The Seventh AAAI Conference on Human
Computation and Crowdsourcing (HCOMP-19)

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

Vikram Mohanty,[1] **Kareem Abdol-Hamid,**[1] **Courtney Ebersohl,**[2] **Kurt Luther**[1,2]
[1]Department of Computer Science, [2]Department of History, Virginia Tech, USA
{vikrammohanty, kkabdolh, courteb, kluther}@vt.edu

Abstract

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 we term the "last-mile problem." We propose 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. We evaluated Second Opinion with a mixed-methods lab study involving 10 experts and 300 crowd workers who collaborate to identify people in historical photos. We found that crowds can eliminate 75% of false positives from the highest-confidence candidates suggested by face recognition, and that experts were enthusiastic about using Second Opinion in their work. We also discuss broader implications for crowd–AI interaction and crowdsourced person identification.

Introduction

Identifying people in photographs (i.e., person identification) has long been an important task across many domains, allowing law enforcement to apprehend criminals (Keefe 2016), human rights investigators to verify atrocities (Human Rights Watch 2017), scholars to correct the historical record (Fortin 2018), and journalists to uncover scandals (Vozzella and Morrison 2019). Person identification tasks remain challenging, however, due to huge candidate pools and poor-quality source material.

Traditionally, person identification tasks were performed by trained experts, such as forensic specialists, who manually gather evidence, inspect visual clues, and conduct comparisons (White et al. 2017). Increasingly, these experts are supported by software tools that leverage artificial intelligence (AI)-powered face recognition to quickly compare a target face to thousands of candidates in photo databases (e.g., mug shots, historical archives) and return a subset of results ordered by confidence or similarity (Garvie, Bedoya, and Frankle 2016; Mohanty et al. 2019).

Despite the increasing popularity and quality of these tools, face recognition remains an imperfect technology. Photos with low resolution (Haghighat and Abdel-Mottaleb 2017) or pose variations (Pontin 2007) can hurt performance, and skewed training data can result in systemic bias against underrepresented demographics (Klare et al. 2012). Even at high confidence thresholds, automated methods frequently return many false positives (Buolamwini and Gebru 2018; Raji and Buolamwini 2019). Thus, while face recognition is a powerful tool for narrowing down thousands of possible faces to a shortlist of very similar-looking candidates, it offers little help for users seeking to select only the correct match(es) among them. Drawing inspiration from similar challenges in transportation and telecommunications, we term this the "last-mile problem" of face recognition.

In this paper, we introduce and evaluate Second Opinion, a system that augments AI-based face recognition with crowdsourced human insight to help experts with last-mile person identification. Our approach is motivated by two threads of prior work. First, humans often outperform face recognition algorithms, especially on fine-grained analysis tasks (Blanton et al. 2016; Best-Rowden et al. 2014), though the importance of expertise is disputed (Wirth and Carbon 2017). We developed a novel crowd workflow, *seed-gather-analyze*, that applies theories of similarity from cognitive psychology (Gentner and Markman 1997; Tversky 1977) to allow novice crowds to highlight important facial similarities and differences for expert review. Second, prior work has documented an emergent practice of experts seeking feedback from others on proposed person identifications (i.e., second opinions) in the domain of historical portraits, despite a lack of technological support (Mohanty et al. 2019). Building on this key finding, we explore how Second Opinion can provide fast, scalable, and organized feedback to these experts via paid real-time crowds.

We 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 provide 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. We also discuss broader implications for crowd–AI interaction and crowdsourced image analysis.

Related Work

AI-based Face Recognition Recent advances in deep convolutional neural networks (Wen et al. 2016; Sun et al. 2015; Taigman et al. 2014) have led to rapid commercialization of AI-based face recognition technologies, with cloud-based APIs underpinning many real-world applications. For example, Uber has used Microsoft Azure's Face API to verify drivers' identities (Microsoft 2019), and Amazon's Rekognition has been used by law enforcement to identify suspects from surveillance footage (Harwell 2019).

These real-world uses have revealed socio-technical challenges outside the lab that result in poor performance, such as high false positive rates and biased results. In one experiment, Amazon's Rekognition wrongly flagged 28 members of the US House of Representatives as people charged with crimes (Singer 2018). UK police wrongly identified people 92% of the time (2,297 false positives) as potential criminals during the 2017 Champions League final (Press Association 2018). Recent investigations of commercial face recognition algorithms from Microsoft, Face++, IBM, Amazon, and Kairos (Buolamwini and Gebru 2018; Raji and Buolamwini 2019) have shown significant error rates in recognizing women and people with darker skin tones. These problems, coupled with more generalized public anxiety about surveillance and privacy concerns, have pressured companies to suspend sales of face recognition technology (Knight 2018).

Human Cognition and Face Recognition Multiple studies have compared face recognition algorithms to human baselines, with some of them showing human recognition skills as superior (Blanton et al. 2016; Best-Rowden et al. 2014; Kemelmacher-Shlizerman et al. 2016; Zhao et al. 2003). Other research (O'Toole et al. 2007; Valeriani and Poli 2019) suggests that a combination of human and algorithmic decision-making can yield the best results. For example, (Phillips et al. 2018) showed that fusing the scores of a single forensic facial examiner and the best performing algorithm yielded better face identification results than fusing only the scores of multiple algorithms or multiple forensic examiners.

(Abudarham, Shkiller, and Yovel 2019) showed that humans use the same critical feature set to represent both familiar and unfamiliar faces, and that these features are also used by a deep neural network face recognition algorithm. All these studies suggest an improved accuracy in fusing scores from humans and algorithms, motivating the hybrid approach we developed for Second Opinion.

Beyond face recognition, cognitive psychology offers broader theoretical insights into how humans perceive and reason about similarities and differences in visual material. Structure mapping theory (Gentner and Markman 1997) suggests that feature differences are more *alignable*, or salient, in high-similarity pairs than low-similarity pairs (e.g., it is easier to compare two types of trucks, versus comparing a truck to a boat). Further, the *extension effect* suggests that the shared presence of certain unique and significant features of "high-diagnostic value" in two objects increases the similarity between them (Tversky 1977). Building on these ideas, we designed Second Opinion's interfaces to help experts and crowds compare faces by focusing on alignable differences organized by typical facial features, as well as unique similarities of high-diagnostic value.

Crowdsourced Image Analysis Crowdsourcing has been widely used to support visual analysis tasks, from identifying objects (Bigham et al. 2010; Noronha et al. 2011) to analyzing video data (Lasecki et al. 2015; Song et al. 2019). Of particular relevance here, (Kohler, Purviance, and Luther 2017) asked crowds to search satellite imagery for a target area using an expert-drawn aerial diagram. Extending this idea, Second Opinion asks experts to highlight unique facial features to help focus crowd attention in the novel context of person identification. Other recent work explores crowdsourcing feedback on graphic designs (Luther et al. 2015). Similarly, Second Opinion employs both a crowd interface (for gathering structured feedback) and an expert interface (for visualizing aggregated crowd results). Unlike this prior work, we investigate how such crowd systems can help experts search and compare multiple images.

Crowds have been used in conjunction with computer vision to for visual analysis tasks such as annotating bus stops and sidewalk accessibility issues (Hara et al. 2015), analyzing scientific imagery (Su, Sui, and Zhang 2018), describing screenshots for software testing (Liu et al. 2018), and finding lost pets (Barrenechea et al. 2015). Inspired by these efforts, we explore how crowds can augment face recognition in the novel context of last-mile person identification.

Crowdsourcing has proven to be an efficient method for assigning semantic attributes to describe how objects look (Kovashka and Grauman 2015). Flock (Cheng and Bernstein 2015) used crowdsourcing to nominate features and labels to train hybrid crowd-machine learning classifiers. Tropel (Patterson et al. 2015) showed that limited examples can be used for crowd workers to annotate additional examples and create visual classifiers. We show that Second Opinion allows crowds to perform fine-grained facial analysis, generating tags that could be useful for training machine learning approaches.

A few studies have explored crowdsourced image analysis in the context of person identification. (Lasecki et al. 2015) investigated using blurring effects to de-identify people in videos analyzed by crowd workers; in the baseline case (without blurring), 90.9% of workers correctly identified a person they had seen earlier from a photo lineup of 5 candidates. In our prior work (Mohanty et al. 2019), we developed Civil War Photo Sleuth (CWPS)[1], a free, public website that combines crowdsourcing and AI-based face recognition to identify unknown soldiers in photos from the

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

American Civil War era. CWPS allows users to upload an unidentified soldier photo, filters results by military service, and uses Microsoft Face API to return a shortlist of potential matches with high facial similarity from a database of over 25,000 identified reference photos.

CWPS does not assist with last-mile person identification, though (Mohanty et al. 2019) observed "many examples of users posting screenshots of potential matches on social media and requesting feedback from fellow history enthusiasts," demonstrating emergent needs for second opinions. Building on this foundation, we use CWPS to generate a shortlist of 5 high-similarity candidates for each mystery photo, and evaluate how well Second Opinion helps crowds and experts select the correct matches among these.

System Description

Second Opinion is a web-based software tool for solving the last mile problem of person identification. It is designed to assist expert users in seeking a "second opinion" from online crowd workers regarding potential matches among very similar-looking photos. Then, users perform a fine-grained analysis of facial similarity for the candidates and finally, decide on the correct match.

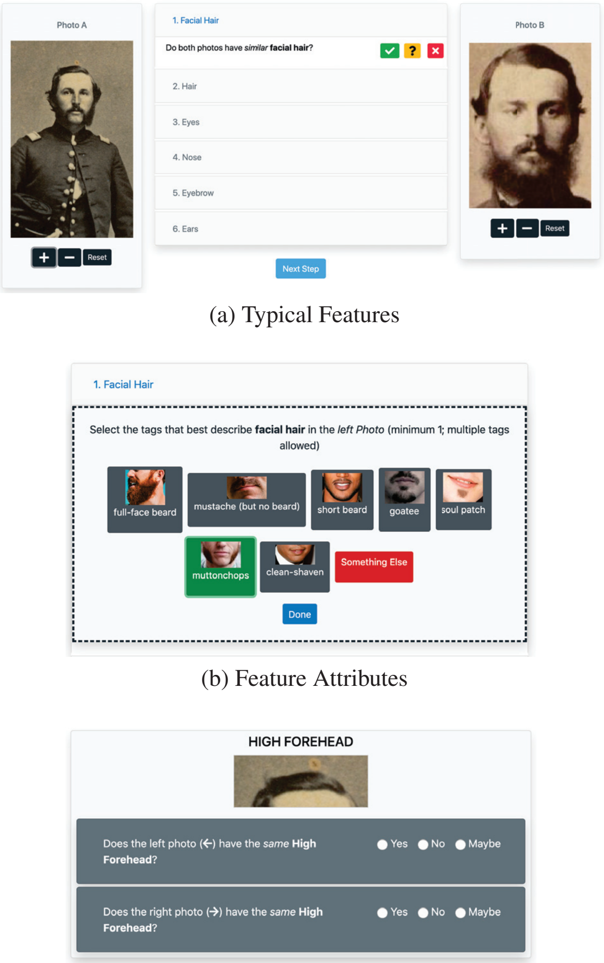
Given a "mystery photo" of an unknown person and the task of correctly matching the person to identified photos of very similar-looking people, we built a workflow that supports collaboration between experts and crowds in near real-time. Second Opinion broadly consists of two separate interfaces, a Crowd Interface and an Expert Interface. The expert initially selects some unique identifying features of the person in the mystery photo. Each crowd worker then compares a similar-looking photo with the mystery photo and answers some questions related to features selected by the expert and the system. Upon receiving responses for all the candidate photos, the system aggregates these responses and represents them using different visualizations. The expert analyzes crowd responses to make a decision about the correct matching photo from the search set. Given our focus on collaboration between experts and crowds, and to avoid biasing their decisions, Second Opinion does not display the face recognition similarity scores.

We describe Second Opinion's novel crowd workflow in three phases: *seed*, *gather*, and *analyze*.

Phase 1: Seeding

As discussed above, unique features of high-diagnostic value increase similarity between objects. In the seeding phase, the Expert Interface instructs the expert to identify several of these unique features, such as a *scarred chin, strong jawline, eye patch*, etc. For example, a distinctive birthmark on the mystery person's left cheek lets the expert quickly rule out any candidates lacking it.

Using a feature selection interface, the expert selects 1–3 unique facial features of the mystery person that would make it easy to identify him or her. The expert marks the feature by cropping the relevant region on the person's face and assigning a short descriptive label to it.



Do both photos have *similar* **facial hair**?

(a) Typical Features

Select the tags that best describe **facial hair** in the *left Photo* (minimum 1; multiple tags allowed)

(b) Feature Attributes

Does the left photo (←) have the *same* **High Forehead**? ○ Yes ○ No ○ Maybe

Does the right photo (→) have the *same* **High Forehead**? ○ Yes ○ No ○ Maybe

(c) Unique Features

Figure 1: Crowd Interface.

Phase 2: Gathering

Once the expert finishes seeding the unique features, the system gathers annotations from crowd workers regarding the presence or absence of these and other identifying features. To create a near real-time experience for the expert, the system recruits workers via LegionTools (Gordon, Bigham, and Lasecki 2015), a toolkit that enables real-time recruiting and routing of large numbers of crowd workers from Amazon Mechanical Turk. We implemented a quality control measure where workers must correctly annotate a facial feature with a known gold-standard answer to proceed to the Crowd Interface.

In the Crowd Interface, the crowd workers compare the mystery person to a similar-looking candidate on the basis of facial similarity. The interface shows the mystery person's photo on the left, the candidate being compared on the right, and an interactive annotation area in the center. The annotation area has three main sections, which the crowd workers attend to sequentially.

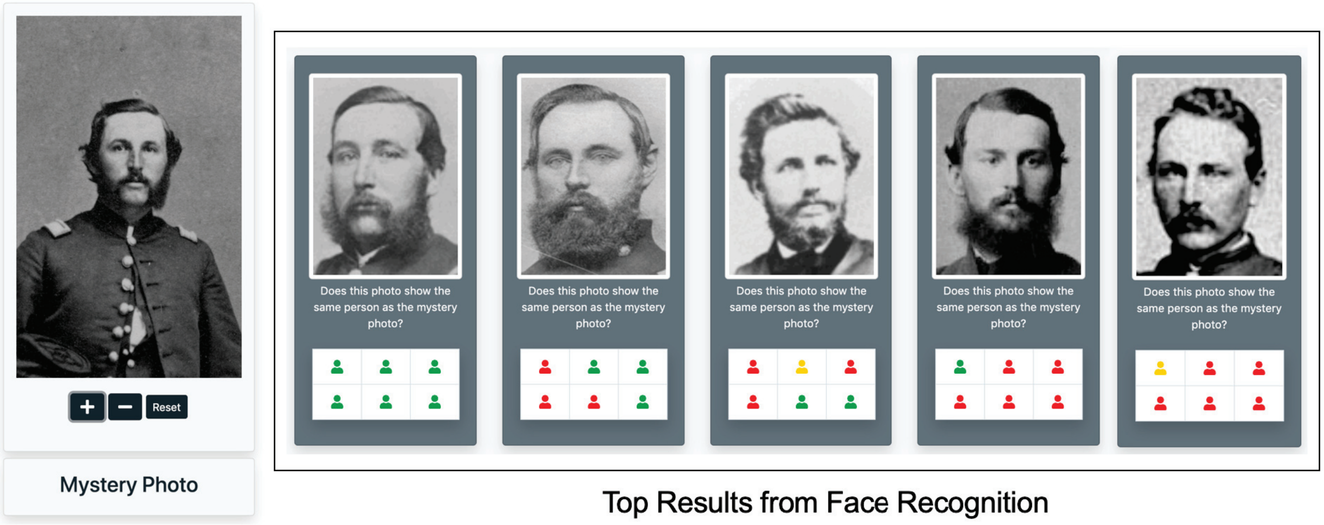

+
−
Reset
Mystery Photo
Does this photo show the same person as the mystery photo?
Does this photo show the same person as the mystery photo?
Does this photo show the same person as the mystery photo?
Does this photo show the same person as the mystery photo?
Does this photo show the same person as the mystery photo?
Top Results from Face Recognition

Figure 2: Expert Interface: Overview Page.

Typical Features Building on the idea that alignable differences become especially salient in a similar-looking candidate pool, the system aims to capture differences for a predetermined set of facial features which are typical to face verification tasks. These typical features include *facial hair, head hair, eyebrows, eyes, nose,* and *ears*, a subset of the generalized visual attributes used by (Kumar et al. 2011). Crowd workers compare whether each of these features is similar or different in two photos (see Figure 1a) and provide a *similar*, *different*, or *don't know* response.

If the worker chooses *different*, the interface asks the worker to toggle one or more specific feature attributes describing the differences from a pre-populated menu of choices (see Figure 1b). For example, a worker might describe the head hair as *curly* in one photo and *straight* in the other. Each attribute is illustrated by a modern-day full-color example photo, collected and verified by the authors. We employed this design choice of visual cues (instead of a text-only description) and toggle buttons (instead of free-flow text input) to streamline the rapid gathering of responses from the crowd.

Unique Features Since unique features of high-diagnostic value increase similarity between objects, the system aims to capture whether the unique identifying features of the mystery photo are present in the candidate photo. The Crowd Interface displays the associated cropped image and the label for all the expert-nominated unique feature(s) from the seeding phase (see Figure 1c). For each unique feature, the workers then compare whether both photos have the same unique facial feature using a 5-point Likert scale (-2 = Definitely different, 2 = Definitely same).

Overall Similarity Finally, the workers answer whether they believe both photos show the same person or not using a 5-point Likert scale (-2 = Definitely different, 2 = Definitely same). This is the overall similarity score for the photo being compared to the mystery photo.

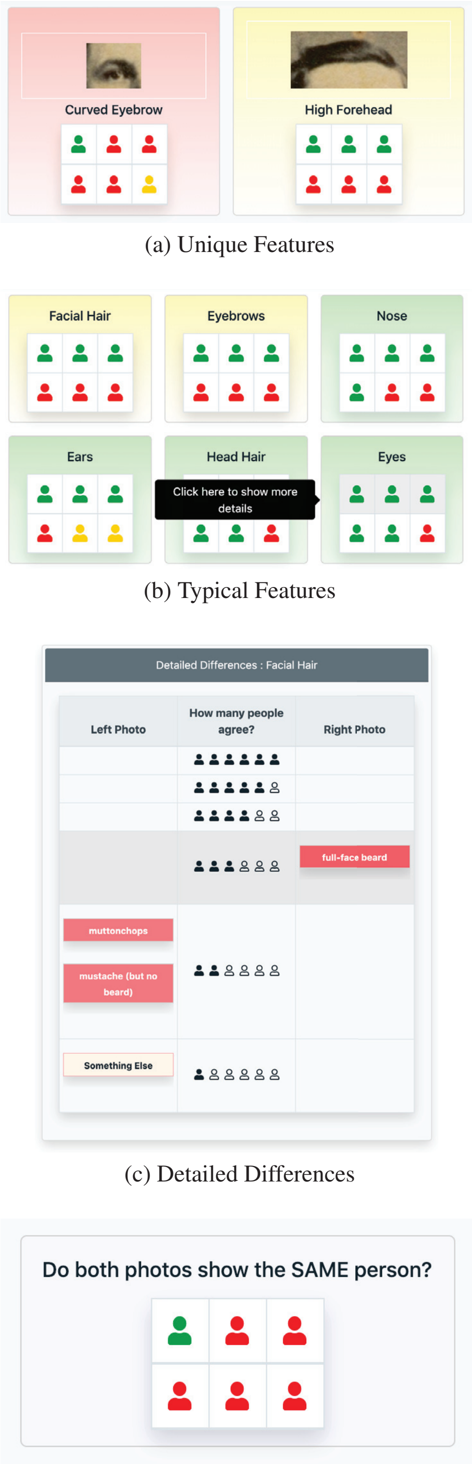
Phase 3: Analyzing

Once the system gathers all the crowd responses, it represents them as visualizations in the Expert Interface for the expert to analyze. The expert first sees an overview page (see Figure 2) where the top candidates are sorted by overall similarity, i.e., the mean aggregate of the overall similarity scores of each photo by multiple workers. Each crowd worker's overall decision is shown using a colored person-shaped icon, with red indicating different, green indicating same, and yellow indicating an undecided worker. The system uses this same color key everywhere in the Expert Interface.

The expert can perform a fine-grained facial similarity analysis for any one of the photos by clicking the photo, which takes them to the details page for that photo (see Figure 3). The layout is analogous to the Crowd Interface, with the mystery photo on the left and the candidate on the right. The middle section has three different visualizations — unique features, typical features, and overall similarity — arranged top to bottom.

Unique Features The expert can visualize the crowd responses for each unique feature that was seeded into the system initially (see Figure 3a). The person-shaped colored icons below each feature show many workers voted that the similar-looking photo has the same unique feature as the mystery photo, or a different one.

Typical Features Similarly, the expert can analyze visualizations of the crowd responses for each of the six typical features predetermined by the system (see Figure 3b). If any crowd responses suggest differences for a typical feature, the expert can click the feature to reveal a details table showing differences in the feature attributes (see Figure 3c). The details table has three columns: the left column showing



(a) Unique Features

(b) Typical Features

(c) Detailed Differences

(d) Overall Similarity

Figure 3: Expert Interface: Details Page. These visualizations represent the fourth candidate photo (from left) in Figure 2.

the attributes that describe the typical feature in the mystery photo (left), the right column showing the attributes that describe the typical feature in the similar-looking photo (right), and the middle column showing worker agreement for each attribute, sorted highest to lowest.

Overall Similarity Finally, the expert can use the overall similarity visualization to see how many crowd workers thought both photos showed the same person (see Figure 3d). This is the same visualization shown for each candidate on the overview page.

Study Setup

We conducted an exploratory, mixed-methods evaluation study of Second Opinion to understand how well crowds could augment automated face recognition techniques, as well as how experts would respond to this type of crowd feedback.

Photo Dataset

Because all datasets are biased, and the composition of the reference dataset affects the performance of the face recognition results, we sought to evaluate Second Opinion using a diverse and representative test dataset. We intentionally selected test photos with respect to several criteria known to be affected by sampling bias and/or face recognition challenges:

- **Army** Due to a naval blockade early in the Civil War, Southern photographers were unable to get regular access to development supplies, so fewer Confederate portraits exist compared to Union portraits (Coddington 2008). The CWPS database currently has 17,430 Union and 1,889 Confederate identified soldiers.
- **Rank** Higher-ranked soldiers (i.e., commissioned officers) of both armies were more likely to be photographed, and photographed more often, than lower-ranked soldiers (i.e., enlisted soldiers or non-commissioned officers) (Zeller 2019). Further, some evidence suggests higher-ranked soldiers had more elaborate and diverse facial hair styles (Adolphus 2013). There are currently 11,925 identified low-ranked soldiers and 10,896 high-ranked ones in the CWPS database.
- **Race** Only 10% of the Union Army was black, and few portraits of black Union soldiers survive today (Coddington 2012). Further, as mentioned above, face recognition struggles to identify non-white faces in modern photos (Buolamwini and Gebru 2018). The CWPS database currently has only 584 identified soldiers from the US Colored Troops, some of whom are white officers.

We selected 5 photos with verified identities across these criteria: a high-ranked white Union officer (UH), a low-ranked white Union soldier (UL), a low-ranked black Union soldier (AF), a high-ranked white Confederate officer (CH), and a low-ranked white Confederate soldier (CL).[2] We refer to these 5 photos as *mystery photos*, as our expert participants will try to identify them.

[2] Black soldiers were generally excluded from high ranks in the Union Army. There were no black soldiers in the Confederate Army.

We searched these mystery photos on CWPS. For each, we gathered the top 4 most similar (incorrect) results from the same army as the mystery soldier, plus a different photo of the mystery soldier. Therefore, for each of the 5 mystery photos, we had 4 incorrect matches (distractors) and 1 correct match. Finally, we cropped the photos to only show the heads.

Participants and Procedure

Experts For expert users, we recruited 10 participants with extensive prior expertise in Civil War photo identification (mean = 15 years, min = 2, max = 40) from different online communities like Facebook groups and CWPS. All 10 participants were white males and the average age was 41 (min = 20, max = 69). We anonymize the experts with identifiers E1–E10.

We randomly assigned 2 experts per mystery photo, while verifying that the experts had not been exposed to their assigned mystery photo before. We identify each user session by appending 1 or 2 to the mystery photo identifier, e.g., UH_1, UH_2, UL_1, UL_2, etc. We achieved theoretical saturation in collecting new insights after 10 user sessions.

Each expert began the study by completing a consent form and pre-survey describing their demographics and Civil War photography experience. The expert then saw all 6 photos (mystery photo, matching photo, and 4 distractors), unlabeled and in random order, and completed a survey (A) about their initial impressions of which photo(s) matched the mystery photo. We told experts that there could be any number of matches, from 0 to all 5.

Next, the expert used Second Opinion to provide 1–3 unique identifying features for the mystery photo. While waiting for the crowd responses, the expert answered several semi-structured interview questions about how they traditionally decided among close matches in Civil War photo identification.

Once all crowd responses were received, the expert proceeded to the overview page. The expert completed the walk-through tutorial and a survey (B) about the overview page. Next, the expert was invited to click on any of the candidates to view the details page. The expert completed another walk-through tutorial of the details page, reviewed the results, and completed a survey (C1–C5) about the details page. The expert then repeated this step for the remaining 4 candidates.

Finally, the expert completed a survey (D) and answered some semi-structured interview questions about their final thoughts on which photo(s) matched the mystery photo, as well as their overall impressions of Second Opinion. The researchers did not provide help with the system or task aside from the built-in tutorials. The entire study was completed online via Zoom video conferencing, which recorded the expert's webcam video and audio, as well as their shared computer screen. The audio recordings were fully transcribed.

Crowd Workers We recruited crowd workers on Amazon Mechanical Turk. Each worker analyzed 1 photo pair using the Crowd Interface, followed by a demographics post-survey. The system recruited workers without any location or qualification constraints and paid them between $1.20 and $2.00 per task, which included the waiting time (where they were free to do other tasks), as well as the Second Opinion task (which pilot tests showed took around 5 minutes to complete). To avoid any learning effects or collusion, the Second Opinion system randomly assigned each worker to only 1 pair of photos. Second Opinion hired 6 workers for each photo pair. Since each session had 5 photo pairs, we hired 30 unique workers per session, totalling of 300 workers across all 10 user sessions.

Data Analysis

We analyzed qualitative data (interview transcripts, open-ended survey responses, observation notes) with respect to the guiding themes in our research questions. We analyzed quantitative data (system logs, Likert-scale survey responses) using statistical software.

Findings

AI Face Recognition Baseline Performance

Face recognition is unable to consistently address the last-mile problem by itself. To reduce false positives, CWPS discards any results for which Microsoft Face API gives a similarity score lower than 0.50 (out of 1.0), so in all 10 cases, both correct and incorrect matches scored ≥ 0.50. Reviewing the similarity scores for our test dataset, we observed that automated face recognition assigned its highest similarity score to the correct match for only 3 of the 5 mystery photos (i.e., 6 of the 10 sessions).

We further observed that face recognition had assigned very similar confidence scores to the incorrect matches for all sessions, with a small average delta of 0.012 between adjacently ranked incorrect matches. The average delta between the top 2 highest-ranked photos was larger, at 0.080. Despite this, there was no universal threshold to consistently separate the correct match from the incorrect ones.

For example, in sessions CH_1 and CH_2, face recognition assigned the following similarity scores: Photo A (0.614), Photo B (0.609), Photo C (0.601), Photo D (0.601) and Photo E (0.597). Thus, 2 incorrect photos (A and B) scored highest, whereas the correct match (Photo C) tied with an incorrect photo (Photo D) for third place. Further, the most- and least-similar photos were separated by a score of only 0.007. The CH sessions illustrate face recognition's inability to solve the last-mile problem by itself.

Expert Performance and Attitudes

Experts primarily used Second Opinion to validate their original decisions. In interviews, experts emphasized how Second Opinion helped them feel more confident in their initial impressions. In some cases, experts felt more strongly that their proposed identifications were correct; E4 said that the system "reaffirmed that who I thought it was is who I think it is." In other cases, experts felt better about their reasons for having instinctually ruled out a candidate. E7 said, "The process I used for identifying that person ultimately gave me more confidence because it forced me to rule out some people in a more systematic way than just saying, 'I don't think it looks like that person."'

The experts also quantitatively reported how easy it was to identify the images before and after using the Second Opinion system on a 5-point Likert scale. Experts found the mystery photos moderately easy to identify from the start (M = 3.70). After using Second Opinion, experts found the photos very easy to identify (M = 4.50). Six of the 10 experts increased their scores, indicating that Second Opinion made the identification task easier for them.

These subjective experiences align with expert performance results. Experts performed well in guessing the identifications before using Second Opinion, with a recall of 100% and precision of 83.33%. These results underscore the expertise of our users, as well as the possibility of a ceiling effect with our test dataset. Further, no expert changed their decision after using Second Opinion. While the system did not detract from the already-high performance, it did not close the gap, either.

Experts saw advantages of Second Opinion over their traditional last-mile methods. In general, the experts found both the overview page (M = 4.89) and details page (M = 4.90) highly informative. The experts also found the visualizations for the typical features (M = 4.80), unique features (M = 4.90), and the overall similarity score (M = 5.0) very easy to understand.

All 10 experts mentioned that they would prefer to use a system like Second Opinion for future identifications whenever they are unable to make a decision. In E3's words, " I can't think of any situations I wouldn't use it in."

The experts also emphasized the value of Second Opinion in structuring and organizing feedback from others, especially in contrast to current practices like Facebook comments, which some considered vague and unstructured. In the words of E2:

> I've been using social media and forums to do exactly what we did today, but when you use the hive mind that's Facebook or an internet forum, it's just a far less organized way to dissect the reactions and opinion of others. When you have it formatted in this way, I think it's much more useful.

Experts said that Second Opinion's crowd responses made them notice things they would have otherwise missed. E7 said, "It's useful to see the majority of people whose opinions correspond with yours and it's useful to see the one person who forces you to re-examine some of those assumptions." E10 emphasized that the system helps "open up brand new information (you) could easily have overlooked."

Experts had mixed opinions about the importance of the crowd's speed and composition. Most experts received all crowd responses within 5–20 minutes. Even though we optimized Second Opinion for a near real-time collaboration, few experts expressed a strong desire for fast crowd responses; 24 hours was often mentioned as a reasonable time frame. Instead, many emphasized giving the crowd sufficient time to perform the task well. In E9's words, "I would rather have accuracy than speed."

	Individual		Aggregate		
	#	Total	Std. Mean	Wgt. Mean	Total Cases
True Positives	52	60	10	10	10
False Positives	138	240	25	9	40
True Negatives	91	240	11	30	40
False Negatives	5	60	0	0	10

Table 1: Individual and aggregate (standard and weighted) crowd performance for *overall similarity* across all expert sessions.

	False Positives		True Negatives	
	Standard Mean	Weighted Mean	Standard Mean	Weighted Mean
UH_1	1	0	3	4
UH_2	3	0	1	4
UL_1	2	1	2	3
UL_2	3	0	0	3
CH_1	2	0	1	4
CH_2	4	3	0	1
CL_1	2	1	2	3
CL_2	2	2	0	2
AF_1	3	1	1	3
AF_2	3	1	1	3
Total	**25**	**9**	**11**	**30**

Table 2: Aggregate (standard and weighted) crowd performance for ruling out incorrect matches per expert session.

Some experts, however, mentioned specific circumstances where they would require a fast decision. E5 said that "time is of the essence" when deciding whether to bid on unidentified Civil War portraits in online auctions, while E1 noted, "Sometimes I'll be in an antique store and I'll come across a photograph that I think might be a Civil War soldier and I want to run it through the database real fast."

Experts diverged somewhat on how much expertise Second Opinion's crowd should have, echoing the diverse groups they currently consult. E2 preferred a crowd of fellow experts who are "more aware of the photograph process at that time and the variation in which different features and colors can manifest themselves." E3 valued the fresh perspective offered by outsiders: "We might be focusing on certain things that someone who knows nothing about the Civil War might just say, 'Hey, look at this.' And, 'Why I didn't see that before?"' However, most experts argued for a mixture of both experts and novices, which could offer the best of both worlds.

Crowd Performance

Individual workers demonstrate high recall but low precision for overall similarity. Table 1 shows that out of 60 crowd workers who were shown a correct match, 52 correctly identified the correct match, for a recall rate of 86.67%. Five out of these 60 workers misidentified the correct match, while 3 were undecided. However, out of the 240 crowd workers who were shown an incorrect match, only 91 workers correctly filtered them out, for a precision of 37.68%.

Weighted mean aggregation of crowd judgements can substantially increase both precision and recall. Given the high recall but relatively low precision of the individualized crowd responses, we conducted post-hoc analyses to consider whether aggregated crowd responses would yield better precision. We used two forms of aggregation for this analysis: standard mean and weighted mean. For the standard mean approach, we computed the average of all worker scores. For the weighted mean approach, due to the lower precision of individual crowd workers, we assigned a higher weight of 4.0 to the negative worker judgements and a lower weight of 1.0 to positive worker judgements.

When we applied the standard mean approach to the crowd judgements from the study, recall improved to 100%, while precision dropped to 28.57%. This still indicates a high number of false positives, at 25. However, when we applied the weighted mean, the crowd's recall again improved to 100%, and precision increased to 52.63%. The false positive cases dropped from 25 in the standard mean condition, to only 9 in the weighted mean condition. From Table 2 and Table 1, we observe that the weighted mean approach ruled out 75% of the false positive cases (30 out of 40) by assigning them a negative score.

While weighted aggregation substantially improved both precision and recall compared to the method shown to experts, it had a slightly detrimental effect for rankings. Using the standard mean aggregation, the crowd assigned the highest score to the correct match in 8 out of 10 sessions, while in the weighted mean case, the correct match had the highest score in 7 out of 10 sessions. Both approaches are improvements over the AI baseline, which ranked the correct photo first in 6 out of 10 sessions.

In our study, the visualizations on the details page of the Expert Interface showed the individual worker decisions, while the standard mean approach was used to sort (rank) the photos on the overview page. All 10 experts evaluated the overall crowd performance for every photo pair comparison using a 5-point Likert scale (1 = Definitely Incorrect, 5 = Definitely Correct). Experts reported a moderately high subjective impression of crowd accuracy (M = 3.7), even without experiencing the weighted aggregate.

Crowd workers are generally good at analyzing facial features. Out of 150 cases, we observed that crowd workers answered correctly the gold-standard question about the expert-nominated unique features in 120 cases, while answering incorrectly in 18 cases or "don't know" in 12 cases, for an accuracy of 80%.

Experts also evaluated the quality of the crowd responses, reporting moderately high scores with respect to unique features (M = 4.0) and typical features (M = 3.92).

Crowd workers are better at identifying faces of their same race. In sessions AF_1 and AF_2, the mystery photo was a black Union soldier. CWPS's face recognition returned a white soldier as the second most similar-looking photo, along with 4 black soldiers. In session AF_1, crowd workers completely rejected the white soldier, giving it the lowest score of the 5 candidates. However, the same behavior was not seen in session AF_2, in which the white soldier ranked second.

Because prior work suggests people are worse at identifying people of other races (the so-called "cross-race effect" (Meissner and Brigham 2001)), we conducted a follow-up analysis investigating how workers' self-identified race could be related to their performance in identifying mystery photos of various races. We found that workers correctly identify or reject candidate photos of their own race more accurately (58.4%) than candidates of other races (41%), and the difference is significant ($p < 0.01$, two-tailed Fisher's exact test). This finding suggests that the cross-race effect also applies to crowdsourced person identification.

Discussion

Combining Strengths of Crowds, Experts, and AI Our findings supported prior work showing that AI-based face recognition is inconsistent in picking the correct match and prone to false positives, making it unsuitable for solving the last-mile problem on its own. The strengths of face recognition may be best applied in rapidly narrowing down a large candidate pool to a shortlist of similar-looking candidates. Crowdsourcing, on the other hand, offers an effective but expensive approach for narrowing down a large candidate pool. Our evaluation showed that aggregating crowd scores with the weighted mean approach could filter out 75% of face recognition's false positives, while still maintaining a 100% recall rate. This approach also modestly improved upon face recognition's rankings.

(Gray and Suri 2019) argue that crowdsourced human insight and creativity remain essential to advancing AI technologies like face recognition, a phenomenon they call "the paradox of automation's last mile." Thus, improving crowd–AI collaboration is an important research challenge for the foreseeable future. While a crowd–AI hybrid would likely outperform either approach by itself, our findings suggest that using a workflow like seed-gather-analyze to augment expert performance may be the most effective and responsible path forward. Experts bring considerable skill to person identification tasks, AI can substantially narrow down possibilities, and crowds can further reduce false positives and help experts notice details they may have otherwise missed. Given the high-stakes decisions that frequently result from person identification (e.g., making an arrest), assuming and integrating expert participation from the beginning can improve accountability and accuracy. Maximizing these benefits requires experts to carefully consider the hybrid results and be open to changing their minds, challenges we discuss below.

Influencing Expert Decisions Experts were enthusiastic about Second Opinion, but it did not change their decisions, partly because the photos may have been too easy, and experts already picked the right ones. This outcome is better than the system causing experts to change initial good decisions into bad ones, but also made it difficult for us to assess the impact of the system on experts.

One way to assess the impact of Second Opinion could be to make the testing conditions more realistic for the experts. In the real-world scenario offered by CWPS's workflow, users are often presented with a significantly larger pool of search results than only five photos, which may or may not include a correct match. This makes the real-world scenario more difficult than the current study conditions. Under these conditions, Second Opinion has the potential to more strongly influence the expert's decisions.

Furthermore, prior work (Mohanty et al. 2019) shows that CWPS users sometimes match photos which do not show in the top-50 search results, indicating that top-5 search results from face recognition might not necessarily be an expert's top-5 choices. Allowing experts the freedom to select their own shortlist from the face recognition results for Second Opinion analysis may feel more natural than the study setup in this paper. This approach also has the potential to capture the expert's initial confusion in selecting the match(es) and contrast it with the final decision after using Second Opinion.

Our post-hoc analysis showed that weighted mean aggregation of crowd judgments substantially improved both precision and recall. However, the visualizations for the experts did not reflect these scores. We believe that integrating these weighted scores in the overview page and details page would better assist the expert in filtering out the false positives. This change would also enable measuring the impact of Second Opinion on expert decisions in a more targeted manner.

Additionally, most experts believed the ideal Second Opinion crowd would be a mix of experts and novices. Experts may be more likely to change their minds if they are aware of each crowd worker's expertise and can contextualize their feedback accordingly. CWPS is an open online community, and both novices and experts with a shared interest in Civil War photos are part of the user base. Integrating Second Opinion with CWPS would create a feedback channel from within the community that could encourage experts to pay closer attention, but more work is needed to understand the most effective ways to assemble these crowds and represent their expertise.

Finally, confirmation bias is a powerful force that even seasoned investigators struggle to overcome (White 2010). Tools like Second Opinion must be part of a broader commitment to evidence-based conclusions.

Considering Race to Improve Accuracy Sessions AF_1 and AF_2, involving the mystery photo of the black Union soldier, suggested how human and AI-based face recognition process visual cues regarding race differently. The AI approach, which focuses on ratios of facial landmarks, ranked a white soldier highly among the results. Crowd workers, who may have considered additional features such as skin tone and hair texture, ranked the white soldier lowest in one session. Which method is most effective — or appropriate — likely depends on the context. During the Civil War, racial categorization was often based on physical characteristics, and military units were racially segregated, so the crowd's inclusion of additional criteria may simulate problematic historical practices that nevertheless effectively narrow down search results.

We also found that crowd workers identified photos significantly less accurately when the person in the photo was of a different race, consistent with prior research on the cross-race effect (Meissner and Brigham 2001). One implication of this finding would be to improve accuracy by recruiting a more racially diverse crowd (Barbosa and Chen 2019) and seek to align workers with same-race photos. Alternatively, it may be possible to reduce the cross-race effect by forewarning workers about it (Hugenberg, Miller, and Claypool 2007).

Conclusion

Identifying people in photographs is an important but challenging task across many domains. While experts increasingly make use of AI-based face recognition to narrow down possibilities, they must still solve the last-mile problem of selecting the correct match(es) among a shortlist of high-similarity. We present and evaluate Second Opinion, a software tool that augments face recognition and expert practice with crowdsourced human intelligence.

Our contributions include: (1) the seed-gather-analyze crowd workflow, inspired by cognitive psychology, for helping experts and crowds focus on key facial similarities and differences; (2) the Second Opinion software tool demonstrating this workflow; and (3) a mixed-methods evaluation with 10 experts and 300 crowd workers showing that crowds can reduce face recognition's false positives by 75%, and that experts felt enthusiasm for the system over their current practices. Our work opens doors for exploring how experts, crowds, and AI can collaborate on complex image analysis tasks and other last-mile problems.

Acknowledgements

This research was supported by NSF IIS-1651969 and IIS-1527453 and a Virginia Tech ICTAS Junior Faculty Award.

References

Abudarham, N.; Shkiller, L.; and Yovel, G. 2019. Critical features for face recognition. *Cognition* 182:73–83.

Adolphus, F. 2013. The Confederate Soldier at Fort Mahone, Battle of Petersburg, April 2, 1865 http://adolphusconfederateuniforms.com/the-confederate-soldier-of-fort-mahone.html. *Adolphus Confederate Uniforms*.

Barbosa, N. a. M., and Chen, M. 2019. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 543:1–543:12. New York, NY, USA: ACM.

Barrenechea, M.; Anderson, K. M.; Palen, L.; and White, J. 2015. Engineering crowdwork for disaster events: The human-centered development of a lost-and-found tasking environment. In *2015 48th Hawaii International Conference on System Sciences*, 182–191.

Best-Rowden, L.; Bisht, S.; Klontz, J. C.; and Jain, A. K. 2014. Unconstrained face recognition: Establishing baseline human performance via crowdsourcing. In *IEEE International Joint Conference on Biometrics*, 1–8.

Bigham, J. P.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R. C.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 333–342. ACM.

Blanton, A.; Allen, K. C.; Miller, T.; Kalka, N. D.; and Jain, A. K. 2016. A comparison of human and automated face verification accuracy on unconstrained image sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 161–168.

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.

Cheng, J., and Bernstein, M. S. 2015. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, 600–611. ACM.

Coddington, R. S. 2008. *Faces of the Confederacy an album of Southern soldiers and their stories*. Johns Hopkins University Press.

Coddington, R. S. 2012. *African American faces of the Civil War: an album*. The Johns Hopkins University Press.

Fortin, J. 2018. She Was the Only Woman in a Photo of 38 Scientists, and Now She's Been Identified https://www.nytimes.com/2018/03/19/us/twitter-mystery-photo.html. *The New York Times*.

Garvie, C.; Bedoya, A.; and Frankle, J. 2016. Unregulated Police Face Recognition in America https://www.perpetuallineup.org/. Technical report, Georgetown Law Center on Privacy & Technology.

Gentner, D., and Markman, A. B. 1997. Structure mapping in analogy and similarity. *American psychologist* 52(1):45.

Gordon, M.; Bigham, J. P.; and Lasecki, W. S. 2015. Legiontools: a toolkit+ ui for recruiting and routing crowds to synchronous real-time tasks. In *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, 81–82. ACM.

Gray, M. L., and Suri, S. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.

Haghighat, M., and Abdel-Mottaleb, M. 2017. Low resolution face recognition in surveillance systems using discriminant correlation analysis. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, 912–917.

Hara, K.; Azenkot, S.; Campbell, M.; Bennett, C. L.; Le, V.; Pannella, S.; Moore, R.; Minckler, K.; Ng, R. H.; and Froehlich, J. E. 2015. Improving public transit accessibility for blind riders by crowdsourcing bus stop landmark locations with google street view: An extended analysis. *ACM Transactions on Accessible Computing (TACCESS)* 6(2):5.

Harwell, D. 2019. Oregon became a testing ground for amazon's facial-recognition policing. but what if rekognition gets it wrong? https://www.washingtonpost.com/technology/2019/04/30/amazons-facial-recognition-technology-is-supercharging-local-police/. *Washington Post*.

Hugenberg, K.; Miller, J.; and Claypool, H. M. 2007. Categorization and individuation in the cross-race recognition deficit: Toward a solution to an insidious problem. *Journal of Experimental Social Psychology* 43(2):334–340.

Human Rights Watch. 2017. Libya: Videos Capture Summary Executions https://www.hrw.org/news/2017/08/16/libya-videos-capture-summary-executions#.

Keefe, P. R. 2016. The Detectives Who Never Forget a Face https://www.newyorker.com/magazine/2016/08/22/londons-super-recognizer-police-force. *New Yorker*.

Kemelmacher-Shlizerman, I.; Seitz, S. M.; Miller, D.; and Brossard, E. 2016. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4873–4882.

Klare, B. F.; Burge, M. J.; Klontz, J. C.; Bruegge, R. W. V.; and Jain, A. K. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7(6):1789–1801.

Knight, W. 2018. Google says it won't sell face recognition for now—but it will be hard to slow its use https://www.technologyreview.com/f/612606/google-will-stop-providing-face-recognition-but-it-will-be-hard-to-curb-its-use/. *MIT Technology Review*.

Kohler, R.; Purviance, J.; and Luther, K. 2017. Supporting image geolocation with diagramming and crowdsourcing. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing*, HCOMP '17, 98–107. AAAI Press.

Kovashka, A., and Grauman, K. 2015. Discovering attribute shades of meaning with the crowd. *International Journal of Computer Vision* 114(1):56–73.

Kumar, N.; Berg, A.; Belhumeur, P. N.; and Nayar, S. 2011. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(10):1962–1977.

Lasecki, W. S.; Gordon, M.; Leung, W.; Lim, E.; Bigham, J. P.; and Dow, S. P. 2015. Exploring privacy and accuracy trade-offs in crowdsourced behavioral video coding. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1945–1954. ACM.

Liu, D.; Zhang, X.; Feng, Y.; and Jones, J. A. 2018. Generating descriptions for screenshots to assist crowdsourced testing. In *2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 492–496. IEEE.

Luther, K.; Tolentino, J.-L.; Wu, W.; Pavel, A.; Bailey, B. P.; Agrawala, M.; Hartmann, B.; and Dow, S. P. 2015. Structuring, aggregating, and evaluating crowdsourced design critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, 473–485. ACM.

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

Microsoft. 2019. Uber boosts platform security with the face api, part of microsoft cognitive services https://customers.microsoft.com/en-us/story/uber.

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

Noronha, J.; Hysen, E.; Zhang, H.; and Gajos, K. Z. 2011. Platemate: crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 1–12. ACM.

O'Toole, A. J.; Abdi, H.; Jiang, F.; and Phillips, P. J. 2007. Fusing face-verification algorithms and humans. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 37(5):1149–1155.

Patterson, G.; Van Horn, G.; Belongie, S.; Perona, P.; and Hays, J. 2015. Tropel: Crowdsourcing detectors with minimal training. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

Phillips, P. J.; Yates, A. N.; Hu, Y.; Hahn, C. A.; Noyes, E.; Jackson, K.; Cavazos, J. G.; Jeckeln, G.; Ranjan, R.; Sankaranarayanan, S.; et al. 2018. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences* 115(24):6171–6176.

Pontin, M. W. 2007. Better face-recognition software https://www.technologyreview.com/s/407976/better-face-recognition-software/. *MIT Technology Review*.

Press Association. 2018. Welsh police wrongly identify thousands as potential criminals https://www.theguardian.com/uk-news/2018/may/05/welsh-police-wrongly-identify-thousands-as-potential-criminals. *The Guardian*.

Raji, I. D., and Buolamwini, J. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products.

Singer, N. 2018. Amazon's facial recognition wrongly identifies 28 lawmakers, a.c.l.u. says https://www.nytimes.com/2018/07/26/technology/amazon-aclu-facial-recognition-congress.html. *The New York Times*.

Song, J. Y.; Lemmer, S. J.; Liu, M. X.; Yan, S.; Kim, J.; Corso, J. J.; and Lasecki, W. S. 2019. Popup: Reconstructing 3d video using particle filtering to aggregate crowd responses. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, 558–569. ACM. event-place: Marina del Ray, California.

Su, W.; Sui, D.; and Zhang, X. 2018. Satellite image analysis using crowdsourcing data for collaborative mapping: current and opportunities. *International Journal of Digital Earth* 1–16.

Sun, Y.; Liang, D.; Wang, X.; and Tang, X. 2015. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*.

Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1708.

Tversky, A. 1977. Features of similarity. *Psychological review* 84(4):327.

Valeriani, D., and Poli, R. 2019. Cyborg groups enhance face recognition in crowded environments. *PloS one* 14(3):e0212935.

Vozzella, L., and Morrison, J. 2019. Investigators could not determine if Virginia Gov. Ralph Northam is in racist yearbook photo https://www.washingtonpost.com/local/virginia-politics/investigators-could-not-determine-if-virginia-gov-ralph-northam-is-in-racist-yearbook-photo/2019/05/22/c53e7362-7bec-11e9-a5b3-34f3edf1351e_story.html?utm_term=.c5a2b3d36362. *Washington Post*.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Computer Vision – ECCV 2016*, 499–515. Springer International Publishing.

White, D.; Norell, K.; Phillips, P. J.; and O'Toole, A. J. 2017. Human factors in forensic face identification. In *Handbook of biometrics for forensic science*. Springer. 195–218.

White, M. D. 2010. Tunnel Vision in the Criminal Justice System https://www.psychologytoday.com/us/blog/maybe-its-just-me/201005/tunnel-vision-in-the-criminal-justice-system/. *Psychology Today*.

Wirth, B. E., and Carbon, C.-C. 2017. An easy game for frauds? effects of professional experience and time pressure on passport-matching performance. *Journal of experimental psychology: applied* 23(2):138.

Zeller, B. 2019. Searching for photos of Civil War Soldiers https://www.civilwarphotography.org/index.php/resources/searching-for-photos-of-civil-war-soldiers/. *Center for Civil War Photography*.

Zhao, W.; Chellappa, R.; Phillips, P. J.; and Rosenfeld, A. 2003. Face recognition: A literature survey. *ACM computing surveys (CSUR)* 35(4):399–458.