatagiri, Jacob Thebault-Spieker, Rachel Kohler, John Purviance, Rifat Sabbir Mansur, and Kurt Luther. 2019. GroundTruth: Augmenting Expert Image Geolocation with Crowdsourcing and Shared Representations. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 107 (November 2019), 30 pages. https://doi.org/10.1145/3359209

1 INTRODUCTION

High-stakes settings like investigative journalism and human-rights advocacy increasingly leverage photo and video evidence from social media in their investigations [15, 32, 75]. For instance, in 2017, Europol launched the Stop Child Abuse: Trace An Object campaign [5] that relies on volunteer crowds to help identify the origin of objects in the backgrounds of imagery (photos and videos) involving child abuse. Similarly, Bellingcat, an online open-source investigative community, uses

Authors' addresses: Sukrit Venkatagiri[1]: sukrit@vt.edu; Jacob Thebault-Spieker[2]: jthebaultspieker@vt.edu; Rachel Kohler[3]: rkohler1@vt.edu; John Purviance[4]: jpurviance@vt.edu; Rifat Sabbir Mansur[2]: rifatsm@vt.edu; Kurt Luther[1]: kluther@vt.edu. [1–4] Department of Computer Science and Center for Human-Computer Interaction, Virginia Tech — [1] Arlington, VA, USA and [2] Blacksburg, VA, USA; [3] BNSF Railway Company, Fort Worth, TX, USA; [4] Bloomberg L.P, New York, NY, USA.

2573-0142/2019/11-ART107 $15.00
https://doi.org/10.1145/3359209

social media imagery to investigate the credibility of claims made by and about governmental or terrorist activity [36].

Uncertainty in image provenance raises questions about whether the imagery has been altered or is being reused. This uncertainty a ects the credibility of the imagery and subsequently harms its e cacy as a form of evidence, whether in the court of law or of public opinion. If these images are to be used as evidence, veri cation is critical [32, 75].

A key step in this veri cation process is *image geolocation*, a complex sensemaking process that involves identifying the exact location where photo or video imagery was taken [66, 76]. If an expert investigator succeeds in geolocating an image, then it can reliably be used to make claims about an event that happened at a particular place and time. Initially, an expert inspects the image for clues, such as familiar landmarks, license plates, street signs, etc., to narrow their search. When these clues are not conclusive, they use inference and experience to conduct a manual, brute-force search through large swathes of satellite imagery, looking for the location depicted in the image [36]. This task may take hours to days, and may not prove fruitful. It also does not scale easily [7, 19], meaning that successful geolocation is limited by experts' time and attention. Computer vision attempts at automating this process [33, 85, 87] are insu ciently accurate, placing photos within within 200km of the correct location less than 30% of the time. Further, they have constraints that may not generalize well for many real-world contexts [14].

An alternative approach that has seen success is leveraging the powerful and adaptive capabilities of geographically distributed online crowds [1, 5, 78]. However, crowds often lack expertise in knowing where to look or how to assess relevance, which can lead to false positive rates as high as 64.1% [46]. Undirected crowds can also lead to vigilantism [83, 93] and misidenti cation [63]. The question then arises, can crowds e ectively augment expert work practice to geolocate images?

In this paper, we propose an approach that combines experts' deep domain knowledge and experience with the speed and scale of crowds. To enable this approach, we extend Heer's idea of shared representations *between* humans and intelligent agents [35], and use it to facilitate crowd-supported expert image geolocation. Heer describes shared representations as a common language through which both humans and intelligent agents can work in tandem to achieve a shared objective, balancing the complementary strengths and weaknesses of each. This approach is relevant to crowd-supported expert work practice because it augments but does not replace experts, while still promoting e ciency and correctness, and it requires neither perfect accuracy nor exhaustive modeling of the user's tasks to be useful [35].

We explore this approach through GroundTruth, a system we developed to help experts geolocate images with a crowd. GroundTruth consists of three shared representations as system components: (1) an expert-created *aerial diagram* to help share context with the crowd, focus their attention, and overcome their spatial reasoning limitations; (2) a *gridded map overlay* specified by experts that generates microtasks for crowd workers, indicating where they should search, while providing the expert an overview of crowd progress; and (3) a *heatmap* displaying expert and crowd decisions which quickly and at-scale indicates to the expert where their own time and attention is best spent.

We conducted a mixed-methods evaluation of GroundTruth involving a think-aloud protocol, log analysis, and semi-structured interviews with 11 experts working with 567 crowd workers. We nd that GroundTruth e ectively merges the bene ts of both expertise and crowdsourcing, demonstrating the feasibility of crowd-supported expert image geolocation using shared representations. Experts worked with crowds in real-time to narrow the search area substantially, and frequently succeeded in geolocating the image. Experts were also excited by the idea of incorporating GroundTruth into their toolset since it provides features that are not currently available in other tools. Finally, we re ect on challenges and successes in designing shared representations highlighted through our evaluation.

In summary, our work makes four contributions:

(1) Our paper makes a technical contribution by introducing shared representations in crowd-supported visual search, allowing visual traits and context to be easily communicated between experts and novice crowds performing a complex sensemaking task: image geolocation.
(2) GroundTruth makes a system contribution as an operationalization of shared representations that enables expert investigators to geolocate images with the help of crowds. This is done using three shared representations: an aerial diagram, a gridded map overlay, and a heatmap displaying experts' decisions and real-time crowd feedback.
(3) A mixed methods evaluation with expert investigators, who drew their own aerial diagrams and worked with crowds in real time to geolocate images. Experts expressed an overall preference for GroundTruth over current expert work practice and tools. Our evaluation finds that shared representations support new collaborative work dynamics.
(4) We further develop implications for enriching expert crowd collaboration in investigative work, and applying shared representations to complex tasks beyond visual search, which are currently only the purview of experts.

2 BACKGROUND

To provide context for our system and evaluation, we first summarize expert image verification and geolocation workflows. Although there are empirical studies of verification in general [24, 74], image geolocation has seen less scholarly attention [45, 60]. Therefore, we draw on practitioner accounts [10, 32, 36, 75] to fill gaps in the limited scholarly record. We also relate these existing practices to sensemaking theory [67] and identify challenges that motivated our design rationale and system development.

2.1 Current Expert Practice in Image Geolocation

Experts in many investigative fields, such as journalism, human rights, and military intelligence, perform image geolocation as a key step in the broader task of verifying photos and videos shared on social media. The goal is to identify the precise location where the image was captured, to help support or refute claims about its provenance and meaning.

Experts perform image geolocation using a largely manual process characterized by iteratively narrowing down possibilities to find a needle in a haystack. They start with a photo that they want to geolocate, obtained through social media or by received from clients. Next, they examine the surrounding context and metadata of the image, researching the user who posted it and the claims made about it. Most social media platforms scrub metadata, including geotags, for uploaded content as a privacy measure, which is why experts focus their attention on the actual visual content of the images [75]. Experts look for road signs, business names, phone numbers, unique landscape or architectural features, or other clues that could point to certain locations or rule out others. If this step does not sufficiently narrow down the location, experts may resort to a brute-force approach of manually searching satellite imagery in candidate locations for potential matches. Experts often draw an aerial diagram of the ground-level photo of interest to ease visual comparison [36]. This translation process requires expertise in spatial reasoning and mental rotation which experts develop over time [60].

2.2 Expert Image Geolocation as Sensemaking

The process of image geolocation, and image verification more generally, can be understood as a sensemaking task, in which the goal is to gather and analyze large amounts of diverse, unstructured information to arrive at a theory or conclusion [23, 67, 69]. One influential model by Pirolli and

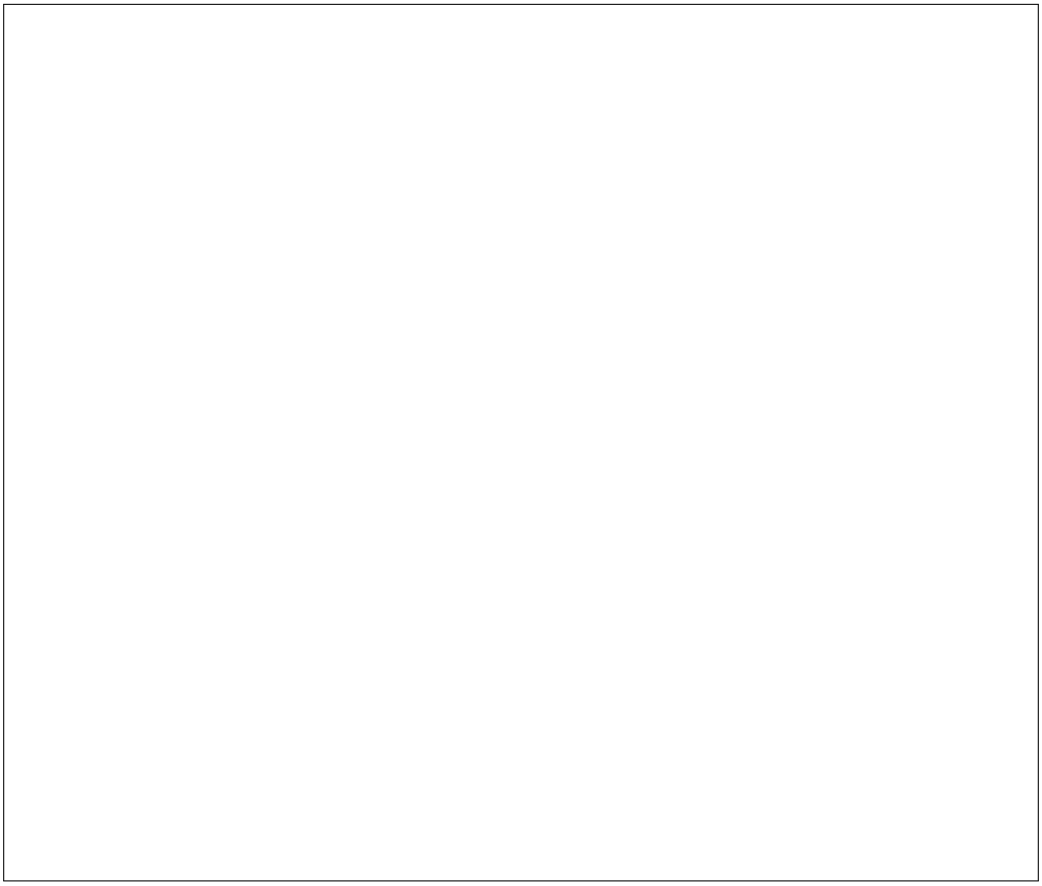

Fig. 1. Expert image geolocation as sensemaking. The orange dotted arrow indicates one possible set of steps that an image verification expert would undertake to debunk or verify visual media through image geolocation. We relate these steps to Pirolli and Card's sensemaking loop for intelligence analysts (from [67]). See Section 2.2 for an example.

Card [67] characterizes sensemaking as a 16-step, iterative process with two key subloops, one focused on gathering, searching, and extracting information (foraging); the other on building a mental model that best fits the information (synthesis), with structure and effort increasing in later steps. In our example below, we relate the steps of geolocation back to the sensemaking steps as defined by Pirolli and Card [67] (see Fig. 1). Half the steps (15, 12, 9, 5, 3) are top-down, moving from theory to data, while the other half (2, 5, 8, 11, 14) are bottom-up, creating a dual search cycle between adjacent steps.

Applying this model to an image geolocation example, suppose an expert investigates a photo that they encountered on social media purporting to show evidence of a violent protest in an unidentified German city. This claim may provide they with an initial top-down hypothesis. However, as they inspect the visual content, clues in the road signage and building architecture instead suggest Austria, providing a bottom-up challenge to the initial hypothesis. On one building, the expert notices a distinctive but unfamiliar logo (step 5: read and extract). They run a reverse image search (3: search for information) which yields a large number of results (1: external data source). Sorting through these (2: search and filter), they find a match for a business with multiple offices in three

cities in Austria (6: search for relations). They then draw an aerial diagram of the photo of interest (8: schematize) and manually search satellite imagery from each city (9: search for evidence), one at a time. Four hours later, they locate the matching building in one of the cities. This discovery, along with other contextual evidence (11: build case), allows them to debunk the original claim of Germany (14: tell story), based on which they writes a news report (16: presentation).

This paper focuses on the latter steps of the geolocation task, when the expert must search satellite imagery using a manual, brute-force process. Looking through large areas can take hours or even days, with no guarantee of success. Fatigue, coupled with the time-crunched nature of this work, is a major challenge because constant vigilance is required to find a needle in a haystack. There is some evidence that image geolocation experts (along with other journalists and investigators [19, 93]) seek crowdsourced assistance, but little technological support exists, and their success rates are unknown. To address this research gap, we investigate how experts respond to crowdsourced support in image geolocation.

3 RELATED WORK

3.1 Collaborative Sensemaking and Visual Search

3.1.1 Collaborative Sensemaking.

Collaboration has the potential to speed up sensemaking by dividing up foraging and search tasks, as well as providing multiple perspectives on schematizing and theorizing about connections, among other benefits [83]. However, collaboration also creates coordination challenges. For example, collaborators may have different skills and backgrounds, be geographically separated, need to externalize their thoughts for others, and have access to different parts of the dataset [22, 24, 27, 38]. Most of these efforts focus on small groups, i.e., dyads [27, 29, 38] or triads [72, 92].

However, less work has explored scaling up collaborative sensemaking to a larger number of people, such as crowd workers, where coordination challenges are amplified [86]. Some projects have focused on crowdsourcing specific sensemaking steps as microtasks, such as searching and filtering [66], reading and extracting [16], or schematizing [18, 43, 56]. Others crowdsourced the entire sensemaking loop to perform complex tasks like solving mysteries [54, 80] or writing articles [31, 44]. While most of these efforts focus on how either novice crowds alone or crowd AI hybrids can complete sensemaking tasks, our system differs by exploring how crowds can augment an expert's sensemaking process. Further, while the majority of crowd sensemaking research focuses on text data [e.g., 9, 16, 31, 43, 56, 80], we focus on visual data.

3.1.2 Crowdsourced Visual Search.

Prior work on crowdsourced visual search has largely focused on satellite imagery. The focus is often on counting or annotating things (e.g., clouds [94], building damage [39], tanks [68]). Sometimes the focus is searching for a specific needle in a haystack, like Genghis Kahn's tomb [55] or the boat of missing computer scientist Jim Gray [32]. These projects, as well as ours, divide up an area of satellite imagery into smaller cells and ask crowds to conduct a visual search for objects, buildings, etc.

In our prior work [46], we asked crowds to help geolocate an image by searching through satellite imagery using either the ground-level photo or an aerial diagram. Likewise, our system here tasks crowds with searching through satellite imagery using aerial diagrams to support experts. While the previous study used a controlled experiment and perfect diagrams to determine an upper-bound on crowd performance, this paper explores the real-world feasibility of crowds using expert-drawn diagrams, bigger search areas, and more diverse test photos.

Furthermore, all of the above projects focus on crowd performance and did not involve actual expert investigators. In contrast, our work here focuses on an expert-driven system, GroundTruth, that aggregates and visualizes the crowd's results and incorporates it into an expert's work flow in

real time. Also in contrast to prior work, our evaluation focuses on how these experts respond to crowd feedback and integrate it into their own investigative activities.

3.2 Requester Crowd Interaction Models

Early crowdsourcing focused on a relatively simple model where a requester posted a task and returned later, after the task was completed, to review results. The introduction of the waiting room or retainer model allowed requesters to receive results in near real-time [11, 49, 50]. This fast turnaround time enabled new, richer modalities of interaction between requesters and crowds, such as clarification of ambiguous task instructions [13, 58], workflows [47], or conversations with multiple exchanges [51]. Like this last class of systems, GroundTruth leverages real-time crowdsourcing (via LegionTools [25]) to quickly return crowd results to the requester, a key requirement for image verification tasks that are often time-sensitive.

Most similar to our work, a subset of real-time crowdsourcing systems support what we term crowd-augmented expert work. This model assumes that (1) requesters are experts in some domain (e.g., animation, innovation, design), (2) requesters are simultaneously performing the same (or a superset) of tasks as the crowd, and (3) the expert's own work is shaped and redirected based on the crowd results streaming in real-time. Crowdboard [8] explored whether online crowds could augment expert ideation in a hybrid physical virtual studio. Apparition [48, 52] and SketchExpress [53] enabled crowds to prototype user interfaces and animations drawn or described by expert designers. Inspired by these prior works, GroundTruth extends the crowd-augmented expert work paradigm to visual search tasks, specifically image geolocation. Our focus on visual search, a type of analytic task, complements prior work in this area largely focused on creative and expressive tasks, which pose distinct challenges [21].

3.3 Visualization and Shared Representations

A variety of technological solutions have been explored for improving coordination in collaborative sensemaking. One theme has been to use tools like visualizations and tabletop displays to help collaborators externalize their ideas in ways that are easily shared and aggregated with others [22, 27, 29, 92], though these tools are geared towards lengthy sessions with 2–3 collaborators, not crowds and microtasks. Likewise, as crowdsourcing workflows have become more complex, new tools have helped requesters to monitor and interpret crowd work, often with dashboards and visualizations [13, 42, 47, 57, 70, 90]. Inspired by these, GroundTruth also leverages visualizations to aggregate and display crowd results to the requester.

However, unlike these works, in which requesters typically wait extended periods of time to review work performed entirely by the crowd, GroundTruth's crowd-augmented expert workflow has crowds provide feedback in real-time, while an expert performs the same (and superset of) tasks. Most similar to our work, Crowdboard [8] provided two versions of an interactive workspace: "studio" for in-person experts, and "web" for online crowds; while Apparition [48] provided a "shared canvas" which showed or hid editing tools depending on roles assigned to experts and crowds. While these interfaces proved effective in their respective task domains, it is unclear how to adapt the underlying principles for an analytic task involving visual search of geographic regions.

Consequently, we adapt a concept from mixed-initiative systems, "shared representation" [35], that suggests concrete design principles for collaborative analytic work, such as data analysis and sensemaking. We detail how we adapted these principles for crowd-augmented expert image geolocation in the following section.

4 GROUNDTRUTH

4.1 Design Rationale for Shared Representations

Heer's [35] shared representations, adapted from Horvitz's guidelines for mixed-initiative systems [37], enable people and agents of varying abilities to work together to achieve a common objective. They enable a user to direct tasks, performing a superset of the agents work. In parallel, the agent can support the user's foraging tasks. This allows for the merging of the agent's scale and speed with the user's deep domain knowledge and skills.

Building on these ideas, we explored whether shared representations could inform the design of a system to support image geolocation in which expert work was augmented by human crowd workers rather than AI agents. To do so, we employed an iterative design process with expert investigators over the span of eighteen months, along with pilot studies. From these efforts, we developed three types of shared representations, each motivated by one of Heer's principles [35].

(1) Heer's first principle encourages augmentations that provide significant value, promoting efficiency, correctness, and consideration of alternate possibilities that a user might not have otherwise considered. We adapted this idea for expert crowd interaction as a shared lens. In GroundTruth, this takes the form of an expert-drawn aerial diagram. The aerial diagram is a shared lens into what the expert believes is relevant in a satellite image search, and indicates to the crowd what to look for. It bootstraps existing expert practice of aerial diagramming and overcomes the crowd's limited spatial reasoning skills, allowing them to support experts at-scale and promote consideration of alternatives.

(2) Heer's second principle emphasizes automated suggestions that augment, but do not replace, user interaction and blend into the interactive experience in a nondisruptive manner and can be directly invoked or dismissed by the user. We adapted this idea as a shared environment between experts and crowds that, for GroundTruth, takes the form of a gridded map overlay. The grid structures the search area, telling the crowd where to look, and divides it into smaller cells (microtasks) for them to easily provide feedback. The grid lets experts direct investigations, and work with the crowd in a nondisruptive, dismissable manner. That is, the expert performs tasks that are the same as and a superset of the crowd's, working alongside them in real time.

(3) Heer's third principle encourages augmentations that require neither perfect accuracy nor exhaustive modeling of the user's task to be useful. We adapted this idea for expert crowd interaction as shared analysis. In GroundTruth, this takes the form of a heatmap. The heatmap enables shared analysis between experts and crowds to locate the ground-level photo (and aerial diagram) within cells of satellite imagery. The heatmap aggregates crowd feedback to prioritize expert attention and allows experts to easily exclude cells as they conduct their search in parallel. Although crowds prioritize some false positives, they also rule out many irrelevant cells, providing value to experts despite imperfect accuracy.

4.2 System Description and Scenario

GroundTruth consists of two different interfaces for the three shared representations (aerial diagram, grid, heatmap). The expert interface allows an expert to define and manage a geolocation task, where the expert uploads the ground-level photo and aerial diagram, and specifies the search space (for both them and crowd workers) by drawing the grid. The crowd worker interface allows crowd workers to perform geolocation microtasks specified by the expert, using the aerial diagram.

We now describe the expert and crowd worker interfaces in detail using the following fictional scenario based on a real-life event [82]. Noor is an investigative journalist who works for an online intelligence group that investigates war crimes. Late last night, the Turkish Air Force bombed the