Photo Sleuth: Identifying Historical Portraits with Face Recognition and Crowdsourced Human Expertise

VIKRAM MOHANTY, Virginia Tech, USA
DAVID THAMES, Virginia Tech and Google, USA
SNEHA MEHTA and KURT LUTHER, Virginia Tech, USA

Identifying people in historical photographs is important for preserving material culture, correcting the historical record, and creating economic value, but it is also a complex and challenging task. In this article, we focus on identifying portraits of soldiers who participated in the American Civil War (1861–65), the first widely photographed conflict. Many thousands of these portraits survive, but only 10%–20% are identified. We created Photo Sleuth, a web-based platform that combines crowdsourced human expertise and automated face recognition to support Civil War portrait identification. Our mixed-methods evaluations of Photo Sleuth one month and 11 months after its public launch showed that it helped users successfully identify unknown portraits and provided a sustainable model for volunteer contribution. We also discuss implications for crowd-AI interaction and person identification pipelines.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing systems and tools**; • **Computing methodologies** → *Computer vision tasks*; • **Applied computing** → *Arts and humanities*;

Additional Key Words and Phrases: Crowdsourcing, online communities, face recognition, person identification, crowd-AI interaction, history

ACM Reference format:
Vikram Mohanty, David Thames, Sneha Mehta, and Kurt Luther. 2020. Photo Sleuth: Identifying Historical Portraits with Face Recognition and Crowdsourced Human Expertise. *ACM Trans. Interact. Intell. Syst.* 10, 4, Article 33 (October 2020), 36 pages.
https://doi.org/10.1145/3365842

1 INTRODUCTION

Identifying people in historical photographs provides significant cultural and economic value. From a cultural perspective, it can help recognize contributions of marginalized groups, as in the recent social media campaign to identify Sheila Minor Huff, the only female African American scientist visible in a group portrait of attendees at a 1971 biology conference [25]. Identification

The reviewing of this article was managed by special issue associate editors Oliver Brdiczka, Polo Chau, Gaelle Calvary, and Minsuk Kahng.
This research was supported by NSF IIS-1651969 and IIS-1527453 and a Virginia Tech ICTAS Junior Faculty Award.
Authors' addresses: V. Mohanty, S. Mehta, and K. Luther, Virginia Tech, 900 N Glebe Rd, Arlington, VA 22203, USA; emails: {vikrammohanty, sudo777, kluther}@vt.edu; D. Thames, Virginia Tech, Blacksburg and Google, Google, 601 N 34th St, Seattle, WA 98103, USA; email: davidcthames@vt.edu.

2160-6455/2020/10-ART33 $15.00
https://doi.org/10.1145/3365842

Corrected Version of Record v.1.1. Published October 22, 2020

can also correct the historical record, as in the case of James Bradley, author of *Flags of Our Fathers*, who was convinced by visual evidence that his father was not pictured in the iconic photo of US Marines at Iwo Jima during World War II, as he once believed [68]. Additionally, identification can generate significant economic value, as when a photo purchased at flea market for $10 was estimated to be worth millions of dollars following its identification as a ca. 1875 portrait of American outlaw Billy the Kid [24].

However, identifying people in historical photos is complex and challenging, and researchers lack adequate technological support. The current research practices employed by historians, genealogists, archivists, collectors, and other experts for identifying portraits are largely manual and often time-consuming. These practices involve manually scanning through hundreds of low-quality photographs, military records, and reference books, which can often be tedious and frustrating, and lacks any guarantee of success. Automated face recognition algorithms can support this effort, but are not widely used by historical photo experts and are often insufficient for solving the problem on their own. Many studies have compared face recognition algorithms to a human baseline, with mixed results [9, 11, 35, 86]. Further, historical photographs add unique challenges, as they are often achromatic, low resolution, and faded or damaged, which may result in loss of useful information for identification.

In this article, we present Photo Sleuth,[1] a web-based platform that combines crowdsourced human expertise and automated face recognition to support historical portrait identification. We introduce a novel person identification pipeline in which users first identify and tag relevant visual clues in an unidentified portrait. The system then suggests filters based on these tags to narrow down search results of identified reference photos. Finally, the user can carefully inspect the narrowed search results, sorted using automatic face recognition, to make a potential identification. This pipeline also bootstraps crowdsourced user contributions to grow the site's database of reference images in a sustainable way, increasing the likelihood of a potential match in the future. Photo Sleuth focuses on identifying portraits from the American Civil War (1861–65), the first major conflict to be extensively documented through photographs. An estimated 3M soldiers fought in the war and most of them had their photos taken at least once. After 150 years, millions of these portraits survive in museums, libraries, and individual collections, but the identities of most have been lost.

We publicly launched Photo Sleuth in 2018 and conducted a mixed-methods evaluation of its first month of usage, including interviews with nine active users, content analysis of uploaded photos, and expert review of user identifications [58]. We found that the system transformed users' research practice and helped them identify dozens of unknown portraits. Additionally, Photo Sleuth's pipeline encouraged users to voluntarily add hundreds of identified portraits to aid future research, suggesting a sustainable model for long-term participation. We also conducted a benchmarking study, as well as a follow-up longitudinal analysis after 11 months of usage. Our primary contributions are:

- a novel person identification pipeline combining crowdsourcing and face recognition;
- a web-based tool and online community, Photo Sleuth, demonstrating this approach;
- a mixed-methods evaluation of Photo Sleuth after one month of deployment with real users;
- a follow-up longitudinal evaluation of Photo Sleuth after 11 months of deployment, validating the sustainable participation model;
- a benchmarking study showing how well Photo Sleuth's face recognition and tagging features perform on a gold-standard dataset developed for this article.

We also discuss implications for crowd–AI interaction and person identification pipelines.

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

2 RELATED WORK

2.1 Person Identification in Photographs

In recent years, commercial computer vision-based face recognition algorithms are finding use in many real-world applications, such as Uber using Microsoft Azure's Face API to verify their drivers [57] and C-Span using Amazon's Rekognition service to index their videos based on who is speaking or who is on camera [4].

Kumar et al. [40] propose the use of generalizable visual attributes (i.e., labels to describe the appearance of an image) for the face, such as gender, age, jaw shape, and nose size, to search faces and verify whether two faces show the same person. Some deep learning approaches such as DeepFace [76], DeepID2 [75], and FaceNet [69] have shown near-perfect face verification accuracy on the Labeled Faces in the Wild (LFW) dataset. Schroff et al. [69] also propose a method to automatically cluster all faces of a particular person. Photo Sleuth, however, does not depend on a training set. Instead, it exploits the strengths of existing face recognition algorithms in a hybrid pipeline by integrating additional relevant information from visual clues in a photograph into the search process to enhance accuracy.

Multiple studies have compared face recognition algorithms to human baselines, and some show that human performance is superior [9, 11, 35, 86]. A recent study shows state-of-the-art face recognition algorithms performing at the level of professional face examiners and suggests that optimal face recognition can be achieved by fusing human and machines [64]. However, these algorithms have also been seen failing to filter out false positives. Recently, Welsh police relying on face recognition technology wrongly identified people at a soccer game as criminals 92% of the time [5]. Amazon's Rekognition wrongly identified 28 members of Congress as people charged with a crime [71]. The workflow of Photo Sleuth prevents face recognition *per se* from making the final decision, instead deferring to human judgment.

Crissaff et al. [20] propose an image manipulation system called ARIES for organizing digital artworks, allowing users to compare images in complex ways and use feature-matching to explore visual elements of interest. Bell and Ommer [7] use computer vision algorithms to retrieve similar images for a query search image of a historical painting. Srinivasan et al. [74] propose using automated face recognition techniques for addressing ambiguities in portrait subjects and understanding an artist's style. Google released an app [27] in which users could find their painting doppelgangers from museums worldwide. Inspired by these recent efforts, Photo Sleuth helps users retrieve the identities of unknown photos of soldiers from the Civil War era by building and searching a digital archive of historical photographs.

Civil War portrait identification has not yet been studied through an HCI or AI lens, but a survey of historical scholarship [19, 77, 79], practitioner articles [47, 48, 51], and media accounts [55, 67, 80] offers some insight into the key tasks and challenges. It is estimated that at least 4M Civil War–era portraits survive today, of which 10%–20% are already identified [3]. Civil War portrait identification or "photo sleuthing" typically requires extensive skill and domain expertise, from identifying obscure uniform insignia and weapons [55], to weighing probabilities [51], to consulting a wide range of reference works [48], to systematically reviewing thousands of potential matches [47]. Photo Sleuth attempts to ease the sleuthing process by bringing together a large repository of soldier portraits and military service records, and the visual clues one would typically use in this process, in a workflow designed for both novices and experts.

2.2 Crowdsourced History and Image Analysis

2.2.1 Crowdsourced History. Research on crowdsourcing systems with applications to historical research has largely been limited to transcription projects (e.g., References [15, 30, 84]). While

person identification is a more complex task than text transcription and requires more historical domain knowledge, we draw inspiration from the approaches these projects take to designing interfaces that help crowd workers visually inspect historical primary sources.

A smaller body of research considers how members of online communities can work together to synthesize complex historical information and even conduct original research. Rosenzweig [66] contrasted the solitary tradition of professional historical research and the collaborative nature of Wikipedia articles about history. Willever-Farr et al. [82] found that genealogists on Ancestry.com are more likely to engage in cooperative research (sharing data) but not collaborative instructions (sharing techniques). A follow-up study [83] of Ancestry.com and Find A Grave showed that contributors are conscious of information quality and inaccurate information, and show skepticism towards open editing practices. These studies drew our attention to the complexities of facilitating original historical research in a public online platform and guided us to design a pipeline that foregrounded attribution and accountability to reward high-quality contributions and discourage the spread of misinformation.

2.2.2 Crowdsourced Image Analysis. Research involving crowdsourced image analysis often focuses on identifying everyday objects, transcribing text, or other tasks requiring only basic knowledge. These projects have yielded impressive results by leveraging crowdsourced visual analysis in well-defined settings where workers know what to look for, e.g., identifying everyday objects [10, 12, 60], analyzing video data [41–43], or performing recognition tasks quickly and at scale [8, 39]. Investigating photographs, however, requires crowds to make sense of unfamiliar historical and cultural contexts without prior knowledge about subjects in the photos, necessitating a different approach.

Various techniques have been employed for crowdsourcing analysis of unfamiliar visual material in a systematic way, such as combining crowds with computer vision to annotate bus stops and sidewalk accessibility issues in Google Street View images [31, 32], providing tutorials to non-expert volunteer crowds for analyzing scientific imagery in GalaxyZoo [45], and asking volunteer crowds to compare photos of missing and found pets to reunite them with their owners after a disaster [6].

Platforms such as Flock [16] and Tropel [63] use crowdsourcing to build hybrid crowd-machine learning classifiers. Due to scale and complexity issues, a person identification task cannot be seen as a multi-label or extreme classification problem. Since these approaches require a user to define the prediction task and example labeled data, they cannot be directly applied to a person identification task.

3 SYSTEM DESCRIPTION

Photo Sleuth is an online platform we developed to identify unknown people in Civil War–era portraits. The website allows users to upload photos, tag them with visual clues, and connect them to profiles of Civil War soldiers with detailed records of military service. This person identification problem can be seen as *finding a needle in the haystack*. Our novel pipeline (Figure 1) has three components: (a) building the haystack, (b) narrowing down the haystack, and (c) finding the needle. Additional screenshots of the Photo Sleuth user interfaces are available in the Appendix.

3.1 Building the Haystack

3.1.1 System Database. Photo Sleuth's initial reference database contains over 15,000 identified Civil War soldier portraits from public sources such as the US Army Military History Institute [78], the US Library of Congress, and the US National Archives, as well as other private sources. This is just a small proportion of the 4M photos estimated to exist today [3]. Therefore, a more

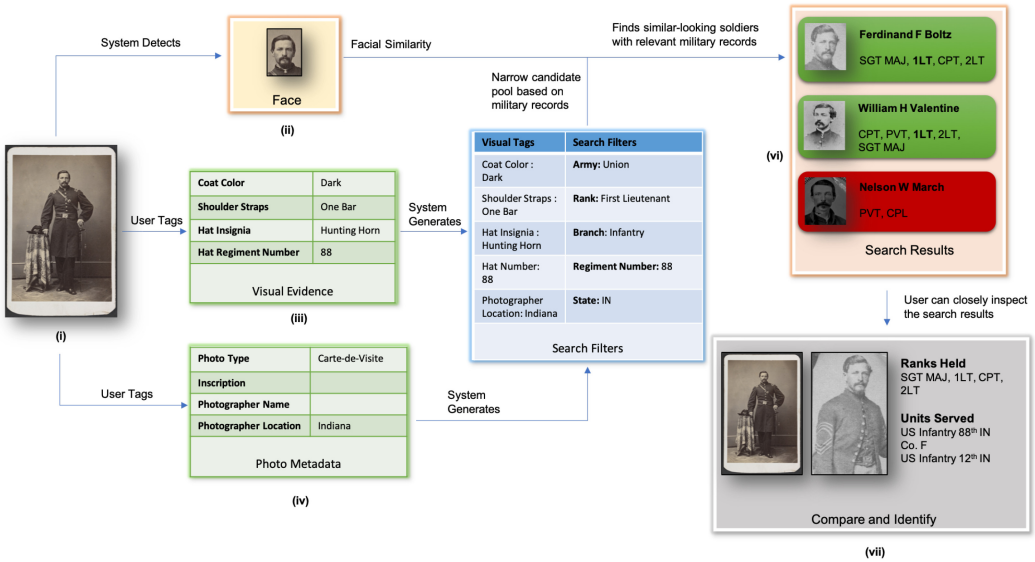

Fig. 1. System Workflow. (i) The user uploads a Civil War soldier portrait. All uploaded photos, identified or not, are added to the reference database for future searches. (ii) The system automatically detects the face in the uploaded photo. (iii) The user looks for visual clues in the photo (e.g., uniforms, insignia) and tags them. (iv) The user tags the photo for metadata, such as original source, photo format, and inscriptions. (v) Photo Sleuth converts the user-tagged visual clues into search filters for matching military service records and other biographical details. (vi) The system runs face recognition on the narrowed candidate pool from the previous step to find similar-looking soldiers with matching military records, sorting the results by facial similarity. (vii) The user can browse the search results and make a careful assessment, considering all relevant context, before deciding on a match.

comprehensive archive with more reference photos and identities would boost Photo Sleuth's goal of identifying a soldier and necessitates *building a haystack*.

3.1.2 Photo Upload and Primary Sources. A user begins the identification process by uploading a photograph with a mandatory front view and an optional back view. The user is also encouraged to provide the original source of the photo. We use Microsoft Azure's Face API [56] to detect a face in the photograph at the time of uploading. Photo Sleuth does not yet support photos with multiple faces.

3.1.3 Photo Metadata. Next, the user tags metadata related to the photograph, if available, such as the photo format, inscriptions on the front and back view of the photo, and the photographer's name and studio location. This metadata can offer insights into the subject's hometown, military unit, or name, both improving the search filters and providing useful source material for researchers.

3.1.4 Visual Tags. Our system then gathers information about visual evidence e.g., *Coat Color*, *Chevrons*, *Shoulder Straps*, *Collar Insignia*, or *Hat Insignia*. These visual tags are mapped on to the soldier's military service information, which provides a useful search parameter. More tags improve the relevance of the candidate pool, and thus reduce the number of false positives.

3.1.5 Bootstrapping and Ownership. Photo Sleuth adds the photo along with this information into the reference database, irrespective of identity, while displaying authorship credentials to the

user. These photos enrich the database for potentially identifying future uploads. Previous work suggests attribution is an important incentive for crowds conducting original research [53, 54]. By storing this information, a future feature of the platform would be to inform the users when their uploaded photos are identified by some other users.

3.2 Narrowing Down the Haystack

3.2.1 Search Filters. A major challenge in person identification tasks is the size of the candidate pool. Larger pools mean greater possibilities for false positives. Photo Sleuth reduces the likelihood of wrong identifications by generating search filters based on the visual evidence tagged by the user. These search filters are based on military service details that would otherwise be unknown to a novice user and are built using domain expertise. The military records used by the filters come from a variety of sources, including the US National Park Service Soldiers and Sailors Database [61]. We scraped the full military service record for every identified soldier portrait in our database, along with, in many cases, vital records and biographical details. This allows for users to filter by visual clues that would only be applicable for a snapshot of a soldier's career.

For example, if the user tagged *Hat Insignia* with a hunting horn, the system would recommend the "Infantry" branch filter, whereas *Shoulder Straps* with two stars would suggest the "Major General" rank filter. These filters narrow down the search pool to all soldiers who might ever have held these positions, including promotions, demotions, and transfers. Our system shows all search filters to the users, allowing expert users to make manual refinements. Photo Sleuth's interface also scaffolds domain knowledge to prevent users from applying search filters that might contradict each other.

3.2.2 Facial Similarity. Photo Sleuth augments the above search filters with facial similarity filtering via Microsoft Azure's Face API. Our initial tests with identified Civil War photos showed that this API yields near-perfect recall at a 0.50 similarity confidence threshold; i.e., retrieved search results at this level almost always include the correct results. However, its poor precision means many other similar-looking false positives also show up in the search results.

The search filters create a reduced search space in which face recognition looks for similar-looking photos of the query image. This complementary interaction between military records and facial similarity ensures that the most accurate information is retained in the search space.

3.3 Finding the Needle

3.3.1 Search Results. The search results page displays all the soldier portraits who satisfy the search filters and have a facial similarity score of 0.50 and above with the query photo, sorted by similarity. The user has the option to hide as-yet unidentified photos. The search results show military record highlights next to the names and photos. The user can then closely investigate the most promising search results before making the final decision of the soldier's identity. The user can also add new names and service records to the database if that soldier profile has not yet been added. To prevent misinformation being spread and promote cross-verification, all users are required to follow the entire workflow, even for photos whose identities they believe they already know. In such cases, the user is asked to provide the source of identification.

3.3.2 User Review. Users who find a potential match among the search results can closely inspect the two photos via a "Comparison" interface. The interface provides separate zoom/pan controls and also displays the service records of the reference photo to provide a broader context of who the soldier might be. Notably, the system hides the facial similarity confidence scores for verifying two faces to avoid biasing the user. If the user is confident about the photo being a match,

they can click on an "Identify" button to link the query photo to the soldier's profile and receive an "identifier" attribution. The user can also undo these identifications, if desired.

3.4 Implementation Details

Civil War Photo Sleuth is a web application built on the Python/Django framework with a PostgreSQL database for data storage and Amazon S3 for image storage. It is hosted on the Heroku cloud platform.

4 INITIAL EVALUATION (1-MONTH)

We released Photo Sleuth to the public on August 1, 2018. We recruited users via a launch event at the National Archives Building in Washington, DC and via advertising in history-themed social media groups. Within one month of its launch, 612 users registered free accounts on the website. A majority (360) registered in the first three days of the launch, followed by a steady stream of 5–10 registrations per day. During that period, users uploaded 2,012 photos, with 931 photos added in the first three days, followed by an average upload of 30 photos per day.

4.1 Log Analysis

We examined website logs for user-uploaded photos between August 1 and September 1, 2018. Users categorized uploads into front and back views. Uploads that did not have a face detected or with only a back view were excluded from our analysis. We then separated the remaining photos into *identified* and *unidentified* ones.

We further analyzed the logs to identify users who had uploaded or identified at least one photo. We also analyzed uploaded photos for which users had associated one or more visual tags, and identified the most commonly tagged categories for these photos. We give details of these log analyses below.

4.1.1 Categorizing Identified Photos. From the logs, we found that users performed 691 soldier identifications in the first month and matched 850 photos to these identities. To clean the data, we first excluded accidental duplicate uploads. Next, we checked all photos for duplicate identities (i.e., different photos of the same soldier under the same name but saved as separate identities) and grouped them together as a single identity. Last, all the photos that did not have a full name but had some demographic or military information were separated out as *partial identities*. The final pool consisted of 648 photos (560 uploaded by users, 88 already in the system) sharing 479 soldier identities between them.

Our pipeline does not automatically distinguish whether the identities of soldiers in photos are known prior to uploading or not. We therefore categorized these identified photos into two categories:

Pre-identified: Photos uploaded by users with their identities known prior to uploading
Post-identified: Photos matched by users to an existing identified photo in the database using Photo Sleuth's photo matching workflow

To determine *pre-identified photos*, we considered soldier identities with only one photo, since they had not been matched to any other photo in the database. We also grouped together all soldier identities matched with multiple photos in this category if none of the photos for an identity came from Photo Sleuth's reference archive. The remainder of the photos, i.e., soldier identities with multiple photos where at least one photo came from Photo Sleuth's archive of reference photos, were labeled *post-identified photos*.

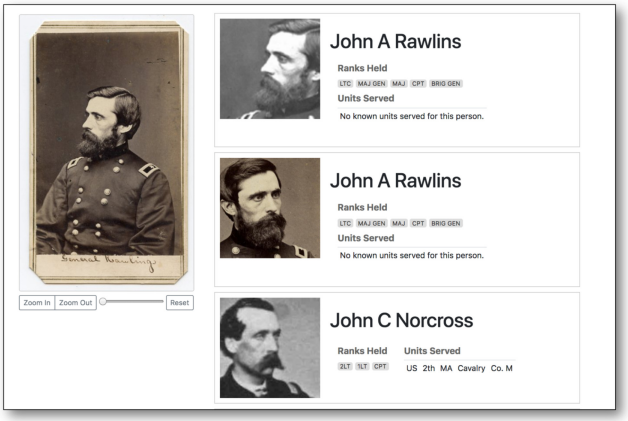

Fig. 2. Search results for an identity with both an inscription and replica. The uploaded photo can be considered a replica of the reference archive version displayed as the top search result.

4.1.2 Grouping Unidentified Photos. We filtered the unidentified photos (with faces detected) by removing 28 duplicate uploads. Photos with no names and no military information from the previous filtering process were also added to the original set of unidentified photos.

4.2 Content Analysis

Based on the above-mentioned categories, we performed a more targeted, in-depth analysis of how users identified the photos using Photo Sleuth.

4.2.1 Sources of Identification. We first analyzed the sources of information users drew upon when adding identified photos. We analyzed both front and back views of all pre-identified photos for the presence of a Civil War–era inscription or autograph of the matched soldier's name. If no name inscription was present, we checked if the user had provided an alternative source of identification.

4.2.2 Supporting Face Recognition. We considered two factors to understand the extent to which face recognition supported a user's identification decision. One was the presence of prior name *inscriptions* in the front or back views of the photo (Figure 3), as this would prompt an easy decision on the user's behalf to match the photo with a search result displaying the same name.

The second consideration was the possibility of an exact duplicate. One of the most popular photo formats during the 1860s was the *carte de visite*, where a subject would receive a dozen or more identical copies of their portrait on small paper cards they could collect in albums and exchange with friends and family. If multiple copies survive today, it is possible one of them is already identified, and a user could upload an unidentified version of the same photo that may differ only slightly due to cropping or age-related damage. We refer to such photos as *replicas* (Figure 2). In such cases, we would expect face recognition to return search results featuring an identified reference copy of the photo with a high similarity score, making it a top result for the user to quickly recognize.

Considering these factors, we analyzed front and back views of all photos in the post-identified category for the presence of the soldier's name inscriptions, similar to our analysis of pre-identified photos. Then, we examined whether any of the user-uploaded photos was a *replica* of an identified reference photo of the matched soldier.

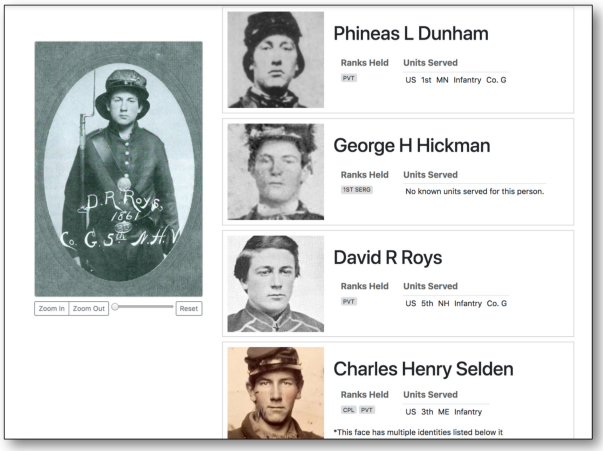

Fig. 3. Search results for an identity with only an inscription and no replica. The inscription on the photo says "DR Roys," which might have prompted the user to match "David R Roys" from the search results.

Based on our findings, we divided the soldier identities with post-identified photos into four sub-categories: (a) inscription and replica, (b) inscription but no replica, (c) replica but no inscription, and (d) no replica and no inscription. For example, if Capt. John Smith had five user-uploaded photos matched to his identified reference photo and any one of the user-uploaded photos had a name inscription and none of them was a replica, we grouped Smith in the *inscription but no replica* category. Similarly, if none of the photos was a replica and none of them had an inscription, we would place the identity in the *no replica and no inscription* category, and so on for the other categories.

4.2.3 Backtracing User Behavior. For a randomly chosen small sample in each of the above defined sub-categories, we backtraced (reconstructed) the identification workflow to re-match a post-identified photo. Backtracing helped us visualize the user's experience when presented with the search results under the original conditions.

4.3 User Interviews

We also conducted in-depth, semi-structured interviews [70] with nine Photo Sleuth users. These participants were active contributors to the site, each adding at least 10 photos to the site during the first month. They also had extensive prior experience identifying Civil War photos (mean=20 years, min=8, max=40), representing a mix of collectors, dealers, and historians. Eight participants were male (one female) and the average age was 54 (min=25, max=69). We anonymized participants with the identifiers P1–P9. All interviews were conducted over phone/video calls and were audio-recorded, fully transcribed, and analyzed with respect to the themes described in Section 5.

4.4 Expert Review

To assess the quality of user-generated identifications, an expert Civil War photo historian (and a co-author of this article) reviewed all post-identified photos added by users and evaluated them whether they were correctly identified or not. We establish ground truth in terms of whether a Soldier X's photo was identified as Soldier X or some other Soldier Y. The expert used the same four sub-categories as defined above to provide a fine-grained assessment of users' identifications.

We captured the expert's responses using a four-point Likert scale (1 = definitely not, 2 = probably not, 3 = possibly yes, and 4 = definitely yes).

5 FINDINGS

Using the methods above, we evaluated Photo Sleuth along three themes: *adding photos*, *identifying photos*, and *tagging photos*.

5.1 Adding Photos

5.1.1 Users Added Photos with Both Front and Back Views. From our logs analysis, we found 2,012 photos uploaded in the first month, of which 1,632 photos were front views and 380 photos were back views. Of the 612 users who had registered for the website in the first month, 182 users (excluding the authors) uploaded at least one photo to the system.

There were three power users who each uploaded more than 200 photos, and 11 users (excluding the authors) uploaded more than 30 photos each. On average, a Photo Sleuth user uploaded 13 photos (median = 3 photos) to the website.

5.1.2 Users Added Both Identified Photos and Unidentified Photos. Our log analysis showed that the number of identified photos (560) is similar to unidentified ones (602). There were also 121 partially identified photos. If we consider only identified photos, 441 were pre-identified (i.e., their identities were already known by the uploader), whereas 119 photos were post-identified (i.e., their identities were discovered using Photo Sleuth's workflow). These post-identified photos were matched to 88 identities with a prior photo in the reference archive.

Additionally, 107 users added at least one *un*identified photo, while 105 users had added at least one identified photo. Fifty-three users added both identified and unidentified photos.

Interviewees expressed a variety of motivations for adding pre-identified photos. Most commonly, participants mentioned trying to help other users identify their unknown photos, but they recognized this generosity could also help themselves. P2 felt it was only fair to contribute, given the identifications he was able to make from others' contributions: "As a way of giving back, I think I'm obligated to now." For P6, the motivation was anticipated reciprocity: "I'm just trying to help other people out like I want me to be helped out." P8 was motivated by curiosity to learn more about his own images: "I just uploaded to see if maybe there's a collector out there that had the same image maybe of a different pose or a different backdrop, different uniform." Some participants made an intentional choice to add identified photos first, waiting to add their unidentified ones later. P4 explained: "As your database of identified people increases, then there's a greater [chance] ...later on when I [upload] an unidentified image then I'll get a hit, where if I do that today, my odds are much less."

Other interviewees explained why they did *not* add pre-identified photos. One concern was bootlegging, i.e., unscrupulous users printing out scans from Photo Sleuth and reselling them as originals. P8 said, "Look on eBay. Look at all the fakes ...Look at the Library of Congress. You can download a file format ...the TIF format where it's high resolution. Then if you have a good printer, you print it out and you can make fake easy as that." A second concern was reuse without attribution. P3 said, "I have a lot of identified images that probably would help other people identify some of their guys. But I'm worried about putting them on there, only because I don't want them using my stuff unless they get permission from me first."

5.1.3 Users Provided Attribution for Most Identified Photos. Users matched 441 pre-identified photos to 386 unique soldier identities. Based on our content analysis, we found that users, while adding pre-identified photos, generally referred either to the name inscription on the photos

Table 1. Types of Post-identified Photos

	Replica	No Replica
Has Inscription	(Easiest) 17 positive matches	(Medium) 20 positive matches; 1 negative match
No Inscription	(Medium) 13 positive matches	(Hardest) 25 positive matches; 12 negative matches

(173 cases) or to the original source of identity (177 cases), to support their identity claims about a soldier's photograph. Users did not attribute a source in only 36 cases.

5.2 Identifying Photos

5.2.1 Users Identified Unknown Photos Using the Website's Search Workflow. Based on our log and content analysis, we found that users successfully used the system's search workflow to identify unknown photos. In the first month, 119 user-uploaded post-identified photos were matched to 88 existing soldier identities with a prior photo in the database. In some cases, more than one photo was matched to an identity.

Participants who added unknown portraits to the site described their success rates in enthusiastic terms. P1 remembered, "I was a half dozen in, and all of a sudden I got a hit on one of them." P5 described his experience: "I started running that whole pile of images that I had trying to find IDs on 'em, and I wanna say I found maybe ten to fifteen percent hits on images that I had squirreled away, that [Photo Sleuth] were able to compare to and bring up either the exact same image or an alternative that was clearly the same person." P2 noted, "Out of those thirty or forty or fifty that I posted on there, I've successfully identified I think at least three. That's a pretty good success rate considering there were hundreds [of] thousands of people fighting in the war."

Participants also favorably compared Photo Sleuth to traditional research methods. P5 lamented that US state archives often lacked searchable databases or digitized imagery, and aside from Photo Sleuth, "there's really nothing else out there as far as trying to find identifications for unidentified images." P8 emphasized that Photo Sleuth "saves a ton of time because now I don't have to just go through every single picture that's available ...When I first get an image, that's usually what I do—books, go online, search different areas, old auction houses ...But I kind of don't have to do that anymore because Photo Sleuth helps a lot."

Participants also recognized how the public nature of the system would affect their future collecting positively and negatively. P5 used the metaphor of a double-edged sword: "If I can find a match, it's good for me, but then it also may give somebody else that match, and then it becomes a bidding war whether I'm gonna pay more for it on eBay than that person is."

5.2.2 Users Decided on a Match Based on Additional Clues in the Photo Beyond Face Recognition. Based on our content analysis, we found that the post-identified photos included additional information that supported identification beyond face recognition. Of the 88 soldier identities that users matched during the first month, a significant proportion had additional helpful clues, such as the presence of an inscription (Table 1). Additionally, participants told us that they considered other contextual information besides facial similarity, such as military service records, when making an identification. In P9's words, "Without more information besides the face, I'm not gonna say it's one hundred percent."

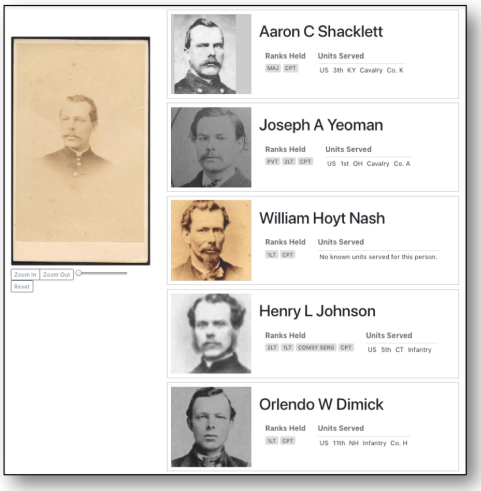

Fig. 4. Search results in which the top result does not show the matched identity. This photo was correctly matched to "Orlendo W Dimick."

5.2.3 Users Checked Multiple Search Results Carefully before Confirming a Match. During the backtracing process for post-identified photos, we observed that the matched identity did not always appear as the top search result (Figure 4). Out of 119 post-identified photos matched, 11 did not have their identities in the top 50 search results, while 19 had their identities in the top 50 but not the top search result. This suggests users confirmed a match only after carefully analyzing the search results beyond the top few ones.

Interviewees compared the automated face recognition to their own capabilities. P5 and P1 noted that, as human researchers, they were more likely to be distracted by similarities and differences in soldiers' facial hair, whereas the AI focused on features that remained constant across facial hairstyles. P1 also gave an example of how the AI challenged his assumptions by finding a matching soldier from a location he had not initially included: "I'm convinced I never would have figured that one out without the site."

Some participants mentioned drawbacks of the face recognition. P3 and P4 emphasized the differentiating value of ear shape, a feature that the AI does not consider. P8 observed that the AI often failed to recognize faces in profile (side) views, whereas he had no trouble. P4 felt he could outperform the AI on individual comparisons, but fatigue limited the number of images he could consider: "I still think that my eye could make the match better, but you just lose energy about it."

Participants also expressed a desire to solicit a second opinion from the community on the possible matches. We saw many examples of users posting screenshots of potential matches on social media and requesting feedback from fellow history enthusiasts. One potential benefit, described by P2, was consensus: "If a person posts a photograph and it's supposedly identified, you'll sort of see Facebook's hive mind kind of spin into action and in the comments, and if there's some dissent then I think there reasonably is doubt. But if everybody just says, 'Yes, duh,' that's the person." Another potential benefit, P8 said, was noticing details one might have missed: "It's always better to have a second opinion or a second pair of eyes to point out things that maybe you were focused on that you didn't really see."

5.2.4 Users are Generally Good at Identifying Unknown Photos. The expert analyzed all 88 identities matched with the post-identified photos and provided responses assessing identifications

done by the users using a four-point Likert scale for all 119 photos matched. Based on the expert's response, we consider the matches to be either *positive matches* (Likert-scale ratings of 3–4) or *negative matches* (ratings of 1–2).

As shown in Table 1, for the first and third categories in post-identified photos, i.e., when at least one replica was present for an identity, the expert validated responses for all 30 identities to be positive matches. For the second category, in which there is an inscription but no replica, only one out of 21 identities was validated as a negative match. We considered the final category of identities that did not have any inscriptions nor replicas to be the most difficult one. Out of 37 identities in this category, the expert assigned 12 identities to be negative matches and 25 as positive matches. Thus, the expert considered the vast majority of the identifications done by users in all categories to be positive matches.

5.3 Tagging Photos

5.3.1 Users Tagged Both Unidentified and Identified Photos. Based on the logs, we found that users had provided one or more tags for at least 401 of the 602 unidentified photos they added to the website. Out of the 560 identified photos (both pre-identified and post-identified) added by users, 445 photos had one or more tags associated with them. Further, 115 of the 182 users who uploaded photos also tagged a photo with at least one or more tags.

Because adding tags was optional, we asked participants why they did or did not provide tags. Some participants (P5, P3, P6) added tags because they believed the tags would help retrieve more relevant search results. For this reason, P8 skipped tags that were not linked to search filters: "If he's a straight-up civilian and there's nothing to go off of, I'll just bypass [tagging] and just hoping the face recognition brings something." In contrast, P2 thought the overhead was minimal: "It's probably just about as easy to put in the correct information as not." Other participants, like P4, added tags because they thought they would help future users, but not necessarily themselves.

5.3.2 Users Added Uniform Tags More Often Than Others. From the logs, we observed that users on an average added five tags per (tagged) photo, which was also the median count. We found that users provide tags related to both the photo's metadata (*Photo Format, Photographer Location,* etc.) and the visual evidence in the photos like (*Coat Color, Shoulder Straps,* etc.). *Coat Color* and *Shoulder Straps* were the most commonly tagged visual evidence, which the system uses to reduce search results by filtering military records by army side and officer rank, respectively.

6 LONGITUDINAL STUDY (11-MONTH)

We argued in our initial evaluation of the first month of Photo Sleuth's deployment that the site's person identification pipeline offered a sustainable model of volunteer contribution. To support this argument, we conducted a follow-up longitudinal analysis 11 months after the public launch (August 2018–July 2019), focusing on longer-term patterns of usage and growth. Below, we use site log data to provide updated statistics and illustrate trends in user registrations and photo uploads and identifications. We also present two illustrative case studies of successful photo identifications.

6.1 User Registrations

Since Photo Sleuth was launched on August 1, 2018, 12,322 users have registered in total, or an average of 1,120 per month. Figure 5 shows the number of users registering every month since its launch.

Photo Sleuth received major press coverage at two points since the launch: October–November 2018 [21, 22, 46, 62, 73] and February-March 2019 [14, 28, 81]. As a result, the website saw large spikes in user registrations during those months. Approximately 3,000 users registered during the

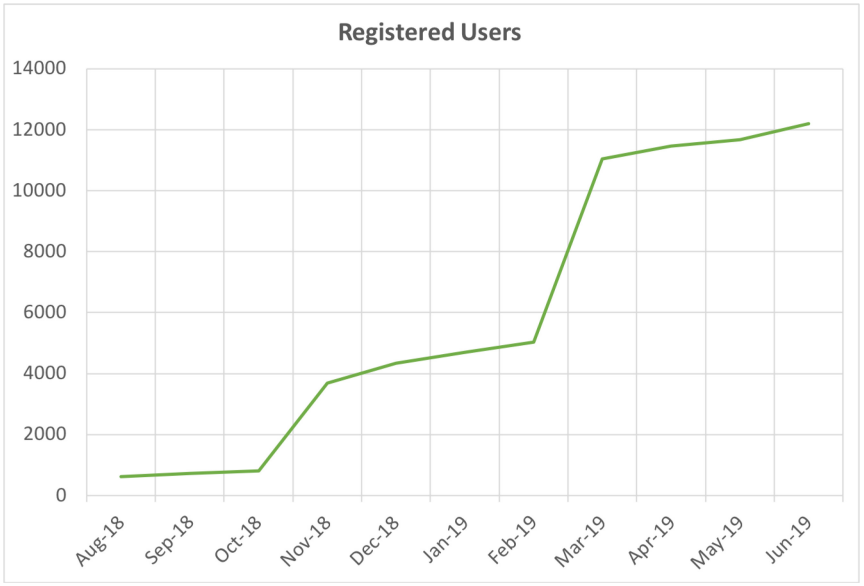

Fig. 5. User registration: Monthly data.

first press coverage, with an additional 6,000 registrations during the second press coverage. Outside of these spikes, registrations grew by an average of about 300 users per month. We attribute the slowest growth period, in the first few months following the launch, to limited awareness of the site outside a core group of Civil War photo enthusiasts.

6.2 Photos Added and Identified

As of June 30, 2019, Photo Sleuth has 28,111 Civil War photos. Of this total, 20,827 were system-added photos and 7,284 were user-added photos.

Out of the 12,322 users, 1,663 have added at least one photo. Six power users have each uploaded over 100 photos, with the highest being 316 photos by a single user. There are 22 users who have each uploaded over 30 photos.

Since the day Photo Sleuth was made public, users added an average of 21 photos per day. These include 6,471 front view photos and 887 back view photos. Figure 6 shows the number of photos added every month by users since it was launched. Similar to user registrations in Figure 5, we also observed spikes in photo uploads during the press coverage months. Nearly 2,000 photos were added between October–November 2018, and an additional 1,000 photos in February–March 2019.

Surprisingly, although the number of user registrations in the first press coverage (3,000) was half that of the first (6,000), the trend was reversed for photo uploads (2,000 uploads in the first, 1,000 in the second). We speculate that first coverage, which included genealogy-focused publications such as *Family Tree Magazine* and *Vintage News*, reached an audience more likely to have Civil War–era photos to upload, while the second coverage, which included national publications such as *Popular Mechanics* and *Fox News*, reached a broader audience.

Of the 12,322 users, 830 have identified at least one photo. Users matched 2,979 photos to 2,819 soldier identities. Out of these, 2,548 were pre-identified photos requiring users to create a new soldier profile, and 269 others were post-identified photos matched by users to 203 unique soldier identities from the original reference database. The remaining 162 photos required manual review to determine if they were pre- or post-identified. Due to the large increase in photos after

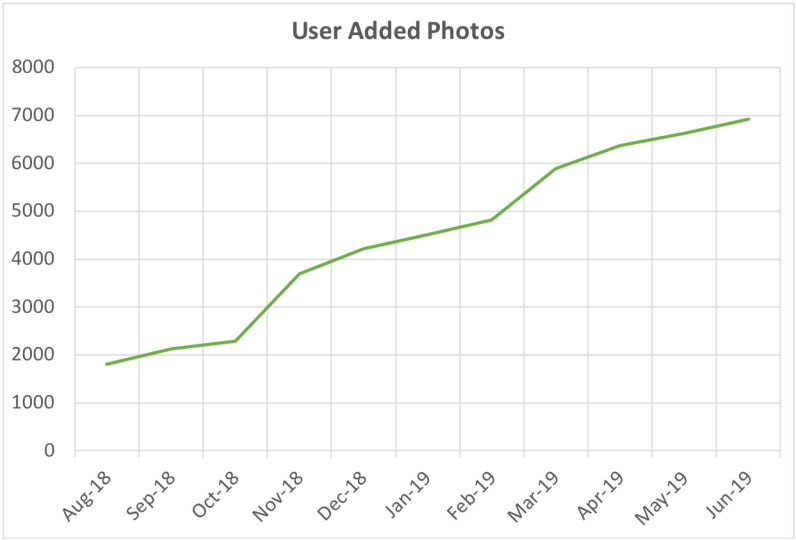

Fig. 6. User-added photos: Monthly data.

11 months, we did not perform a content analysis to categorize these 162 photos, nor arrange an expert evaluation of the identification accuracy of the 269 post-identified photos, as we did for the initial one-month evaluation.

6.3 Notable Case Studies

With over 100 expert-verified positive matches in the first month alone, and dozens more proposed identifications in the subsequent 10 months, Photo Sleuth has already made a substantial real-world impact in supporting historical photo identifications. In this section, we present two notable case studies of confirmed successful identifications, one from a public collection and one from a private collection, to illustrate this impact.

Public Collection. One of the most prominent public collections of Civil War soldier portraits is the Liljenquist Family Collection at the US Library of Congress, a collection of over 2,500 museum-quality portraits of Union and Confederate soldiers, as well as Civil War–era women and children [1]. Most of the portraits are hard images (i.e., tintypes and ambrotypes) depicting lower-ranked soldiers representative of the typical war experience, and many are unidentified. In preparing a demonstration of Photo Sleuth for its public launch event, a co-author of this article selected a tintype of an unidentified Union assistant surgeon from the Liljenquist Collection and ran it through Photo Sleuth's pipeline, applying visual tags for dark coat (Union army) and shoulder straps with one bar (first lieutenant), as well as a filter for assistant surgeon, due to the soldier's medical staff sword. The top-ranked search result was Francis M. Eveleth, who served as assistant surgeon and, later, full surgeon in two Maine infantry regiments [50]. The matching photo in the US Army Military History Institute's MOLLUS-MASS Collection showed a different view of the soldier. The carte was autographed by Eveleth himself, providing an airtight reference. Additional views of Eveleth were located in the Maine State Archives. After learning of this identification at the Photo Sleuth launch event, Library of Congress staff confirmed the match and updated the unidentified tintype's database entry with the soldier's name and service record, crediting Photo Sleuth for the discovery [2].

Private Collection. One Photo Sleuth user, David Morin, is a collector of New Hampshire–themed Civil War photos. Morin purchased a carte de visite of an unidentified Union officer, because the backmark indicated the photographer was based in New Hampshire. Morin uploaded the carte to Photo Sleuth and added visual tags for the dark coat (Union army) and shoulder straps with no insignia (second lieutenant). His search yielded five results. The top-ranked result was William H. Baldwin, a Union officer who served as second lieutenant, first lieutenant, and captain in a New York regiment of engineers [52]. The matching photo showed a different view of the soldier in the US Army Military History Institute's MOLLUS-MASS Collection. While Morin noticed the strong facial similarity between the two photos, he was initially skeptical, because the mystery photo was taken in New Hampshire, while Baldwin served in a New York regiment. In follow-up research, he found the connection—Baldwin was born and grew up in New Hampshire, later moving to New York. Several months later, a co-author of this article was demonstrating the Photo Sleuth software at a Civil War photo collectors event when he recognized a familiar carte for sale in a dealer's display case—coincidentally, it was another copy of Morin's mystery photo. The reverse of this copy, however, included a period inscription with the name "William H. Baldwin." Thus, Morin's original hypothesis that the mystery soldier was Baldwin was confirmed by another copy of his photo with an inscription bearing that name.

This case study suggests several lessons about effective use of Photo Sleuth. First, while his use of the software's face recognition, tags, and filters produced the correct match as the top-ranked result, additional research was needed to confirm the match and make sense of contradictory contextual information; i.e., the New Hampshire versus New York locations. Second, this case shows that even reasonable filters must be handled with care. If Morin had applied the "New Hampshire" state filter based on his mystery photo's backmark, the search results would have excluded the correct match, Baldwin, who never served with a New Hampshire regiment. Third, this example speaks to the benefits of Photo Sleuth's community, as when the co-author noticed an identical, identified copy of the mystery photo tentatively identified by Morin, providing an airtight confirmation.

7 BENCHMARKING STUDY

In addition to the previous user studies examining in-the-wild performance, we also sought benchmark data on how well the main components of Photo Sleuth's person identification pipeline, i.e., the AI-based face recognition and the visual tags and filters, performed against controlled gold-standard tests. Such data would shed light on the relative strengths and importance of these main components under different conditions, as well as the role of crowdsourced human expertise in augmenting them. Therefore, we conducted a benchmarking study with two analyses, each addressing one of the following research questions:

- **Study 1:** How well does Photo Sleuth perform at identifying photos with just the face recognition algorithm? (Hypotheses 1–3)
- **Study 2:** How useful are Photo Sleuth's visual tags and filters in augmenting the identification process? (Hypotheses 4–5)

7.1 Test Dataset

7.1.1 Selection Criteria. For our studies, we required a gold-standard test dataset of soldier photographs containing multiple soldier examples, such that each soldier has two different photos (i.e., *photo pairs*). Such pairs allow us to study how well the system performs on ground-truth examples. Furthermore, to test the robustness of the pipeline on different types of photos, and explore challenges posed by potential bias in our reference database caused by users as well as historical context, we required diversity within the dataset along several dimensions, including:

Format of the photographs. As mentioned in Section 4.2.2, the carte de visite was a popular photographic format during the Civil War that involved printing multiple copies of a photo on paper cards. Photo Sleuth users seeking to identify an unknown carte de visite may come across an identical, possibly identified copy, or *replica*, in another collection. Another type of replica involved a photographer making a period reproduction of a photo when the original negative was lost or when sitting for a new photo was not practical, i.e., if the soldier had died in battle and his family desired memorial copies [49].

Another portrait format we wished to explore was sketches. While photography was common during the Civil War, technological limitations did not yet allow photos to be directly printed in books and newspapers. Instead, artists were often hired to create engraved copies of the photos (lithographs) that could be mass-produced. Since lithographs were hand-drawn and some artists intentionally made deviations for aesthetic reasons, they were not perfect duplicates of the original photos. We examined whether these images, which we call *sketches*, could also be used to identify unknown soldier portraits.

Military service of the soldiers. The military service of soldiers affects the likelihood of their portraits being identified in complex ways. Fewer Confederate photographs exist compared to Union ones, since Confederates were not able to get access to developmental supplies for photographs, due to a naval blockade and lack of local production facilities [17]. High-ranked soldiers were photographed more often than low-ranked soldiers, in part because officers tended to be wealthier [85]. However, there were many more low-ranked soldiers than high-ranked ones (e.g., each regiment had 1,000 privates but only 1 colonel).

Race of the soldiers. Race may also impact the likelihood of identifying photos. Of the 2M soldiers in the Union Army, the majority were white; approximately 180,000 were black; and a much smaller number were other races. There were no black soldiers in the Confederate Army [44]. Therefore, reference databases have far more white photos, and few portraits of black Union soldiers survive [18]. Additionally, race interacts with military service. The Union Army was racially segregated by regiment, and black soldiers were generally excluded from officer ranks. Therefore, clues about an unknown soldier's race can shed light on his military service, and vice versa, to help narrow down possibilities.

In Civil War–era America, a person's race was classified by a combination of ancestry and physical characteristics (e.g., skin color). Throughout this article, we use the term "race" to refer to the race categories assigned to soldiers in contemporary military records. In practice, these categories tend to be correlated with differences in skin color depicted in the photos, but we avoid imposing our own assumptions about a soldier's race based on physical appearance.

7.1.2 Building the Dataset. No existing dataset was readily available for our benchmarking purposes. Furthermore, the 75 user-identified photos reviewed by the expert in the initial one-month study did not have enough diversity along the above-mentioned dimensions. This prompted us to build a new dataset for conducting our benchmarking studies.

To build this dataset, we consulted with Civil War photo experts and conducted research to locate multiple photo pairs for each of the three dimensions above, i.e., format, military service, and race. To maximize the quality of the dataset, we applied several selection criteria. To minimize age-related differences caused by the passage of time, we included only wartime photos (ca. 1861–65), excluding pre- and post-war portraits. We also required at least one of the two photos in each pair to depict the solider wearing his uniform so we could test the visual tags and filters. To ensure our ground-truth data about the identities and military records of soldier photo pairs was accurate, we collected materials only from established sources, e.g., published books, government databases, libraries, and museums.

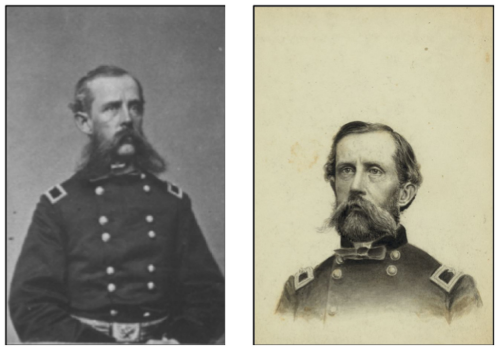

Fig. 7. An example pair for sketches (Left - database photo, Right - query sketch) *Absalom Baird* US Army Military History Institute (left) Library of Congress (right).

(a) Union (H) *Edward D. Townsend* Library of Congress (top) Library of Congress (bottom)

(b) Union (L) *Thomas Clark* Maine State Archives (top) Kurt Luther Collection (bottom)

(c) Confederate (H) *John B. Gordon* National Archives (top) Georgia Historical Society (bottom)

(d) Confederate (L) *Edmund Ruffin* Library of Congress (top) National Archives (bottom)

(e) African American *Lewis H. Douglass* Howard University (top) Museum of African American History (bottom)

Fig. 8. One photo pair example from each category (H - High Rank, L - Low Rank).

For the format dimension (Figure 7), we collected five photo pairs of replicas and five pairs of soldier sketches (i.e., one sketch and one actual photo per pair). For the military service dimension (see Figure 8), we collected five pairs of high-ranked white Union soldiers, five pairs of low-ranked white Union soldiers, five pairs of high-ranked white Confederate soldiers, and

five pairs of low-ranked white Confederate soldiers. For the race dimension (Figure 8(e)), we collected five pairs of black Union soldiers that can be compared against all of the white soldiers in the previous dimension. To summarize, we collected 10 photo pairs for the format dimension, 20 photo pairs for the military service dimension, and 5 photo pairs for the race dimension, for a total of 35 photo pairs (70 individual photos) representing 35 unique soldiers.

We added all photo pairs to the Photo Sleuth database along with the verified military service records. Some soldier profiles were already present in the database, while others were added by us. As mentioned earlier, Photo Sleuth requires that a face is successfully detected in the uploaded photo to allow the user to proceed with the identification process. Thus, we ensured faces were detected for all photo pairs in our test dataset. (Face *detection* simply means the algorithm detected a face in the image, not that it recognized or identified the face as belonging to a specific person.) Otherwise, to avoid biasing the finding, we did not have any prior knowledge about how well Photo Sleuth would perform in *identifying* the photo pairs.

Limitations. Datasets for benchmarking face recognition performance on modern-day photos often contain thousands of photo pairs (e.g., Reference [69]). Our effort stands apart for focusing on the unique challenges of historical photo pairs in general, and the context of the American Civil War specifically. Historical circumstances severely limit the availability of such images. For example, our research suggests that fewer than 10 photo pairs depicting identified wartime black Union soldiers in uniform are extant, of which our dataset brings together, likely for the first time, at least half. The scarcity of these images constrains what benchmarking and generalizable claims are possible. Despite these constraints, we argue that such efforts are nevertheless worthwhile to enrich our understanding of the past, especially the contributions of marginalized groups such as black Union soldiers who have seen comparatively less scholarly attention.

7.2 Study 1: Benchmarking AI-based Face Recognition

7.2.1 Hypotheses. Based on our findings from the initial one-month study and the historical context described above, we tested three hypotheses about Photo Sleuth, focusing on the *format* and *race* dimensions.

H1: Face recognition performs better at identifying replicas than photos with different views.
We expected face recognition to return the exact identical copy in the database as the top search result, making it easier for users to identify the replica. Our findings from the initial study showed users identifying 100% of replica photos correctly, thus supporting the assumption. However, we did not investigate how well a standalone face recognition algorithm performs on replicas. Our hypothesis stems from the belief that maximum overlap of facial landmarks can be detected in case of replicas, compared to other non-replica photos.

H2: Face recognition identifies photographs better than sketches, drawings, or lithographs.
Over the past 11 months, we observed users uploading old sketches, paintings, and lithographs of soldiers to Photo Sleuth, raising the question of how our person identification pipeline performs in these cases. Sketch-to-photo matching is an important aspect of criminal investigation often employed by law enforcement agencies [33]. However, facial recognition algorithms are far from perfect when it comes to matching sketches to photos, partly due to the of lack of real-world databases as training data [59]. Sketches mostly comprise high-frequency information, while lacking details related to texture and pigmentation, which is often insufficient for good face recognition performance [59, 72]. Thus, we hypothesize that face recognition will perform better with photos.

Table 2. Face Recognition Results for Replicas in Study 1

Soldier Name	Ranking of Correct Match	Number of Search Results	Confidence Score	Confidence Delta
Fielder A. Jones	1	134	0.90095	0.25572
Allen G. Shepherd	1	130	0.88784	0.26395
Hollis O. Dudley	1	919	0.89727	0.21225
Edgar M. Blanche	1	929	0.78837	0.07768
Thomas Barnstead	1	942	0.74837	0.01548
Average	**1**	**611**	**0.84456**	**0.16502**

H3: Face recognition identifies white soldiers better than black soldiers.
Prior work has found face recognition algorithms showing substantial accuracy disparities between darker- and lighter-skinned subgroups, such as low error rates for lighter-skinned males [13, 65]. Moreover, partly due to historical circumstances, the Photo Sleuth database has significantly fewer identified black soldiers (<584) compared to white soldiers (>20,000). This disparity might tilt the face recognition results in favor of white soldiers, and thus forms the basis for this hypothesis.

7.2.2 Tasks. For all the soldiers in each category, we searched one photo on Photo Sleuth without selecting any visual tags or search filters and examined the search results returned by AI-based face recognition. The search space was the entire database, since no filters were selected for this study. This included all 11 months of user-uploaded data. We used Photo Sleuth's default face recognition confidence threshold of 0.50 for our study; i.e., only search results with a similarity confidence score greater than 0.50 would show up in the search results.

7.2.3 Metrics and Data Analysis. For each test photo, we recorded: (1) the ranking of the correct match in the search results, (2) the confidence score from face recognition, and (3) the total number of search results. We also recorded the confidence score for the next search result that followed the correct match in rankings, and calculated the difference between their scores (*confidence delta*).

7.2.4 Findings: Study 1.

Replicas. Results for replicas are shown in Table 2. For all five replica pairs, face recognition returned the correct match as the top-ranked search result. We do not observe similar perfect rankings for the non-replica photos (cf. Tables 4, 5), where in some cases the correct match does not show up as the top-ranked search result, if at all. Replicas also have the highest average confidence scores and confidence deltas compared to non-replicas (cf. Tables 4, 5). Given that the replicas show up as the top-ranked search result and have the highest average confidence deltas, we can infer that face recognition detects the maximum similarity in facial features here compared to other non-replicas. Our findings, therefore, support H1.

However, even though replicas are near-identical copies—i.e., a best-case scenario for face recognition—confidence scores ranged widely, from only 0.74837 to 0.90095. Further, in the cases of Thomas Barnstead and Edgar Blanche, the confidence deltas were less than 0.02, meaning that incorrect matches nearly beat them. Thus, even though automated face recognition performed very well with rankings, the low average and wide range for confidence scores, and low confidence deltas, suggest that identifying even replicas is nontrivial for automated techniques.

The number of search results for replicas also varied widely, from about 130 results in two cases to about 900 in the other three cases. The wide range in number of search results may be attributed to the uniqueness of the faces being queried or the quality of the images, as seen through the lens

Table 3. Face Recognition Results for Sketches in Study 1

Soldier Name	Ranking of Correct Match	Number of Search Results	Confidence Score	Confidence Delta
Absalom Baird	66	918	0.67579	0.01417
John M. Palmer	657	888	0.5575	0.00005
James Slack	–	21	–	–
Bryan Grimes	7	893	0.64735	0.00181
John G. Mitchell	11	920	0.66305	0.00143
Average	**185**	**728**	**0.63592**	**0.00437**

Table 4. Face Recognition Results for Race (Black Union Soldiers) in Study 1

Soldier Name	Ranking of Correct Match	Number of Search Results	Confidence Score	Confidence Delta
William Matthews	7	28	0.54476	0.00521
Daniel S. Lathrop	1	9	0.79925	0.21152
William H. Dupree	1	89	0.75159	0.13030
James Monroe Trotter	1	68	0.68694	0.09754
Lewis Douglass	1	5	0.83395	0.23734
Average	**2**	**40**	**0.72330**	**0.13638**

of the face recognition algorithm. Faces with more distinctive features will generally return fewer search results, and vice versa. Likewise, in lower-quality images, face recognition may detect fewer facial landmarks, increasing the number of potential matches in search results.

Sketches. Results for sketches are shown in Table 3. Face recognition does not return the correct match as the top-ranked search result for any of five sketches. In the case of James Slack, face recognition was unable to retrieve the correct match at all. For sketches, the high average rank of 185 for the correct match suggests that, practically speaking, the correct match will be lost among the search results. Sketches also show the lowest average confidence delta among all categories (see Tables 2, 4, 5). All of these findings indicate that face recognition shows poor performance in identifying sketches compared to photographs, thus supporting H2.

Race. Face recognition showed similar results for white versus black soldiers. It was able to find all the correct matches in case of black soldiers, while it missed out on 4 (out of 20) white soldiers. Except for these 4 and John Singleton Mosby (ranked 148), all other white soldiers showed up as the top-ranked search result. For black soldiers, all except 1 (William Matthews) showed up as the top-ranked search result. The average confidence scores and confidence deltas are also similar for black and white soldiers, with the similar confidence deltas being especially notable, given the substantially larger pool of search results for white soldiers (>20,000 vs. <500 black soldiers) and corresponding larger number of potential false positives. Therefore, H3 is not supported, as face recognition identified both black and white soldiers equally in this context.

This outcome conflicts with the findings of prior studies, which have shown commercial face recognition algorithms performing poorly on faces of darker-skinned males compared to lighter-skinned males [13, 65]. However, the smaller sample size in this study and other contextual factors (e.g., fewer black soldiers in the reference database) may have influenced this outcome. Therefore, it is not sufficient to infer the overall performance of face recognition algorithms towards different skin colors in unconstrained scenarios.

Table 5. Face Recognition Results for Race (White Union and Confederate Soldiers) in Study 1

Soldier Name	Ranking of Correct Match	Number of Search Results	Confidence Score	Confidence Delta
Frank H. Peck	1	886	0.81276	0.12698
Orrin E. Smith	–	208	–	–
Edward D. Townsend	1	3	0.58057	0.02039
Joseph B. Carr	1	933	0.77012	0.07267
Christopher C. Andrews	1	860	0.81754	0.10053
Boston Corbett	1	185	0.91801	0.31785
Thomas Clark	1	903	0.82801	0.1572
Joseph Lyman	1	65	0.7411	0.12549
Jasper Warren	1	913	0.78134	0.10538
Augustus Weissert	1	68	0.74524	0.11792
M. J. Thompson	–	3	–	–
John B. Gordon	1	921	0.7588	0.00505
Joseph E. Johnston	1	846	0.78024	0.06523
John Singleton Mosby	148	919	0.62263	0.00018
J. B. Richardson	1	939	0.8772	0.15942
John L. Rapier	1	103	0.69646	0.08295
Edmund Ruffin	–	36	–	–
William Anderson Roberts	1	243	0.78994	0.13453
William Toffier	–	437	–	–
Columbus Rush	1	58	0.73237	0.10053
Average	**10**	**477**	**0.76577**	**0.10577**

7.3 Study 2: Benchmarking the Visual Tags and Filters

7.3.1 Hypotheses.

H4: Adding visual tags for the soldier in a mystery photo improves the ranking of the correct match in search results.
Visual tags are mapped on to a soldier's military service information and activate relevant search filters upon selection, which narrows down the candidate pool in which face recognition performs its search. By adding these tags, we expect that the candidate pool will comprise soldiers most likely to have held the same military positions as the mystery photo, and therefore the correct match might show up at a higher rank among search results compared to no visual tags.

H5: Adding visual tags for the soldier in a mystery photo reduces the number of false positives in search results.
Because the search pool, upon adding tags, would filter out soldiers who may not have held the same military position as the soldier in the mystery photo, we expect there is a lower likelihood of false positives in the search results.

7.3.2 Tasks. Similar to the tasks in the first study, we searched for one photo of each soldier. In this second study, however, we also applied visual tags for each search in two to three iterative rounds. In the first round, we added the first visual tag (coat color) to filter the database for the army of the soldier. In the second round, we added the second visual tag (shoulder straps or chevrons) to filter the database for the rank of the soldier. In an optional third round, we add a third visual

tag, if present. This is usually a hat insignia or letters on the hat to filter the branch or regiment. For black soldiers, we set the third search filter to be US Colored Troops to select the military unit in which the African American soldiers served. As with the first study, our search space here included 11 months of user-uploaded data, and we used Photo Sleuth's default face recognition confidence threshold of 0.50.

In this study, we are investigating visual (uniform) tags that are related to military records and unrelated to the facial features examined in previous study. Therefore, for conducting this study, we used photos only from the five categories based on military credentials.

7.3.3 Metrics and Data Analysis. Similar to the previous study, we recorded the ranking and the number of search results. For this study, we also record the change, if any, in ranking and number of search results in each successive round of tags.

7.3.4 Findings: Study 2.

Ranking. We noted in the previous study that for most photos, face recognition without any tags already returned the correct match as the top-ranked search result (see Tables 5, 4). After adding the visual tags for both army and rank, the correct matches still show up as the top-ranked search result for these photos (see Table 6). In these cases, the visual tags complement face recognition by acting as corroborative evidence for the correct match. However, there was one exception (Joseph E. Johnston) where the correct match disappeared from the search results after applying the second round (rank) tag. On further investigation, we found out that the Photo Sleuth database had missing military service information for that soldier. This error points to double-edged sword of crowdsourced additions to the reference database. While users can greatly scale up the inclusion of new photos and biographical profiles, the effectiveness of the search filters is beholden to the accuracy of this unverified information.

Face recognition did not pick the correct match as the top-ranked search result for two soldiers: John Singleton Mosby and William Matthews. For these soldiers, adding the army tags improved the rankings, from 148 to 28, and seven to three, respectively. Adding the rank tags further improved the rankings, from 28 to 12, and three to two, respectively. In case of Matthews, we applied a third-round filter (US Colored Troops) because he was a black soldier, improving the ranking from two to one. Thus, for both cases where face recognition did not return the correct match as the top-ranked result, each successive round of tags and filters, without exception, improved the ranking, moving from 148 to 12 (Mosby) and seven to one (Matthews). Here, these tags complement face recognition by improving the ranking and providing corroborative evidence for the correct match.

To summarize, AI-based face recognition performed well without any tags, returning the correct match as the top result in all but two cases. Visual tags retained (but did not improve) these top rankings, providing corroborative evidence. In the two cases where the correct match was not the top ranking, each successive round of tags and filters improved the ranking, culminating in either a successful number-one ranking (Matthews) or substantially closing the gap (Mosby). Thus, our findings support H4.

False Positives. Face recognition, without any tags, generally returned a large number of search results and false positives (mean = 389). Table 7 shows that adding the army tags narrowed down the number of false positives (mean = 228) while retaining the correct match. The second-round rank tags reduced them further (mean = 81). In the cases where additional (third-round) tags were present, the number of false positives dropped even further (mean = 26). As illustrated in Figure 9, this trend was evident across all categories. As we add more visual tags, the number of false

Table 6. Impact of Visual Tags on Rankings in Study 2

Soldier Name	Category	Ranking of Correct Match			
		No. Tags	Army Tag	Army + Rank Tags	Army + Rank + Additional Tags
Frank H. Peck	UH	1	1	1	N/A
Orrin E. Smith	UH	–	–	–	N/A
Edward D. Townsend	UH	1	1	1	N/A
Joseph B. Carr	UH	1	1	1	N/A
Christopher C. Andrews	UH	1	1	1	N/A
Boston Corbett	UL	1	1	1	1
Thomas Clark	UL	1	1	1	N/A
Joseph Lyman	UL	1	1	1	1
Jasper Warren	UL	1	1	1	1
Augustus Weissert	UL	1	1	1	N/A
M. J. Thompson	CH	–	–	–	N/A
John B. Gordon	CH	1	1	1	N/A
Joseph E. Johnston	CH	1	1	–	N/A
John Singleton Mosby	CH	148	28	12	N/A
J. B. Richardson	CH	1	1	1	N/A
John L. Rapier	CL	1	1	1	N/A
Edmund Ruffin	CL	–	–	–	N/A
William Anderson Roberts	CL	1	1	1	N/A
William Toffier	CL	–	–	–	N/A
Columbus Rush	CL	1	1	1	N/A
William Matthews	AF	7	3	2	1
Daniel S. Lathrop	AF	1	1	1	1
William H. Dupree	AF	1	1	1	1
James Monroe Trotter	AF	1	1	1	1
Lewis Douglass	AF	1	1	1	1

UH – Union high-ranked, UL – Union low-ranked, CH – Confederate high-ranked, CL – Confederate low-ranked.

positives declines while retaining the correct match. The only exception, also mentioned above, was Joseph Johnston, due to missing profile data.

However, the different slopes of the lines after each round suggests some tags are more powerful than others. Confederate army tags showed the steepest downward slopes for both ranks, due to the smaller pool of identified reference images. In contrast, black soldiers had the flattest slopes at all rounds, suggesting that the tags had limited impact. Given that only the Union Army had black soldiers and few black soldiers held high rank, the low impact of the army and rank tags here is not surprising. However, there were few black soldiers in the database to begin with, and the initially low number of false positives before any tags were applied suggests that face recognition *per se* effectively rules out many white false positives on the basis of facial features. Applying all the tags further narrowed down the false positives for all the black soldiers. However, it does not appear to be a significant reduction compared to other categories that initially had larger numbers of false positives before tags were applied.

We also observed that the army tags narrow down the search results to a greater extent for Confederate soldiers than Union ones. This might be because of fewer Confederates in the database compared to Union soldiers. We also observe that tags for low ranks in both armies

Table 7. Impact of Visual Tags on False Positives in Study 2

Soldier Name	Category	Number of Search Results			
		No. Tags	Army Tag	Army + Rank Tags	Army + Rank + Additional Tags
Frank H. Peck	UH	886	804	353	N/A
Orrin E. Smith	UH	208	200	132	N/A
Edward D. Townsend	UH	3	3	1	N/A
Joseph B. Carr	UH	933	849	291	N/A
Christopher C. Andrews	UH	860	767	119	N/A
Average	**UH**	**578**	**525**	**179**	**N/A**
Boston Corbett	UL	185	177	54	21
Thomas Clark	UL	903	881	209	N/A
Joseph Lyman	UL	65	65	64	18
Jasper Warren	UL	913	866	209	138
Augustus Weissert	UL	68	67	65	N/A
Average	**UL**	**427**	**411**	**120**	**59**
M. J. Thompson	CH	3	0	0	N/A
John B. Gordon	CH	921	162	91	N/A
Joseph E. Johnston	CH	846	188	59	N/A
John Singleton Mosby	CH	919	160	83	N/A
J. B. Richardson	CH	939	197	92	N/A
Average	**CH**	**726**	**141**	**65**	**N/A**
John L. Rapier	CL	103	16	15	N/A
Edmund Ruffin	CL	36	12	12	N/A
William Anderson Roberts	CL	243	34	33	N/A
William Toffier	CL	437	52	51	N/A
Columbus Rush	CL	58	9	9	N/A
Average	**CL**	**175**	**25**	**24**	**N/A**
William Matthews	AF	28	15	9	1
Daniel S. Lathrop	AF	9	9	5	5
William H. Dupree	AF	89	89	19	12
James Monroe Trotter	AF	68	61	45	9
Lewis Douglass	AF	5	5	5	4
Average	**AF**	**40**	**36**	**17**	**6**
Overall Average		**389**	**228**	**81**	**26**

UH – Union high-ranked, UL – Union low-ranked, CH – Confederate high-ranked, CL – Confederate low-ranked, AF – African American.

(24 Confederate and 120 Union) filter out false positives to a greater extent than the ones for higher ranks (65 Confederate and 179 Union). Adding the USCT filter for black soldiers also reduces false positives. All of these findings show the number of false positives declining with the addition of visual tags, supporting H5.

7.4 Summary of Findings

Our benchmarking studies showed that AI-based face recognition, in general, had good recall at Photo Sleuth's 0.50 confidence threshold for Civil War photos across diverse examples with respect

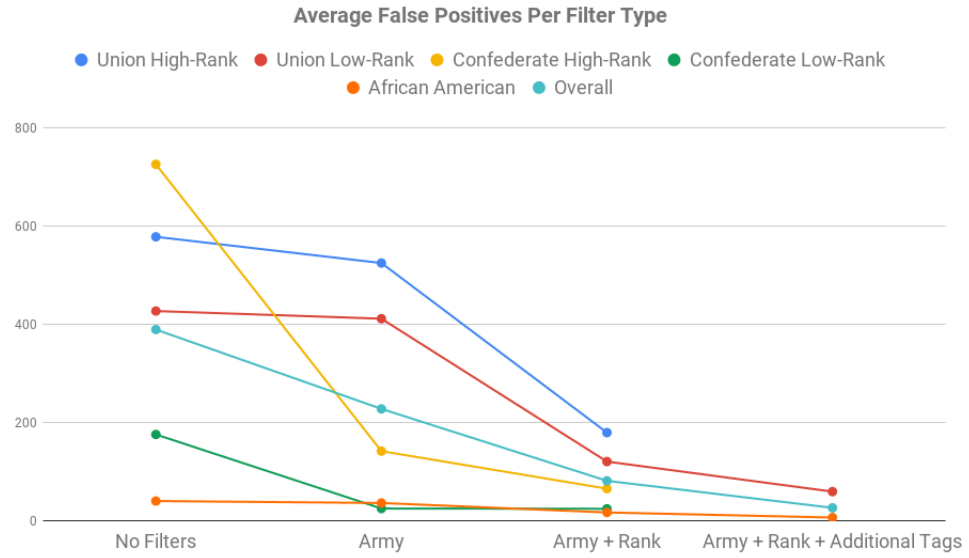

Fig. 9. Average number of false positives by tag type in study 2.

to format, race, and military credentials. Aside from a handful of exceptions, it was able to retrieve the correct match within its search results and with high rankings (often top one). We also found it was able to identify replicas better than photos with different views, though with lower confidence scores and deltas than expected, and struggled to identify sketches or drawings.

Face recognition also showed low precision, returning a large number of search results that are mostly false positives. This issue is complemented by Photo Sleuth's visual tags and filters, which consistently retained correct matches, reduced false positives, and improved rankings when the top result was incorrect. Tagging and filtering provided corroborative evidence and allowed users to focus on the highest-probability matches. While they proved especially powerful for some rare photo categories, such as Confederate soldiers, they were less useful for black soldiers, for which face recognition already reduced most false positives by excluding white soldiers.

8 DISCUSSION

8.1 Fostering Original Research while Preventing Misinformation

Prior work pointed to problems with misinformation in online history communities [82, 83], a concern also voiced by our study participants. In Photo Sleuth, we made design decisions explicitly to support accuracy and limit the spread of misinformation. One such decision was to give users the option to provide the original sources of identification to add credibility to their identification claims. Although this feature was optional, users took advantage of it for all but 36 of 386 pre-identified photos.

A second design decision to promote accuracy was requiring all users to go through the entire pipeline, even if they believed they already knew the pictured soldier's name. A third, related design decision was asking users to separate the visual clues they could actually observe in the image (e.g., tagging visible rank insignia) from their interpretation of the clues (e.g., activating search filters for certain ranks). Both of these design decisions encouraged tagging of more objective visual evidence, with 401 of 602 unidentified photos and 445 of 560 identified photos receiving tags in the first month. These interfaces allowed for clearer delineations between fact and opinion, and left room for reasonable disagreement.

In the first month, users post-identified 75 unknown historical portraits, including 25 in the most difficult category (no inscription and no replica). This is promising evidence of the success of our approach—in P6's words, traditionally, "it's really rare that you can identify a non-identified image." However, 13 of the 88 post-identifications were judged by our expert as negative matches, indicating potential misinformation. In future work, we are exploring allowing users to express more nuanced confidence levels in their identifications, based on the the expert's four-point Likert scale, as well as capturing user disagreements.

8.2 Building a Sustainable Model for Volunteer Contributions

We observed substantial volunteer contributions to Photo Sleuth in its first month, even without typical incentive mechanisms such as points and leaderboards. In interviews, participants described a variety of motivations for adding and tagging both unidentified and identified photos, ranging from generating income to preserving history. Given the promising results from our initial one-month evaluation and longitudinal 11-month evaluation suggesting high-quality, steadily growing contributions from over 12,000 registered users, we are optimistic that we have built a sustainable model for volunteer participation.

Our workflow leverages network effects so the more people use it, the more beneficial it becomes to all. Users, when uploading and tagging known and unknown photographs, are enhancing the reference archive. These photos, along with their visual tags and metadata, are bootstrapped into the system for future searches and identifications, allowing the website to continuously grow. These users are also publicly credited for their contributions. Designing crowdsourcing workflows that align incentive mechanisms for enriching metadata and performing searches, as well as publicly recognizing contributions, can help build a sustainable participation model.

Keeping up with the needs of the ever-growing user base, we are committed to the ongoing maintenance and improvement of the Photo Sleuth platform. In the near future, we plan on providing a public RESTful API to facilitate interchange with the digitized collections of libraries and museums and expand our database. Our goals for the future feature enhancements and database expansions aim to promote a self-sustaining platform that actively supports historical research.

8.3 Combining the Strengths of Crowds and AI

We deliberately decided not to allow the Photo Sleuth system *per se* to automatically identify any photos. Although this feature is one of our most persistent user requests, examples from popular media show the danger of a fully automated approach [5, 71]. Instead, the system suggests potential matches largely driven by objective user tagging and hides quantitative confidence levels. The face recognition algorithm influences results in a more subtle way, by filtering out low-confidence matches and sorting the remainder. We believe this approach improves accuracy, but at the cost of increased requirements for human attention per image. Because Photo Sleuth helps users quickly identify a much more relevant set of candidates compared to traditional research methods, participants did not seem to view this attention requirement as a major drawback.

This human-led, AI-supported approach to person identification is further emphasized in our design decision to attribute individual users as responsible for particular identifications. This approach aims to promote accountability through social translucence [23] and to recognize the achievements of conducting original research, as recommended by prior work [53, 54]. It also aligns with traditions of expert authentication in the art history and antiquarian communities.

Our benchmarking study further illustrated the complementary strengths of humans and computational approaches. Face recognition generally showed high recall on our test dataset, often returning the correct match ranked first out of over 20,000 candidates. However, sometimes face recognition did not rank the correct match first, or entirely excluded it from the results. Even when

face recognition ranked the correct match first, confidence scores were as low as 0.58057 for non-replicas and 0.74837 for replicas, and confidence deltas were often very small. Further, it often returned hundreds of false positives. Taken together, these results suggest that with respect to Photo Sleuth's face recognition, (1) top-ranked results are not always correct matches, (2) low-ranked results (i.e., ranked 100+) are sometimes correct matches, (3) low-confidence results (i.e., <0.60) are sometimes correct matches, (4) correct and incorrect matches often have very similar confidence scores (i.e., within <0.001 for non-replicas), and (5) most results—often hundreds—will be false positives. Fortunately, Photo Sleuth's human-authored visual tags connected to military records address many of these shortcomings. Correctly applied tags consistently filtered out false positives while retaining correct matches and sometimes improving rankings. However, more work is needed to help users tackle this "last-mile" challenge of selecting the correct match among a small number of strong candidates.

Relatedly, we observed users posting screenshots of Photo Sleuth on social media to solicit second opinions from the community. This suggests a potential benefit of the wisdom of crowds not yet supported by our system, but also the potential dangers of groupthink. In future work, we are exploring ways to capture discussions directly within Photo Sleuth's Comparison interface, drawing inspiration from social computing systems supporting reflection and deliberation around contentious topics [37, 38]. Leveraging the collective interests and expertise of the Photo Sleuth community for providing feedback regarding facial similarity, and filtering out incorrect candidates, can contribute towards solving this "last-mile" challenge.

8.4 Enhancing the Accuracy of Person Identification

Prior work on person identification has mostly been limited to studies of face recognition algorithms. These studies often focus on face verification evaluations, i.e., comparing two photos and providing a confidence score about how similar or different they are. The algorithm gives the final verdict on a potential match based on a confidence threshold. Such approaches are usually evaluated on fixed datasets and are therefore prone to false positives. Even though human-machine fusion scores are shown to outperform individual human or machine performances, none of these systems proposes a hybrid pipeline where human judgment complements that of a machine, or vice versa.

Photo Sleuth addresses accuracy issues in person identification by enhancing face recognition with different layers of contextual information, such as visual clues, biographical details, and photo metadata. Users provide visual clues along with the face, which help the system in generating search filters based on military service records. This ensures that the facial recognition runs on a plausible subset of soldiers satisfying the clues. We also show how users consider photo metadata such as period inscriptions and historical primary sources to correctly match a person with an identity. Since the final decision of identification is reserved for the users, they can make an informed decision based on the contextual information along with facial similarity.

Our benchmarking study revealed face recognition's poor performance in finding the correct matches using soldier sketches. We observed large numbers of false positives and low ranking of the correct matches in the search results, if found at all. These sketches were hand-drawn representations of how the soldiers looked in real life, and therefore might fail to capture important identifiable facial features or introduce noise and distortions. Collectively, these factors can greatly influence the performance of face recognition, as our findings showed, posing a substantial risk of missing the correct matches in the search results.

The unreliability of sketches as an input for face recognition has greater consequences in high-stake scenarios such as law enforcement, where such practices appear to be commonplace. There are at least half a dozen police departments in the US that permit, if not encourage, the use of

composite sketches in face recognition systems to identify suspects [26]. Multiple studies, along with accounts from law enforcement officials, have found face recognition yielding marginal success with composite sketches [26, 29, 34, 36]. Furthermore, forensic artists draw individuals they have never seen before by listening to the memory recollections of victims and witnesses, which opens up the possibility for additional errors to reduce the reliability of sketches. Therefore, great caution should be exercised if using sketches as an input, and conclusions should draw heavily from all available contextual information and human judgment at every step of the investigation.

In future work, Photo Sleuth's pipeline could be adapted for other historical or modern person identification tasks by incorporating a domain-specific database and tagging features in a context-specific manner. For example, to identify criminal suspects in surveillance footage or locate missing persons from social media photos, an initial seed database of identified portraits with biographical data could be fed to the system. The user interface could be tuned, with the guidance of subject matter experts, to support tagging relevant photo metadata and visual clues such as distinctive tattoos, clothing styles, and environmental features. These tags could similarly be linked to search filters to narrow down candidates after face recognition. Especially in high-stakes domains like these examples, where both false positives and false negatives can have life-altering impacts, it would be critical for experts in law enforcement or human rights investigation to oversee the person identification process.

9 CONCLUSION

Photo Sleuth attempts to address the challenge of identifying people in historical portraits. We present a novel person identification pipeline that combines crowdsourced human expertise and automated face recognition with contextual information to help users identify unknown Civil War soldier portraits. We demonstrate this approach by building a web platform, Photo Sleuth, on top of this pipeline. We launched Photo Sleuth as a free public website, and through two evaluations of real-world usage, we show that Photo Sleuth's pipeline has enabled identification of dozens of unknown photos and encouraged a sustainable model for long-term volunteer contribution. We also present results from a benchmarking study. Our work opens doors for exploring new ways for building person identification systems that look beyond face recognition and leverage the complementary strengths of human and artificial intelligence.

APPENDIX

A SCREENSHOTS OF PHOTO SLEUTH

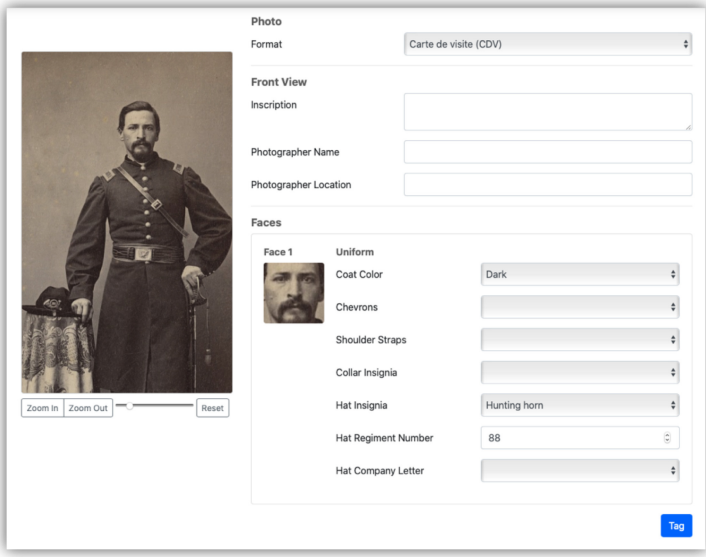

Fig. 10. Tagging a photo.

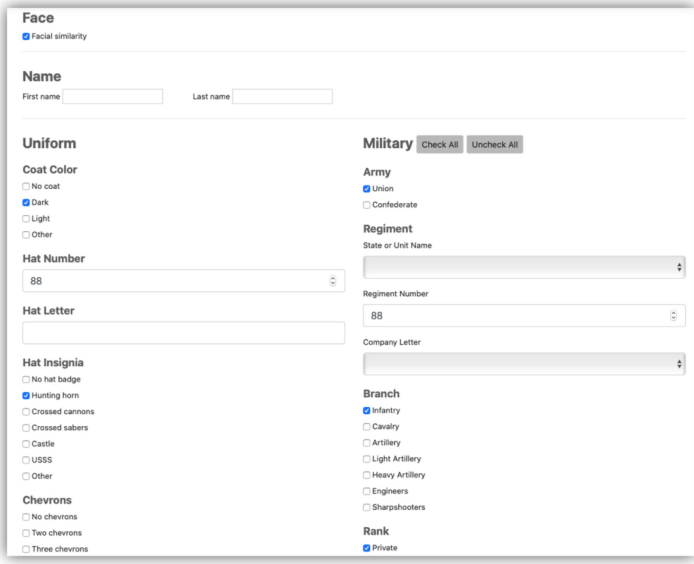

Fig. 11. Suggesting search filters.

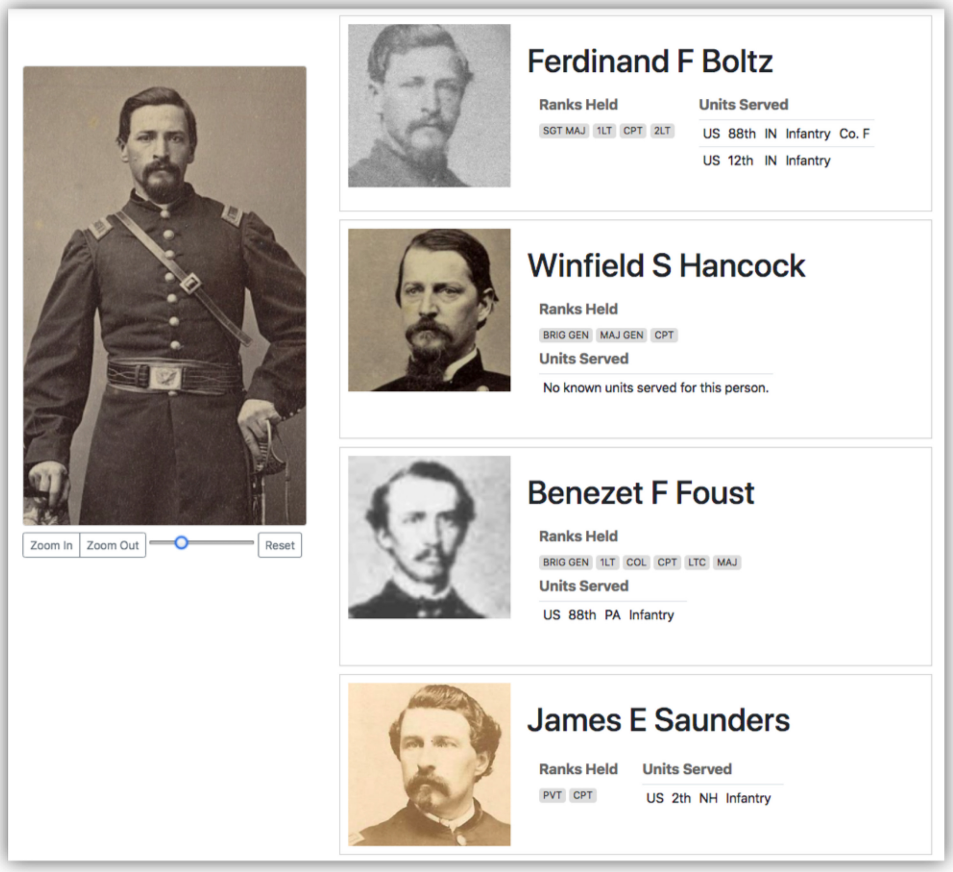

Fig. 12. Reviewing search results.

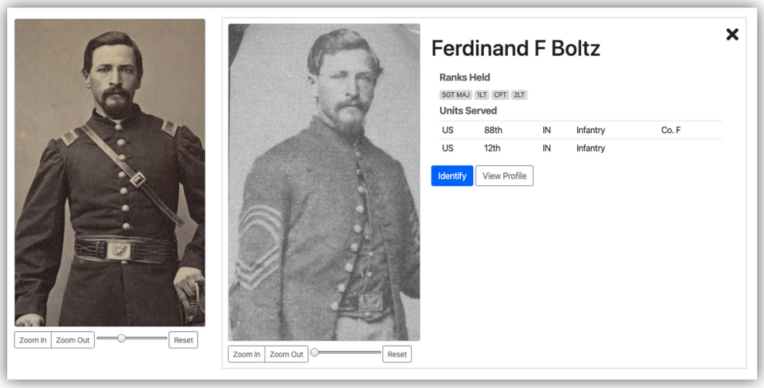

Fig. 13. Comparing profiles.

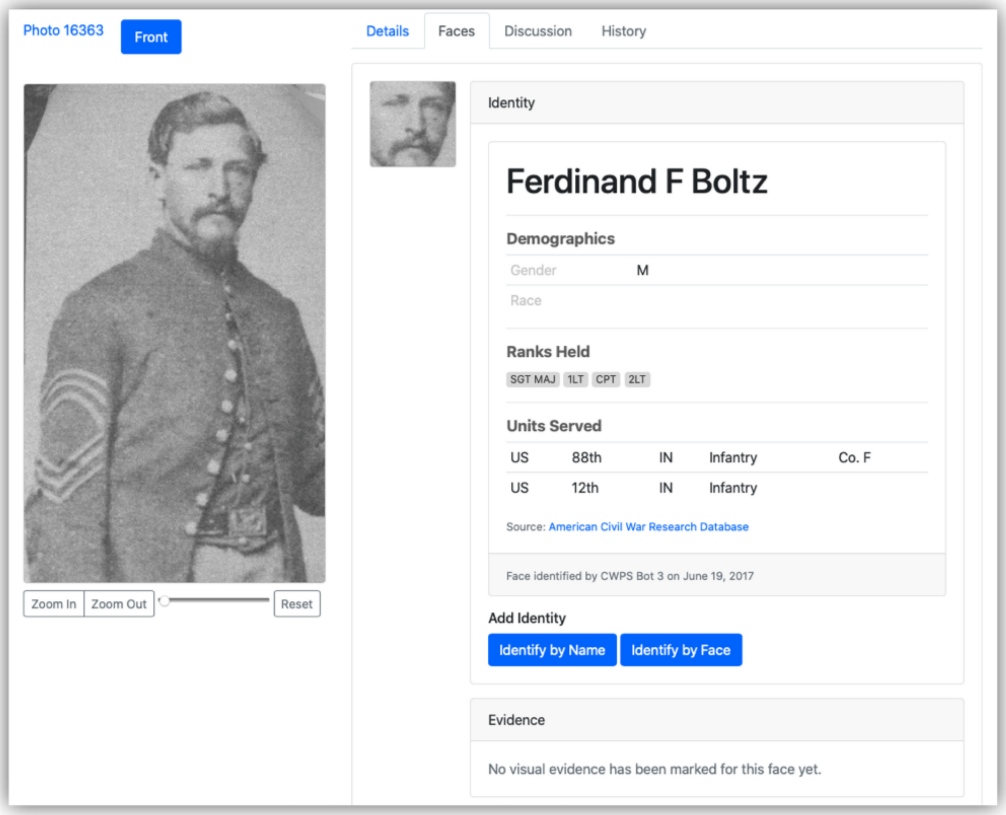

Fig. 14. An Identified soldier profile.

ACKNOWLEDGMENT

We wish to thank Ron Coddington, Paul Quigley, Nam Nguyen, Abby Jetmundsen, Ryan Russell, Natalie Robinson, and our study participants.

REFERENCES

[1] 1860. Liljenquist Family Collection of Civil War Photographs—About this Collection. Retrieved from www.loc.gov/pictures/collection/lilj/.

[2] 1861. Surgeon Francis M. Eveleth of 7th Maine Infantry Regiment and 1st Maine Veteran Infantry Regiment in assistant surgeon uniform with Ames medical sword. Retrieved from https://www.loc.gov/pictures/item/2012649100/.

[3] 2013. American Battlefield Trust - An Interview with Ron Coddington, Editor of Military Images Magazine. Retrieved from https://www.battlefields.org/learn/articles/military-images-magazine.

[4] Amazon. 2018. Amazon Rekognition Customers—Amazon Web Services (AWS). Retrieved from https://aws.amazon.com/rekognition/customers/.

[5] Press Association. 2018. Welsh police wrongly identify thousands as potential criminals. *The Guardian* (5 May 2018). Retrieved from https://www.theguardian.com/uk-news/2018/may/05/welsh-police-wrongly-identify-thousands-as-potential-criminals.

[6] M. Barrenechea, K. M. Anderson, L. Palen, and J. White. 2015. Engineering crowdwork for disaster events: The human-centered development of a lost-and-found tasking environment. In *Proceedings of the 48th Hawaii International Conference on System Sciences*. 182–191. DOI: https://doi.org/10.1109/HICSS.2015.31

[7] P. Bell and Björn Ommer. 2016. *Digital Connoisseur? How Computer Vision Supports Art History*. Artemide, Rome.

[8] Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. 2011. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. ACM, 33–42.

[9] L. Best-Rowden, S. Bisht, J. C. Klontz, and A. K. Jain. 2014. Unconstrained face recognition: Establishing baseline human performance via crowdsourcing. In *Proceedings of the IEEE International Joint Conference on Biometrics*. 1–8. DOI: https://doi.org/10.1109/BTAS.2014.6996296

[10] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White et al. 2010. VizWiz: Nearly real-time answers to visual questions. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology*. ACM, 333–342.

[11] Austin Blanton, Kristen C. Allen, Timothy Miller, Nathan D. Kalka, and Anil K. Jain. 2016. A comparison of human and automated face verification accuracy on unconstrained image sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 161–168.

[12] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. 2013. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2117–2126.

[13] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT'18)*. 77–91. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a.html.

[14] Christopher Carbone. 2019. AI could help identify Civil War veterans in your family. Retrieved from https://www.foxnews.com/tech/ai-could-help-identify-civil-war-veterans-in-your-family.

[15] Tim Causer and Melissa Terras. 2014. Crowdsourcing Bentham: Beyond the traditional boundaries of academic history. *Int. J. Hum. Arts Comput.* 8, 1 (Apr. 2014), 46–64. DOI: https://doi.org/10.3366/ijhac.2014.0119

[16] Justin Cheng and Michael S. Bernstein. 2015. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 600–611.

[17] Ronald S. Coddington. 2008. *Faces of the Confederacy: An Album of Southern Soldiers and Their Stories*. The Johns Hopkins University Press, Baltimore, MD.

[18] Ronald S. Coddington. 2012. *African American Faces of the Civil War: An album*. The Johns Hopkins University Press, Baltimore, MD.

[19] Ronald S. Coddington and Michael Fellman. 2004. *Faces of the Civil War: An Album of Union Soldiers and Their Stories* (1st ed.). The Johns Hopkins University Press, Baltimore, MD.

[20] Lhaylla Crissaff, Louisa Wood Ruby, Samantha Deutch, R. Luke DuBois, Jean-Daniel Fekete, Juliana Freire, and Claudio Silva. 2018. ARIES: Enabling visual exploration and organization of art image collections. *IEEE Comput. Graph. Applic.* 38, 1 (2018), 91–108.

[21] Alexandra Dantzer. 2018. Unknown Civil War Faces Are Being Identified through Facial Recognition App. Retrieved from https://www.thevintagenews.com/2018/12/01/civil-war-photo-sleuth/.

[22] Erica X. Eisen. 2018. Historians are using facial recognition software to identify people in Civil War photographs. *Slate Mag.* (Nov. 2018). Retrieved from https://slate.com/technology/2018/11/civil-war-photo-sleuth-facial-recognition.html.

[23] Thomas Erickson and Wendy A. Kellogg. 2000. Social translucence: An approach to designing systems that support social processes. *ACM Trans. Comput.-hum. Interact.* 7, 1 (Mar. 2000), 59–83. DOI: https://doi.org/10.1145/344949.345004

[24] Jacey Fortin. 2018. A photo of Billy the Kid bought for $10 at a flea market may be worth millions. *The New York Times* (16 Nov. 2017). Retrieved from https://www.nytimes.com/2017/11/16/us/billy-the-kid-photo.html.

[25] Jacey Fortin. 2018. She was the only woman in a photo of 38 scientists, and now she's been identified. *The New York Times* (19 Mar. 2018). Retrieved from https://www.nytimes.com/2018/03/19/us/twitter-mystery-photo.html.

[26] Clare Garvie. 2019. Face recognition in flawed data. Retrieved from https://www.flawedfacedata.com.

[27] Google. 2018. Google App Goes Viral Making an Art Out of Matching Faces to Paintings. Retrieved from https://www.npr.org/sections/thetwo-way/2018/01/15/578151195/google-app-goes-viral-making-an-art-out-of-matching-faces-to-paintings.

[28] David Grossman. 2019. AI could help you identify Civil War vets in your family tree. *Pop. Mech.* (Mar. 2019). Retrieved from https://www.popularmechanics.com/military/a26625006/civil-war-photo-sleuth-search/.

[29] Patrick J. Grother and Mei L. Ngan. 2014. *Face Recognition Vendor Test (FRVT) Performance of Face Identification Algorithms*. NIST Interagency/Internal Report (NISTIR)-8009.

[30] Derek L. Hansen, Patrick J. Schone, Douglas Corey, Matthew Reid, and Jake Gehring. 2013. Quality control mechanisms for crowdsourcing: Peer review, arbitration, & expertise at familysearch indexing. In *Proceedings of the Conference on Computer Supported Cooperative Work*. ACM, 649–660.

[31] Kotaro Hara, Shiri Azenkot, Megan Campbell, Cynthia L. Bennett, Vicki Le, Sean Pannella, Robert Moore, Kelly Minckler, Rochelle H. Ng, and Jon E. Froehlich. 2015. Improving public transit accessibility for blind riders by crowdsourcing bus stop landmark locations with Google Street View: An extended analysis. *ACM Trans. Access. Comput.* 6, 2 (2015), 5.

[32] Kotaro Hara, Vicki Le, and Jon Froehlich. 2013. Combining crowdsourcing and Google Street View to identify street-level accessibility problems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 631–640.

[33] Drew Harwell. 2019. Oregon became a testing ground for Amazon's facial-recognition policing. But what if Rekognition gets it wrong? *Washington Post* (30 Apr. 2019). Retrieved from https://www.washingtonpost.com/technology/2019/04/30/amazons-facial-recognition-technology-is-supercharging-local-police/.

[34] Anil K. Jain, Brendan Klare, and Unsang Park. 2011. Face recognition: Some challenges in forensics. In *Proceedings of the Face and Gesture Conference*. IEEE, 726–733.

[35] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. 2016. The MegaFace benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4873–4882.

[36] Scott Klum, Hu Han, Anil K. Jain, and Brendan Klare. 2013. Sketch based face recognition: Forensic vs. composite sketches. In *Proceedings of the International Conference on Biometrics (ICB'13)*. IEEE, 1–8.

[37] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. 2014. Integrating on-demand fact-checking with public dialogue. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW'14)*. ACM, New York, NY, 1188–1199. DOI: https://doi.org/10.1145/2531602.2531677

[38] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Andrew Ko. 2012. Is this what you meant?: Promoting listening on the web with reflect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*. ACM, New York, NY, 1559–1568. DOI: https://doi.org/10.1145/2207676.2208621

[39] Ranjay A. Krishna, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A. Shamma, Li Fei-Fei, and Michael S. Bernstein. 2016. Embracing error to enable rapid crowdsourcing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, New York, NY, 3167–3179. DOI: https://doi.org/10.1145/2858036.2858115

[40] Neeraj Kumar, Alexander Berg, Peter N. Belhumeur, and Shree Nayar. 2011. Describable visual attributes for face verification and image search. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 10 (2011), 1962–1977.

[41] Gierad Laput, Walter S. Lasecki, Jason Wiese, Robert Xiao, Jeffrey P. Bigham, and Chris Harrison. 2015. Zensors: Adaptive, rapidly deployable, human-intelligent sensor feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1935–1944.

[42] Walter S. Lasecki, Mitchell Gordon, Danai Koutra, Malte F. Jung, Steven P. Dow, and Jeffrey P. Bigham. 2014. Glance: Rapidly coding behavioral video with the crowd. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. ACM, 551–562.

[43] Walter S. Lasecki, Mitchell Gordon, Winnie Leung, Ellen Lim, Jeffrey P. Bigham, and Steven P. Dow. 2015. Exploring privacy and accuracy trade-offs in crowdsourced behavioral video coding. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1945–1954.

[44] Kevin M. Levin. 2019. *Searching for Black Confederates: The Civil War's Most Persistent Myth*. The University of North Carolina Press, Chapel Hill.

[45] Chris J. Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu et al. 2008. Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Month. Not. Roy. Astron. Societ.* 389, 3 (2008), 1179–1189.

[46] Melissa Locker. 2018. Online sleuths are using face recognition to identify Civil War soldiers in old photographs. *Fast Company* (Dec. 2018). Retrieved from https://www.fastcompany.com/90275255/online-sleuths-are-using-face-recognition-to-identify-civil-war-soldiers-in-old-photographs.

[47] Kurt Luther. 2015. Blazing a path from confirmation bias to airtight identification. *Milit. Images* 33, 2 (2015), 54–55. Retrieved from http://www.jstor.org/stable/24864385.

[48] Kurt Luther. 2015. The photo sleuth's digital toolkit. *Milit. Images* 33, 3 (2015), 47–49. Retrieved from http://www.jstor.org/stable/24864403.

[49] Kurt Luther. 2017. Officer identified using classic research and CWPS. *Milit. Images* 35, 4 (2017), 12–13. Retrieved from http://www.jstor.org/stable/26214051.

[50] Kurt Luther. 2018. A new era in photo sleuthing begins. *Milit. Images* 36, 4 (2018), 8–9. Retrieved from https://www.jstor.org/stable/26483958.

[51] Kurt Luther. 2018. What are the odds? Photo sleuthing by the numbers. *Milit. Images* 36, 1 (2018), 12–15. Retrieved from http://www.jstor.org/stable/26240155.

[52] Kurt Luther. 2019. Civil War photo sleuth: An update. *Milit. Images* 37, 2 (208) (2019), 8–9. Retrieved from https://www.jstor.org/stable/26590745.

[53] Kurt Luther, Scott Counts, Kristin B. Stecher, Aaron Hoff, and Paul Johns. 2009. Pathfinder: An online collaboration environment for citizen scientists. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*. ACM, New York, NY, 239–248. DOI: https://doi.org/10.1145/1518701.1518741.

[54] Kurt Luther, Nicholas Diakopoulos, and Amy Bruckman. 2010. Edits & credits: Exploring integration and attribution in online creative collaboration. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2823–2832.

[55] Ramona Martinez. 2012. Unknown No More: Identifying a Civil War Soldier. Retrieved from http://www.npr.org/2012/04/11/150288978/unknown-no-more-identifying-a-civil-war-soldier.

[56] Microsoft. 2018. Face API—Facial Recognition Software. Microsoft Azure Retrieved from https://azure.microsoft.com/en-us/services/cognitive-services/face/.

[57] Microsoft. 2018. Uber boosts platform security with the Face API, part of Microsoft Cognitive Services. Retrieved from https://customers.microsoft.com/en-us/story/uber.

[58] Vikram Mohanty, David Thames, Sneha Mehta, and Kurt Luther. 2019. Photo sleuth: Combining human expertise and face recognition to identify historical portraits. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI'19)*. ACM, New York, NY, 547–557. DOI: https://doi.org/10.1145/3301275.3302301

[59] Shruti Nagpal, Mayank Vatsa, and Richa Singh. 2016. Sketch recognition: What lies ahead?*Image Vis. Comput.* 55 (2016), 9–13.

[60] Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z Gajos. 2011. Platemate: Crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. ACM, 1–12.

[61] NPS. 2018. Soldiers and Sailors Database—The Civil War (U.S. National Park Service). Retrieved from https://www.nps.gov/subjects/civilwar/soldiers-and-sailors-database.htm.

[62] Annie Palmer. 2018. Facial recognition software helping identify Civil War soldiers. *Daily Mail* (Nov. 2018). Retrieved from https://www.dailymail.co.uk/sciencetech/article-6399039/The-facial-recognition-software-identify-thousands-faces-Civil-War-photographs.html.

[63] Genevieve Patterson, Grant Van Horn, Serge Belongie, Pietro Perona, and James Hays. 2015. Tropel: Crowdsourcing detectors with minimal training. In *Proceedings of the 3rd AAAI Conference on Human Computation and Crowdsourcing*.

[64] P. Jonathon Phillips, Amy N. Yates, Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G. Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa, David White, and Alice J. O.Toole. 2018. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proc. Nat. Acad. Sci.* 115, 24 (2018), 6171–6176. DOI: 10.1073/pnas.1721355115

[65] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, New York, NY, 429–435. DOI: https://doi.org/10.1145/3306618.3314244

[66] Roy Rosenzweig. 2006. Can history be open source? Wikipedia and the future of the past. *J. Amer. Hist.* 93, 1 (June 2006), 117–146.

[67] Michael E. Ruane. 2014. Facebook helps identify soldiers in a forgotten Civil War portrait. *Washington Post* (7 Mar. 2014). Retrieved from https://www.washingtonpost.com/local/facebook-helps-identify-soldiers-in-a-forgotten-civil-war-portrait/2014/03/07/a4754218-a47a-11e3-8466-d34c451760b9_story.html.

[68] Michael S. Schmidt. 2018. "Flags of our Fathers" author now doubts his father was in Iwo Jima photo. *The New York Times* (3 May. 2016). Retrieved from https://www.nytimes.com/2016/05/04/us/iwo-jima-marines-bradley.html.

[69] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 815–823.

[70] Irving Seidman. 2006. *Interviewing as Qualitative Research: A Guide for Researchers in Education And the Social Sciences* (3rd ed.). Teachers College Press.

[71] Natasha Singer. 2018. Amazon's facial recognition wrongly identifies 28 lawmakers, A.C.L.U. says. *The New York Times* (26 July 2018). Retrieved from https://www.nytimes.com/2018/07/26/technology/amazon-aclu-facial-recognition-congress.html.

[72] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. 2006. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proc. IEEE* 94, 11 (2006), 1948–1962.

[73] Meilan Solly. 2018. Facial recognition software is helping identify unknown figures in Civil War photographs. *Smithsonian* (Nov. 2018). Retrieved from https://www.smithsonianmag.com/smart-news/facial-recognition-software-helping-identify-unknown-figures-civil-war-photographs-180970863/.

[74] Ramya Srinivasan, Conrad Rudolph, and Amit K. Roy-Chowdhury. 2015. Computerized face recognition in Renaissance portrait art: A quantitative measure for identifying uncertain subjects in ancient portraits. *IEEE Sig. Proc. Mag.* 32, 4 (2015), 85–94.

[75] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation by joint identification-verification. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 1988–1996.

[76] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1701–1708.

[77] Alan Trachtenberg. 1985. Albums of war: On reading Civil War photographs. *Representations*9 (1985), 1–32. DOI: https://doi.org/10.2307/3043765

[78] USAHEC. 2018. MOLLUS-MASS Civil War Photograph Collection. Retrieved from http://cdm16635.contentdm.oclc.org/cdm/landingpage/collection/p16635coll12.

[79] Sarah Jones Weicksel. 2014. To look like men of war. *Clio* 40, 2 (2014), 137–152. Retrieved from https://www.cairn-int.info/article-E_CLIO1_040_0137--to-look-like-men-of-war.htm.

[80] Charlie Wells. 2012. Unknown soldier in famed Civil War portrait identified. *New York Daily News* (22 Aug. 2012) Retrieved from http://www.nydailynews.com/news/national/unknown-soldier-famed-library-congress-civil-war-portrait-identified-article-1.1142297.

[81] Sarah Wells. 2019. The computer scientist who wants to put a name to every face in Civil War photographs. *Smithsonian* (Mar. 2019). Retrieved from https://www.smithsonianmag.com/innovation/computer-scientist-who-wants-to-put-name-to-every-face-in-civil-war-photographs-180971754/.

[82] Heather Willever-Farr, Lisl Zach, and Andrea Forte. 2012. Tell me about my family: A study of cooperative research on ancestry. com. In *Proceedings of the iConference*. ACM, 303–310.

[83] Heather L. Willever-Farr and Andrea Forte. 2014. Family matters: Control and conflict in online family history production. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 475–486.

[84] A. C. Williams, J. F. Wallin, H. Yu, M. Perale, H. D. Carroll, A. F. Lamblin, L. Fortson, D. Obbink, C. J. Lintott, and J. H. Brusuelas. 2014. A computational pipeline for crowdsourced transcriptions of ancient Greek papyrus fragments. In *Proceedings of the IEEE International Conference on Big Data (Big Data'14)*. 100–105. DOI: https://doi.org/10.1109/BigData.2014.7004460

[85] Bob Zeller. 2019. Searching for photos of Civil War soldiers. Center for Civil War Photography. Retrieved from https://www.civilwarphotography.org/index.php/resources/searching-for-photos-of-civil-war-soldiers/.

[86] Wenyi Zhao, Rama Chellappa, P. Jonathon Phillips, and Azriel Rosenfeld. 2003. Face recognition: A literature survey. *ACM Comput. Surv.* 35, 4 (2003), 399–458.

Received July 2019; revised October 2019; accepted October 2019