Designing Human-AI Collaborative Systems for Historical Photo Identification

Vikram Mohanty
Virginia Tech, Arlington, USA
vikrammohanty@vt.edu

Abstract

Identifying people in historical photographs is important for preserving material culture, correcting the historical record, and creating economic value, but it is also a complex and challenging task. My dissertation addresses this challenge by leveraging the combined strengths of crowdsourcing and AI. Specifically, I study 1) how crowdsourced human expertise and facial recognition can be combined to support historical photo identification, 2) how novice crowds and face recognition can assist experts in picking the correct match from a set of similar-looking candidates, 3) how to design a quality assessment framework for verifying historical photo IDs, and 4) how the underlying AI model affects user behavior in the context of historical photo identification.

Introduction

Identifying people in historical photos is important as it can generate significant cultural and economic value and correct and preserve historical records. However, the task of identifying historical photos can be complex and challenging, due to issues such as a lack of centralized datasets and large search pools, and researchers lack adequate technological support. The current research practices employed by historians, antiques dealers, and collectors for identifying portraits largely involve manually scanning through hundreds of low-quality photographs, military records, and reference books, which can often be tedious and frustrating, without any guarantee of success.

AI-based face recognition algorithms can help support this effort, but are not widely used by historical photo experts due to lack of software products geared towards the history domain. These algorithms are often insufficient for solving the problem on their own, as prior studies have shown mixed results when comparing these algorithms to human baselines. Further, these algorithms are prone to false positives and gender and racial biases (Buolamwini and Gebru 2018). Historical photos pose other unique challenges as they are often achromatic, low resolution, and faded or damaged, which might result in loss of useful information for identification.

While face recognition is a powerful tool for narrowing down a large search pool of potential matches to a shortlist of very similar-looking candidates, it is less helpful for users seeking to select the correct match(es) among them. Drawing inspiration from similar challenges in transportation and telecommunications, I term this the *last-mile problem* of face recognition.

Much like misinformation in modern times, historical photos are also prone to being misidentified, with non-trivial consequences ranging from sowing conspiracy theories to reaping undeserved monetary benefits. This problem can be exacerbated by a lack of experience in historical photo ID verification and the use of imperfect AI such as facial recognition.

Similar to other AI technologies, facial recognition models are constantly evolving and being updated to provide better accuracy and address racial and gender biases. However, they are generally evaluated against established benchmarks using modern datasets. Little is known about how users perceive different facial recognition models and evaluate their accuracy in a real-world context, specifically in the domain of historical photo identification.

My dissertation attempts to address these problems through these specific research questions:

- **RQ1.** How can we combine the complementary strengths of crowdsourced human expertise and automated face recognition to support historical person identification?
- **RQ2.** How can we support experts in solving the last-mile problem of person identification?
- **RQ3.** How can we support accurate assessment and validation of historical photo identification quality?
- **RQ4.** How does the underlying AI model (i.e., facial recognition model) influence user behavior in the context of historical photo identification?

Completed Work

Photo Sleuth: Combining Human Expertise and Face Recognition to Identify Historical Portraits

To answer **RQ1**, I developed Photo Sleuth[1], a web-based platform that combines crowdsourced human expertise and automated face recognition to support historical portrait identification (Mohanty et al. 2019b; Mohanty, Thames, and

[1] www.civilwarphotosleuth.com

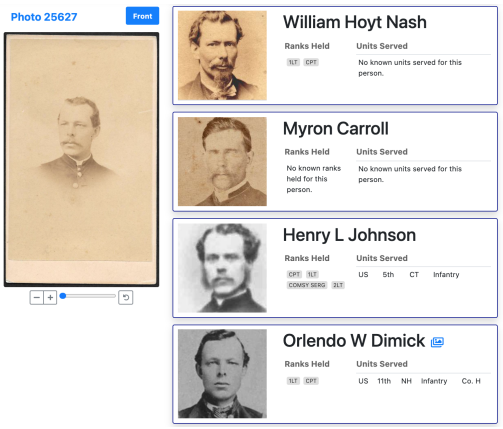

Figure 1: Search Results on Photo Sleuth

Luther 2018). I introduced a novel *haystack model* person identification pipeline in which users first identify and tag relevant visual clues in an unidentified portrait. The system then suggests filters based on these tags to narrow down search results of identified reference photos. Finally, the user can carefully inspect the narrowed search results, sorted using automatic face recognition, to make a potential identification (see Figure 1). This pipeline also bootstraps crowdsourced user contributions to grow the site's database of reference images in a sustainable way, increasing the likelihood of a potential match in the future.

Photo Sleuth initially focuses on identifying portraits from the American Civil War (1861–65), the first major conflict to be extensively documented through photographs. An estimated three million soldiers fought in the war and most of them had their photos taken at least once. After 150 years, millions of these portraits survive in museums, libraries, and individual collectors, but the identities of most have been lost.

Photo Sleuth was publicly launched in 2018 and I conducted a mixed-methods evaluation of its first month of usage, including content analysis of uploaded photos and expert review of user identifications. My findings showed that the system helped users identify dozens of unknown portraits. Additionally, Photo Sleuth's pipeline encouraged users to voluntarily add hundreds of identified portraits to aid future research, suggesting a sustainable model for long-term participation. In a follow-up paper (Mohanty et al. 2020), I conducted a longitudinal evaluation of Photo Sleuth after 11 months of deployment, validating the sustainable participation model, and a benchmarking study showing how well Photo Sleuth's face recognition and tagging features perform on a gold-standard dataset.

Second Opinion: Supporting Last-Mile Person Identification with Crowdsourcing and Face Recognition

As AI-based face recognition technologies are increasingly adopted for high-stakes applications like locating suspected criminals, public concerns about the accuracy of these technologies have grown as well. These technologies often present a human expert with a shortlist of high-confidence candidate faces from which the expert must select correct match(es) while avoiding false positives, which I term the *last-mile problem* in person identification.

To address **RQ2**, I developed Second Opinion, a web-based software tool that employs a novel crowdsourcing workflow inspired by cognitive psychology, *seed-gather-analyze*, to assist experts in solving the last-mile problem (Mohanty et al. 2019a). I evaluated Second Opinion in a mixed-methods, exploratory study where 10 experts, aided by 300 novice crowd workers, performed last-mile person identification tasks with top-5 candidates returned by AI-based face recognition. We found that a weighted aggregation strategy allows crowds to reduce face recognition's false positives by 75% while including the correct match 100% of the time, and also provides a modest improvement in ranking. Additionally, we found that experts were enthusiastic about the system and felt it helped them notice new details and build confidence in their decisions, though challenges remain in convincing experts to fully consider the crowd results.

Work in Progress

DoubleCheck: Designing Community-based Data Validation and Assessability for Historical Person Identification

As Photo Sleuth grows in size, currently serving over 17,000 registered users and hosting over 40,000 historical photos, it also faces the problem of misidentifications, partly due to automation bias. This problem draws attention to **RQ3**, and necessitates building systems that can support users in making accurate assessments of historical photo IDs and validating them. To address this problem, I propose DoubleCheck, a holistic quality assessment framework for supporting historical photo ID verifications, based on the concepts of information provenance and stewardship for assessable designs (Forte et al. 2014). I developed DoubleCheck on top of the existing Photo Sleuth platform to support its users in accurately assessing and validating photo IDs. As part of DoubleCheck, I modified Photo Sleuth's architecture for capturing accurate provenance information, and incorporated historical domain knowledge to determine source trustworthiness. I also built a validation workflow that allows users to compare two photos for facial similarity and express their fine-grained opinions on the ID of a photo. Finally, I built an automated quality assessment engine that determines whether an ID is verified or not, based on community opinions and source trustworthiness. DoubleCheck also visualizes stewardship at three different levels, by showing: 1) whether the community and AI consider a source matched via facial similarity to be reliable or not, 2) whether the community considers an ID to be reliable or not, and 3) whether an ID has been verified or not.

DoubleCheck was publicly released on Photo Sleuth in the last quarter of 2020. I conducted a mixed-methods evaluation of four months of usage, which included interviews with potential users of different expertise levels, and

log analysis of provenance and stewardship behaviors on the platform. My preliminary findings showed that users were able to assess historical photo IDs better using the DoubleCheck framework. The stewardship visualizations boosted the confidence of users' assessment. Users provided a wide range of provenance information, and found the organization of sources to be useful for their assessment. Users validated hundreds of different IDs on the platform, and found the workflow to be useful for validating photo IDs with careful deliberation.

Understanding User Perception of Facial Recognition Models

Prior work has shown that more accurate AI models do not necessarily translate into better performance with human users. Researchers have proposed design guidelines for human-AI interaction, where they have suggested a cautious approach while updating an AI system's behaviors and to inform users about any changes in the AI's capabilities. As facial recognition developers continue to update the models, it is unclear how it affects user experience for applications built on top of these models. In this study (addressing **RQ4**), I plan to investigate whether the basis of these guidelines hold true in the context of a face recognition-based system, if the underlying model was to be updated.

Prior work on search engines and search results has shown that more search results led to poorer decision-making (i.e., "a paradox of choice") and fewer results yielding higher subjective satisfaction in choice and greater confidence in its correctness compared to more results (Oulasvirta, Hukkinen, and Schwartz 2009). Similarly, research suggests that even slightly higher retrieval latency by web search engines can lead to dramatic decreases in users' perceptions of result quality and engagement with the search results. Drawing inspiration from this prior work, I investigate the impact of underlying face recognition models in the context of Photo Sleuth, specifically to understand:

1. the factors (i.e., number of search results, time taken to load the results, order of the results, content of the search results, and query photos) responsible for users noticing a change in the model.
2. how users perceive and interact with different facial recognition models in terms of identification confidence, response latency (time taken to analyze a result), and the number of results they analyze; and how these metrics correlate with the number of search results and time taken to load the results.
3. how users rate the accuracy of different facial recognition models, and which model they prefer for their task of historical photo identification.

Expected Contributions

My dissertation aims to contribute: 1) a web-based platform that combines crowdsourcing and face recognition for supporting historical person identification and addressing the last-mile problem, 2) a quality assessment framework for supporting historical photo ID verification, and 3) design implications for updating underlying models in an AI-infused application. My work opens doors for exploring new ways for building person identification systems that look beyond face recognition and leverage the complementary strengths of human and artificial intelligence. At the same time, it also pushes research on building assessable online information systems and engaging online communities and AI for historical photo identification and verification.

Goals for HCOMP Graduate Consortium

I am hoping to gather feedback on the framing, organization and direction of my dissertation proposal. I am also interested in getting a fresh perspective on both the gaps and wider applicability of my research. As I enter the final year of my program and plan my final study (i.e., **RQ4**), I am also interested in refining the details of this study. A major goal for me is to connect to senior researchers in this domain and seek for career mentoring. I am also interested in gathering thoughts about how Photo Sleuth can be adapted for other types of person identification tasks in a sustainable and ethical manner. I will be applying for industry research positions in the future, and I hope to learn from successful experiences of others as I plan for the next steps of my career.

References

Buolamwini, J., and Gebru, T. 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.

Forte, A.; Andalibi, N.; Park, T.; and Willever-Farr, H. 2014. Designing information savvy societies: an introduction to assessability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2471–2480.

Mohanty, V.; Abdol-Hamid, K.; Ebersohl, C.; and Luther, K. 2019a. Second opinion: Supporting last-mile person identification with crowdsourcing and face recognition. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, 86–96.

Mohanty, V.; Thames, D.; Mehta, S.; and Luther, K. 2019b. 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.

Mohanty, V.; Thames, D.; Mehta, S.; and Luther, K. 2020. Photo sleuth: Identifying historical portraits with face recognition and crowdsourced human expertise. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10(4):1–36.

Mohanty, V.; Thames, D.; and Luther, K. 2018. Are 1,000 features worth a picture? combining crowdsourcing and face recognition to identify civil war soldiers. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018)*.

Oulasvirta, A.; Hukkinen, J. P.; and Schwartz, B. 2009. When more is less: the paradox of choice in search engine use. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 516–523.