Are 1,000 Features Worth A Picture? Combining Crowdsourcing and Face Recognition to Identify Civil War Soldiers

Vikram Mohanty, David Thames, Kurt Luther
Department of Computer Science and Center for Human-Computer Interaction
Virginia Tech, Arlington, VA, USA

Abstract

We introduce a web-based platform called Civil War Photo Sleuth for helping users identify unknown soldiers in portraits from the American Civil War era. Our system employs a novel person identification pipeline by leveraging the complementary strengths of crowdsourced human vision and face recognition algorithms.

Introduction

The American Civil War (1861–1865) was the first major conflict to have been extensively photographed, with the images being widely displayed and sold in large quantities. Around 4,000,000 soldiers fought the war, and most of them were photographed at least once. After 150 years, thousands of these photographs have survived, but most of the identities of these soldiers are lost. Identifying people in historical photos is important for preserving material culture (Martinez 2012), correcting the historical record (Schmidt 2016), and recognizing contributions of marginalized groups (Fortin 2018), among other reasons.

The current research methods employed by historians, genealogists, and collectors largely involve manually scanning through hundreds of low-resolution photographs, military records, and reference books, and can often be tedious and frustrating. To help these researchers identify Civil War portraits, we built a web platform called *Civil War Photo Sleuth*.

Identifying a Civil War soldier photo, like many person identification tasks, requires a combination of face recognition and analysis of visual clues about the person's body, clothing, accoutrements, and other contextual details.Automated face recognition algorithms are helpful for this process, but not sufficient for several reasons. Studies of these algorithms often compare with a human baseline, and many show humans outperforming these algorithms (Blanton et al. 2016; Kemelmacher-Shlizerman et al. 2016; Zhao et al. 2003). Furthermore, historical photos creates real-world challenges for algorithms because they are often achromatic, low resolution, and faded or damaged, impeding the detection of facial landmarks. These algorithms also ignore relevant facial features like scars or other skin characteristics, as well as distinctive non-facial features like ear shape or facial hair styles.

Still, these algorithms can often narrow down the search space from a large candidate pool to a smaller one that contains the correct matching photo (i.e., no false negatives) at the cost of many similar-looking photos (i.e., many false positives). This brings us to the "last mile" of person identification, i.e. helping a user pick the correct match from a set of very similar-looking photos suggested by the algorithm.

This paper attempts to address the "last mile" problem by leveraging the strengths of the human vision system via crowdsourcing to complement those of face recognition algorithms. Specifically, we address the following questions:

- How well can crowds identify a person from a set of very similar-looking photos?
- How can the complementary strengths of crowds and face recognition algorithms be combined to support person identification?
- How can an interface design help a user interpret the complementary information provided by crowds and algorithms to correctly identify a person?

In this project, we scope these questions to focus on identifying American Civil War soldiers using the *Civil War Photo Sleuth* platform.

Related Work

Our work draws on concepts from both artificial intelligence and cognitive science to create a novel person identification pipeline. (Kumar et al. 2011) introduce the concept of describable visual attributes for face recognition. We use these attributes, which have the advantage of being generalizable and human-interpretable, to help novice crowds systematically distinguish facial features.

The Feature Contrast Model (FCM) (Tversky 1977) proposes that similarity between two objects increases with addition of common features and deletion of distinctive features (i.e., features belonging to one object and not the other). In addition, the extension effect suggests that features shared by some objects in the candidate pool, but not all, have higher diagnostic value and increase the similarity between the objects having these features (Tversky

1977). (Gentner and Markman 1997) show that when comparing two objects, differences are more salient in high-similarity pairs than low-similarity ones. If the differences are *alignable*, meaning that matching relations have matching arguments and any element in one representation corresponds to only one matching element in the other representation, then they decrease similarity more than *non-alignable* differences. Our novel pipeline leverages these cognitive science concepts to help crowds and algorithms work together.

Flock (Cheng and Bernstein 2015) is an interactive machine learning platform that uses crowdsourcing for nominating features and labels to train hybrid crowd-machine learning classifiers. Tropel (Patterson et al. 2015) creates visual classifiers with limited training examples using crowdsourcing. A person identification task, however, cannot be seen as a multi-label classification problem because of scalability and complexity issues. Since both Flock and Tropel require a user to define the prediction task and example data with labels, we cannot directly apply these approaches to a person identification task.

Civil War Photo Sleuth

Base System

The proposed person identification pipeline is built on the foundation of Civil War Photo Sleuth[1], a website we developed for sharing and discussing Civil War-era portraits. The site is equipped with basic features such as the ability to upload and tag photos. It allows a user to connect the photos to profiles of Civil War soldiers with detailed military records. Currently there are over 15,000 identified Civil War soldier portraits and military service records aggregated from multiple public sources like the US Military History Institute (USAHEC 2018) and the US National Park Service Soldiers & Sailors Database (NPS 2018).

User Tags The person identification process begins with the user uploading an unidentified portrait to Civil War Photo Sleuth, which simultaneously adds the photo to the reference database to support future photo identifications. Thereafter, the user adds tags based on visual clues about the uniform, insignia, equipment and weapons. Our initial user base is targeted towards history enthusiasts with a degree of familiarity with these categories.

Filter Suggestions The system then draws upon encoded domain knowledge of Civil War portraits to generate search filters based on user-provided tags. For example, if the user tagged a hat insignia of crossed swords and a shell jacket, the system will recommend "Cavalry", and adding the coat color to be dark would add another search filter for "Union Army". These filters leverage military records to significantly narrow down the search results pool.

Face Recognition The current prototype employs Microsoft's Face Recognition API (Microsoft 2018) to scour through this filtered search pool and generate a set of candidates with faces highly similar to the unknown soldier (see Figure 1). Our initial tests show that when the similarity confidence threshold parameter is set to 0.50 with Civil War portraits, the API yields poor precision but near-perfect recall. This implies that the correct result is almost always present in the search results, bringing us to the "last mile" problem.

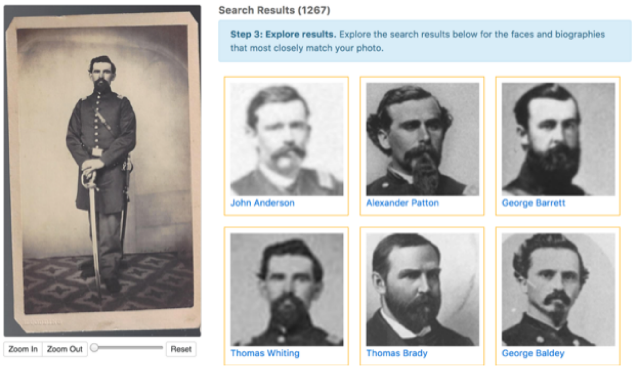

Figure 1: Photo Sleuth software prototype showing real face recognition results from the Microsoft Face API. Note that due to image quality differences, even an exact duplicate (Thomas Whiting) is not the top result.

Novel Crowdsourced Pipeline

Our crowdsourcing pipeline helps the user to find the correct match from a pool of similar-looking people by performing fine-grained photo analysis.

Feature Selection We perform fine-grained pairwise analysis by capturing information about features according to the cognitive science models described above. Based on these models, we classify these features into two types:

1. **Alignable Differences.** Building on the idea that differences are salient in this similar-looking candidate pool, the system captures information for a pre-determined feature list. We modified a subset of attributes used in (Kumar et al. 2011) for the purposes of this project. We term these attributes as *high-level features*. Examples include "hair", "eyebrow", "eyes", "nose", "mouth", "chin/jaw". For each of these high-level features, we add possible common *low-level tags* in an ad hoc manner; e.g., "hair" can be "receding", "straight", "short", etc.
2. **Unique Similarities.** Since unique features of high-diagnostic value increase similarity between objects, the system allows users to input such features that may be uniquely distinctive for the unknown photo. The pipeline captures information about the presence of these features in the search pool. Examples can include "no right hand","muttonchops facial hair style","baldness", etc.

Crowd Interface The system launches crowdsourcing tasks using Amazon Mechanical Turk such that three crowd workers consider each pair of photos. This crowd interface shows the crowd worker the unknown photo and another photo from the search pool and asks which of the high-level features are similar or different in both the photos. For the

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

features that were selected as different, the interface asks for low-level tags to be associated with both the photos.

For example, if a crowd worker selects "hair" as different in both the photos, the system asks the crowd worker which of the low-level tags for "hair" e.g. "curly", "receding", "straight", "full", "long", etc. are associated for which photo. Since the crowd worker thinks that "hair" is different in both the photos, it can be assumed that at least one of the tags will be different in order to justify the decision. The worker does the same comparison for all the different features.

The system then asks the crowd worker about the presence of the uniquely distinctive features in the other photo. After comparing all the features, the crowd worker makes an overall judgement about the similarity of the people in both photos using a four-point Likert scale.

Search Interface The search interface shows the user 1) the final aggregated crowd scores next to each photo in the original search pool and 2) the search results sorted by the these scores. The user can also perform a fine-grained analysis of one photo at a time by checking the distribution of aggregated differences along with the tags, as provided by the crowd workers. The user also sees the presence/absence of the high-diagnostic valued features.

The system also provides the option of filtering search results by the high-level features, and sorting by smallest differences compared to the unknown soldier. This is in accordance with the theory of having few features that can be counted as alignable differences and the presence of high diagnostic-valued similar features for finding "more similar" objects from a set of very similar objects.

Preliminary Results

We conducted a pilot study to measure how crowds perform on a pairwise photo comparison task. Aggregated crowd scores for four unknown soldiers suggested that the initial search pool of six photos for each soldier (that included the correct matching photo and five similar-looking photos) could be narrowed down further to a smaller pool of three photos for each soldier, with the score for the correct matching photo being the highest among all in two of those cases. These results support our hypothesis that crowds can further filter the initial pool of similar-looking photos.

We further validated the use of prior high-level features by asking crowds to nominate features that justified their comparison decision in a pair-wise analysis. We collected 216 feature responses, and our post hoc analysis found that they fell into 17 feature categories. If only facial features were considered from these categories, then they overlapped with the high-level feature list.

This justifies the use of a prior system-provided feature list since there is no apparent loss of information. There is also a speed trade-off with a prior feature list as we can employ a "yes/no" line of questioning rather than free-flow text inputs for capturing feature-related information.

Future Work

We are currently planning several studies to address the original research questions. The first study will examine how well the aggregated crowd scores work. We will compare with the ground truth and check the average rank of the correct matching photo when the search results are sorted by these scores. Further, we will evaluate the performance of crowd scores by seeing if a threshold can be established that narrows down the original search pool. We will measure how the "crowd + face recognition" system differs from the original search pool.

A second study will examine whether the crowd decisions and the feature responses are correlated. Here we plan to find the effectiveness of alignable differences and unique similarities in contributing to the final decision, and correlate with the ground truth. We will also perform a qualitative evaluation of the responses.

A third study will evaluate the user's interaction with the overall system. We compare the success rate of the user correctly identifying matches by using only the face recognition search results as opposed to the "crowd + face recognition" system. In addition, we will also compare the percentage of search results the user has to scour through in both the systems before making a final decision. Further, we will evaluate the effectiveness of the feature information by checking how often the user refers to fine-grained pair-wise analysis. Here we check the number of cases in which the user uses the distribution of differences to make a final decision, and how often it is correct. We plan a similar evaluation for the presence of unique similarities.

Finally, we will evaluate how our proposed system compares against the user's current, manual identification methods, in terms of the time taken and success rate for correctly finding a match.

Conclusion

Civi War Photo Sleuth's hybrid crowd + face recognition pipeline attempts to address the "last mile" problem in person identification, on a dataset of historical photographs that presents both cultural value and technical challenges. Since this pipeline has the flexibility of being data-agnostic, our hybrid approach may also generalize to other domains where person identification is relevant, like journalism and criminal investigation. At the same time, our work opens doors to exploring new ways to leverage the strengths of the human vision system for complementing the power of an AI system in complex image analysis tasks.

Acknowledgements

We wish to thank Ron Coddington and Paul Quigley for historical expertise, and Sneha Mehta, Nam Nguyen, and Abby Jetmundsen for early prototyping. This research was supported by NSF CAREER-1651969.

References

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.

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.

Fortin, J. 2018. She Was the Only Woman in a Photo of 38 Scientists, and Now Shes Been Identified. *The New York Times*.

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

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.

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.

Martinez, R. 2012. Unknown No More: Identifying A Civil War Soldier. *NPR.org*.

Microsoft. 2018. Face API - Facial Recognition Software | Microsoft Azure https://azure.microsoft.com/en-us/services/cognitive-services/face/.

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

Patterson, G.; Van Horn, G.; Belongie, S. J.; Perona, P.; and Hays, J. 2015. Tropel: Crowdsourcing detectors with minimal training. In *HCOMP*, 150–159.

Schmidt, M. S. 2016. Flags of Our Fathers Author Now Doubts His Father Was in Iwo Jima Photo. *The New York Times*.

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

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

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