Extracting, Assimilating, and Sharing the Results of Image Analysis on the FSA/OWI Photography Collection

Paul Rodriguez
SDSC
10100 John Hopkins
La Jolla, CA.
prodriguez@sdsc.edu

Sandeep Puthanveetil Satheesan
NCSA
1205 W Clark St
Urbana, IL 61801
sandeeps@illinois.edu

Jeffrey Will
Valparaiso University
1900 Chapel Drive
Valparaiso, IN 46383
1-219-464-6875
jeff.will@valpo.edu

Elizabeth Wuerffel
Valparaiso University
1709 Chapel Drive
Valparaiso, IN 46383
1-219-465-7908
liz.wuerffel@valpo.edu

Alan Craig
XSEDE
P.O. Box 5020
Champaign, IL 61825
a-craig@illinois.edu

ABSTRACT

This paper reports on the continued work on image analysis of the Farm Security Administration – Office of War Information Photography Collection team, supported through an XSEDE grant (Extreme Science and Engineering Discovery Environment) and Extended Collaborative Support Service (ECSS). The team is refining existing algorithms, developing new algorithms and executing them on the Comet supercomputer to analyze the FSA-OWI corpus from 1935-1944, held by the Library of Congress (LOC). The project spans many fields within the humanities and beyond, including photography, art, visual rhetoric, linguistics, American history, anthropology, and geography, as well as appealing to the general public. Progress includes refining image, metadata, and lexical semantics analysis, as well as developing a search, retrieval, and sorting interface through Clowder, which will serve as the public portal. Methods and tool refinement for this project are suitable for use on other large image corpora.

CCS Concepts

•Applied computing → Arts and humanities, Optical character recognition • Computing methodologies → Image processing, Object recognition • Information systems → Data mining, Information extraction, Web interfaces, Web services.

Keywords

Data Mining; Image Analysis; Photography; Art History; Humanities; Linguistics; Computer Science; Interdisciplinary

PEARC17, July 09-13, 2017, New Orleans, LA, USA

ACM 978-1-4503-5272-7/17/07…$15.00

http://dx.doi.org/10.1145/3093338.3093365

1. INTRODUCTION

Supercomputing power offers the potential to humanities scholars and artists the ability to sift through image-heavy information expediently. Questions that were previously limited by time and scale of work become, with supercomputing power and image analysis tools, much more achievable in a practical timeframe.

The collection analyzed in this effort includes images from the Resettlement Administration (1935-1937), the Farm Security Administration (FSA, 1937-1942) and the Office of War Information (OWI, 1942-1944), as seen in Fig. 1. In this paper, the collection will be referred to as the "FSA collection" for brevity.

The FSA is a collection of primarily black and white photographs made of American life, with the initial intent of educating the public on the plight of farmers and sharecroppers and rural rehabilitation efforts. The collection grew to include photographs of towns and cities, and, in the 1940s, the war effort. Of the hundreds of thousands of images made, nearly 175,000 are digitized and accessible to the public through the Library of Congress.

Our previous work focused on the development of infrastructure in order to extract features from the collection [1]. This involved identifying key questions to ask of the data, identifying image features and processes that can extract such information, and implementing programs to use the Comet supercomputer resources at San Diego Supercomputer Center. We also ran through a sample of images and did preliminary analyses in order to clarify features and possible data mining use cases. Here, we report on our progress using the entire corpus. We describe refined algorithms, methods, and evaluations to attain greater accuracy with facial detection, Stryker hole punch detection and extraction, metadata processing, and lexical semantics analysis of captions. We also found that the collection includes many images that depict signs and text, as can be seen in Fig. 1. Therefore, the team also deployed and enhanced optical character recognition (OCR) tools for text detection.

Finally, we also continued to develop Clowder [2] as a future public portal. Here, researchers and the general public will be able to access data from the image analysis and perform a Boolean search, with the potential for visual data mining.

Figure 1. Image from the FSA collection and held by the LOC. This image contains numerous forms of text, including hand painted window signage and various posters. Title: "Barber and shop, South Omaha, Nebraska." Photographer: John Vachon. Date Created: 1938.

2. MOTIVATION

Photographers and patrons of the arts would be unlikely to consider such famous images as Dorothea Lange's Migrant Mother (see Fig. 2) as data. But with such a sizeable collection of photographs from the FSA/OWI that encompass a relatively brief time period, it is hard not to see the beauty in both the image itself and the overall corpus as an incredible source of data.

The object (the film negative), became a digital artifact when the Library of Congress archived and scanned the collection. Our ability to easily access and analyze the images in this collection of digital artifacts offers the public and researchers the ability to encounter it as both art and big data. Because the Library of Congress included critical metadata such as extensive captions, dates, and locations, with the collection, it is not simply big data, but is, in fact, heading toward smart big data. With increasingly refined open access algorithms for image processing and the included metadata from the Library of Congress, the corpus allows us to take advantage of this smart big data. Christof Schöch argues that smart big data is what scholars in the humanities need because it can not only "adequately represent a sufficient number of relevant features of humanistic objects of inquiry to enable the level of precision and nuance scholars in the humanities need, but it can also provide us with a sufficient amount of data to enable quantitative methods of inquiry [3]."

Large image collections, such as the Farm Security Administration and Office of War Information 1935-1944 photo project of 175,000 images, are ripe for supercomputing image analysis. As more image analysis tools are developed, "reading" the image for valuable information, followed by sorting and processing the gathered data allows artists, art historians, and humanities scholars entrance into collections in a new and more expedient way. The process of a researcher identifying the face of a person in a photograph, while a simple task, becomes overwhelming when faced with a collection that enters into the tens of thousands or hundreds of thousands of images.

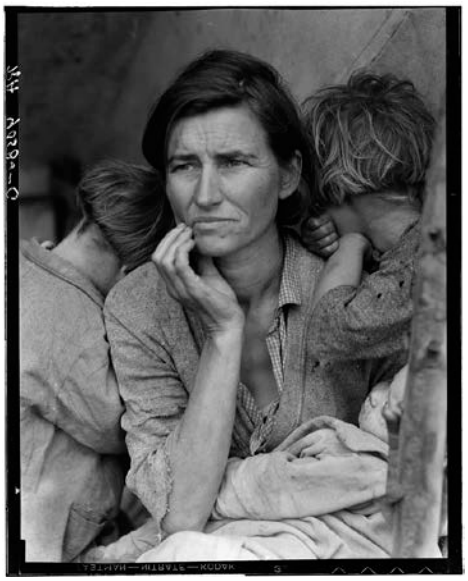

Figure 2: Dorothea Lange's famous image known as "Migrant Mother," from the FSA collection and held by the LOC. Title: Destitute pea pickers in California. Mother of seven children. Age thirty-two. Nipomo, California. Photographer: Dorothea Lange. Date Created: 1936

Such a task would take months to accomplish. But with existing algorithms and supercomputing power, the process can be expedited and allows the researcher and artist to begin to ask more complicated questions more quickly. Which FSA photographers tended toward photographing people? Were there photographers that had their photographs rejected by editors more frequently than others? Were images more likely or less likely to be "killed" (IE – The image is rendered useless by having a hole punched in it by Roy Stryker. See Section 3.2) if they included faces of people? Is there a correlation between photographers and what they photographed? Where they photographed? What they wrote in the caption? Which photographers wrote more descriptive captions? Did some use more emotionally engaged rhetorical approaches? These questions are not impossible to ask without computational support; they would, however, involve an incredibly time-consuming manual process.

3. FEATURE EXTRACTION

3.1 Image Content Extractions

3.1.1 Face Detection. Using Python with OpenCV 3.0 on Comet, faces and face-related information (eyes, profile) was extracted for all images. The batch job ran on Comet in about 12 hours for all images on one compute node.

We conducted multiple iterations of tests, to establish the tradeoffs in true vs. false positives for a sample set of about 2K (~1%) images at different resolutions. The LOC makes available small, large, larger, and largest sizes of the images. We performed a summary of resolution sizes and found that larger and largest sizes are not consistent nor always available. However, the 'large' size was more consistent in that it was usually near 1024x1024 pixels and it was always available. After running through several parameters, it was found that a minimum face size=75x75, and

minimum neighborhood size of 2 (i.e. face detected over shifts of 2 pixels), scale factor of 1.2, seemed to give a good trade off of false positives to false negatives. In a sample of 91 images, 90 faces were true positives, 7 false positives for an accuracy of about 93% (i.e. TP/(TP+FP)). There were about 5 faces that were subjectively obvious false negatives, as well as 10-15 other head shots with faces that were obscured, or not facing the camera, and thus not obviously showing a face (see Fig. 3).

Figure 3. Face detection results showing a true positive, a false negative (likely because of the hat and shadow), and 2 false positives (one of which is easily eliminated by excluding small faces, and one of which coincidently has face-like arrangement of dark and light areas). Image id: fsa1997000948.

3.1.2 OCR Detection. Many images have text, such as prominent signs, address numbers, the 'Ohio' on the side of a railroad car; an exit sign; etc. We ran tests with OCR in MATLAB™ and found that it produces a large number of false positives and misses obvious text in some images. We improved sensitivity using OCR over different image sizes (0.4 to 1.4 scaling) with and without binarization (but always with contrast enhancement).

We improved specificity by using a classifier in combination with OCR as follows. We took the scale/binarization combination that has the most characters detected, where there were at least 3 characters together, after removing all special characters. We noticed anecdotally that patterned lines were sometimes recognized as letters, so some of these ('iIlLxXuUhHmMvV1235') were deemed as potentially 'bad' characters. For each text-box we created variables for the count of bad characters, count of double 'bad' characters, number of capital letters, whether the box was vertical or horizontal, the number of letters/digits recognized, the number of 'CVC' occurrences (consonant-vowel-consonant), the OCR confidence score, as well as resize parameter and whether or not image was binarized. We gathered these features for all images that had some OCR result on a 1% random sample of the 171k images. Out of ~1800 images there were 307 text boxes found in 149 images. We labelled these by hand to be True/False occurrences of text and fed these to a RandomForest decision tree model to get a prediction for each text box. Out of the 307 boxes identified by OCR alone, 173 were true positives, 134 were false positives, for total accuracy of 56% (173/307). (Of course, false negatives are ill defined given that there are a large number of possible boxes one could extract from an image.) The OCR + RandomForest correctly classifies 146 of 173 of the true positives, 104 of 130 of the true negatives (which in this case are boxes that have already been extracted), for 250 of 307 (81%) total accuracy, which is a significant improvement.

We then extracted OCR variables for all 171K images, and using the RandomForest model made a final prediction for the presence of text in an image as True if at least 1 text box in an image was predicted 'True'. As a rough estimate of false negatives of images, we quickly viewed about 800 left over images from the sample set and found about 40 had subjectively legible text, 20 (2.5%) of which seemed obvious for OCR, and only 5 (0.6%) of which had subjectively prominent signage, and which might be clearly considered false negatives. The OCR pipeline ran in about 48 hours on one Comet compute node with all cores.

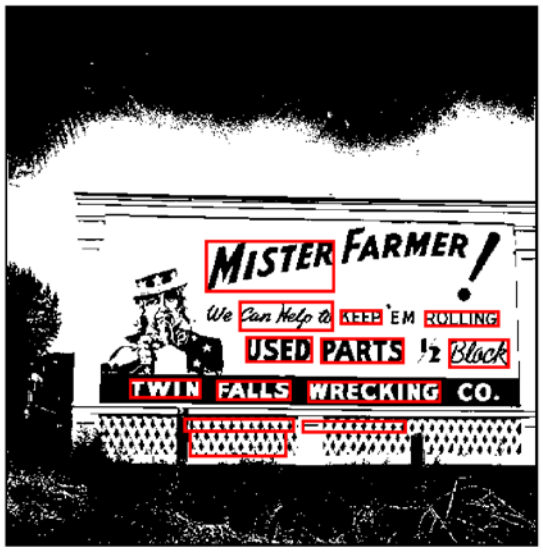

Figure 4. OCR results shown by red boxes around identified text locations in a binarized image. It shows several true positives, false negatives, and false positives for the *presence* of text. False positives often occur for repeating visual patterns. Identified characters are often noisy, especially for unusual fonts or slanted text, as seen in the following set of strings: [fof; NbN; O0bH; WRECKING; FALLS; TWIN; TMIM; PARTS; USED; ROLLING; KEEP; zgm; M1511]. Image id: fsa2000050760.

3.2 Image Properties Extraction

We also processed each image and derived the presence of a Stryker hole, which indicates when the original curator (with last name Stryker) applied a hole punch in the photo. The Stryker hole punch is of particular interest given that it was a method to reject images that curator Roy Stryker deemed unsuitable for public consumption. The Stryker hole punch detection feature allows quicker access to this subset of images, enabling researchers to more quickly answer questions regarding whether there were content or compositional reasons for killing the image, or if

certain photographers were disproportionately edited. An extractor also removes the black circle from the image, thus eliminating skewing image property metrics. The parameters for hole size, darkness, and radius were evaluated and we were able to get 100% correct on a 100-image sample set. We also gathered information about mean and variance of gray values, not including the border and hole, as well as a histogram of gray scale values. This processing took about 96 total hours on one compute node on Comet for all 171021 images.

3.3 Metadata Extraction

The image metadata consists of a variety of information. The creator and creation-date information was easily picked off the LOC json file. However, the creation dates are highly inconsistent. Often the dates are given as a range of years, or a full date, or partial date, or not at all. Thus, we only did minimal post-processing of month names.

The location information was contained in a field along with subject information, such as:

United States - Cedar Rapids, IOWA - Small Farms

Although the first segment is always the country, the other fields are sometimes rearranged, not properly delimited, or missing values, including County names, or replaced the state name with a regional name (e.g. 'Midwest'). We processed the geolocation using available longitude/latitude databases for cities and states. We programmed some flexibility by applying entity resolution on the state or region first, before identifying the city. Our final results out of the 171021 images were non-US locations: 22.2% (38100); US only 12.4% (21233); US and state only (no city) 28.7% (49137); US and regional 0.3% (573); US, city and state 36.2% (61978). This process ran in about 4 hours on one Comet compute node.

3.4 Caption Textual Analysis

Each image caption has one or more title sentences (e.g. Fig. 4). One aspect that makes this work especially relevant to image analysis for humanities is that the metadata was not written to identify what is in a photograph, but rather to give some framing and cultural context. In other words, the image captions often do not directly describe objects in the image. The objective of this processing is to extract words from the caption, and provide part of speech and semantic categories for user review.

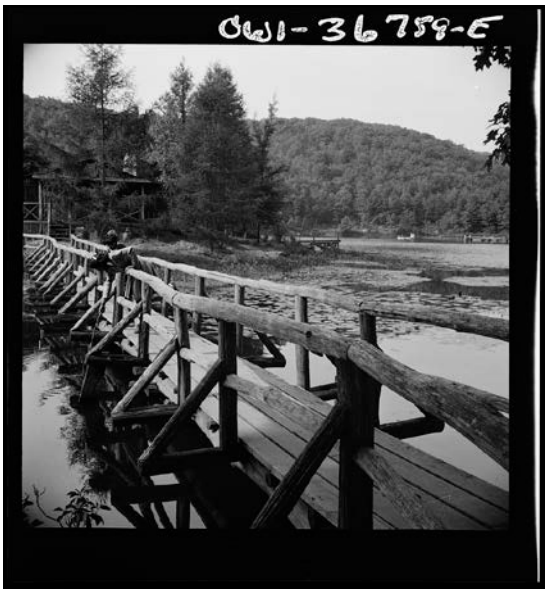

Figure 5. This figure (owi2001034959) had the following caption: "Arden, New York. Interracial activities at Camp Gaylord White and Ellen Marvin, where children are aided by the Methodist Camp Service. The view from the bridge".

All captions were run through Python scripts that perform several steps. First, captions that are untitled, or untitled with a reference to another caption were removed and cross referenced. Out of 171021 images, 59.4% (101510) have titles, 10.7% (18367) were untitled but were marked as related to another titled image, and 29.9% (51144) were untitled. Then, raw captions were separated into sentences. Words that would be appropriate for frequency count analysis (i.e. no stop words) were put aside. It would be straightforward to perform sentiment analysis on these words as well as compare counts as function of time or creator. Location information and named entities were removed and put aside. This left more basic, but still complete, phrases and sentences. These are run through a script to extract part of speech and word sense information. This script uses the Stanford Natural Language Processing module for parsing and dependency relations, Natural Language Toolkit 3.0 for part of speech tagging, and WordNet ontology for semantic interpretations [5-8]. There was no attempt to perform word-sense disambiguation, only the most frequent sense was listed. However, two scores were derived to help indicate if that sense was highly likely, as shown below in Table 2.

Table 1 shows the named entities extracted, and the category of those entities for the title in Fig. 4. The category is more useful for considerations of word sense and what entities are depicted in the image. Table 2 shows the lexical-semantic features extracted and part of the full data schema (other fields include position in sentence, lower level semantic categories, and actual frequency proportions for a word sense from the WordNet corpus). The processing for the corpus took about 192 hours on 1 compute node on Comet with 8 cores – although it's embarrassingly parallel and was processed using many nodes in about 2 days.

Entity Words	Category
Camp Gaylord White	LOCATION
Ellen Marvin	PERSON
Methodist Camp Service	ORGANIZATION

Table 1. Named Entities found and categorized using Stanford NLP tools.

Word	POS	Semantics	P(W|S)	P(S|W)
CHILDREN	Noun	Entity; Agent; Person	0.71	0.69
BRIDGE	Noun	Entity; Structure; Bridge	1.0	0.36
VIEW	Noun	Abstraction; Orientation; Position	0.60	0.45
etc…	…	…	…	…

Table 2. Some of the words from captions with their semantic information and probabilities in the WordNet dataset. P(W|S) is the probability, based on frequency proportions, of using that word for the given semantic sense, and similarly for P(S|W).

One way this data could be used is to select an image of a certain kind of content. For example, selecting all captions that include a semantic category of 'FOOD' yields 50 images out of a 1000 image test set. All but one or two depict eating, farming activities, dinner gatherings, or food preparation. As another example, in order to find images of indoor scenes one could try selecting semantics that include some kind of 'Structure', combined with the occurrence of the preposition 'in' (and excluding 'in front '). In a test set of 1000 images there were 126 images that satisfy this query, 23 of which were not indoor scenes.

4. DATA ACCESS and PUBLIC GATEWAY

One of our main challenges for this project is to present the data for mining and analysis in a way that can include visual presentation. As an interim first result, all the features described above were extracted, identified by the LOC image index as a primary key field, and setup as individual tables in SQLITE on Comet. Currently, the features are organized as tables for Face, Image Properties, OCR, Subject, Creator, Location, Raw captions, Caption Words, Caption Semantics, as well as supporting tables for Image file names, Duplicate titles, and Named Entities. These tables are available now to the project team. In fact, the example photograph below (Fig. 7) was produced using our results so far and a combination of SQL, Unix shell scripting, and Matlab programming. These tools are obviously not viable for a general audience. Moreover, the database is useful but does not easily lend itself to combine with a visual analysis, where the results could be an album of images, or a graph in which the axes represent visual properties and the plotted points are themselves images. A better solution would include a way to visualize results with the option of accessing the images themselves, which we describe next.

We are developing a public facing web gateway for those who want to browse and search the metadata generated from our work. We are using Clowder [2], which is a cloud-based scalable research data management software (see Fig. 5) for the data management needs of this project. The data is extracted off of an Sqlite database on Comet. In fact, some of the programs that generated the feature data were derived from Brown Dog extractors [9]. In addition to data management, Clowder and Brown Dog tools will be used to set up this public facing gateway.

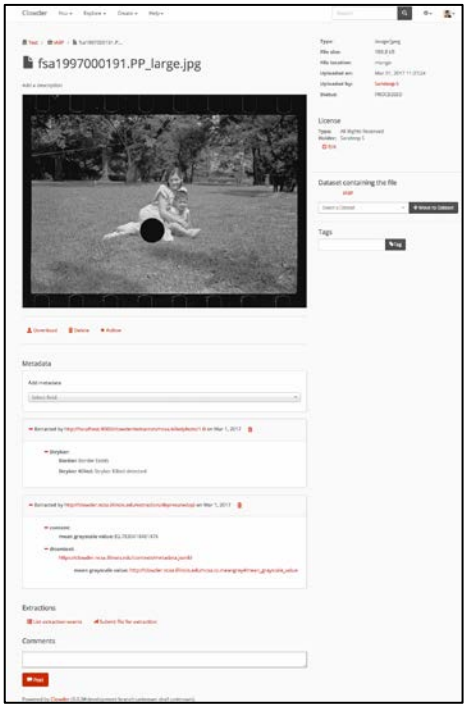

Figure 6. A screenshot showing the file page of the Clowder instance used by this project. The metadata extracted about the presence of Stryker hole punch and mean grayscale value can be seen at the bottom of the photograph.

5. FUTURE WORK

This work motivates a number of extensions for the future. New and increasingly refined open access algorithms for image processing will allow this corpus to be mined for a wealth of new information. Increasingly precise extractors such as gradient histograms that indicate the many edges in an image would provide greater insight into the visual content in the corpus. Likewise, the specificity of analyzing image captions could be improved by searching specifically for the structure and preposition in the same clause, and using sense scores. This could be further combined with visual property information for a more robust analysis.

Further development of the public portal on Clowder will provide the general public and researchers the ability to interact with advanced visualization and search features to browse through the image analysis results with ease. The advanced search feature (see Fig. 6) provides capabilities for users to build custom queries like *"Find all photographs from the Farm Security Administration collection that were not Stryker killed* [10]*, whose mean grayscale value is greater than 85, and contain eyes."* Advanced visualization will include Venn diagrams, charts, and other

illustrations that will give users a high-level overview of the extracted metadata and a different perspective of the FSA/OWI collection. For example, Fig. 7 depicts the results of a query that looks for photographs with faces, Stryker killed photograph, the intersection, and their matching non-killed photos. For visual review, one could zoom into the album results – notice that in one case the family portrait was rejected in favor of a mother/child image, in the other case the mother was rejected in favor of the family portrait. This kind of difference is only available upon visual inspection, but leads to new queries and new image subsets one could review. The advanced search feature is already available in Clowder and could be enhanced to fulfill requirements of this and future projects. We are in the design phase for developing tools for integrating the generated metadata in Clowder with visualizations.

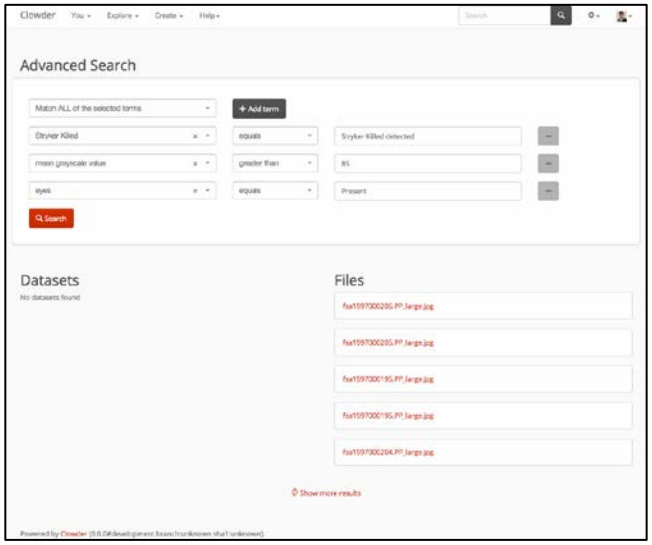

Figure 7. A screenshot of the advanced search page of Clowder where a user can generate custom queries and search the generated metadata.

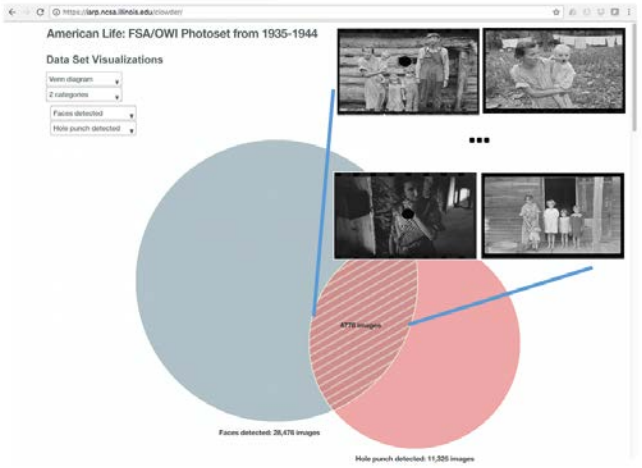

Figure 8. Mockup showing a possible Venn diagram based visualization for exploring the generated metadata and mining subsets of photos.

6. CONCLUSIONS

This work demonstrates not only the challenges in analyzing an image database of such a large scale, but identifies key techniques for achieving this goal. Though this work started with identification of key high-level questions to gain insight on the history and culture of the subject material, the iterative process of implementing extractors, evaluating results, and refining questions leads to additional investigative questions. The infrastructure presented here provides an agile platform to not only interact with present extractors, but to add extractors based on evolving needs without the need to change the underlying structure. The development of a unified public user interface allows artists and humanists to explore a dataset that would normally be overwhelming without the mental burden of understanding the underlying image processing algorithms.

A key contribution to this work is not only the tangible results for the present Farm Security Administration – Office of War Information Photography Collection, but the extension of these techniques to other image databases, several of which are available through the Library of Congress.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grant numbers ACI-1053575 (XSEDE) and ACI-1261582 (Brown Dog), and ACI-1341698 (Comet).

PI and Co-PIs thank Sandeep Puthanveetil Satheesan, Paul Rodriguez, Alan Craig, Kenton McHenry, and Marcus Slavenas for their assistance with image analysis, which was made possible through the XSEDE Extended Collaborative Support Service (ECSS) program.

REFERENCES

[1] Slavenas M, Wuerffel E, Rodriguez P, Will J, Craig A. (2016) Image Analysis and Infrastructure Support for Mining the Farm Security Administration – Office of War Information Photography Collection. *Proceedings of the XSEDE16 on Diversity, Big Data, and Science at Scale - XSEDE16.* (2016).

[2] Open source data management for research: https://clowder.ncsa.illinois.edu/. Accessed: 2017-03-01

[3] O Schöch, Christof. 2013. Big? Smart? Clean? Messy? Data in the Humanities. Journal of Digital Humanities. 2, 3, pp.2-13.

[4] Bradski, G. 2000. The OpenCV Library. Dr. Dobb's Journal of Software Tools.

[5] Bird, S., Klein, E., and Loper, E. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

[6] Finkel, J. R., Grenager, T., and Manning, C. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 363-370.

[7] Klein, D. and Manning, C. D. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423-430.

[8] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[9] Padhy, S. et al. 2016. An Architecture for Automatic Deployment of Brown Dog Services at Scale into Diverse Computing Infrastructures. *Proceedings of the XSEDE16 on Diversity, Big Data, and Science at Scale - XSEDE16.* (2016).

[10] Benson, A.C., 2010. Killed negatives: the unseen photographic archives. *Archivaria*, 68, pp.1-37. Van