Photo Steward: A Deliberative Collective Intelligence Workflow for Validating Historical Archives

Vikram Mohanty
Center for Human-Computer Interaction and
Department of Computer Science, Virginia Tech
Arlington, VA, USA
vikrammohanty@vt.edu

Kurt Luther
Center for Human-Computer Interaction and
Department of Computer Science, Virginia Tech
Arlington, VA, USA
kluther@vt.edu

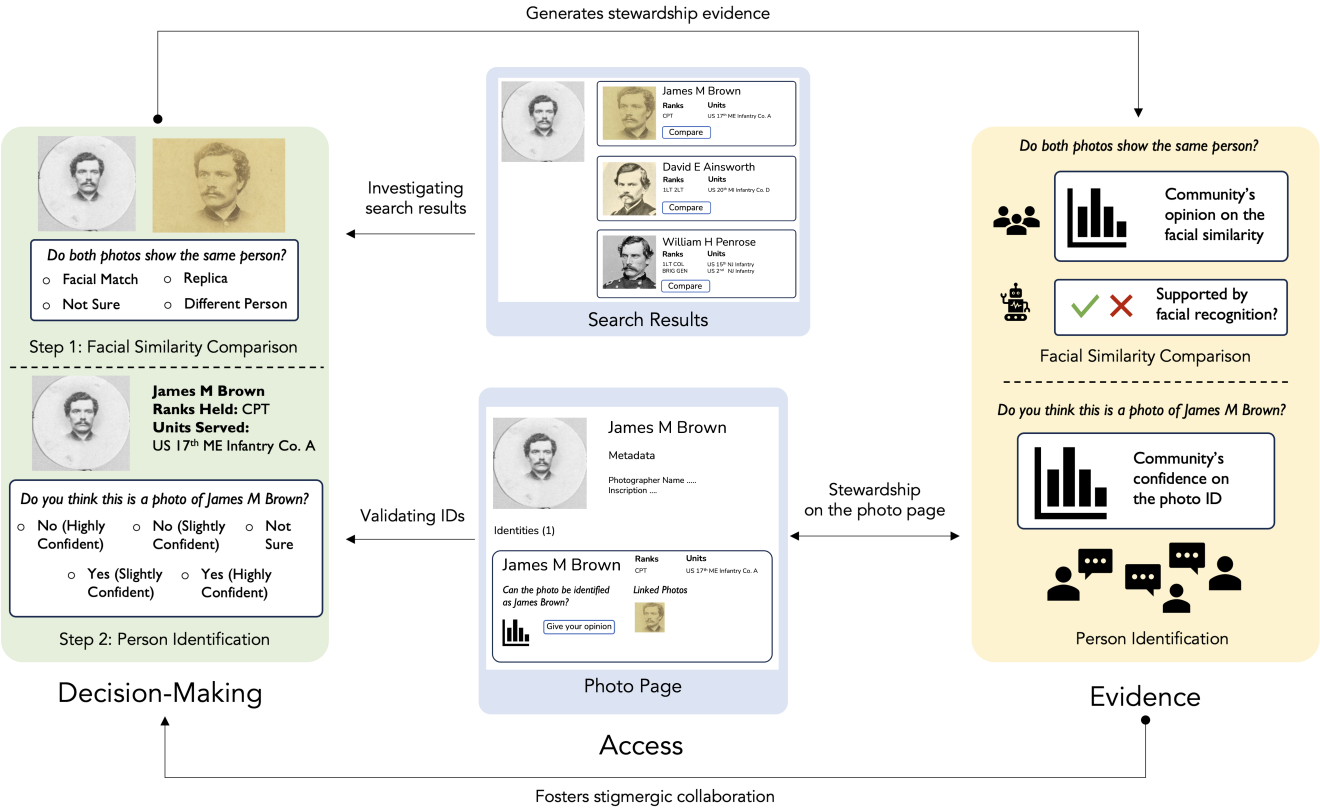


Figure 1: Workflow of Photo Steward. (1) Decision-Making: The user compares photos for facial similarity and make decisions on the photo ID using Photo Steward's deliberative validation interface. (2) Access: The user accesses the validation interface from Civil War Photo Sleuth's search results and photo page. (3) Evidence: The community's responses from the validation interface feed into stewardship visualizations that are visible on the Photo Page, which subsequently foster a form of stigmergic collaboration among the users.

ABSTRACT

Historical photographs of people generate significant cultural and economic value, but correctly identifying the subjects of photos can be a difficult task, requiring careful attention to detail while synthesizing large amounts of data from diverse sources. When photos are misidentified, the negative consequences can include financial losses and inaccuracies in the historical record, and even the spread of mis- and disinformation. To address this challenge, we introduce Photo Steward, an information stewardship architecture that leverages a deliberative workflow for validating historical photo IDs. We explored Photo Steward in the context of Civil War Photo Sleuth (CWPS), a popular online community dedicated to identifying photos from the American Civil War era (1861–65) using facial recognition and crowdsourcing. While the platform has been

CI '23, November 06–09, 2023, Delft, Netherlands

ACM ISBN 979-8-4007-0113-9/23/11.
https://doi.org/10.1145/3582269.3615600

successful in identifying hundreds of unknown photographs, there have been concerns about unverified identifications and misidentifications. Our exploratory evaluation of Photo Steward on CWPS showed that its validation workflow encouraged users to deliberate while making photo ID decisions. Further, its stewardship visualizations helped users to assess photo ID information accurately, while fostering diverse forms of stigmergic collaboration.

CCS CONCEPTS

• **Human-centered computing** → *Collaborative and social computing design and evaluation methods*; *Interactive systems and tools.*

KEYWORDS

crowdsourcing, human-AI interaction, online deliberation, community stewardship, information assessability, online communities, history, person identification, facial recognition, stigmergic collaboration

ACM Reference Format:
Vikram Mohanty and Kurt Luther. 2023. Photo Steward: A Deliberative Collective Intelligence Workflow for Validating Historical Archives. In *Collective Intelligence Conference (CI '23), November 06–09, 2023, Delft, Netherlands.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3582269.3615600

1 INTRODUCTION

The task of correctly identifying individuals in historical photos holds great cultural and economic importance [3, 19, 39, 61]. This identification process is analogous to solving a complex mystery. It often involves corroboration of multiple research processes such as investigating visual clues in a photo, finding relevant reference resources, and comparing multiple low-resolution reference photos [31, 32, 37]. Historical photos pose many challenges, including low-resolution images, scattered reference materials, limited domain expertise, and lack of suitable verification tools. These hurdles often result in misidentifications which can have negative consequences, ranging from distorting historical narratives [60] and fueling conspiracy theories [12] to spreading disinformation [16] and unwarranted financial gains from inaccurate representations [21]. As online platforms such as Ancestry.com, Find-a-Grave, and FamilySearch democratize historical and genealogical research, the risk of misidentification is further amplified due to factors such as inadequate experience, confirmation bias, and automation bias introduced by imperfect automated tools [44, 75].

To address these challenges, we introduce *Photo Steward*, a deliberative workflow that leverages collective intelligence to validate historical photo identifications (IDs). Photo Steward's architecture builds upon the concept of information stewardship [18, 72], which involves community-driven validation of content, as seen in online communities like iNaturalist and Wikipedia. We designed and applied the Photo Steward architecture to Civil War Photo Sleuth (CWPS),[1] an AI-infused online platform for identifying historical photos. CWPS has over 20,000 registered users and over 25,000 identified Civil War portraits, and faces the problem of historical photo misidentification [43]. Photo Steward provides a validation workflow that promotes careful deliberation during facial similarity comparison and photo ID verification, while enabling users to share opinions. It also visualizes community opinions on the reliability of photo IDs and facial similarity, fostering information stewardship at multiple levels.

We publicly released Photo Steward on CWPS and conducted an exploratory evaluation of twelve months of usage, including interviews with users of different expertise levels and log analysis of stewardship behaviors on the platform. We found that Photo Steward's stewardship visualizations helped users to find additional evidence (e.g., external sources, uniform clues matching service records, etc.) for assessing the reliability of photo IDs. Users validated hundreds of different IDs on the platform, and found the workflow to be useful for deliberating on facial similarity comparison and fine-grained photo ID decisions.

We also discuss the implications of community participation, deliberative human-AI interaction, and assessable designs for historical photo identification.

2 RELATED WORK

2.1 Misinformation in a Historical Context

Numerous online communities, forums, and websites have emerged in recent years for archiving and documenting history [57], generating family histories [74, 75], identifying and sharing historical photos [43], trading antiques [1, 8], and facilitating discussions around history [20]. Much like popular social media platforms such as Facebook and Twitter, these history-based platforms are also prone to the problem of misinformation, albeit in a more specialized historical research context.

Prior work has shown that erroneous family history trees were being disseminated across Find-a-Grave and Ancestry, two popular genealogy research communities, as a result of the platform's low bar to entry and inexperienced contributors' over-reliance on inaccurate automated features [75]. Mohanty et al. [43] showed that despite successful identifications on Civil War Photo Sleuth (CWPS), several photos were misidentified in the first month, particularly for photos without period inscriptions or duplicate views (12 misidentifications out of 37). Multiple factors — the correct candidate not present in the search pool, or the user incorrectly assessing facial similarity [53] and picking the wrong match — can lead to incorrect IDs. A follow-up benchmarking study of the underlying face recognition algorithm [44] highlighted its low precision (i.e., it retrieves over hundreds of search results), raising the possibility of errors due to automation bias, i.e., the tendency of users to over-rely on automation for making a decision [47, 50, 64].

Identifying historical photos is a complex investigative process, often involving the corroboration of multiple evidence pieces [31, 33] and can be seen analogous to "finding a needle in a haystack". As a result, historical photo IDs run a high risk of getting misidentified even with the best of intentions. Multiple Civil War photos have also been misidentified in the collections of professionally managed museums and archives, such as the US Library of Congress [36] and the Abraham Lincoln Presidential Library [34]. At the same time, historical photo IDs also have the potential to generate significant monetary value [1, 3, 8, 10], and such financial incentives might also lead to falsified identifications [21]. Validating these historical photo IDs, which are a result of complex, subjective original research,

[1] www.civilwarphotosleuth.com

becomes tricky without the lack of domain expertise and access to investigative tools.

We addressed these challenges in Photo Steward by designing a stewardship architecture that allows users to share their expertise with others. To address the impact of automation, we introduced a two-step validation workflow for the users to deliberate on decisions while interacting with the AI's recommendations.

2.2 Data Validation in Online Communities

Multiple online platforms have leveraged the strengths of crowdsourced contributions for validating the quality of data generated on those sites. Elliott discusses how stigmergic collaboration, where indirect coordination within a community stimulates subsequent actions, plays a role in maintaining articles on Wikipedia [13]. This concept was observed by Wiggins et al. in their study of iNaturalist, an online platform for identifying species, wherein community stewardship behaviors were seen as users agreed on organism identifications to influence the platform's quality grade status [72]. Prior work has shown that stewardship visualizations on Wikipedia (i.e., article quality) [18] and iNaturalist (i.e., ID research grade status) [72] have a positive impact on users' assessment of the information.

Along these lines, we also built DoubleCheck [42], a quality assessment framework that builds upon the concepts of provenance and stewardship for verifying historical photo IDs. DoubleCheck focused on displaying quality indicator badges for historical photo IDs by capturing accurate provenance information and combining the source trustworthiness information with community opinions on the ID. In this work, we focus solely on the underlying stewardship architecture that helped facilitate the community opinions. Both DoubleCheck and Photo Steward were evaluated in the same lab study, but there is no overlapping data.

Visualizations displaying (surrogate) quality metrics, such as popularity among expert users, social reputation, and content coverage, have been effective in helping users assess the credibility of websites and search results [62]. Prior work has also shown that visualizing the history of edits for a Wikipedia article can have a significant impact on users' perceived trustworthiness of the article [54, 67]. Similarly, Chevalier et al. [7] showed that visualizing the number of contributors, length of the article and discussion, and the history of edits helped users assess the quality of Wikipedia articles faster. On the other hand, Towne et al. [68] found that being exposed to editor conflicts in the discussion of a Wikipedia article lowered the perception of the article's quality, even though the users reported that the transparency raised their perceptions of the page and Wikipedia in general. Morris et al. [45] found that Twitter users relied on the author information for making assessments about the credibility of information in a tweet.

Drawing from this prior work, we designed Photo Steward's stewardship visualizations to highlight the role of collective intelligence, while fostering stigmergic collaboration on CWPS to validate the quality of photo identifications.

2.3 Background: Civil War Photo Sleuthing

The American Civil War (1861–65) was one of the first major conflicts to be extensively photographed. Over 3 million soldiers fought in the war, with many of them having been photographed at least once. Over 150 years, many of these photos have survived in museums, libraries, and personal collections, but only 10–20% are are identified [69, 77]. Civil War photography has garnered a lot of interest among historians, collectors, dealers, genealogists, archivists, and other experts, who often try to identify unknown photos for personal, cultural, and economic reasons. However, the identification process is complex and challenging, which often involves identifying visual clues in a photo and manually scanning through hundreds of low-resolution photos, military records, and reference books for corroborating evidence [31, 33, 38].

3 ENHANCING CIVIL WAR PHOTO SLEUTH: DESIGN OPPORTUNITIES

Civil War Photo Sleuth (CWPS) is a free, public website where users can identify unknown portraits from the American Civil War era using a person identification pipeline that combines crowdsourced human expertise and face recognition [43]. Drawing analogies to *finding a needle in a haystack*, Mohanty et al. propose a 'haystack model' to describe CWPS's person identification pipeline. In this pipeline, a user begins the identification process by first tagging a photo for uniform clues, which then generates search filters based on service records, and then facial recognition returns facially similar-looking results from a pool of potential candidates, ordered by similarity to the query photo, that satisfy the search filters (see Figure 2).

The CWPS haystack model is designed to prevent misidentifications by placing human decision-making at the forefront and treating AI as a supportive tool. It avoids automatically selecting the best match or displaying the algorithm's inconsistent confidence levels [41, 44]. Instead, the user carefully inspects search results for potential matches based on facial similarity and corresponding biographical details. Once a photo is identified, CWPS links the face and identity together and displays the ID on the photo page.

Despite these measures, the open participation model of CWPS, which lacks verification, has raised concerns about the trustworthiness of proposed identities and the potential increase of "false positives" as the site grows [22]. To address these concerns, we enumerate three **design goals** which draw upon prior work on Civil War photo identifications and CWPS system designs, evaluations, and critiques [22, 30, 33, 35, 41, 43, 44], as well as our own observations and experiences using the publicly available version of the website. We provide further details in Appendix A.

Design Goal 1: Decouple facial similarity comparison from the overall task of person identification. The current CWPS workflow conflates facial similarity and person identification into a single decision-making process (see Figure 2-C). Facial similarity, while important, can conflict with the identity suggested by personal details like biographical information and service records. The facial recognition algorithm's low precision [44] adds to the complexity, with the possibility of users interacting with false positives. In order to discourage over-reliance on facial similarity, we propose this design goal of separating both these tasks, allowing users to deliberate on the facial similarity and other person identification attributes separately, thereby minimizing inaccuracies.

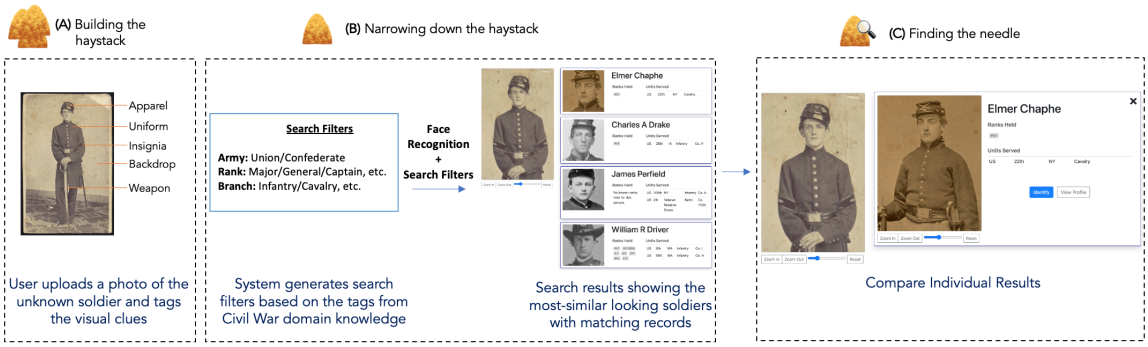


Figure 2: CWPS Haystack Model: Person Identification Pipeline [43]

Design Goal 2: Support fine-grained, deliberative decision-making. With the current CWPS interface only permitting binary feedback during photo identification, there is a heightened risk of misrepresentation and misidentifications (see Figure 2-C). Historical photo identification is intricate, often demanding careful corroboration of numerous evidence pieces, with user confidence varying according to the evidence at hand [33]. To more effectively facilitate this complex process, we advocate for a design that enables users to express their level of certainty in their identification decisions. This design goal seeks to foster more accurate and nuanced user feedback, as well as reflective and deliberate decision-making [28, 29].

Design Goal 3: Encourage community contribution and transparency for validating photo IDs. Identifying individuals in photos can be considered a subjective process and often benefits from multiple perspectives [41]. Currently, CWPS lacks mechanisms for community feedback and transparency about the process of photo identification, leaving potential misidentifications unchecked (see Figure 8 in Appendix). To support accurate original historical research, the platform should encourage community participation in validating identifications and promoting transparency around the roles of community contribution and facial recognition [18, 72]. This approach also encourages collective responsibility, facilitating stigmergic collaboration [13, 14], where user contributions guide future validation efforts.

4 SYSTEM DESCRIPTION: PHOTO STEWARD

We developed Photo Steward, an information stewardship architecture that integrates a deliberative workflow for the community to validate historical photo identifications, which we then integrated into CWPS. Photo Steward's architecture has three main components (see Figure 1): 1) a deliberative *decision-making* interface for facial similarity comparison and photo identification, 2) new *access* points for validating photo identities, and 3) stewardship *evidence* for fostering stigmergic collaborations.

4.1 DECISION-MAKING: Deliberating on facial similarity and photo identification

As part of Photo Steward, we introduce a multi-step "Validation Interface" (see Figures 3 and 4) to replace CWPS's single-step comparison interface. Photo Steward's validation interface allows users to deliberate while interacting with the facial recognition results. Meeting *Design Goal 1*, the validation workflow separates the task of facial similarity comparison from the overall goal of identifying the photo.

To inform our design, we draw on evidence-based decision-making [11], a model primarily used in healthcare, policymaking, and judicial sectors, which advocates for justifying decisions (photo IDs in this case) by gathering available evidence (facial similarity as visual evidence here). In the first step, the user compares the query photo to all other photos with the same identity for facial similarity. After deliberating on the facial similarity evidence, the user then votes on whether the query photo fits the target identity in the second step (which is the user's primary goal).

The validation interface is divided into four columns (from left to right): 1) the task description, 2) the query photo, 3) the evidence that is being weighed, and 4) the biographical information. The query photo and evidence are positioned in the two middle panels for easy side-by-side comparison. The *task description* panel displays the rating question for both the facial similarity comparison and the identification steps. Here, we used structured feedback to capture both the user's facial similarity comparison and their confidence on the photo ID, in an effort for encouraging users to exercise personal deliberation on all available evidence before making a decision on the ID. The interface updates the *task description* and *evidence* column depending on which task the user is performing.

To investigate the identity of a *query photo*, the user opens the validation interface which loads all the photos and biographical information available for the *target identity*.

4.1.1 *Validation Step 1: Facial Similarity Comparison*.

For the first step of the validation process, the interface displays the *target photo* in the evidence column next to the query photo for easy facial similarity comparison (see Figure 3).

The user's task is to determine whether both photos show the same person (regardless of whether the identity is known). Users can select from the following options: *No (Different Person)*, *Not*

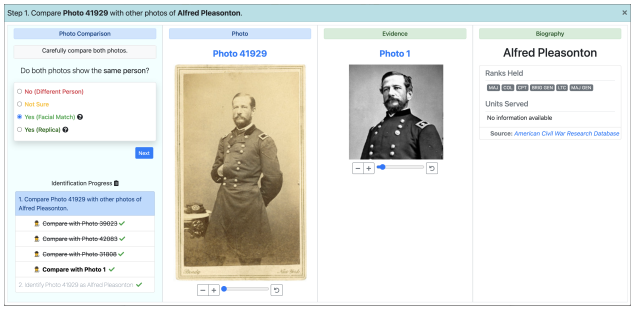


Figure 3: Step 1 of Photo Steward's Validation Interface. Users can compare two photos and answer whether they show the same person or not. They have the option of selecting whether the two photos are a facial match (i.e., same person, different views) or a replica (i.e., same person, same view). Here, the user is comparing whether Photo 41929 and Photo 1 show the same person or not. If multiple faces are available for the same ID, they appear one after the other in the order in which they were uploaded to CWPS.

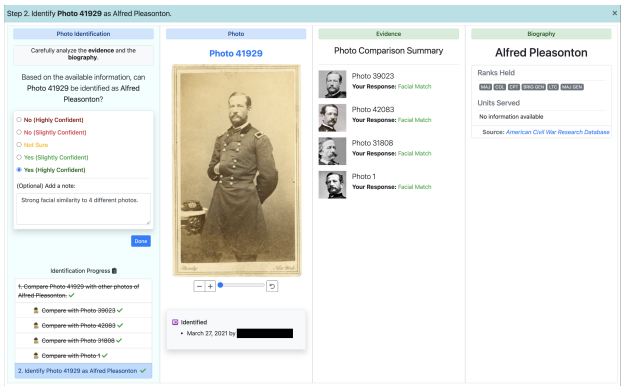


Figure 4: Step 2 of Photo Steward's Validation Interface. Users vote on whether the query photo can be identified as the target identity by expressing their confidence. They can also add an optional note to justify their decision. The evidence panel displays a summary of the user's responses from the first step, where the faces are displayed next to the user's facial similarity comparison with the query photo. The faces are ordered in the way they appear for comparison, i.e., the order in which they were uploaded to CWPS.

Sure, *Yes (Facial Match)* and *Yes (Replica)*. Mohanty et al. found that photos correctly identified on CWPS were either *facial matches* (i.e., same person, different view) or *replicas* (i.e., same person, same view) [43], which informed the design of this input scale. Since facial similarity does not have any standard scale and users may perceive the similarity or dissimilarity of two faces differently [40, 76], we chose not to capture any further granularity in their responses for *facial match*, *replica* or *different person* as this might lead to inconsistent data collection.

Capturing these responses in a structured way allows users to deliberate on the task of facial similarity; this becomes more critical as users are also interacting with the results of a low-precision facial recognition algorithm [44]. In this step, the user compares facial similarity of the query photo with all available photos of the target identity, one photo at a time.

4.1.2 *Validation Step 2: Fine-Grained Photo Identification*.

In this step, the user analyzes the biographical information and incorporates the facial similarity evidence from the previous step to make a decision on the photo's identity.

The validation interface displays information in the same four-column layout (see Figure 4), with the evidence column now displaying a summary of the user's responses about facial similarity between the *query photo* and the *target photo(s)*. The biography column shows the name and the service records for the user to analyze.

The user now decides whether the query photo can be identified as the target identity (see Figure 4). The instruction above nudges the user to factor in the prior photo comparison evidence and the biography information. Meeting *Design Goal 2*, users indicate their confidence about the task question by selecting one of the five options displayed in radio buttons: *No (Highly Confident)*, *No (Slightly Confident)*, *Not Sure*, *Yes (Slightly Confident)* and *Yes (Highly Confident)*. This scale, which offers more nuance than a binary decision, serves as a proxy measure for the ID's reliability given the complexities of quantifying accuracy in a historical photo ID investigation. These options reflect the varying degrees of confidence users have based on the quality and quantity of corroborating evidence, such as reputable sources, facial similarity to additional photos, and expert opinions. Users also have the option to elaborate their decision rationale in a free-text note.

4.2 ACCESS: Expanding Validation Opportunities for Photo Identifications

Photo Steward provides stewardship capabilities for the CWPS user community by allowing them to access and use the validation interface at different stages of the photo identification process from multiple gateways (see Figure 1). On the "Search Results" page, it can be used for identifying a photo from a pool of potential similar-looking candidates, or ruling out some potential candidates. After a photo has been identified, users can also access it on the "Photo Page" to either validate an existing ID or dispute an incorrect one, and collaborate with other users in a stigmergic manner (*Design Goal 3*).

4.2.1 *Search Results Page: Matching and Ruling Out Candidates*.

While identifying a *query photo*, users can now inspect potential matches on the search results page with the help of the validation interface. The "Compare" button on a search result brings up the validation interface, loading all the target information for the corresponding search result. The *target identities* in the validation interface will update as the users check new search candidates for matches. The interface allows users to make two types of decisions, depending on their confidence response: 1) either of the "Yes" responses will match the photo with the target identity with varying degrees of confidence, and 2) either the "No" or "Not Sure"

responses will rule out the search candidate as a potential match for the current user.

4.2.2 ***Photo Page: Validating and Disputing Existing IDs.*** After a photo has been identified, Photo Steward allows users to review opinions from other users (described in Section 4.3) and contribute their own for a given photo ID on CWPS's photo page, fulfilling *Design Goal 3*. By clicking the "Give Your Opinion" button, users launch the validation interface featuring the *query photo*, the linked *target identity*, and *target photos* of the same ID, a new feature previously absent from CWPS. The two-step process mirrors that on the search results page, enabling community deliberation on the validity of an ID. Users can validate the facial match among photos linked to the same ID, express their agreement or disagreement on an ID with varying confidence levels, and optionally add a note explaining their decision. Thus, each vote contributes to a stigmergic collaboration, enhancing the reliability of photo IDs on CWPS. Consistent with CWPS's open participation model, Photo Steward allows any registered user to share their opinion on an identification.

4.3 EVIDENCE: Visualizing Information Stewardship

As part of Photo Steward, we designed stewardship visualizations to help users assess the reliability of 1) facial matches (i.e., photos that were matched to each other by the user), and 2) photo IDs. The CWPS community's opinions on facial similarity comparison and photo IDs, captured through the validation interface, feeds into these reliability visualizations. These visualizations not only promote user accountability through social translucence [15], but also serve as deliberative evidence for subsequent stigmergic user collaborations *(Design Goal 3)*.

4.3.1 ***Reliability of Facial Similarity.*** For each photo pair that has been compared, the system aggregates the community's decisions for the visual match type and generates a distribution, which is displayed in the form of an interactive horizontal bar chart on the photo page. This chart appears next to the corresponding photo matched to the query photo (see Figure 5). Users can click the "View Details" button or an individual bar to see how each user voted. When multiple photos have been matched to the query photo, the matched photos appear one below the other, with each having its own visualization next to it. The bar charts are stacked vertically above each other to allow users to easily see and compare the reliability of every match.

To complement the community stewardship visualization, we also added an AI stewardship badge that indicates whether the particular match is supported by facial recognition (see Figure 5). On the search results page, CWPS retrieves those search results that have a facial similarity score greater than 0.50, so we use the same threshold here. However, the badge intentionally does not display the exact similarity scores (which have been found to be inconsistent [41, 44]) to avoid a false perception of precision, and cautions users to carefully analyze all the context and evidence, as there is a possibility of false positives with face recognition.

4.3.2 ***Reliability of Photo Identifications.*** Similar to the facial similarity visualization, the system aggregates the community's

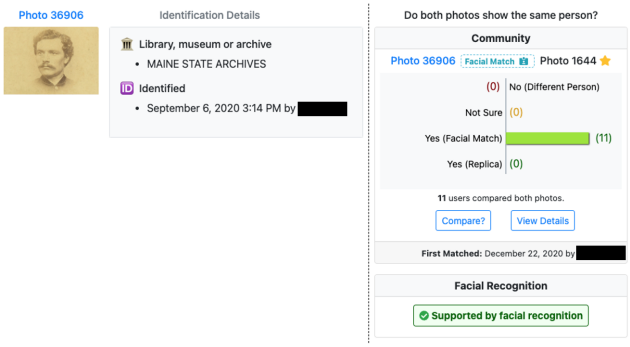


Figure 5: Facial Similarity Reliability Visualization on the Photo Page. The photo matched to the query photo is displayed here, whereas the query photo is displayed on top of the page. Other photos matched to the query photo are displayed vertically one below the other. Users can bring up the query photo and the matched photo side-by-side by clicking the "Compare" button. Each matched photo has its own visualization next to it, and the responses are visible to everyone.

confidence levels for an identification and generates a distribution, displayed in the form of an interactive horizontal bar chart. This visualization is displayed for every proposed identity on the photo page, below the biography subsection (see Figure 6a). If users propose multiple IDs for a given photo, they are displayed one below the other; each ID will have its own visualizations. The community's confidence votes are aggregated to order these IDs. Users can click the "View Details" button or the individual bars to view a modal dialog box with each user's confidence scores and optional text comments (see Figure 6b).

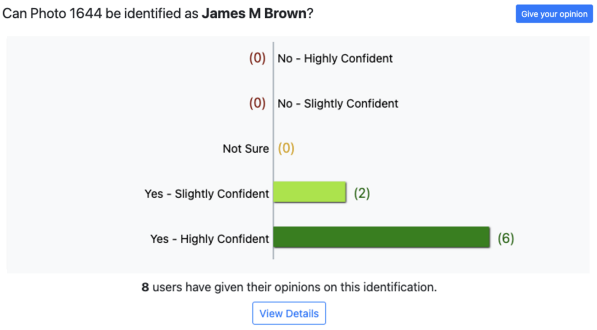
This visualization reflects the community's expertise, and is intended for users to quickly assess the reliability of an identification. Prior work on crowdsourced person identification [41] suggests that airtight identifications are likely to show consensus from the community, whereas potential misidentifications are likely to reflect disagreement from the community. Further, the comments may reflect the voters' decision rationale and any external research they conducted before giving their decision on the photo's identification, allowing users to build on the work of others in making their own assessment [17].

4.4 Summary

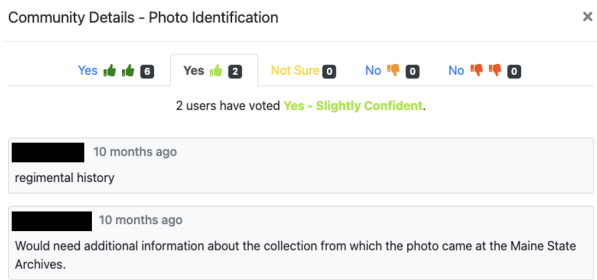
Photo Steward augments the CWPS platform with an information stewardship architecture to support community validation of historical photo IDs in a stigmergic manner. We summarize all the changes in the Appendix (see Table 6).

5 EVALUATION

We obtained permission to publicly launch Photo Steward on CWPS in December 2020. We conducted a mixed-methods, exploratory evaluation study to understand how well users with different expertise levels could validate and assess Civil War photo IDs using



(a) Community's confidence on the ID



(b) Community's notes justifying their votes.

Figure 6: Photo ID Reliability Visualization on the Photo Page for a given ID. If multiple IDs are present, they are displayed vertically one below the other, ordered in terms of aggregate votes. Each ID will have a separate visualization listed under the respective IDs.

CWPS with Photo Steward. Specifically, we wanted to understand 1) how users validated photo IDs using Photo Steward, 2) how the stewardship visualizations (i.e., ID and facial similarity reliability visualizations) impacted users' assessment of an ID. The study was approved by our university's IRB.

5.1 Log Analysis

To understand the community's stewardship behaviors, we analyzed website logs of all user activities for a year after new features were launched, which included 5843 voting instances on 5672 photos for 5355 unique IDs. Our analysis included categorization of user deliberations as 'pre-identified' (i.e., the user knew the ID of the photo ahead of time and therefore, might be looking for a specific target on the search results page) or 'post-identified' (i.e., the user did not know the ID of the photo and therefore, may not have a specific target while analyzing the search results), coding of user comments (see Appendix C), and comparison of community's facial similarity comparisons against facial recognition scores. Details of these analysis methods are provided in Appendix B.

5.2 Lab Study

In order to understand how well Photo Steward supports diverse users in validating the quality of photo identifications, we also conducted an exploratory lab study.

5.2.1 ***Participants.*** We recruited 15 participants representing the three major expertise levels: 5 history students, 5 amateur experts (experienced users of CWPS), and 5 expert historians. Participant details can be found in Appendix D. We anonymize these groups with the following identifiers, respectively: S1–S5, C1—C5, and H1–H5.

5.2.2 ***Dataset.*** For the study, we created a dataset of 10 different photos identified on CWPS. Three of these photos had an ID conflict, i.e., multiple identities were proposed. For two of these photos, one ID was correct and the other one was incorrect. The community had already researched both photos, voted on the correct ID, and left credible evidence in the comments. Both IDs were linked to additional photos as well. The third photo was one of the seeded photos on CWPS, but was originally misidentified. We added another false ID, making both IDs for the third photo incorrect. All photos had multiple photos matched to them; eight of them were linked via facial matches, while two of them had replicas.

5.2.3 ***Procedure.*** The entire study was conducted online via recorded Zoom sessions, with at least one researcher attending each session. Each participant first completed a consent form and a pre-survey describing their demographics and Civil War photography experience.

As part of the study, participants reviewed three randomly assigned photos from the dataset one-by-one in the original CWPS system first, followed by the same photos on the Photo Steward version. Participants used a think-aloud protocol while using the two systems; after the completion of the task with each system, they were asked a few semi-structured questions about their experience. Finally, the participants completed a summative post-survey of standard usability questions (e.g., ease of use, usefulness of features, instruction clarity, preferred system, etc.) (see Appendix E)

We maintained this sequence (original CWPS first, CWPS with Photo Steward second) for all the participants, rather than using a randomized sequence, for two reasons. First, we did not want participants' assessments to be biased in favor of Photo Steward after seeing additional features in the new interface. This design allowed us to observe if the original interface misled the participants towards incorrect assessments, and if, subsequently, the Photo Steward interface helped correct them. Second, in a randomized sequence, Photo Steward would expose the participants to new information in the form of prior user votes and responses, and therefore, may confound how they assess the information on the original CWPS version.

5.2.4 ***Data Analysis.*** The first author fully transcribed and analyzed the interviews and think-aloud recordings using an inductive qualitative thematic approach [4] . The transcript sections were first divided according to the interface in question (i.e., original CWPS or Photo Steward), followed by an open coding of the transcripts using MAXQDA 2020 [65]. The first author iterated and settled on a total of 28 codes (e.g., change in opinions, comparison interface, source trustworthiness, etc.) for 634 coded segments across all the transcripts. These codes were then organized into themes as described in Section 6 after discussing with the co-author.

Votes / ID	# of IDs	Note Present	Negative Votes
1	5650	511	61
2	157 (Agreement: 119) (Disagreement: 38)	83	12
3+	36 (Agreement: 21) (Disagreement: 15)	31	4

Table 1: Distribution of User Votes.

5.2.5 ***Limitations.*** We conducted a qualitative lab study to understand how users with different backgrounds and expertises validated photo IDs using Photo Steward and hit theoretical saturation. However, there are a couple of limitations with the study: 1) limited insights on the role of expertise, and 2) the task sequence could have order effects. Further, the large-scale analysis of Photo Steward logs provided us with insights of its usage amongst users. However, it lacked an expert-prepared gold standard dataset, which hindered our ability to conduct specific performance analyses as part of this study.

6 FINDINGS

Using the methods above, we evaluated how well Photo Steward's stewardship architecture supported CWPS users in validating photo identifications, compared to the original version of CWPS.

6.1 Validation Interface

Users found Photo Steward's validation interface to be useful for comparing different photos. While assessing the IDs with the original interface, participants would go back and forth between different photos to compare whether they are the same person or not. Some participants opened the photos in two different browser windows and kept them side-by-side. While using the validation interface in the new system, participants appreciated being able to see the photos side-by-side at the same time.

H1 said, *"As an historian using this, this is really great to see them both together. It just makes a comparison a lot easy for me to do. I mean, this is the same gentleman, he's got a little dark facial hair. It looks a little bit different there and the photo on the right, but the facial match is definitely there. "* This was also echoed by C3, who said, *"This, I really find extremely useful, especially when I'm trying to do facial recognition. I can zoom in and have them side by side here. [...] Where in the past, I would have to go back and forth between tabs or cut and paste them into a different document to look at them side by side."*

From our logs, we found that 223 users had compared 2319 unique photo pairs for facial similarity, with 156 pairs receiving comparisons from at least 2 different users. The facial similarity responses were distributed as follows: *763 replicas, 1232 facial matches, 283 unsure, and 280 different people.*

Users preferred the ability to provide granular feedback for photo IDs using Photo Steward's validation interface. All participants expressed preference for the fine-grained confidence levels, including the ability to dispute an ID, in Photo Steward's validation interface, appreciating how it more accurately mirrored the inherent uncertainty present when assessing photo IDs. S3 said, *"I definitely like the five levels. I think it leaves more room for interpretation. Like sometimes it's kind of hard to just say yes or a hard no because so much goes into it. Especially because a lot of this stuff was so long ago, there's so many unanswered questions."* H4 initially defended the original interface's binary vote, but changed her mind after experiencing Photo Steward's confidence levels: *"[W]hat I had said has this very black and white feel to it, you're wrong or you're right. I like these degrees of disagreement or agreement. I think that's way more helpful broadly."*

Mean Confidence	1 Vote / ID	2 Votes / ID	3+ Votes / ID
-2 (No - Highly Confident) to -1 (No - Slightly Confident)	45	3	2
-1 (No - Slightly Confident) to 0 (Not Sure)	16	3	2
0 (Not Sure)	99	8	4
0 (Not Sure) to 1 (Yes - Slightly Confident)	760	15	5
1 (Yes - Slightly Confident) to 2 (Yes - Highly Confident)	4730	128	23

Table 2: Distribution of Confidence Levels.

The usage of the voting feature was reflected in our logs, which showed 5843 voting instances from 328 unique users (see Table 1). Table 2 shows that while users utilized the full range of confidence levels, including when they were unsure (mean confidence = 0) or slightly confident (mean confidence = 0 to 1) about the ID, the vast majority of the votes were highly confident ones (mean confidence = 1 to 2). A small proportion of votes (77) were cast for disputing an ID.

Users justified their voting decisions through notes covering an extensive range of topics. From our logs, we found that 155 users had left 682 notes for 600 different photos. However, as Table 1 shows, around 10% of the votes had a note. Table 3 shows the different topics covered by the notes. We observe that users' voting patterns are significantly influenced by the availability and quality of evidence, with clear facial similarity, period inscriptions, personal anecdotes, visible clues in the photograph, and added biographical context often leading to high confidence "Yes" votes, while lack of information typically results in "Not Sure" votes.

Users most frequently left a comment attributing *facial similarity* (after comparing it in the first step of the validation process) to be the reason for their decision (e.g., *"Identical to the other CDV"*). In some instances, they would expound on it by discussing facial features: *"The eyes, nose, cheek bones, shape of face, all look similar to George Pickett, although possibly reversed based on hair part"*. In

Category	Sub-Category	Total Number of Notes	No (Highly Confident)	No (Slightly Confident)	Not Sure	Yes (Slightly Confident)	Yes (Highly Confident)
Photo Comparison	High-Level Comparison	**206**	9	5	9	20	**163**
	Describing Facial Features	37	4	3	9	8	13
Word-of-Mouth	Descendant	55	1	0	0	8	46
	Ownership	47	1	0	0	0	46
	Familiarity	18	0	0	0	2	16
Visual Evidence	Uniform	73	7	1	8	17	40
	Inscription	**145**	4	3	3	6	**129**
	Other Visual Clues	17	1	0	1	2	13
External Information	External URL	47	1	0	1	4	41
	Other Sources	**172**	9	1	1	21	**140**
	Lack of Information	21	0	2	13	6	0
Providing Additional Information	Biographical Information	46	11	2	3	3	27
	Additional Context	**117**	5	0	2	11	**99**

Table 3: Distribution of Note Topics. The table also displays how the notes are distributed for different user confidence levels.

many instances, we found users inferring biographical information (service records, location, etc.) from visual evidence in the photo, be it uniform or backmarks (e.g., *"Initials MN on chinstrap (brass letters). Signature on verso is made out to Marlin's oldest sister. Style of insignia is consistent with other 1862 recruits for Co. B 1st USSS"*).

Interestingly, we also observed a large number of *word-of-mouth* evidence notes for justifying the user's decisions, such as claiming to be a descendant, or owning the original copy of the photo, or having seen the photo somewhere. For example, one user noted, *"He is my great-great grandfather and this photo has been passed down through the generations to me and was identified by his son John Albert Johnson, my father's grandfather."* Users also left external URLs and source details in the notes as evidence. Sometimes, they provided additional context (e.g., *"This image came with a group of 7th Iowa images. The majority were of Company G., but there is only one person in the entire 7th Iowa Infantry that could be identified by the first or last name of 'Nelson.'"*).

The validation workflow encouraged users to exercise careful deliberation while making photo ID decisions. Users felt that the questions in each step of the validation process helped them to carefully weigh in all the evidence and deliberate while voting on the ID (Q4, mean = 4.60, SD = 0.49). C3 explained why the two steps were necessary: *"It's two separate things. One is asking, do you think that this face is the same face? Then the second is, do you think that this face matches this name? I think that that is a necessary question for both of those scenarios. I don't think it's redundant, I think it's necessary."* H2 appreciated the thoughtfulness that the two-step process encouraged, saying, *"It could be the same guy, but it might be a different guy, but now that you know the other interface kind of forces me to slow down a little bit and think more carefully, because it's asking specific questions about things."* A couple of participants, however, expressed initial confusion over the separation and order of these two tasks.

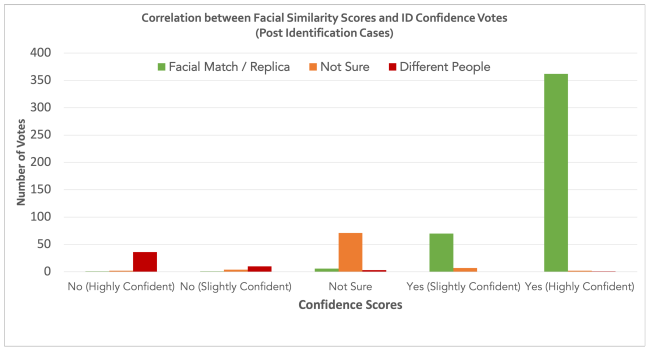


Figure 7: Deliberation in case of post-identification voting. The table shows the distribution of confidence votes (second step of the validation process) against the user's facial similarity comparison (first step of the validation process). The comparison scores were computed by aggregating all the facial similarity comparisons in the first step (replica = 2, facial match = 1, not sure = 0, and different people = -1).

Our logs affirm the deliberative aspect of the two-step validation process. We observed that when users compared one or more photos in the first step, they factored in the facial similarity evidence for their ID vote decision. Figure 7 shows that when the cumulative comparison scores are positive (i.e., majority of the comparisons were a facial match or a replica), the confidence scores are generally positive. Similarly, when the cumulative comparison scores are negative, the confidence votes are also largely negative. When the users are unsure about the facial similarity comparison, it reflects in their final ID vote. Notably, when there were mismatches between

comparison scores and ID votes, users provided reasons in the notes, citing visual evidence and an unsure photo comparison.

Photo Steward's stewardship architecture fostered diverse forms of meaningful stigmergic collaboration amongst users. Although most IDs received only a single vote, about 3% (193 IDs) received multiple votes (see Table 1). Among these, 140 IDs saw total consensus among voters on confidence scores, with half involving an explanatory note. When the initial vote lacked a note, subsequent voters often added information like external sources or context. Almost all IDs where voters agreed positively were supported by metadata such as period inscriptions, scholarly sources, or additional comparison photos, underscoring the credibility of these agreements.

When users disagreed on an ID, they would almost always justify in notes why they differed from the previous voters' opinions (43 out of 53 disagreement instances). From analyzing the notes, we found that the subsequent voters provided additional information about external sources or service records to disagree (e.g., *"Please see Military Images magazine, MI, Volume XVI, Number 3, November - December 1994 for the correct identification of this image. The correct identification via the Michael McAfee collection is Richard Cramer, 4th Michigan Infantry"*). Among 18 instances of conflicting IDs for the same photo, seven saw resolution through a process of voting one ID over another, often accompanied by external evidence in notes or citing facial similarity.

User agreement was particularly strong in facial similarity comparisons. Of the 2200 unique photo pairs compared, 141 received multiple user reviews, with 114 seeing total agreement. The remaining cases typically involved user errors in distinguishing between 'facial match' and 'replica', or uncertainty about facial similarity.

6.2 Stewardship Visualizations

Stewardship visualizations reflecting community insights enhanced the accuracy of photo ID assessments. Participants found Photo Steward's bar chart visualization showing the five confidence levels to be simple and easy to understand (Q1, mean = 4.93, SD = 0.25). S1 said, *"I'm a visual person. Bar graphs or charts like statistical data helps me put things into a better perspective or gives me an idea of what I am working with versus just something more plain [like] the other interface."*

When participants were uncertain about an ID, they saw the additional evidence and justification that the community members had provided along with their vote as essential to taking the vote seriously. S5 said, *"If there's other users giving comments like 'I've used this source,' and you know you get information from a source where they found it, I'm probably gonna agree with them. If they're just voting yes without anything else, then I am probably more likely to go out and find sources for the photo myself and make my own decision."*

In three instances, students (S1, S3, and S5) initially made incorrect assessments on the original interface but rectified these after using Photo Steward. The community's consensus and evidence in Photo Steward were pivotal in these corrections. S5, confronted with ID conflicts, stated after using Photo Steward: *"You know that helps a lot with what people are thinking and presumably these people have also gone to the Maine State Archives or something and verified and then given their opinion so that helps."* Similarly, historian H2 and collector C3 initially made incorrect choices but amended their decisions after analyzing community-based evidence on Photo Steward. H2 remarked: "*So I believe it's the first person, of course, because it's based on the Maine State Archives. And then you've got the piece on there that said that you looked and found a different man. With that name in the actual regimental history. I say that's fairly accurate information.*" This highlights the critical role of community consensus in resolving photo ID conflicts.

On the other hand, when participants were confident about an ID, seeing the community's opinions affirmed their own assessments, for better or worse. In the words of H5: *"I really liked the very clear community consensus, and the ability to be able to see the identities of the people who were looking at these images. It was something that (gave) a boost of confidence in terms of my final decision."* Overall, the participants found the community opinions to be useful for assessing the IDs (Q2, mean = 4.93, SD = 0.25).

Users gave additional weight to the opinions of members they were familiar with and desired more contextual information about all contributors. Participants examined the bar chart visualization details and gave higher weight to the opinions of prominent names from the Civil War photography community while assessing the ID. C3 said, *"Here's <name redacted>, period inscription with valediction, the uniform matches his service record. <name redacted> said the same thing. These two uploaders I hold in very high regard to their opinions on this site. If they're saying that they think highly confident of this identification, that gives me a lot of confidence as well."* H5 became further unsure about an ID after seeing a fellow historian's vote: *"And, you know, sort of knowing <name redacted>, someone who I know deals with primary source material a lot, and sort of being on the fence with it as well, leaves me in that unsure position."*

Participants also sought additional details about community members, proposing indicators of professional status or active participation in the CWPS community. H1 emphasized his appreciation of amateur experts and opposition to gatekeeping, yet he suggested having some kind of credential indicator next to the username would be helpful: *"It will be great if, you know, somebody was an academic historian or a published Civil War author, if there was some way to just say some little tidbit next to <name redacted> 'Oh, saying, hey, I'm from <university redacted>, Professor or, you know, author of whatever.'"* Others preferred indicators of community activity levels or personal connections (i.e., descendants) to the identified individual in the photo.

Our log analysis showed revealed the presence of certain active voters, with 9 out of 328 unique voters voting on more than 50 photos each, and 2 voting on over 2000 photos each (mostly ones that they had uploaded and identified). If we consider only votes on photos identified by someone else, we have 8 users who have voted on more than 15 IDs each, with 1 user voting on over 150 IDs.

The community opinions made the platform feel more engaging, but users had mixed thoughts about the ideal number of votes per photo. Participants, in general, felt the community opinions made Photo Steward more engaging compared to the original system. H5 said, *"I thought it appeared more user-centric and*

User Comparison	Mean	Median	SD	Count
Replica	0.88	0.91	0.16	659
Facial Match	0.52	0.65	0.31	1015
Different Person	0.55	0.59	0.16	265
Not Sure	0.51	0.58	0.23	261

Table 4: Face Recognition Similarity Score Stats vs. User Comparisons (Facial Similarity).

user-friendly and it felt more participatory." C3 saw the benefits of increased engagement for making IDs more reliable: *"I think for this crowdsourcing project that we're building on this database [it] is very important to have those comments, those feedback, that we see in the voting system. It only makes this ID stronger and makes the project and the database a more trustworthy and reliable source."*

Users wanted to see more community opinions and comments for IDs that had few votes. S1, on seeing only one vote for an ID, said, *"That doesn't make me feel as confident because that's not too many for me to give an answer."* S3 said it would have been easier to assess some photos if *"there was more input from other people."* Users had mixed opinions about how many votes they wanted to see for verifying an ID. In general, they wanted to see consensus among the community for an identity and at least three votes. C3 said, *"I usually shoot for like three to four [votes] as the lowest where I take some good quality out of those votes. [...] If there's only one or two, and especially if there's two that are split, that is not as reliable to me."*

Participants' concerns about spreading voters too thin were borne out in the log data (see Table 1). We analyzed the logs to check how often CWPS users vote on the photos they are browsing. We found 1784 instances (out of 5843 voting instances) where the number of "lurkers" for a given photo page exceeded the number of voters on that photo.

Users found the face recognition badge and community's opinions complementary for assessing the reliability of matched photos. From our logs, we found that 1408 photo pair comparisons were supported by both users and facial recognition, by far the most common outcome (see Table 5). This information would be visible to the larger user community in the form of stewardship visualizations (see Figure 5). Interestingly, we also see 266 cases where a comparison is supported by users, but disputed by facial recognition (similarity confidence score < 0.50). That outcome was approximately as common as when users disputed a comparison but face recognition supported it (239 pairs) or when a user was unsure but face recognition supported it (219 pairs). However, it was far less common for facial recognition to dispute a comparison when a user also disputed it (26 pairs) or was unsure (42 pairs).

When we analyze the face recognition's confidence scores in more detail, Table 4 shows that there is a clear separation between replicas and other types of user comparison scores. Face recognition confidence scores for photo pairs that users labeled as replicas were much higher in terms of both mean and median (0.88 and 0.91, respectively, versus scores in the 0.50s and 0.60s for all others). While these fine-grained scores are not displayed to users — they intentionally see only the face recognition badge — the very close mean and median confidence scores for "facial match" versus "different person" illustrate the difficulty of automatically identifying non-replica matches and offer support for a hybrid human-AI approach (cf. Section 4.3.1).

Participants found the community's opinions to be helpful for assessing whether two photos were facial matches or not. S4 said, *"I mean, I think it's cool to see what the community is saying, because I do feel when it comes to saying 'Is this the same person in both these pictures?', that's really the best way to do it if you don't have any [other] information."* C1 said that he found the community opinions for facial matches especially helpful because he is *"face blind"*: *"I can't really identify the face-to-face, but the hair and the mustache and all the stuff that, in addition, it helps me with that for sure. It's good because other than having to find somebody close to me and be like, 'Do you think this is the same people?' [I] have that community right there."*

Participants had mixed opinions about facial recognition technology, but most found the badge indicating whether it supported the two photos being a match to be a useful data point. S3 said, *"The facial recognition saying they are similar — I would go ahead and trust that but I don't know if I would trust it enough to make a verification on my own."* In general, participants found the strengths of facial recognition and the community to be complementary in determining whether two photos showed the same person, and liked seeing both results together. H4 felt the community and the technology had separate roles:

> *This is facial recognition, and this is the historical background. I trust facial recognition, but it makes me feel better to have that historical background. I think the human eye can be tricked by different hairstyles and different beards. Just to have this outside historical verification to say like, 'Okay, maybe you or I was tricked, but the machine was not,' I think that's really helpful.*

S1 relied on both the community and facial recognition to make a decision on a facial match: *"I think both give me kind of an idea. Okay, there's this facial recognition technology being used, but also there's other users that are leaning towards that this is the same person."* H5 got a similar boost of confidence: *"This is certainly reassuring seeing not only the AI match, but also in terms of the community — seeing that seven users have said that this is a facial match as well. I'd be quite convinced by this."*

7 DISCUSSION

7.1 Leveraging Collective Intelligence for Validating Person Identification

Prior work has raised concerns about misinformation in online history communities [43, 75]. To address these problems on CWPS, we built Photo Steward for supporting community-based validation of photo IDs. Users found Photo Steward's stewardship visualizations not only helpful for affirming their own assessment, but also for discovering new knowledge and correcting their decisions, if need be. These visualizations, combined with the validation workflow, exhibit a form of stigmergic collaboration, where users build on prior knowledge left by the community and leave their own assessment for other users [14, 17, 27, 56].

	# of Photo Pairs	# of Photo Pairs compared by 1 user	# of Photo Pairs compared by 2 users	# of Photo Pairs compared by 2+ users
Supported by users and face recognition	1408	1281	109	18
Disputed by users and face recognition	26	24	0	1
Users unsure, face recognition disputes	42	42	0	0
Users unsure, face recognition supports	219	217	2	0
Users dispute, face recognition supports	239	238	1	0
Users support, face recognition disputes	266	257	8	1

Table 5: User Comparisons (Facial Similarity) vs. Face Recognition.

Photo Steward allowed users to express how confident they are about an ID in a fine-grained manner, in contrast to the binary agreements or disagreements observed on iNaturalist by Wiggins et al. [72]. The CWPS community preferred this nuanced form of stewardship as users are likely to have different degrees of confidence based on the evidence available for identifying a photo, thus demonstrating the effectiveness of Design Goal 2. As S3 pointed out, users often experience difficulty in making a binary decision about individuals who lived 150 years back due to the lack of surviving documentation.

Beyond its basic usefulness, Photo Steward's full potential can best be realized through sustained community participation, but most IDs on CWPS only received one vote. To address this challenge, we can leverage different crowdsourcing and online community strategies. For example, we can draw the community's attention towards IDs that are *"more of a puzzle,"* as H5 suggested, similar to Twitter's Birdwatch promoting tweets for fact-checking [52]. Designing nudges to encourage lurkers to vote on the IDs they are viewing can further help in these efforts. Organizing community events can help foster interest and participation in collaboratively verifying IDs, drawing inspiration from crowdsourcing events like CrowdSolve, where experts and novices collaborate on solving missing persons cold cases [70]. Incentive mechanisms such as leaderboards and challenges [46] can drive extrinsic motivation within the community for verifying the IDs. Finding users who are more likely to vote on an ID, based on their skills and interests, can also be an effective collaboration strategy [71]. In future work, we plan to integrate these strategies and introduce explicit "calls to action" [51, 58] on the home feed, guiding the community's attention towards IDs that require validation and fostering more sustainable, collaborative participation in historical photo identification.

7.2 Exercising Deliberation in Human-AI Teams

We found that Photo Steward's validation workflow was effective not only for voting on the IDs and comparing the photos side-by-side, but also encouraged users to deliberate on their decision, drawing parallels to other social computing systems that support reflection and deliberation (e.g., [28, 29]). This deliberative intervention was non-trivial as users on CWPS follow an identification pipeline which is powered by facial recognition, an AI algorithm that is far from perfect [44]. Users are trying to find the correct match, if present at all, from a pool of potential candidates, which are largely comprised of similar-looking false positives — akin to *finding a needle in a haystack*. Further, the task of comparing photos of people is by no means an easy task for humans, even in a modern context [53]. While Photo Steward can not completely curb automation bias, an issue that has been previously observed in multiple online history communities [44, 75], its multi-step, validation workflow with structured feedback interventions encouraged users to deliberate over AI suggestions before making a decision.

Photo Steward's workflow also compartmentalizes the tasks that AI is good at – such as quickly retrieving similar-looking candidates from a large search pool – from the tasks where the AI makes more errors – such as verifying whether two faces show the same person or not [5, 55]. Decoupling facial similarity comparison from the person identification task (Design Goal 1) allows the users to now focus on the face verification task. In doing so, Photo Steward's workflow supports effective human-AI teaming in the context of person identification by allowing the user to make a granular assessment for the face verification task instead of the AI, while also ensuring that an AI-retrieved, similar-looking potential candidate is being compared against.

As imperfect AI algorithms get deployed in high-stakes scenarios such as medical imaging, law enforcement, etc. [6], it becomes more critical to reduce automation bias and encourage more deliberative decision-making. Amershi et al. recommend granular user feedback while interacting with AI systems as part of their "Guidelines for Human-AI Interaction" [2]. Similarly, other forms of design interventions, such as counterfactual AI explanations [63], chatbots [26], and community opinions [59] can also be explored for encouraging deliberative decision-making with AI assistance.

Prior work in human-face recognition teams has shown that algorithmic suggestions can have a significant biasing effect on a user's decision [23]. Our findings showed that Photo Steward was able to encourage users to exercise deliberation while interacting with results retrieved by facial recognition. At the same time, Table 4 also showed that users can differ from the algorithm's suggestions, thus necessitating a deeper dive analysis of this dissonance as part of future work.

7.3 Assessing Quality in Crowdsourced Original Historical Research

We found that Photo Steward's stewardship visualizations helped users assess the reliability of photo IDs on CWPS, which were a result of Design Goal 3. However, crowdsourced identifications always run the risk of groupthink [24, 25], which can eventually mislead users into believing and amplifying misidentifications, a concern also raised by H4. Public deliberation of modern photo IDs on social media can have profound negative consequences for false targets, as exemplified by the Boston Marathon bombing [49, 66] and the recent US Capitol riot [48]. This raises the question: are Photo Steward's stewardship visualizations sufficient for assessing the quality of photo identifications made on CWPS?

Prior work on crowdsourced scholarship suggests an answer. Rosenzweig [57] analyzed Wikipedia as a source of historical scholarship, noting its policy against original research, and advocating for it as a tool for teaching the limitations of information sources and critical analysis of primary and secondary sources. Motivated along similar lines, Forte et al [18] proposed the *assessability* framework for designing assessable participatory information systems, based on information provenance and stewardship. The concept of provenance, extensively used in history and archival studies, describes information that makes it possible to trace the ownership or origins of the content, while stewardship refers to the processes that were used for maintaining the content, including its authorship. In the case of Wikipedia, Forte et al. found that visualizing provenance (i.e., citation types) and stewardship (i.e., article quality) had a significant impact on assessments of articles and Wikipedia as an information source.

While Photo Steward enables information stewardship on CWPS, there is an opportunity for incorporating provenance into the CWPS platform to make it a truly *assessable* online platform. A significant proportion of the notes left by users on Photo Steward qualified as provenance information, namely comments about period inscriptions, family trees, external sources and URLs. The challenges of assessing IDs on CWPS are, however, different from assessing information on Wikipedia, primarily because CWPS supports original research unlike Wikipedia's no original research policy [73]. This was also the reason why we designed Photo Steward to be a review system rather than a single editable output such as Wikipedia; original research such as historical photo identifications is often times an evolving investigation rather than a final decision. To assess the reliability of original photo IDs made on CWPS, users may want to factor in the provenance of the reference photos that were used in the identification process. In such cases, Photo Steward's stewardship visualizations (i.e., facial similarity reliability) can further help the user in assessing whether the reference photos can be used as reliable provenance or not.

8 CONCLUSION

Photo Steward attempts to help users assess and validate photo IDs better on CWPS. We present an information stewardship architecture, and adapt it for the task of historical person identification. We demonstrate the effectiveness of Photo Steward on CWPS, an existing online platform, where users found the stewardship visualizations, which included the community opinions and the AI verdict, useful for making accurate assessments of photo IDs on the platform. Further, users found Photo Sleuth's multi-step, structured validation workflow to help them deliberate before making decisions about the photo's identity. This work opens doors for exploring new ways to leverage collective intelligence and AI in creating assessable online information systems for historical archives.

ACKNOWLEDGMENTS

We wish to thank Ron Coddington, Paul Quigley, Liling Yuan, and our study participants. This research was supported by NSF IIS-1651969 and a Virginia Tech ICTAS Junior Faculty Award.

REFERENCES

[1] 2021. Heritage Auctions: World's Largest Collectibles Auctioneer. https://www.ha.com/
[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
[3] Brakkton Booker. 2015. $2 photo found at Junk Store has Billy the kid in it, could be worth $5M. https://www.npr.org/sections/thetwo-way/2015/10/15/448993361/-2-photo-found-at-junk-store-has-billy-the-kid-in-it-could-be-worth-5-million
[4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
[5] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*. 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html
[6] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.
[7] Fanny Chevalier, Stéphane Huot, and Jean-Daniel Fekete. 2010. Wikipediaviz: Conveying article quality for casual wikipedia readers. In *2010 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 49–56.
[8] H Jason Combs. 2005. The Internet's Impact on the Antiques Trade. *Material Culture* (2005), 26–41.
[9] Anthony DeBartolo. 1975. Appraisers can tell when old photo really may be golden. https://www.chicagotribune.com/news/ct-xpm-1985-12-27-8503300067-story.html
[10] Anthony DeBartolo. 2021. Appraisers can tell when old photo really may be golden. https://www.chicagotribune.com/news/ct-xpm-1985-12-27-8503300067-story.html
[11] Harley D Dickinson. 1998. Evidence-based decision-making: an argumentative approach. *International Journal of Medical Informatics* 51, 2-3 (1998), 71–81.
[12] For The Inquirer Edward Colimore. 2019. Did John Wilkes Booth get away with murdering President Abraham Lincoln? https://www.inquirer.com/news/john-wilkes-booth-lincoln-conspiracy-photo-recognition-20190415.html
[13] Mark Elliott. 2006. Stigmergic Collaboration: The Evolution of Group Work: Introduction. *m/c journal* 9, 2 (2006).
[14] Mark Elliott. 2016. Stigmergic collaboration: A framework for understanding and designing mass collaboration. In *Mass collaboration and education*. Springer, 65–84.
[15] Thomas Erickson and Wendy A Kellogg. 2000. Social translucence: an approach to designing systems that support social processes. *ACM transactions on computer-human interaction (TOCHI)* 7, 1 (2000), 59–83.
[16] Dan Evon. 2020. Did Joe Biden's Great-Grandfather Own Slaves? https://www.snopes.com/fact-check/joe-biden-slaves-great-grandfather/
[17] Kristie Fisher, Scott Counts, and Aniket Kittur. 2012. Distributed sensemaking: improving sensemaking by leveraging the efforts of previous users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 247–256.
[18] Andrea Forte, Nazanin Andalibi, Thomas Park, and Heather Willever-Farr. 2014. Designing information savvy societies: an introduction to assessability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2471–2480.
[19] Jacey Fortin. 2018. She Was the Only Woman in a Photo of 38 Scientists, and Now She's Been Identified. *The New York Times* (Mar 2018). https://www.nytimes.com/2018/03/19/us/twitter-mystery-photo.html
[20] Sarah A Gilbert. 2020. " I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a Public Scholarship Site on Reddit:

A Case Study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.

[21] Jerome S Handler and Michael L Tuite. 2007. *Retouching History: The Modern Falsification of a Civil War Photograph.*

[22] M. Keith Harris. 2019. Civil War Photo Sleuth. *Journal of American History* 106, 2 (2019), 544–546. https://doi.org/10.1093/jahist/jaz498

[23] John J Howard, Laura R Rabbitt, and Yevgeniy B Sirotin. 2020. Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making. *Plos one* 15, 8 (2020), e0237855.

[24] Pan Hui and Sonja Buchegger. 2009. Groupthink and peer pressure: Social influence in online social network groups. In *2009 International Conference on Advances in Social Network Analysis and Mining*. IEEE, 53–59.

[25] Nassim JafariNaimi and Eric M Meyers. 2015. Collective intelligence or group think? Engaging participation patterns in World Without Oil. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 1872–1881.

[26] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator chatbot for deliberative discussion: Effects of discussion structure and discussant facilitation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.

[27] Aniket Kittur, Andrew M Peters, Abdigani Diriye, and Michael Bove. 2014. Standing on the schemas of giants: socially augmented information foraging. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 999–1010.

[28] 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, USA, 1188–1199. https://doi.org/10.1145/2531602.2531677

[29] 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, USA, 1559–1568. https://doi.org/10.1145/2207676.2208621

[30] Kurt Luther. 2016. How Fellow Collectors, Field Photos and Muttonchops Identified an Unknown Officer. *Military Images* 34, 1 (2016), 29–31.

[31] Kurt Luther. 2017. Merrill Carbine Leads to a Soldier's Identification. *Military Images* 35, 2 (2017), 64–65.

[32] Kurt Luther. 2018. Non-Traditional Research Tools—and Serendipity. *Military Images* 36, 3 (2018), 12–13.

[33] Kurt Luther. 2018. What are the odds? Photo sleuthing by the numbers. *Military Images* 36, 1 (2018), 12–15.

[34] Kurt Luther. 2019. What to Do When Gold Standards Go Wrong? *Military Images* 37, 1 (2019), 8–9. https://www.jstor.org/stable/26532101

[35] Kurt Luther. 2020. Real-life accounts on the research trail: How to Trust the Worthiness of an Identification. *Military Images* 38, 3 (213) (2020), 8–11. https://www.jstor.org/stable/26914966

[36] Kurt Luther. 2020. Real-life accounts on the research trail: Lost and Found in the Library of Congress. *Military Images* 38, 2 (212) (2020), 10–13. https://www.jstor.org/stable/26890126

[37] Kurt Luther. 2020. Real-life accounts on the research trail: The Art of Photo Sleuthing. *Military Images* 38, 4 (214) (2020), 8–11. https://www.jstor.org/stable/26925454

[38] Ramona Martinez. 2012. Photo mystery solved, then doubted, then deciphered, thanks to readers. https://www.npr.org/sections/pictureshow/2012/04/17/150801239/photo-mystery-solved-then-doubted-then-resolved-thanks-to-readers

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

[40] Christian A Meissner and John C Brigham. 2001. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law* 7, 1 (2001), 3.

[41] Vikram Mohanty, Kareem Abdol-Hamid, Courtney Ebersohl, and Kurt Luther. 2019. Second opinion: Supporting last-mile person identification with crowdsourcing and face recognition. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 86–96.

[42] Vikram Mohanty and Kurt Luther. 2023. DoubleCheck: Designing Community-based Assessability for Historical Person Identification. *ACM Journal on Computing and Cultural Heritage (JOCCH) (to appear)* (2023).

[43] 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*. 547–557.

[44] Vikram Mohanty, David Thames, Sneha Mehta, and Kurt Luther. 2020. Photo Sleuth: Identifying Historical Portraits with Face Recognition and Crowdsourced Human Expertise. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 4 (2020), 1–36.

[45] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing? Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 441–450.

[46] Benedikt Morschheuser, Juho Hamari, and Jonna Koivisto. 2016. Gamification in crowdsourcing: a review. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE, 4375–4384.

[47] Kathleen L Mosier and Linda J Skitka. 1999. Automation use and automation bias. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 43. SAGE Publications Sage CA: Los Angeles, CA, 344–348.

[48] Greg Myre. 2021. How Online Sleuths Identified Rioters At The Capitol. https://www.npr.org/2021/01/11/955513539/how-online-sleuths-identified-rioters-at-the-capitol

[49] Johnny Nhan, Laura Huey, and Ryan Broll. 2017. Digilantism: An analysis of crowdsourcing and the Boston marathon bombings. *The British journal of criminology* 57, 2 (2017), 341–361.

[50] Raja Parasuraman and Dietrich H Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Human factors* 52, 3 (2010), 381–410.

[51] Junwon Park, Ranjay Krishna, Pranav Khadpe, Li Fei-Fei, and Michael Bernstein. 2019. AI-based request augmentation to increase crowdsourcing participation. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 115–124.

[52] Sarah Perez. 2022. Twitter to show 'Birdwatch' community fact-checks to more users, following criticism. https://techcrunch.com/2022/03/03/twitter-to-show-birdwatch-community-fact-checks-to-more-users-following-criticism/

[53] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. 2018. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences* 115, 24 (2018), 6171–6176.

[54] Peter Pirolli, Evelin Wollny, and Bongwon Suh. 2009. So you know you're getting the best possible information: a tool that increases Wikipedia credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1505–1508.

[55] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products.

[56] Amira Rezgui and Kevin Crowston. 2018. Stigmergic coordination in Wikipedia. In *Proceedings of the 14th International Symposium on Open Collaboration*. 1–12.

[57] Roy Rosenzweig. 2006. Can History Be Open Source? Wikipedia and the Future of the Past. *Journal of American History* 93, 1 (June 2006), 117–146.

[58] Saiph Savage, Andres Monroy-Hernandez, and Tobias Höllerer. 2016. Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 813–822.

[59] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.

[60] Michael S. Schmidt. 2018. 'Flags of Our Fathers' Author Now Doubts His Father Was in Iwo Jima Photo. *The New York Times* (Jan 2018). https://www.nytimes.com/2016/05/04/us/iwo-jima-marines-bradley.html

[61] Jennifer Schuessler. 2017. Found: Oldest Known Photo of a U.S. President (Socks and All). https://www.nytimes.com/2017/08/16/arts/design/john-quincy-adams-daguerreotype-sothebys-auction.html

[62] Julia Schwarz and Meredith Morris. 2011. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1245–1254.

[63] Ruoxi Shang, KJ Kevin Feng, and Chirag Shah. 2022. Why Am I Not Seeing It? Understanding Users' Needs for Counterfactual Explanations in Everyday Recommendations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1330–1340.

[64] Linda J Skitka, Kathleen L Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51, 5 (1999), 991–1006.

[65] Verbi Software. 2019. MAXQDA 2020 [computer software]. VERBISoftware. Availablefrommaxqda.com.

[66] NPR Staff. 2016. How Social Media Smeared A Missing Student As A Terrorism Suspect. https://www.npr.org/sections/codeswitch/2016/04/18/474671097/how-social-media-smeared-a-missing-student-as-a-terrorism-suspect

[67] Bongwon Suh, Ed H Chi, Aniket Kittur, and Bryan A Pendleton. 2008. Lifting the veil: improving accountability and social transparency in Wikipedia with wikidashboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1037–1040.

[68] W Ben Towne, Aniket Kittur, Peter Kinnaird, and James Herbsleb. 2013. Your process is showing: controversy management and perceived quality in Wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1059–1068.

[69] Civil War Trust. 2021. Military Images Magazine | Interview with Ron Coddington. https://www.battlefields.org/learn/articles/military-images-magazine
[70] Sukrit Venkatagiri, Aakash Gautam, and Kurt Luther. 2021. CrowdSolve: Managing Tensions in an Expert-Led Crowdsourced Investigation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–30.
[71] Shaun Wallace, Lucy Van Kleunen, Marianne Aubin-Le Quere, Abraham Peterkin, Yirui Huang, and Jeff Huang. 2017. Drafty: Enlisting Users To Be Editors Who Maintain Structured Data. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5.
[72] Andrea Wiggins and Yurong He. 2016. Community-based data validation practices in citizen science. In *Proceedings of the 19th ACM Conference on computer-supported cooperative work & social computing*. 1548–1559.
[73] Foundation Wikimedia. 2022. No original research. https://en.wikipedia.org/wiki/Wikipedia:No_original_research
[74] 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 2012 iConference*. ACM, 303–310.
[75] 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.
[76] Jeremy B Wilmer. 2017. Individual differences in face recognition: A decade of discovery. *Current Directions in Psychological Science* 26, 3 (2017), 225–230.
[77] Bob Zeller. 2022. Searching for photos of Civil War Soldiers | David Wynn Vaughan. https://www.civilwarphotography.org/searching-for-photos-of-civil-war-soldiers/

A CIVIL WAR PHOTO SLEUTH: DESIGN CHALLENGES AND OPPORTUNITIES

Here, we provide details of three key challenges posed by Civil War Photo Sleuth that might have contributed towards misidentification and subsequently, inaccurate assessments, and how they present design opportunities for Photo Steward.

A.1 Conflating facial similarity with photo identification

A.1.1 Challenges: Mohanty et al. conducted a benchmarking study of CWPS [44], and found the face recognition algorithm to be of low precision; i.e., it retrieved hundreds of search results which may look similar to the query photo but are actually different people (false positives). Low precision increases the chances that users will interact with a lot of false positives. In such cases, one may need to garner additional information (i.e., comparing biographical information) before making a decision. However, it is plausible that automation bias may play a role in non-expert users making a match solely based on facial similarity [47]. As a result, there is strong potential for misidentification (see Figure 8).

A.1.2 Current Workflow: CWPS's compare interface allows users to closely inspect the search results for a potential match, but does not make any distinction between facial similarity comparison and photo identification. Both are conflated into a one-step process, with one "Identify" button for the users to make their decisions (see Figure 2-C). Yet, users may want to indicate agreement with just the facial similarity (i.e., query photo and the search result showing the same person) but not the identity (i.e., name and biographical information), or vice versa.

A.1.3 Design Goal 1: To support accurate investigation of photo identifications, users should be able to deliberate on the different aspects of the decision-making process. Providing users with a decision-making workflow that decouples facial similarity comparison from the overall photo identification task would allow them to focus on these tasks separately, while discouraging them from making decisions solely on the basis of facial similarity.

A.2 Lacking support for fine-grained, deliberative decision making

A.2.1 Challenges: Historical photo identification is a complex task, where experts often corroborate multiple pieces of evidence, including facial similarity comparison, before reaching a decision about the identity of the photo [33]. While confirming an identity, experts may be highly confident if the source, military records, uniform clues, and additional photos of the same person all line up, or slightly confident if they need additional evidence. Conflicting evidence pieces may also affect their confidence levels. Similarly, they may have different degrees of certainty while ruling out an identity for a photo. A lack of support for expressing and displaying granularity in these photo identification decisions can lead to varying degrees of uncertainty being captured and misinterpreted as a confirmation, and eventually propagating misidentifications. Further, while it is safe to assume the vast majority of the Civil War photography community care about the accuracy of the photo

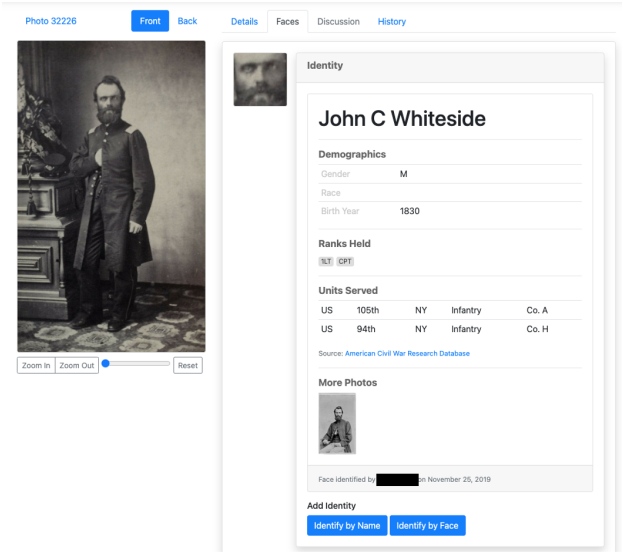


Figure 8: An example of a misidentified photo on CWPS. This photo was identified as John C Whiteside based on facial similarity with the top search result: an identified reference photo of John C Whiteside. However, other visual clues, such as the photographer studio, do not align with Whiteside's biographical information.

IDs, there exists a small risk of financially-driven misidentifications since certain types of identified photos are considered to be more valuable [9].

A.2.2 Current Workflow: The compare interface's "Identify" button (see Figure 2-C) is the only mechanism available on the platform for users to provide (a unary) input on photo identifications. CWPS users currently cannot provide fine-grained feedback on a given photo's identity, either at the time of identifying or afterwards on the photo page.

A.2.3 Design Goal 2 : Users should be able to express how certain or uncertain they are about a photo identification. Interventions for allowing users to provide structured feedback can not only facilitate more accurate, fine-grained responses, but also encourage deliberation on the available evidence before making a decision, borrowing inspiration from other social computing systems that support users reflecting and deliberating on available information [28, 29].

A.3 Limited validation opportunities for the community

A.3.1 Challenges: For humans, deciding whether two photos show the same person is a highly subjective experience. A study by Mohanty et al. [41] showed that participants in a crowdsourcing study often disagree with each other and with facial recognition results in person identification tasks. The same study showed that people often seek a second opinion from peers for validating photo identifications. Without validation, original crowdsourced historical research may result in unresolved cases of conflicting IDs proposed

for the same face, incorrect IDs remaining unchecked, or airtight IDs not being confirmed.

A.3.2 Current Workflow: Once a photo has been identified, CWPS does not offer other users the option to express their opinions on whether two photos show the same person or not, or if the photo has been identified correctly or not (see Figure 8). The photo page does not indicate if (or why) the photos are linked to each other, how they are similar, or which user(s) considered them to be similar. Further, it does not indicate the role of facial recognition in linking them together.

A.3.3 Design Goal 3 : To support accurate original historical research, the platform should encourage information stewardship from the community by allowing members to give their opinions on existing photo identifications [72]. Further, the platform should also be transparent about the role of community stewardship and facial recognition for a given photo identification, which can then act as evidence for aiding subsequent validation efforts by other users, thus supporting a form of stigmergic collaboration [13, 14].

B DETAILS FOR LOG ANALYSIS

To understand the community's stewardship behaviors, we examined website logs for all user activities for one year since we launched the new features (December 2020 – December 2021). During this period, we observed 5843 voting instances on 5672 photos for 5355 unique IDs, where a user voting on whether Photo N can be identified as a Person M or not is considered to be one *voting instance*. For a given voting instance, we analyzed 1) any associated facial similarity comparisons to understand how they *deliberated* on their final voting decision, and 2) the user's confidence vote plus any justification notes left by the user to understand their decision rationale. Of the 5672 query photos, 4297 photos (4377 voting instances) did not have any facial similarity comparisons — only the user's confidence on the ID (plus any notes) was captured.

For the remaining voting instances which had at least one associated facial similarity comparison (i.e., users deliberate through the two-step validation process), we broke them down into *pre-identified* (i.e., the user knew the ID of the photo ahead of time and therefore, might be looking for a specific target on the search results page) and *post-identified* (i.e., the user did not know the ID of the photo and therefore, may not have a specific target while analyzing the search results) cases. As mentioned earlier in Section 4.2, users had the opportunity to validate IDs either on the search results page on the photo page. While a *pre-identified* voting instance almost certainly originated from the search results page, a *post-identified case* could be from either page. Since the logs did not give us the page origin of each vote directly, we triangulated from CWPS's timestamps to determine whether a given voting instance was for a *pre-identified* case or a *post-identified* one.

After collating all the associated facial similarity comparisons for a given voting instance, we had 1064 *pre-identified* and 576 *post-identified* user deliberations to analyze. Each deliberation instance is a user's attempt to identify a query photo as a given target (person) ID, where they first compare facial similarity with all other photos that have been identified as the target ID, followed by the user's confidence on the query photo being the target ID. We analyzed the user responses to see whether the facial similarity comparisons had any impact on the user's confidence.

Users had provided comments in 682 (out of 5843, or 11.7%) voting instances. We coded these comments using an iterative, inductive approach, which resulted in five high-level themes, which can broken down into 13 sub-categories (see Appendix C).

To understand any stigmergic collaboration processes at play, we also analyzed the IDs which had multiple votes to check for agreements and disagreements between the voters. We further analyzed how the community's facial similarity comparisons compare against the facial recognition scores.

C THEMES FOR NOTES ANALYSIS

- **Photo Comparisons**
  - **High-Level Comparison:** The note mentions "replica", "facial similarity", "facial match", "identical", "visual comparison", and other similar terms that describe comparisons with a prior identified photo.
  - **Describing facial features:** The note mention facial features like "eyes", "hairline", "ears", etc. to make comparisons
- **Word-of-Mouth**
  - **Descendant:** The note either mentions that the user is a descendant of the person being identified, or they got the information from the family of the person.
  - **Ownership:** The note either mentions that the user owns a printed version of the photo, compared with a photo in their collection, or they know the owner of the photo.
  - **Familiarity/Self-Reported Research:** The note mentions that the user has seen the photo somewhere, be it in a book, museum, etc.
- **Visual Clues**
  - **Uniform:** The note mentions visual clues that pertain to the uniform of the person (e.g., hat insignia, shoulder straps, etc.) The user may infer the possible service information (i.e., ranks, branches, regiments, etc.) from the uniform clues.
  - **Inscription:** The note mentions the presence of a period inscription on the photo (a highly trustworthy primary source for a person's ID), or an album case, or modern inscriptions such as books, which is generally the name of the person being identified. In some instances, the inscribed text may point to the person's service information.
  - **Other Visual Clues:** The note mentions visual clues in the photo (e.g., backmarks, borders, etc.) beyond the person's face.
- **External Information**
  - **External URL:** The note mentions an external URL, which supposedly has additional information about the photo's ID.
  - **Other sources (e.g., museum, website, book, etc.):** The note mentions an external source (e.g., museum, book, etc.) that supposedly has evidence for the photo's ID, but no URLs are provided. Details about the source may or may not be available.

  - **Lack of Information / Seeking Additional Evidence:** The note mentions the lack of evidence or seeking additional evidence, be it about the source or the service information.
- **Providing Additional Information**
  - **Biographical Information:** The notes mentions additional information about the person's service records, specific regiments, biographical information (name, year, location), etc.
  - **Additional Context:** The note mentions some additional context provided by the user to justify their decision, such as information about the photo collection, or pointing to someone else's research, or some historical context, or incorrect evidence, or if the person is prominent.

D PARTICIPANT DETAILS

Students: Undergraduate and master's students concentrating in history who use Civil War photos for their coursework and research projects, but are not (yet) employed in a professional capacity as historians. We recruited five students via recommendations from our university's history department. None of the students had used CWPS before, or were known to the authors prior to the study. Three students were men and two were women, and all were in the "18 to 30" age group. We anonymize them with identifiers S1–S5.

Amateur Experts: Experienced users of Civil War Photo Sleuth who have added over 50 photos each and have substantial knowledge of Civil War history, but are not professional historians. We recruited five amateur experts from the CWPS contact list. All five users were men, and they were distributed across different age groups (two in "18 to 30", two in "31 to 40", and one in "51 to 60"). We anonymize them with identifiers C1–C5. C1 and C3 are among the most active daily users on CWPS. Only two of the five had used Photo Steward before.

Historians: Expert historians with a graduate degree in history, specializing in American Civil War history, but with little or no previous experience with CWPS. We recruited five historians via recommendations from our university's history department. Three historians were men and two were women. They were distributed across different age groups (two in "18 to 30", two in "31 to 40", and one in "51 to 60"). We anonymize them with identifiers H1–H5. None of them had used Photo Steward before.

E LAB STUDY QUESTIONS

E.1 Semi-Structured Questions

- Is there a way that you would like to capture your thoughts on this ID and share them with others, if possible? If so, what would that look like? If not, can you explain why not?
- What did you think about the community opinions?
- What do you think about the ID quality visualization?
- What did you think about the 2-step process while agreeing/disagreeing on an identity?
- What is your overall opinion of both the interfaces?
- Which interface would you prefer for validating the information? And why?
- What would you change or improve?

E.2 Usability Survey

- **Q1.** The community's opinions about an identity were clear and easy to understand in the 2nd system. (1 = Strongly Disagree to 5 = Strongly Agree)
- **Q2.** The community's opinions about an identity were useful for assessing the information. (1 = Strongly Disagree to 5 = Strongly Agree)
- **Q3.** The process of voting on an identity was clear and easy to understand. (1 = Strongly Disagree to 5 = Strongly Agree)
- **Q4.** Comparing other photos first and then voting on an identity helped me deliberate and make more accurate decisions. (1 = Strongly Disagree to 5 = Strongly Agree)
- **Q5.** I was able to validate the information better using the 2nd system. (1 = Strongly Disagree to 5 = Strongly Agree)

F SUMMARY OF CHANGES

We summarize in Table 6 about how Photo Steward differs from CWPS.

CWPS	CWPS + Photo Steward
Decision-Making: How do users make identification decisions for the query photo when they see the similar-looking search results retrieved by facial recognition?	
A single-step comparison interface that allows the user to compare the query photo with one similar-looking photo and the associated biographical information of the target ID. The single-step workflow conflates the two tasks of facial similarity and person identification.	A multi-step validation interface that allows the user to compare the query photo with all previously identified photos of a similar-looking target ID (search result). The two-step workflow decouples the two tasks of facial similarity and person identification, thus allowing the user to deliberate on the facial similarity comparison between the query photo and the target photo(s) before making a decision on the target identity.
Users make a unary input on the query photo's ID by clicking an "Identify" button in the comparison interface, which will link the target ID to the query photo. The comparison interface does not allow users to rule out candidates.	Users can provide fine-grained decisions for both facial similarity and person identification steps in the validation interface. In Step 1, the user compares the query photo and the target photo for facial similarity by selecting from the following options: No (Different Person), Not Sure, Yes (Facial Match), and Yes (Replica). In Step 2, users indicate how confident they are about the query photo's ID (as the proposed target ID) by selecting the following options: No (Highly Confident), No (Slightly Confident), Not Sure, Yes (Slightly Confident), and Yes (Highly Confident).
Access: What kind of validation opportunities are available for users?	
Users can only access the comparison interface from the search results page if they search for similar-looking candidates using facial recognition. Once the photo has been identified, users cannot access the comparison interface on the photo page. The community cannot weigh in on an ID's reliability on the photo page.	Users can access the validation interface on both the search results page (while identifying the query photo) and the photo page (after the photo has been identified). The community can weigh in on the reliability of the proposed ID(s) and facial matches using the validation interface on the photo page, thus engaging in a form of stigmergic collaboration.
Evidence: What kind of stewardship evidence is presented to the users?	
The photo page displays the proposed ID(s) for the photo without any additional information about the ID's reliability. Similarly, other photos that have been matched to the query photo are also displayed without any reliability indicator. Since community opinions are not captured on CWPS, they are not displayed.	The photo page displays the proposed ID(s) for the photo along with stewardship visualizations of the a) community's confidence on the ID, and b) facial similarity comparison with other photos of the same ID by both the community and AI.

Table 6: Summary of changes: Civil War Photo Sleuth (CWPS) with and without Photo Steward.